



SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

School of Medicine and Surgery

PhD Program in Molecular and Translational Medicine

XXXIV Cycle

TITLE: DISSECTING THE GENETIC ARCHITECTURE OF
AUTOIMMUNITY: A SPOTLIGHT ON PRIMARY BILIARY
CHOLANGITIS

Dr. Alessio Gerussi
Matr. No. 849287

Tutor: Prof. Pietro Invernizzi
Co-Tutor: Prof. Rosanna Asselta
Co-Tutor: Dr. Fabrizio Mafessoni
Coordinator: Prof. Andrea Biondi

ACADEMIC YEAR 2020-2021

A gene is defined as any portion of chromosomal material that potentially lasts for enough generations to serve as unit of natural selection.

Richard Dawkins, The Selfish Gene (1976)

Any living cell carries with it the experiences of a billion years of experimentation by its ancestors.

Max Delbrück, A physicist looks at biology. Trans. Conn. Acad. Arts Sci. 38, 173–190 (1949)

When two genes, like the brown eye and the blue eye gene, are rivals for the same slot on a chromosome, they are called alleles of each other. For our purposes, the word allele is synonymous with rival.

Richard Dawkins, The Selfish Gene (1976)

Complexity can be defined as the degree of interactions, ordered or disordered, between the components of a system.

R. Singh, Genes and genomes and unnecessary complexity in precision medicine, npj Genomic Medicine (2020)

'No Man is an Island'

*No man is an island entire of itself; every man
is a piece of the continent, a part of the main;
if a clod be washed away by the sea, Europe
is the less, as well as if a promontory were, as
well as any manner of thy friends or of thine
own were; any man's death diminishes me,
because I am involved in mankind.*

*And therefore never send to know for whom
the bell tolls; it tolls for thee.*

MEDITATION XVII

Devotions upon Emergent Occasions, John Donne (1624)

TABLE OF CONTENTS

CHAPTER 1: General Introduction	11
1.1 The Genetic Architecture of Complex Traits	12
1.1.1 Missing Heritability of Complex Traits.....	14
1.2 Established Statistical Genetics Approaches	19
1.2.1 Genome-wide association studies	19
1.2.2 Polygenic Risk Scores	21
1.3 Novel methodological approaches	25
1.3.1 Machine Learning applications in Population Genomics	25
1.3.2 Chromosome X-Wide Analysis	32
1.4 Evolutionary perspectives on complex traits	33
1.4.1 Determinants of evolutionary genetics.....	34
1.4.2 Neanderthal and Denisovan DNA	37
1.4.3. The role of archaic variants in present-day humans.....	39
1.4.4. The role of archaic introgression in immunity ..	40
1.5 The Genetics of Autoimmunity	42
1.5.1 Genetic determinants of Autoimmune Diseases	46
1.5.2 The Genetic Architecture of Primary Biliary Cholangitis	49
SCOPE OF THE THESIS	69
References	70
CHAPTER 2	97
An international genome-wide meta-analysis of primary biliary cholangitis: novel risk loci and candidate drugs	98
Abstract.....	102

Introduction	104
Methods	106
Results	113
Discussion	122
Conclusions	126
Table Legends	127
Figure Legends	128
Table 1	129
Table 2A	130
Table 2B	131
Table 3	133
Figure 1A	134
Figure 1B	135
Figure 1C	136
References	137
CHAPTER 3	142
X chromosome contribution to the genetic architecture of primary biliary cholangitis	143
Abstract	147
Introduction	149
Methods	151
Results	157
Discussion	163
Conclusions	167
Table Legends	168
Figure Legends	169
Table 1	174

Table 2	175
Figure 1	176
Figure 2	177
Figure 3	178
Figure 4	179
Figure 5	180
References	181
CHAPTER 4	188
A novel Polygenic Risk Score in Primary Biliary Cholangitis	
.....	189
Abstract	191
Introduction	193
Methods	194
Results	199
Discussion	204
Conclusions	207
Table Legends	208
Supplementary Table Legends	210
Figure Legends	211
Supplementary Figure Legend	214
Table 1	217
Table 2	220
Table 3	221
Table 4	222
Table 5	223
Table 6	224
Table 7	225

Figure 1	226
Figure 2	227
Figure 3	228
Figure 4	229
Figure 5	230
Figure 6	231
Supplementary Table 1	232
Supplementary Table 2	233
Supplementary Figure 1	234
Supplementary Figure 2	235
Supplementary Figure 3	236
Supplementary Figure 4	237
References	238
CHAPTER 5	242
Machine-learning SNP-based prediction for Primary Biliary Cholangitis: a proof-of-concept study	243
Abstract	245
Introduction	247
Methods	249
Discussion	263
Conclusions	268
Table Legends	269
Figure Legends	270
Supplementary Table Legend	272
Table 1	274
Table 2	276
Table 3	280

Table 4	281
Figure 1	282
Figure 2	283
Figure 3	284
Figure 4	285
Supplementary Table 1	286
Supplementary Table 2	287
Supplementary Table 3	288
Supplementary Table 4	293
Supplementary Table 6	295
References	296
CHAPTER 6	301
The archaic mutational load predicts the fate of introgressed fragments in humans	302
Abstract	304
Introduction	305
Methods	306
Results	309
Discussion	314
Conclusions	316
Figure Legends	317
Supplementary Figure Legends	319
Figure 1	321
Figure 2	322
Figure 3	323
Figure 4	324
Supplementary Figure 1	325

Supplementary Figure 2.....	326
Supplementary Figure 3.....	327
Supplementary Figure 4.....	328
Supplementary Figure 5.....	329
Acknowledgements	330
References.....	331
CHAPTER 7: Summary, Conclusions and Future Perspectives.....	336
 Summary.....	337
 Conclusion and application to translational medicine .	341
 References.....	349
Publications	352
Acknowledgements.....	363

CHAPTER 1: General Introduction

1.1 The Genetic Architecture of Complex Traits

The fundamental question in genetics is to understand how genetic variation influences phenotypic variation. “Complex traits” are those traits that, in contrast with monogenic conditions, tend to cluster in families but do not show a Mendelian inheritance pattern, because the genetic liability to manifest the trait is dependent on multiple variants together with environmental factors.

There is established evidence supporting the concept that complex traits are highly polygenic, and thousands of alleles scattered around the genome give their contribution to genetic variance[1]. In addition, differently from Mendelian conditions, genetic variation for complex traits is largely dependent on noncoding variants that affect gene expression[2].

Recently, Boyle and colleagues proposed the so-called *omnigenic model*[3]. This model defines “core genes” as genes that have a biological role in disease and “peripheral genes” those not directly connected to disease. The omnigenic model postulates that variation in peripheral genes is the main contributor to the genetic risk; the activity of peripheral genes influences, via regulatory networks, the function of core genes. Since the number of core genes is largely outnumbered by the group of expressed genes in a disease-relevant tissue, the sum of small effects across peripheral genes exceeds the genetic contribution of core genes. Therefore, authors conclude that causal variants are scattered across the genome and not isolated

in biologically relevant genetic modules, that any variant having the capacity to modulate gene expression occurring on a peripheral gene has a non-zero effect on regulation of the core genes, bringing a small but collectively not negligible effect on disease risk. From an evolutionary point of view, supporters of the omnigenic model argue that it is in accordance with the evidence that most of the evolutionary change is derived by small shifts in allele frequency involving several causal variants across the genome rather than few variants carrying large phenotypic effects[4]. Risk alleles at core genes are kept by natural selection at low frequencies in the population based on their key functional role. Interestingly, rare variants do show a different pattern, since they are more frequently deleterious and have large effect sizes. In addition, rare variants are more frequently related to disease-relevant functional categories than common variants identified by genome-wide association studies (GWAS)[3]. The omnigenic model suggests that the reasons why many variants identified by GWAS do not relate with the phenotype is because they are correlated with peripheral genes and not with core genes. A key consequence of this method is the suggestion to focus more on rare variants rather than carrying on with the GWAS methodology.

While appealing, this theory has received criticism from some researchers in the field because it is considered too simplistic, since experimental evidence does not support such a significant reduction in the complexity of architecture of complex traits[5]. When rare variants of very large effect are associated with

complex traits and are not lethal they more often generate a more severe phenotype, which is different from the “average” phenotype of the disease. Moreover, effect size is not a synonym of biological relevance, since many GWAS variants linked to drug targets of registered medications are common risk variants with small effect size[6]. Rare monogenic diseases may help to pinpoint core genes; yet, for common diseases, the limiting factor is sample size, more than genotyping technology[5]; the UK Biobank represents a good example of such valuable initiatives. Indeed, for precision medicine to become reality, we need to increase the number of variants implicated in the genetic liability of disease, and the plateau has not been reached yet. In fact, while the ultimate goal is to move from the genetic investigation to the identification of druggable targets to improve treatment of affected patients, on the other hand there is the need to tailor available (or repurposed) drugs to individuals. This implies to match risk alleles with phenotypic presentations.

1.1.1 Missing Heritability of Complex Traits

The concept of “missing heritability” has been introduced by Teri Manolio and other researchers to summarize the gap between the numerous variants that have been identified by GWAS and the small effect size that these variants do carry, even when considered altogether and when their additive effects are taken into account; in other words, the missing heritability concept

underscores that heritability estimates derived from GWAS are much lower than those derived from twin studies[7].

To better understand this concept we need to revise how heritability is defined in population genetics.

The total genetic variance (V_G) can be divided into different components: an additive term (V_A), a dominance term (V_D) and an interaction term (epistatic component, V_I).

$$V_G = V_A + V_D + V_I$$

While the generation of an individual is characterized by complex, nonlinear interactions, from a quantitative point of view most of the genetic variance V_G has been reduced to its additive component.

If L is the number of segregating sites that affect the given trait and a_l and a'_l are the effects of the parents' alleles at site l , the additive genetic variance is the sum of the contributions of a_l and a'_l .

$$V_A = \sum_{l=1}^L (a_l + a'_l)$$

The total phenotypic variance (V_P) is the sum of the additive variance V_A and the contribution from the environment (V_E).

$$V_P = V_A + V_E$$

Broad-sense heritability (H^2) is defined as the ratio between V_G and V_P , while Narrow-sense heritability (h^2) is defined as the ratio between V_A and V_P and does not include the contribution of dominance and interaction. It is worth remembering that: the heritability estimate is specific to the population and environment you are analyzing; the estimate works at the population level, and is not an individual parameter; and finally, heritability does not indicate the degree to which a trait is genetic, it measures the proportion of the phenotypic variance that is the result of genetic factors.

The heritability ratio (*π explained*) is the ratio between the heritability due to observed variants, h^2 *known*, derived from genome-wide significant single-nucleotide polymorphisms (SNPs) found in GWAS, and the total heritability, h^2 *all*, determined from concordance studies on homozygous siblings. The concept of “missing heritability” refers to the fact that this ratio is far from the unit and for many diseases is below 0.5. However, some authors disputed this definition of missing heritability, introducing the notion of “phantom heritability” and suggested that the issue may reside in the way the denominator is calculated. Specifically, when heritability is calculated, the assumption is that no genetic interactions (epistasis) is present. After including epistatic effects they proved that the denominator

can be overinflated, exaggerating the missing heritability issue[8].

Whether the effect of dominance and epistasis is negligible for complex traits is still a matter of debate. They are typically observed for mutations with large effect size, while variants with large effect size are uncommonly found in GWAS. Molecular biology studies clearly show that gene-gene interactions are common in nature, although quantitative genetics falls short in proving this, which means showing the non-additive effects to phenotypic variation[5].

The common disease-common variant hypothesis postulates that most of the missing heritability is in common variants that still need to be identified. The most important evidence to support this concept is the finding that when all the SNPs are fitted altogether instead of testing one-by-one and adopting a stringent threshold for multiple testing, the variance of a highly polygenic trait -such as height- goes from less than 10% to around 45%. Many of the variants included in this type of studies would be typically missed by the standard approach due to the high threshold for inclusion. The remaining variance is possibly missing due to incomplete linkage disequilibrium (LD) between causal variants and genotyped SNPs[9]. This hypothesis marries well with the evidence that the increase in sample size allows the discovery of further novel variants, expanding the fraction of explained genetic variability. Indeed, if the variance explained by each SNP is small, larger samples are needed to find SNPs associated with the trait of interest. A brilliant example of this

argument is provided by the largest GWAS performed to date on height, including GWAS data from 5.4 million individuals of diverse ancestries and showing that around 12k SNPs covering ~21% of the genome explain nearly all of the common SNP-based heritability[1]. To conclude, the main issue for the common disease-common variant hypothesis is lack of power in GWAS. On the contrary, the common disease-rare variant hypothesis focuses its attention on rare variants. It postulates that rare variants that are mildly deleterious are key players in the susceptibility to complex traits[10]. Arrays used to perform genotyping studies do not commonly capture these variants, and this would explain missing heritability. Under this model, many genetic variants that are rare and at high penetrance would be involved in susceptibility. For height, 83 height-associated coding variants with MAF > 0,1% but < 5% revealed greater effect size than effects of common variants[11]. On top of improving the capture of SNP chips, novel strategies beyond association studies should be validated, able to look for signatures of negative selection (see for instance the BayesS method described in[12]). Moreover, structural variants and Copy-Number Variants have been ignored by first studies due to technical issues of the arrays, and more recent data including recurrent copy-number variants improved prediction. **Figure 1** recapitulates the feasibility of identifying genetic variants by risk allele frequency and odds ratio.

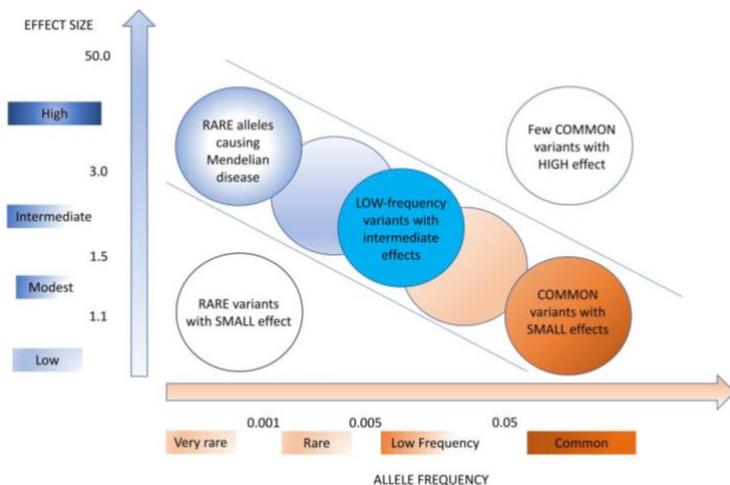


Figure 1. Types of genetic variants by risk allele frequency and strength of genetic effect (adapted from[7])

1.2 Established Statistical Genetics Approaches

1.2.1 Genome-wide association studies

GWAS test the association between genotype and phenotype by statistical comparison of allele frequencies of genetic variants from individuals with same ancestry but different phenotype (cases vs controls). The most common study design employs SNPs and the final result is a list of SNPs that are associated with a trait of interest (“genomic risk loci”)[13].

Despite many criticisms and not negligible intrinsic limitations, GWAS have represented a significant scientific breakthrough:

more than 5700 GWAS have now been completed and have identified thousands of variants in several cohorts. Many of these GWAS have paved the way for further fine mapping studies and helped to identify new therapeutic targets (e.g. anti-IL-12/IL-23 pathway in Crohn's disease, PCSK9 inhibitors for high cholesterol)[13]. Conversely, the determination of direct causality between a variant and a trait is hampered by LD; in other words, it is not immediate to tease apart which is the gene located in the genomic area in proximity with the variant pinpointed by GWAS that is actually playing a role in the biology of the disease. Several strategies to overcome this limitation are under study and this is one of the most active areas of research in genomics at the moment[13–15].

Main steps of the GWAS design include[13]:

- collection of blood sample and clinical information from individuals under study;
- genotyping of each individual;
- quality control;
- imputation based on reference populations and haplotype phasing;
- statistical analysis;
- post-GWAS analysis to characterize results emerged from statistical analysis (fine-mapping strategies).

GWAS had indisputable merits: they allowed to generate hypotheses on pathways not previously even considered, shed a

light on ethnic diversity in the genetic predisposition to diseases, generated a list of variants that can be incorporated in polygenic risk scores, and have boosted the data science field through the institution of many well-established consortia[16]. Yet, GWAS have intrinsic limitations related to their design and should now be put together with other experimental approaches. Indeed, GWAS are penalized by multiple testing burden, do not take into account gene-gene and gene-environment interactions, are unable to pinpoint causal variants but only tag candidate loci, and most polygenic risk scores have fallen short in their prediction power[16]. Moreover, many ethnic groups have not been included for years in GWAS[17]. Several reports are available in the literature clearly showing the benefit in the dissection of the genetic architecture of diseases by including subjects from under-represented geographical regions[18].

1.2.2 Polygenic Risk Scores

Complex traits derive from the interaction of environmental factors and several genomic variants; therefore, they are known also as “polygenic” diseases. To assess genetic liability for a polygenic disease, polygenic risk scores (PRSs) can be generated. These scores provide an estimate about how a subset of variants influence the risk of developing a condition[19].

The current standard to calculate a PRS is to weigh a number n of alleles based on effect sizes β derived from GWAS. A score S of an individual is the weighted sum of individual genotypes X_j for n SNPs.

$$S = \sum_{j=1}^n X_j \beta_j$$

PRS assumes an additive model of heritability. Typically, weights are derived from summary statistics of GWAS or meta-analysis and applied to external cohorts as validation. Genomic variants typically largely exceed the number of individuals that have been genotyped; for these reasons, several methods to select variants to include in PRS have been proposed and studied. The simplest and most commonly used one is the pruning and thresholding method. This method takes into account LD, which causes high levels of correlation between variants with shared evolutionary history in a population (pruning phase). Thresholds can vary from including only SNPs at genome-wide statistically significance ($p < 5 \times 10^{-8}$) or looser thresholds. Other methods have been employed by the genetic community with increasing interest, such as bayesian methods[20].

The use of summary statistics distinguishes PRSs from models leveraging individual data; the latter are promising since they might offer more powerful prediction, in spite of being more time-consuming and less scalable[21]. Most PRSs are not sufficiently accurate to predict alone risk of disease; yet, when clinical covariates are added to the model, the combination of genomic

variants and clinical variables increase the capacity to assess susceptibility of a disease. Nevertheless, a recent study revealed that PRSs for common and less common diseases such as coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease and breast cancer can stratify individuals in groups well separated so that those subjects within the group with the highest values of score have an estimated disease risk similar to that conferred by mutations in monogenic disorders[22]. Although, a PRS offers a relative risk measure, since at current stage they cannot be used to evaluate absolute risk of the disease in a specific individual[20]. Again, PRSs should be seen as an additional tool on top of clinical risk scores rather than an alternative method. The clinical application of PRSs is still unclear, but it can be predicted that PRSs will allow public health measures targeted at higher-risk categories in order to change lifestyle habits and reduce preventable diseases; it is recent evidence that even in healthy subjects high genetic scores identify individuals with detectable abnormalities in several blood analytes that predate onset of overt diseases (**Figure 2**)[23].

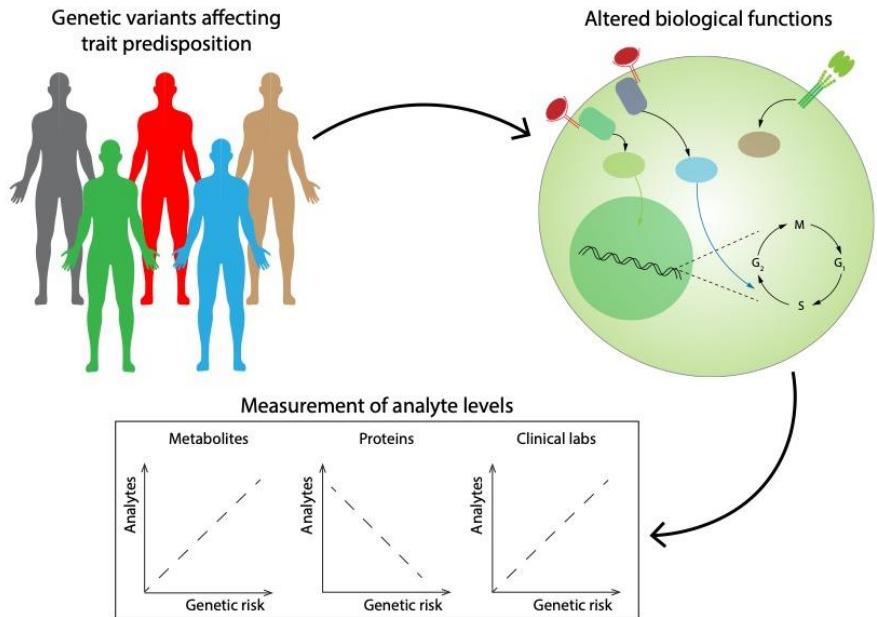


Figure 2. Healthy individuals with high genetic risk scores for a specific trait have already detectable abnormalities in several blood analytes. Genetic risk scores generated from risk variants for several traits have been associated with levels of plasma analytes (standard blood analytes, proteomic and metabolic measurements), revealing that a not negligible level of dysregulation of these analytes can be found already in healthy subjects with high genetic risk. These approaches could be potentially leveraged for early detection of diseases. Adapted from[23]

1.3 Novel methodological approaches

Over the last decade, most of the efforts have been put toward finding new causal variants. Yet, based on the limitations of GWAS and PRSs, it is essential to investigate novel strategies to address the methodological issues that hamper the full exploitation of the enormous amount of data generated from high-throughput genome sequencing[16]. We present here two novel approaches, one devoted to the application of Machine Learning on genomics data, and the other aimed to study genetic variants present on the X chromosome, which can represent a useful advancement for the field.

1.3.1 Machine Learning applications in Population Genomics

1.3.1.1 A primer on Machine Learning

In 1959, Arthur L. Samuel, a computer scientist, firstly introduced the expression “machine learning” in his paper[24], which was focusing on how a machine could learn the game of checkers. Generalizing this idea, machine learning (ML) could be defined as the process in which, by means of an algorithm, a process of interest is modelled, so that the behaviour of upcoming instances of the same process can be predicted. For instance, this could

be the case of a process that analyses historical data concerning the time needed to recover from a disease, and it is later used to predict the recovery time for new patients with the same disease. This is an example of a problem that can be dealt with through ML.

In the above mentioned example, a set of *labeled* data is at disposal: in other words, the number of days needed to recover, which constitutes the value of the *target* (alternatively referred to as the *output*) of the problem at hand, is known for a set of examples: more specifically, for the patients that are tracked in the historical record. If this condition is met, the problem falls into the area of ***supervised learning***.

More specifically, considering that the desired output is a *quantity* (the number of days needed for recovery), the problem is referred to as a *regression*. If it consisted instead, for instance, in diagnosing whether a patient had a high risk of developing the disease or not, considering that the *output* would represent a *quality* (in the mentioned case, a binary one: yes/no), rather than a *quantity*, the problem would be defined as a *classification*.

There are also cases in which a set of *labeled* data is not available: for instance, splitting a set of patients into homogeneous subgroups with respect to a set of features or determining which diagnostic factors are correlated to each other, rather than with respect to a specific (and known) *target*. These examples, respectively referred to as a *clustering* and an *association mining* problem, belong to the ***unsupervised learning*** macro-category.

Similar problems can be treated also with a *statistical approach*. The main difference between *statistics* and *machine learning* is that *machine learning* does not require to make any assumption concerning the statistical distribution of the considered features. Conversely, the outcome of a *statistical* pipeline relies on (and benefits from) the knowledge about the underlying distribution of the considered population, as well as the statistical properties of the chosen estimator.

This requirement makes the statistical approach less affordable in the case of high-dimensional data, as it further increases complexity. Conversely, considering that ML approaches cannot be compared against any reference distribution for evaluation, it is harder to assess their performance. A common procedure to tackle this issue consists in comparing the model predictions against the data themselves.

More specifically, a part of the data (for instance, a subset of the patients whose data are available in the historical record) are used to *train* the model, while another part is not supplied to the training pipeline. The former part is usually referred to as the *training set*, while the latter is commonly denoted as the *test set*. After having completed the modeling phase on the *training set*, the model itself is applied on the *test set*: in other words, it is used to make predictions on a subset of data which looks unknown to the model (and simulates the new data on which it will be applied). More complex scenarios involve also the introduction of a third dataset, referred to as a *validation set*, so that we can

have a dataset for *training*, a dataset for *hyperparameter tuning*

on top of training and another one for *performance evaluation*.

In the example then, referring to the *test set*, a set of predictions about the number of days needed to recover for patients that were not considered in the modeling phase, but for which a ground truth value is known would be produced. Evaluating, for instance, the *root mean squared error* (or, in the case of a *classification*, other indicators such as *empirical accuracy* or Area Under the Curve (*AUC*) would allow then to have an idea about how much the model that has been extracted is effective on previously unseen data, such as the ones that it will be required to process once it is operating.

In case a good performance is measured if the model is re-applied on the training set, but a poor one is observed on the test set, the model is overfitting data: in other words, it is too focused on replicating training data and is not able to generalize its predictive power.

Among the different ML available models, a couple of well-known approaches are artificial neural networks[25] and Support Vector Machines[26]. In the last couple of decades, neural networks, and especially multi-layered ones, have become even more popular, being at the core of deep-learning approaches.

Deep learning refers to a pipeline in which the learning process is modularized: the first layer of modelling can be considered as in charge of learning features that will be used by the following layer, enabling the final one to provide a prediction. This has been shown to be particularly promising in the field of high-

dimensional, unstructured data, such as documents or images[27].

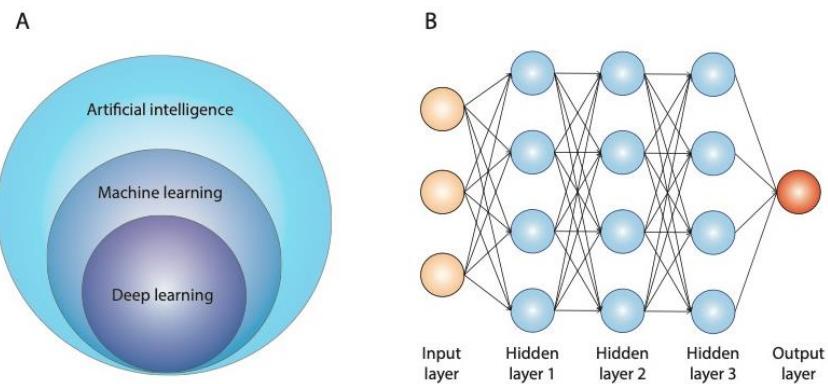


Figure 3. Schematic Representation of the relationship between Artificial Intelligence, Machine Learning and Deep Learning (A). Basic architecture of artificial neural networks (B).

1.3.1.2 Foreseeable Applications

ML represents a potent tool for analyses of data derived from high-throughput sequencing. As for other scientific fields, some of the most relevant applications of ML can be: 1) generation of models for classification; 2) clustering of individuals in groups; 3) feature selection.

ML is considered a complementary tool in population genetics, where several methodological hurdles need to be overcome. Research in population genetics has mostly focused on the

formalization and validation of statistical models that describe patterns of variations and their application to experimental molecular data[28]. While classical population genetics has been mainly characterized by parameter estimation in the context of a predetermined probabilistic model (typically the Wright-Fisher model), the target of ML is optimization of the accuracy of predictions[28]. PRS predictions are based on a linear parametric regression model, with strict assumptions like additive effects, independent effects, normal distribution of the data, and independence of observations[29]. These assumptions are often not valid in complex diseases. For example, thanks to their non-linearity, ML algorithms allow to account for complex interactive effects between associated alleles[30]. Another peculiar and powerful feature of ML is its capacity to handle thousands of dependent variables, each characterized by a massive amount of information; this ability is of interest in the genomics world, where increasing dimensionality of data is an issue[28].

In population genetics the output could be represented by the status (case or control) or a continuous phenotype (such as the value of a blood biomarker of interest), and the features are the individual sample genotype data[29]. Data feature selection is the key step to obtain an accurate ML model[30]. There are a few methods (embedded methods, wrappers) useful to select only informative SNPs as potential predictors[29].

The research question should be clear: does one want to predict outputs or to interpret data? The generative approach builds a model for two classes in a supervised manner, while the

discriminative approach focuses only on separating them via an unsupervised approach.

The main application of supervised ML in population genetics is to build a model to classify cases and controls based on SNPs. Another possible application of supervised ML could be to identify novel predictive features (SNPs) associated with phenotype, possibly looking at biologically-distinct sub-phenotypes of the disease (early vs advanced disease, onset at younger age vs older age, positivity for specific autoantibodies). In this way a predictive model is generated, taking advantage of the different contribution of variables within the training genotype data[29]. After the training phase, the models with the maximum predictive power are selected for validation. This stage is essential to avoid overfitting and is usually achieved by cross validation (dividing original dataset in a training set and a test set). Nonetheless, external replication is still required for the final validation of the model[29].

Unsupervised approaches may be used to cluster patients according to genotype data and investigate whether these novel groups have different clinical presentation, trajectories, and treatment response. For example, based on the availability of other omics, clustering can be extended to genomics and transcriptomic data. After generating clusters with hypothesis-free means, it is mandatory to understand if they hold biological and clinical significance, to create a classification that is really meaningful for clinicians.

1.3.2 Chromosome X-Wide Analysis

Chromosome X (ChrX) constitutes approximately 5% of the nuclear genome. While mutations in genes mapping on this chromosome account for approximately 10% of Mendelian disorders and are easily detected in males, limited signals have been identified as predisposing variants for complex traits. More specifically, only 114 ChrX susceptibility loci (0.8%) at $P \leq 5 \times 10^{-8}$ have been described on a total of approximately 15,000 signals identified by GWAS[31]. Several reasons have been proposed to justify this discrepancy: among others, lack of coverage on arrays, numeric imbalance in variants as compared to autosomes, but also methodological issues in the analytic pipelines classically established[31]. Yet, since at least 2013 genotyping chips have included tens of thousands of SNPs present on ChrX; similarly, it is difficult to believe that the number of variants does matter, as ChrX used to have fewer association signals than the tiny chromosome 21[31]. In a poll distributed among geneticists, many felt that the sex-specific analyses needed when including ChrX in GWAS were more difficult and less powerful.

Considering that ChrX can be regarded as an immunologic chromosome, since it contains the largest number of immune-related genes compared to other chromosomes[32], its analysis in the context of autoimmune diseases, and PBC specifically, is fundamental to uncover genes possibly contributing to the disease.

XWAS is a new software for the analysis of ChrX that accounts for X-specific issues: it allows X-specific quality check, sex-stratified analysis, testing for higher variance between heterozygous and homozygous females (which accounts for ChrX inactivation)[33]. There is evidence that application of XWAS to previously analyzed GWAS can discover new loci on ChrX, shedding a light on a neglected portion of the human genome[33,34].

1.4 Evolutionary perspectives on complex traits

Selective pressures change over time and the adaptive change involves shifting in allele frequency at many loci. The characterization of the polygenic response is central to understanding the evolution of phenotypic differences among populations and species. GWASs have elucidated the genetic architecture of many human traits; the combination of the information about the effects of individual loci on traits with changes in allele frequency across populations and time is essential to learn about polygenic adaptation and to move towards tailored management of patients in the context of precision/personalized medicine.

1.4.1 Determinants of evolutionary genetics

Natural selection, genetic drift, and gene flow are the main determinants of the modification of allele frequencies over time[35]. When one or more of these forces are acting in a population, the population does not meet the Hardy-Weinberg assumptions, and evolution happens. Therefore, the Hardy-Weinberg Theorem represents a null model for the study of evolution.

Natural selection occurs when there is survival and reproductive advantage in individuals with some genotypes as compared to some others with other genetic background; the fittest ones will pass on their alleles to the next generation[35,36].

Charles Darwin argued in *On the Origin of Species* that natural selection will occur if the following conditions are met[37]:

1. Individuals within a population have different phenotypes for the same trait;
2. Phenotypic variation is at least partly genetically driven (i.e. the offspring will somehow resemble their parents as regards the trait);
3. There is variation in fitness when there is trait variation (the relative average net reproduction of individuals with a given genotype as compared to the other individuals).

Assuming that selection is the only determinant that violates Hardy-Weinberg assumptions, we can argue that an allele A will

become fixed in the population if AA individuals are fitter, because they will generate more offspring carrying the A allele. Dominance features among alleles at the locus in question will control the rapidity of fixation. A rare and recessive variant, which is also advantageous, will be slower, because until it reaches a relatively high frequency (thanks to genetic drift for example); it will not be seen by natural selection as it will rarely appear in homozygotes. The effect of a dominant allele is immediately evident because it occurs in heterozygous individuals. Eventually, the recessive advantageous allele will be fixed and, conversely, deleterious alleles will be rare and difficult to purge for natural selection. The type of natural selection described in this paragraph is called directional selection[38].

Another type of selection is balancing selection. Balancing selection preserves genetic polymorphism within populations. It typically derives from heterozygous advantage, i.e. the higher fitness of heterozygotes; the result will be the presence of multiple alleles with stable frequencies.

Genetic drift derives from the sampling error that occurs during transmission of gametes by individuals[38,39]. Allele frequencies will change over time in a population due to chance events — that is, the population will undergo genetic drift. In a population with a constant number of individuals, the smaller the population size (N), the more important the effect of genetic drift. Drift can cause fixation of one allele among multiple neutral alleles, reducing heterozygosity. Motoo Kimura postulated in its neutral theory of molecular evolution that fixation of neutral mutation due

to drift is the main driver of evolution in nature[40]. Due to the prominent role of the size of a population in determining the strength of genetic drift, often its magnitude is modeled as a function of the effective population size (N_e), a quantity which captures the number of individuals effectively participating in producing offspring for the next generation in an idealized population.

The third force is gene flow, i.e. the movement of genes into or out of a population due, for instance, to migration of individuals[41]. If gene flow is restricted, the population will diverge due to selection and drift, potentially leading to speciation.

Mutation can be considered the first mechanism to generate diversity from a mechanistic perspective. A mutation can change one allele into another. Mutations generally occur at small rates (in humans, in the order of 10^{-8} mutations per base pairs per generation) thus affecting allele frequencies only to a negligible extent. However, by generating new genetic diversity, they may produce an allele that is selected against, selected for, or selectively neutral, based on the effect of the mutation on fitness. Hence, selection and drift are mostly responsible for the changes in frequencies of novel mutations. Harmful mutations are purged away from the population by selection and will be found in low frequencies equal to the mutation rate. Advantageous mutations will disseminate within the population through selection.

All these evolutionary determinants play together to shape populations over time. Since all real populations are finite, they

are susceptible to the effects of genetic drift. This means that directional selection, which would fix an advantageous allele if working alone, could be vanished by the effects of drift, especially if selection is weak and/or the population is small. Small populations are at risk of reduced fitness due to the effect of drift, because it will potentially fix deleterious alleles[41].

1.4.2 Neanderthal and Denisovan DNA

The field of evolutionary genetics has been recently revolutionized by the genome sequencing of archaic hominins. Of particular interest for the biomedical field is the study of the role of alleles that persist in the gene pool of present-day humans derived from these archaic hominins after interbreeding between them and ancestors of present-day humans.

The archaic hominins for which we currently have more data are Neanderthals. They are an extinct group of hominins who have been present in Eurasia before anatomically modern humans (ie humans with similar skeletal features to those of present-day) moved from Africa[42]. The first evidence of genetic admixture between Neanderthals and Eurasian modern humans came up in 2010: the genome of three Neanderthal individuals was sequenced and compared to the genome of five present-day humans, revealing a higher overlap between SNPs of Neanderthals and present-day humans in Eurasia than that

present between SNPs of Neanderthals and present-day humans in sub-Saharan Africa. After scanning the human genome for alleged Neanderthal genome segments, it has been estimated that 1-4% of the genome of non-African individuals is derived from Neanderthals in a positive selection process[43]. Later on, two seminal studies have investigated the persistence of Neanderthal genes on a big cohort of present-day humans. Both works are the result of an extensive effort to set tools to detect Neanderthal DNA by using computational methods[44,45].

Denisovans are another extinct group of hominins who lived in an area ranging from Siberia to Southeast Asia[46]. From a genetic perspective, Denisovans share a common origin with Neanderthals[47], and interbred with the ancestors of some modern humans, with about 3-5% of the DNA of Melanesians and Aboriginal Australians and around 6% in Papuans deriving from Denisovans[48]. More recent studies have shown that there are at least three different branches, one contributing an introgression signal in Oceania, another restricted to New Guinea and a third in East Asia and Siberia[48].

1.4.3. The role of archaic variants in present-day humans

Human genome sequencing can identify millions of genetic variants within individuals, and hundreds of millions of variants across populations[49]. Yet, the field of genomics is still facing a challenge to interpret genetic variation, and the development of new methods to prioritize variants that have a significant impact on human traits is essential to leverage sequencing data. As previously mentioned, the resolution of genetic strategies based on standard fine-mapping strategies is somehow limited[41]. There is a rising interest in complementary methods to prioritize variants based on evolutionary information such as sequence conservation, genic effects and regulatory element annotations; these methods have the potential to enhance the characterization of Mendelian phenotypes[41] as well as common traits[41].

Genetic admixture between archaic humans and modern humans, and consequent introgression of archaic alleles, seems to have an enduring impact on phenotypes of individuals. As already stated, around 2-4% of the genome of present-day humans in Europe and Asia is derived from Neanderthals[44]; it is likely that introgressed fragments constituted at least 10% in the generations immediately after first inbreeding events, but evolution has rapidly purged deleterious Neanderthal genes[50]. Most of the archaic alleles that are still present are considered mildly deleterious[51].

In some genomic regions of present-day individuals, researchers have found “desert” areas, i.e. areas with very few Neanderthal genes compared to others, and these deserts were mainly located in ChrX and in genes expressed in testes[44,45]. Adopting a genome-wide overview, coding regions are the most depleted areas; as regards regulatory regions, pleiotropic enhancers are more devoid of Neanderthal variants than tissue-specific ones[52]. Simonti and colleagues found that Neanderthal alleles may explain disease risk for a few human clinical traits such as skin disorders and mood disorders[53]. Archaic alleles might also be implicated in predisposition to diseases, e.g. type 2 diabetes[54] or severe Coronavirus disease 2019[55]. Recent data have shown that the general trend was selection against introgressed functional variants, with some relevant exceptions involving skin and immunity[56].

1.4.4. The role of archaic introgression in immunity

There is a long line of evidence that infections have played a key role in human evolution. Tuberculosis, a dreadful lung infection that has plagued humanity for centuries, has been a powerful driver of negative selection, based on the evidence that homozygotes for the P1104A of *TYK2* have higher risk to develop clinical forms of the disease and that allele frequencies of this

polymorphism have widely fluctuated in Europeans over time[41].

Neandertals have introduced into European genomes a few variants that modulate immune responses[57]. These variants might have contributed to substantial immune advantage for modern humans coming out of Africa. There is evidence that T cell lymphocytes are among the most enriched tissues in Neanderthal DNA[51]. Researchers have identified three Toll-like receptors that carry archaic alleles with a well-defined functional role in present-day humans[58]. There are also proofs that introgressed Neanderthal DNA in modern humans helped them to adapt against viruses[59]. While most of the studies have investigated immunity in general or host-pathogen interactions, there is scanty evidence about the link between archaic variants and autoimmunity. There is evidence that the Neanderthal variant of the glutathione reductase associates with an increased risk to develop inflammatory bowel disease[41]. Risk alleles for autoimmune diseases may be the outcome of an evolutionary trade off[60,61]. Yet, if the environment changes more rapidly than the genetic background, like it has occurred over the last two centuries, maladaptation may occur[61]. The study of these aspects may inform the reason why there is a trend toward an increased incidence and prevalence of these conditions.

1.5 The Genetics of Autoimmunity

Autoimmunity occurs when the immune system activates against endogenous antigens. This phenomenon represents the pathogenetic mechanism behind some chronic diseases defined for this reason as autoimmune diseases (AiDs).

Self-tolerance is the result of a double-step mechanism: the central self-tolerance and the peripheral self-tolerance.

T lymphocytes acquire central tolerance in the thymus; it involves a first positive selection of double positive CD4+ and CD8+ T cells that recognize self-antigens presented by MHC class I or II with medium-high affinity in correspondence of the thymic medulla. These selected cells are then translated to the corticomedullary junction, whereas a negative selection of those lymphocytes that bind self-antigen with high affinity occurs[41]. AutoImmune REgulator transcription factors (AIRE), mainly expressed by medullary Thymic Epithelial Cells (mTECs), play an important role in the induction of T-cell central tolerance (Transcriptional regulation by AIRE: molecular mechanisms of central tolerance) as they upregulate the expression of Tissue-Restricted Antigens (TRAs) in the same cells[41] in order to identify autoreactive T cells that strongly bind these TRAs and thus to help in their negative selection[41]. mTECs develop mainly thanks to the interaction between APC cells CD40 receptor and T cells CD40L[41]. At the end of this double-step selection, autoreactive cells are destroyed or, with the help of mTECs, they can be turned into regulatory T cells (Treg) that play

an important role in the development of the peripheral tolerance[41].

A similar mechanism to the one described for T-cells central tolerance is implemented in the bone marrow for B lymphocytes, so that, after their VDJ (Variable, Diversity, and Joining) portion rearrangement of the genes encoding for their membrane immunoglobulin, autoreactive cells are negatively selected and thus destroyed[41].

The induction of central self-tolerance can be unsuccessful and therefore it is necessary that remaining self-reactive cells can be eliminated also peripherally. A first mechanism of peripheral self-tolerance foresees that autoreactive T cells can be deleted through a Fas-Fas ligand system-induced apoptosis[41]. Another way by which peripheral self-tolerance takes place is based on the cell-mediated immune response mechanism; the naive T cell CD4+ that encounters the class II MCH presented antigen by the APC cell compatible with its TCR can be activated in the secretion of IL-2 only in the case in which there are additional cofactors to stimulate the immune response: traditionally, the B7 molecule on the membrane of the presenting cell must bind the corresponding CD28 receptor on the membrane of the T cell. In this way, IL-2 amplifies the immune response and polarizes the diversification of the T cell CD4+ towards the Th1 lymphocyte, which in turn activates the macrophage response against intracellular pathogens[62]. If a mature naive T cell CD4 + is self-reactive, the listed co-stimulatory signals do not meet and the cell itself undergoes an anergic phenomenon that makes it

functionally unresponsive[62]. Furthermore, Treg cells constitutively express the Cytotoxic T Lymphocyte-associated Antigen 4 (CTLA-4), which determines the CD4+ naive T cell inhibition when bound to B7[62]. This is precisely the mechanism by which Treg cells, whether they are natural CD4+ Foxp3+ Treg cells or thymic induced Treg cells, express their mechanism of action in the context of the peripheral self-tolerance acquisition. The immune response inhibition can occur according to a further mechanism, which involves the binding between the PD1 receptor, expressed on the surface of many immune cells and either of its ligand PD-L1 or PD-L2, which are expressed on various cells of hematopoietic and non-hematopoietic origin[63]. The B and T Lymphocyte Attenuator (BTLA)[64] and the V-domain Ig-containing suppressor of T-cell activation (VISTA)[65] both expressed by CD4+ Th1 cells, similarly to what happens for the other immunomodulatory receptors listed above, play a role in modulating the peripheral immune response against self. Lastly, T cell immunoglobulin and mucin domain 3 (TIM-3) expressed in CD4+ and CD8+ T cells[66], as well as Lymphocyte Activation Gene 3 (LAG3)[67], expressed also on the surface of NK cells and dendritic cells are involved in suppressing proximal T-cell signaling as for modulating the immune response.

A further strategy of immune tolerance to self-antigens is represented by clonal ignorance, according to which there are immunologically privileged sites (brain, eyes, testicles, placenta and fetus)[68], in which immune cells have limited access and do not activate against local self-antigens[69].

If any of the steps involved in the self-tolerance induction fails, then the pathophysiologic basis for the development of AIDs can occur. However, it should be also pointed out that the activation of the immune response against self-antigens can also occur due to peripheral altered interactions among immune cells during the development of the physiological immune reaction against non-self-antigens[70]. The interaction between the CD40 receptor constitutively expressed by APC cells ad lymphocytes B and its CD40L expressed by T cells stimulates the change of the immunoglobulin isotype synthesized by B cells and promotes their differentiation into plasma cells[71] responsible for the antibody-mediated adaptive immune response. Aberrant expression of CD40 on tissues, in which it is normally undetectable, may be responsible for initiating autoimmune reactions[72]. Similarly, excessive T cell expression of the Inducible T Cell Costimulator (ICOS), of CD27 and of CD28 costimulators have been reported in patients suffering from AIDs[73].

Moreover, there is growing evidence that CD4+ pro-inflammatory Th17 cells, which are physiologically involved in the defense against extracellular pathogens, play a pivotal role in the etiopathogenesis of autoimmunity. The IL-17 cytokine family acts, in fact, as chemoattractant for immune cells and it has proven to have detrimental effects on various tissues (Th17 cells in renal inflammation and autoimmunity), as well as it revealed a principal role in systemic autoimmunity.

More recently, it has also been recognized an important role of the innate immune response in the etiopathogenesis of AiDs(Control of adaptive immunity by the innate immune system): endosomal TLR3, TLR7, TLR8 and TLR9 receptors for the recognition of self-nucleic acids, as well as other cytosolic sensors expressed by innate immune cells, such as monocytes, may induce the production of type I interferons, specifically IFN- α , and pro-inflammatory cytokines (Intracellular nucleic acid sensors and autoimmunity). Finally, the role of NK cells in the pathogenesis of AiDs is no less important than those already described: if, on one side, some NK subsets modulate excessive adaptive immune responses that may otherwise lead to autoimmunity, on the other side other NK subset activity is linked to inflammasome activation[74].

AiDs can be distinguished into organ-specific AiDs or systemic AiDs. They overall have a global prevalence of over 5% in the general population[75] and have a variable impact on the quality of life and life expectancy of affected subjects. Most AiDs exhibit higher incidence in females[76].

1.5.1 Genetic determinants of Autoimmune Diseases

AiDs include both monogenic and polygenic disorders. Yet, AiDs with Mendelian inheritance represent only a tiny fraction of the whole group, including paradigmatic conditions such as APS-1 (autoimmune polyendocrine syndrome type 1), IPEX

(immunodysregulation, polyendocrinopathy, and enteropathy X-linked) syndrome, and ALPS (autoimmune lymphoproliferative syndrome).

Common AIDs are polygenic, and environmental factors play a significant role in their etiology. Interestingly, the degree of polygenicity is less than other complex traits such as schizophrenia and height, because a restricted number of regions (the HLA locus among the others) provides a large contribution[77]. GWAS identified hundreds of genomic regions mediating risk for several AIDs. These associations primarily map in non-coding regions: lead GWAS SNPs are more likely to be associated with the expression levels of neighboring genes than is expected by chance, and the same lead SNPs are enriched in regulatory regions marked by chromatin accessibility and modification. Fine-mapping has revealed enrichment of AID-associated variants in enhancer elements active in stimulated T-cell sub-populations, and heritability is strongly enriched in such regulatory regions. In addition, lead SNPs are enriched near binding sites for immune-related transcription factors but non-coding causal variants seem to act by non-canonical modification of regulatory sequence. A lower proportion (8% vs 60%) is represented by promoters, again mostly in CD4+ T-cell subpopulations. Yet, some AID showed preferential mapping to B cells: systemic lupus erythematosus, Kawasaki disease and primary biliary cholangitis[78]. Collectively, these lines of evidence suggest that the majority of disease risk is mediated by changes to gene regulation in specific cell subpopulations[77].

Recently, a novel tool evaluating whether autoimmune risk variants and *cis*-eQTLs of three immune cell subpopulations share a single genetic effects, has shown that only a fraction (about 25%) do show a joint effect. In other words, despite GWAS loci are often enriched in *cis*-eQTLs, it is likely that perturbations induced by different variants are restricted within a particular set of tissues and/or cell subtypes. In many other cases, the expression levels and the disease risk work independently in the same locus[77,78]. This piece of evidence supports novel strategies of data integration aiming at linking GWAS results with results derived from single-cell RNA sequencing techniques[14]. The study of the regulation of gene expression means to investigate how the variation in the regulatory systems generates diversity in traits. Loci involved in genetic predisposition to autoimmunity are above all the Major Histocompatibility Complex (MHC) region, followed by the protein tyrosine phosphatase gene *PTPN22*, expressed in lymphocytes. *PTPN22* encodes a protein involved in signaling; the minor allele is associated with the tryptophan allele and determines a large defect in signaling. Many other candidate loci disclosed by GWAS are related to genes coding for proteins modulating co-stimulatory signals such as *CTLA-4*, *CD2/CD58*, *CD28*, *ICOSLG*, and *TNFSF15*[79]. Besides signaling molecules, there is another remarkable class: cytokines and chemokines. Several genetic variants have been associated with perturbation of cytokines pathways[79]; it is worth mentioning, among the others, IL2 and its receptor, IL23 receptor, IL10 and the tumor necrosis factor (TNF) pathway[79].

AIDs affect females more often than males. Several hypotheses have been made to explain this observation, including hormones, sex chromosomes, behavior, and differences in environmental exposures[80]. Recently the Genotype-Tissue Expression (GTEx) project evidenced that 37% of genes show sex-biased expression in at least one tissue of the human body.

To conclude, human genomics is currently undergoing a shift toward a better understanding of genetic determinants of sexually differentiated traits. AIDs represent one of the disease categories characterized by more marked sex bias and will benefit from the development and improvement of sex-stratified genetic analysis. Ultimately, a net benefit will be reached when sex-informed diagnostic and therapeutics are realized[80].

1.5.2 The Genetic Architecture of Primary Biliary Cholangitis

Primary biliary cholangitis (PBC) is a rare condition characterized by the autoimmune attack of the small bile ducts[81]. The etiology of the disease is still unclear, and it is useful to distinguish the two phases of its pathogenesis: the initiation phase, driven by defects in the biliary homeostasis that, together with aberrant antigen presentation, lead to autoimmune phenomena in the liver, and the progression phase, mostly driven by cholestasis (ie. toxic retention of bile acids)[82]. Both phases are under the

control of genetic and environmental factors. From a therapeutic point of view, only drugs targeting the latter phase of the disease have proved some degree of benefit. Indeed, until recently, only ursodeoxycholic acid (UDCA) was available to treat PBC, and 20-40% of patients did not respond to treatment and were likely to progress to liver cirrhosis and death[83]. Over the recent years many new drugs have been developed; unfortunately, there are still no promising drugs able to target key pathogenic processes in the early phase of the disease course[84]. From a genetic point of view, PBC is a complex trait, meaning that susceptibility to the disease is derived from the interaction of environmental triggers and an unfavourable set of gene variants[85,86]. GWAS have provided a landscape of genetic variants and the progressive evolution of fine mapping strategies has fostered the understanding of pathways involved in the disease[14]. Moreover, GWAS have clearly brought evidence that many genetic variants are shared with other extrahepatic autoimmune conditions, supporting the notion that PBC is a prototypical autoimmune condition affecting small bile ducts[78]. Yet, due to inherent limitations of GWAS design, there is still a significant portion of heritability to be captured and efforts to link gene variants to subsets of cells specifically involved in the pathogenesis are still to be made[16].

1.5.2.1 PBC is an autoimmune disease

In line with other autoimmune conditions, PBC shows striking female preponderance[87]. The female-to-male ratio 9:1 from historical series has been recently challenged[88] but, even if less pronounced, female sex does remain the most represented one among patients. The diagnostic hallmark of PBC is the presence of anti-mitochondrial antibodies (AMAs), which are present in >90% of cases[89]; in AMA-negative cases other autoantibodies, such as PBC-specific anti-nuclear antibodies (gp210 and sp100), are found[90]. Evidence showing the presence of autoreactive cells within the liver is well grounded, and self-antigens have been revealed, most frequently referring to the E2-domain of the pyruvate dehydrogenase complex (PDC-E2)[91]. On histology, PBC is characterized by lymphocytic inflammation around portal tracts and biliary epithelial cells; these infiltrates are mainly composed of autoreactive CD4+ and CD8+ T-cells that are reacting to PDC-E2[92]. Immunosuppressive medications (e.g. Budesonide) have shown some beneficial effect on liver enzymes in PBC[93], but their benefit-risk ratio is considered unfavourable and are typically reserved to cases when also features of Autoimmune Hepatitis are concomitantly present[89]. PBC can recur after liver transplantation, which suggests that effector memory T cells of the recipient are able to attack the donor liver; the persistence of autoimmune memory adds on the evidence that PBC is an autoimmune disease[94].

1.5.2.2 PBC is a cholestatic liver disease

PBC is a disease of the small bile ducts of the liver and it is characterized by intrahepatic cholestasis that follows biliary injury[82]. Loss of tolerance to the E2 subunit of the PDC-E2 complex on cholangiocytes has been experimentally shown, and biliary epithelial cells in apoptosis form remnants (apoptotic bodies) whose clearance is impaired in PBC[91]. Yet, the historical view used to associate these phenomena to the cascade of events linked to progression, while initiation was associated with some environmental triggers like chemical compounds or infectious agents[95]. More recently, it has been proposed that defects in cholangiocyte bicarbonate homeostasis may be involved in the initiation phase as well[82]. Anion exchanger 2 (Ae2) plays a key role in human cholangiocytes, being one of the master regulators of $\text{Cl}^-/\text{HCO}_3^-$ exchange activity and ultimately the bicarbonate umbrella protecting biliary epithelial cells[96]. The correlation between this transporter and PBC has been suggested after finding that its expression in liver tissue and peripheral blood mononuclear cells was reduced in patients with PBC[82]. More evidence supporting a causal role derives from animal models: $\text{Ae2a,b}^{-/-}$ mice develop PBC-like features, showing histological and immunological features of PBC[97].

1.5.2.3. PBC is a complex heritable trait

From a genetic perspective, PBC is a polygenic complex trait. Several common variants, each carrying a small effect size, shape the architecture of the genetic risk for PBC. Environmental agents (eg recurrent urinary tract infections, tobacco, hair dyes) do play an important role in the etiopathogenesis of the disease[98]; Dyson and colleagues recently shown geographical clusters in the UK, which further reinforces the environmental trigger notion[99]. Like other autoimmune conditions, the disease onset is thought to follow the interaction of an environmental trigger(s) with a predisposing genetic background.

Epidemiological evidence of a strong heritable component in PBC points to several elements related to the familiarity for the disease. Indeed, patients with a positive family history are ~1.3 to 9.0% of cases, a value that is higher than the expected risk in the general population[100]. Moreover, the presence of one-degree relative affected by the disease is an independent risk factor for the development in the other first-degree members of the family (Odds Ratio = 6.8-10.7)[101]. Sibling risk is in line with other autoimmune complex disorders, with a lambda value for sibling relative risk of 10.5[102]; the concordance rate in identical twins is 0.63, among the highest reported in autoimmune conditions[103].

1.5.2.4. Genetic studies in PBC

Before the advent of GWAS, numerous investigations specifically pointing to the HLA region or to candidate non-HLA genes were performed. The HLA region analysis was repeated by several research groups and evidenced a relatively small number of alleles, making the results more reliable than candidate studies on non-HLA alleles, which were mostly underpowered and resulted in several signals not validated in following GWAS[104]. Up to now, several GWAS (including Immunochip fine-mapping studies) and one meta-analysis have been performed in PBC with subjects from Europe, North America, Japan, and China included. GWAS have identified numerous HLA and non-HLA variants, contributing to the increase in the understanding of the immunogenetics of the disease and suggesting possible druggable pathways[105–112].

1.5.2.5 HLA variants associated with PBC

The HLA complex on chromosome 6 includes groups of genes encoding HLA class I and HLA class II proteins involved in antigen presentation, and HLA class III proteins such as tumour necrosis factor (TNF) and other immune-related molecules. Genetic variations in this region have been associated with many, if not most, AIDs[113].

In pre-GWAS times, candidate studies found as main risk allele DRB1*0801 in British and Italian individuals, and the DRB1*13 and DRB1*11 in Italians as protective alleles[114,115]. More recent studies have confirmed these findings, showing that HLA-DRB1 genes (alleles *08, *11, and *14) account for most of the DRB1 association signal; DRB1*08 is the strongest predisposing allele, whereas DRB1*11 is the protective one[108].

Further studies have been repeated in non-European populations, including mostly Japanese and Chinese subjects. In Japanese studies, DRB1*0803 was identified as the strongest risk allele, similarly to what reported in studies from Europe and USA, and DQB1*0604 as the protective one[116]. For Chinese individuals, analyses revealed that HLA-DQB1*0301, HLA-DPB1*1701, and HLA-DRB1*0803 could largely explain HLA association with PBC[117]. HLA-DQB1*0301 confirmed its protective role also in this Han Chinese PBC cohort, in line with a previous study from China[117,118].

The strong LD that characterizes this genomic region makes challenging the definition of causality for identified variants. Together with the low prevalence of haplotypes in the general population, non-HLA variants have drawn progressively more interest over time. Yet, HLA associations have a larger contribution in terms of PBC heritability and further efforts to overcome current methodological hurdles should be made. A full list of HLA associations with PBC are reported in a recent review[86].

Despite the crucial role of HLA in the genetic predisposition to the disease, current clinical practice do not include assessment of HLA haplotypes for diagnostic purposes like in autoimmune hepatitis, where specific haplotypes are part of the original revised criteria[119].

1.5.2.6 Non-HLA variants associated with PBC

Many variants outside the HLA region have also been found, and genes directly or indirectly involved in immune regulation have been called in action[86]. At present, GWAS have identified 44 non-HLA PBC predisposition loci at a genome-wide level of significance in PBC. A second international meta-analysis has been recently published, unravelling 20 novel additional loci (see Chapter 2 of this thesis)[120]. Interestingly, no signals related to genes involved in biliary physiology have been found hitherto. This finding is in contrast with the more recent data derived from experimental models suggesting that abnormalities in the biliary epithelium could represent a primer of the disease. While the different methodological approaches (GWAS vs in vitro/in vivo models) could account at least partially for this discrepancy, the current evidence on the genetic architecture of PBC is unequivocally that of an autoimmune disease. This is further reinforced by the observation that a huge amount of loci overlap between PBC and other well-established autoimmune

conditions, such as inflammatory bowel disease, multiple sclerosis and systemic lupus erythematosus, to name just a few[86].

The IL-12/JAK-STAT pathway

Gene variants within the IL-12 signaling pathway emerged as strongly associated with PBC in GWAS. IL-12 is a heterodimeric cytokine encoded by two different genes, IL-12A (which codes for the p35 subunit) and IL-12B (which codes for the p40 subunit)[121]. IL-12 controls growth and function of T and Natural Killer (NK) cells. It is generally produced by antigen-presenting cells (mostly monocytes/macrophages and dendritic cells) and favours the switch toward the Th1 response. T lymphocytes and NK cells are stimulated by IL-12 to produce interferon-gamma (IFN- γ) and TNF- α increasing their cytotoxicity capacity.

The Th17 pathway is also involved in IL-12 signaling: the p40 protein heterodimerizes with IL-12p19, thus forming IL-23. The Th17 pathway involvement is key in the pathogenesis of PBC[122]. The IL-12 receptor is a heterodimeric receptor made by two chains (b1 and b2); cytokines promoting Th1 cell development stimulate IL-12 receptor, while those that promote Th2 cell development act as inhibitors. The IL-12 receptor is expressed on the cell membrane of activated CD4+ T cells and is located at the top of the cascade activating the JAK-STAT pathway. Upon interaction with IL-12, the IL-12 receptor offers

binding sites for two kinases, TYK2 and JAK2 that in turn activate transcription factors like STAT4. IL-12 has been linked to several AIDs, such as Type 1 Diabetes and Systemic Sclerosis[123,124]. The biological role of IL-12 in PBC is further supported by evidence coming from animal models[125]: knockout mice for IL-12p40 subunit show less inflammation and bile duct damage compared to non-knockout mice in dominant negative transforming growth factor beta receptor type II (dnTGF β RII) model[126].

Gene variants at TYK2[109] and STAT4 loci have also been found associated with PBC susceptibility[108,109,112] and other autoimmune conditions[127].

The immunological synapse

T cell activation requires two signals. The first signal derives from the binding of the T-cell receptor (TCR) to peptide-MHC complex; the second signal derives from the interaction of costimulatory molecules[128]. Many variants have been mapped to genes related to the (patho-)physiology of this synapse.

The CD80 gene codes for a membrane protein, known also as B7-1, which is found on the surface of antigen-presenting cells (B cells, monocytes, dendritic cells) and works as receptor for CD28 and CTLA-4[128]. It is essential in priming of naïve T cells[128]. CD80 interacts also with NK cells, promoting their activation. Polymorphisms in the CD80 gene have been

associated with Vitiligo, Systemic Lupus Erythematosus, and Asthma susceptibility[129,130].

CD28 codes for the receptor of CD80 and CD86 (B7-2), which is a tandem protein working together with CD80 in the immunological synapse[128]. Conventional T cells constitutively express CD28; after interaction with CD80 and CD86, CD28 activates T-cells[131]. The interaction of MHC:antigen complex with TCR without CD28:B7 interaction causes anergy. Polymorphisms in CD28 have been associated with several AIDs, including primary sclerosing cholangitis and inflammatory bowel disease[127,132–135].

CTLA4 codes for another membrane receptor of T cells, which is constitutively expressed in regulatory T cells and upregulated in conventional T cells after activation[136]. The main scope of CTLA4 is to slow down immune response, balancing CD28 activity[128]. Polymorphisms in CTLA4 have been reported in several AIDs other than PBC[127,137–140], while germline haploinsufficiency of CTLA4 determines a severe genetic disorder characterized by lymphoproliferation, autoimmunity, and recurrent infections[141].

Chinese GWAS in PBC identified another possible player in the immunological synapse in PBC[112]: CD58. CD58 encodes a protein also known as lymphocyte function-associated antigen-3 (LFA-3), which is critical in the early phase of immune response. LFA-3 interacts with the T-cell specific CD2 adhesion molecule and starts antigen-independent cell adhesion, expansion of naïve T helper cells, and release of IFN-gamma in memory cells.

As regard other AiDs, polymorphisms in the CD58 gene have been correlated with susceptibility to multiple sclerosis[142].

Overall, these results point to key genetic control of T cell function and signaling in the predisposition to PBC.

B cell signals

Autoreactive B cells are found in the PBC inflamed liver and AMAs are hallmarks of PBC, despite lacking an established pathogenic role. While B cells can suppress Tregs activity[143] and there is small evidence supporting some beneficial effect on reduction of AMA titers after treatment with B-cell depletion strategies[144], the role of B cells in PBC is less investigated and clear than that of other immune cells.

C-X-C Motif Chemokine Receptor 5 (CXCR5) encodes a protein involved in B-cell homeostasis. Mature B-cells express CXCR5 and favours their migration to follicles within spleen and Peyer patches. CXCR5 has been implicated in Autoimmune Thyroid Disease and Systemic Lupus Erythematosus (SLE).

POU2AF1 is a transcription factor regulating the expression of OCT1 and OCT2 genes, essential for the response of B-cells to antigens and formation of germinal centers.

Protein Kinase C Beta (PRKCB) codes for a protein part of the family of Protein kinases C, which have a wide range of molecular targets and control several cellular signaling pathways. Among others, B cell activation and differentiation is under control of PRKCB. PRKCB also controls energy

homeostasis and autophagy via the mitochondria axis, so there is some speculation that it may play a role via this pathway[116].

The inflammatory cascade and innate immunity

Genetic studies in PBC have provided a huge contribution to the notion that the innate arm of the immune system is also involved in its pathogenesis[122]. There is evidence that bacterial byproducts from the gut may initiate and favour progression of the cholangiocyte injury within the liver after their translocation through portal vein and chronic stimulation of TLRs[145]. Despite causal genes are difficult to identify by current methods, some variants point to innate immunity and are briefly described in this paragraph.

Interleukin 1 (IL-1) Receptor Like 2 and IL-1 Receptor Like 1 are part of the IL-1 Receptor family. Mainly involved in innate immunity, they mediate activation of NF-kappa-B, MAPK, as well as other pathways[146]. The IL-1 family shares many features with the TLR family and it is mostly involved in innate immunity[146]. IL-1 is implicated in the activation of NK cells[147], which have been called in action in the pathogenesis of PBC[148].

C-C Motif Chemokine Ligand 20 (CCL20) is a chemokine involved in many inflammatory processes[149]; it acts as chemoattractant for lymphocytes and downregulate proliferation of myeloid progenitors. In the inflamed liver, Th17 cells express high levels of CCR6 which is ligated by CCL20. Th17 cells

migrate and accumulate in bile ducts attracted by CCL20-expressing inflamed bile ducts[150].

Interferon Regulatory Factor 5 (IRF5) and Interferon Regulatory Factor 8 (IRF8) code for transcription factors regulating a broad span of processes, ranging from cell growth, differentiation, and apoptosis[151]. As regards immune system activity, these proteins control virus-mediated activation of interferons. Low-grade, constitutive expression of type 1 Interferon is associated with the development of PBC-like features in mice[152]. Dysregulation of the IRF signaling is thought to contribute to several AIDs; for a detailed discussion see[151].

Overall, these variants suggest some pathways of interest; further downstream fine-mapping is therefore warranted, together with functional investigation.

TNF ligands and receptors

TNF ligands and receptors represent a family of proteins which are involved in most inflammatory processes. TNF- α is a pro-inflammatory cytokine that, together with IFN- γ , is secreted by Th1 lymphocytes and disrupts biliary homeostasis favouring loss of tolerance and cholangiocyte apoptosis[122]. In terms of genetic signals, there are several variants related to the TNF family that have been identified in GWAS in PBC. Interestingly, pre-GWAS era gene-candidate studies had already pointed to the role of TNF α in PBC[153].

Tumor Necrosis Factor Receptor Superfamily Member 1A (TNFRSF1A) codes for a receptor binding to the TNF-a. After binding, the receptor trimerizes and activates, playing a role in cell survival and apoptosis, especially in the context of inflammation[154]. Tumor Necrosis Factor Receptor Superfamily Member 14 (TNFRSF14) encodes a protein that controls signaling pathways related to activation and inhibition of T-cells[154]. Tumor Necrosis Factor (Ligand) Superfamily, Member 15 (TNFSF15) codes for a ligand part of the TNF ligand family. This protein is not expressed in lymphocytes, but is an autocrine factor promoting apoptosis in endothelial cells[154]. The possible role of endothelium and its interaction with other players such as cholangiocytes and immune cells has been investigated in[155], showing that endothelial and liver sinusoidal endothelial cells interact with each other as regards adhesion capability and production of TNF-a by liver infiltrating mononuclear cells.

A different perspective and link is provided by another gene, the TNF Superfamily Member 11 (TNFSF11) gene. Also known as RANKL, it encodes for a ligand for osteoprotegerin, which actively participates in bone homeostasis. RANKL production is induced by IL-17A[156], which is in line with the several pieces of evidence behind the role of Th17 cells in osteoporosis in patients with AIDs[157]. In an established animal model of cholestasis (*Abcb4^{-/-}* mice) increased osteoclastogenesis is amended by IL-17 inactivation[158]. In another autoimmune disease of the biliary system, primary sclerosing cholangitis, the

percentage of Th17 cells in peripheral blood of patients negatively correlates with the degree of bone loss[158].

Further, RANKL is involved in dendritic cell survival and regulation of T-cell immune response[159], providing a link between innate and adaptive immunity. It is constitutively expressed by biliary epithelial cells in healthy liver but its expression is higher in patients with PBC and correlates with disease severity[159].

1.5.2.7 X chromosome and PBC

The role of the X chromosome in PBC remains largely unknown, with no association signals being reported at a genome-wide threshold of significance so far.

Sex influences gene expression, and variation at genes on X chromosome is higher than those on autosomes[160]. X-related gene variants associated with major AIDs are highly expressed in tissues related to immunity and have differential expression by sex[34]. Therefore, the study of the X chromosome may not only pinpoint new gene loci but also help to elucidate the biology of sex differences in human diseases.

To balance allele dosage differences in X-linked genes between male and female individuals, dosage mechanisms are at work in mammals. Random inactivation occurs in X chromosome in female cells, so that female subjects are mosaics having the

paternal or maternal alleles active in different cells of the body[161].

Part of the alleged role of X chromosome in female predominance and autoimmunity is related to the possible role of escape genes, i.e. those X-linked genes escaping complete inactivation[87,162]. Incomplete X chromosome inactivation (XCI) refers to the fact that one of the two copies of X chromosomes may not be silenced and this can happen differently in different cells and tissues. A recent, extensive study of the landscape of X chromosome inactivation across human tissues has estimated that incomplete XCI affects at least 23% of X-chromosomal genes and determines sex differences in gene expression introducing an additional layer of phenotypic diversity[163]. The heterogeneity of partial XCI along the X chromosome reflects its evolutionary history[163,164]. Escape genes are mostly located in the X added region (XAR), which is the region of X chromosome that was added since the divergence between eutherian mammals and marsupials[161].

Interestingly, the expression from the inactivated copy at the level of genes escaping inactivation is on average at 33% of expression of the active copy, and only seldom reaches comparable levels[163]. Nevertheless, the relative overexpression of genes escaping from XCI can still have a biological effect. In an earlier study, Tukiainen and colleagues had adopted an XWAS approach and discovered three novel loci associated with metabolic conditions and height but, most importantly, had found that one of them was close to the ITM2A

gene, which is involved in the early development of cartilage, and whose blood expression is different between male and female subjects[165]. Interestingly, other reports had shown that ITM2A shows escaping from XCI in the majority of women[166], but a more recent study did not confirm this observation[167]. Part of the complexity is related to the lack of understanding whether the XCI status of the tagged SNP is relevant for the causal variant or for the target gene.

Another alleged mechanism for sex bias in autoimmunity is haploinsufficiency, which refers to X chromosome aneuploidy[87,162]. A model for this is Turner's syndrome, where X monosomy is due to a germinal defect and affects all cells in the body: individuals with this condition have higher prevalence of AIDs than the general population[168]. Also women with isochromosome-Xq syndrome, where the short arm of the X chromosome is fully deleted, are more likely to develop autoimmune thyroid and inflammatory bowel diseases[162]. Monosomy can also occur in subsets of cells (e.g. lymphocytes), with different patterns and frequencies[169]. While it is clear that the loss of a functional chromosome generates an imbalance, why haploinsufficiency would cause autoimmunity is less clear. The functional link between haploinsufficiency and autoimmunity might be the loss of the pseudo-autosomal region (PAR) 1 of X chromosome; 26 genes are located on this region and some of them have been linked to immunity[162]. In the recent survey on escape genes by Tukiainen and colleagues[163], the PAR1 region resulted in male bias; authors suggested that this

derangement was due to XCI present also in this region, despite being naturally present in both sexes. We have reported that there are several pieces of evidence showing that escape from XCI is not a rare occurrence in females; so, another possible explanation of the role of haploinsufficiency is that it would cause relative loss of non-PAR, X-linked genes, which permanently escape XCI[162,170]. Further evidence about the link between infection, autoimmunity and XCI is provided by Toll-like receptor 7 (TLR7) biology. TLR7 locus is located on X chromosome and TLR7 is an essential innate component of antiviral defense but, at the same time, is also involved in the pathogenesis of systemic lupus erythematosus. In women and men with Klinefelter syndrome, the TLR7 gene escapes silencing in immune cells, with several consequential pro-autoimmune phenomena, such as increased propensity to immunoglobulin G class switch in biallelic B lymphocytes[171].

1.5.2.8 Genetic variants and risk stratification

Most of the genetic studies in PBC have focused on susceptibility to the disease. Very few genotype-phenotype studies have been performed in PBC to investigate the role of genetic variants to assist risk stratification. Little evidence is present in the setting of predictors of recurrence of PBC after liver transplantation. Specifically, regarding non-HLA variants, the association of

rs62270414 (tagging the IL-12 locus) and use of a specific immunosuppressive drug after liver transplant (Tacrolimus) has been associated with higher rates of recurrent PBC[172]. In contrast to primary sclerosing cholangitis and autoimmune hepatitis, a recent analysis of the United Network for Organ Sharing database in the United States found that neither serotyping nor HLA mismatch have impact on graft survival in PBC[173]. There is still no published evidence about association between specific SNPs and treatment response to UDCA or second-line therapies like obeticholic acid or fibrates.

SCOPE OF THE THESIS

Autoimmune disorders are conditions with complex genetic architecture. Being a prototypical autoimmune disease, PBC has been chosen as the disease model of interest.

The general aim of this PhD project was to explore the genetic determinants of autoimmunity, and PBC specifically, through different standard and novel methodological approaches.

The specific aims of this project were:

- to meta-analyze all available GWAS performed in individuals affected by PBC, to increase the number of risk loci;
- to dissect the chromosome X contribution to the genetic architecture of PBC;
- to generate a polygenic risk score to estimate genetic liability to PBC, based on the most-recent meta-analytic data;
- to evaluate the feasibility and accuracy of a ML-based model for genetic risk prediction of PBC;
- to investigate the evolutionary history of genetic variants associated with autoimmunity, and more specifically their enrichment in archaic variants (Neanderthal and Denisova);
- to study whether archaic information, and more specifically the status of immune-associated archaic variant, improves the characterization of fitness of human mutations.

References

- [1] Fisher RA. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Trans R Soc Edinburgh*. 1918;52:399–433.
- [2] Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet*. 2014;94:559–573.
- [3] Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* [Internet]. 2017;169:1177–1186. Available from: <https://pubmed.ncbi.nlm.nih.gov/28622505>.
- [4] Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*. 2010;20:R208-15.
- [5] Wray NR, Wijmenga C, Sullivan PF, et al. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* [Internet]. 2018;173:1573–1580. Available from: <https://doi.org/10.1016/j.cell.2018.05.051>.
- [6] Gandal MJ, Haney JR, Parikshak NN, et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* [Internet]. 2018;359:693–697. Available from: <https://pubmed.ncbi.nlm.nih.gov/29439242>.
- [7] Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* [Internet].

2009;461:747–753. Available from:
<http://www.ncbi.nlm.nih.gov/pubmed/19812666> %0A
[http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2831613.](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2831613/)

- [8] Zuk O, Hechter E, Sunyaev SR, et al. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci.* 2012;109:1193–1198.
- [9] Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010;42:565–569.
- [10] Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet.* 2001;69:124–137.
- [11] Marouli E, Graff M, Medina-Gomez C, et al. Rare and low-frequency coding variants alter human adult height. *Nature* [Internet]. 2017;542:186–190. Available from: <https://doi.org/10.1038/nature21039>.
- [12] Zeng J, de Vlaming R, Wu Y, et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nat Genet* [Internet]. 2018;50:746–753. Available from: <https://doi.org/10.1038/s41588-018-0101-4>.
- [13] Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nat Rev Methods Prim* [Internet]. 2021;1:59. Available from: <https://doi.org/10.1038/s43586-021-00056-9>.
- [14] Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms

Underlying Complex Diseases. *Front Genet.* 2020;11:1–21.

- [15] Soskic B, Cano-Gamez E, Smyth DJ, et al. Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nat Genet* [Internet]. 2019;51:1486–1493. Available from: <https://doi.org/10.1038/s41588-019-0493-9>.
- [16] Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* [Internet]. 2019;20:467–484. Available from: <https://doi.org/10.1038/s41576-019-0127-1>.
- [17] Gurdasani D, Barroso I, Zeggini E, et al. Genomics of disease risk in globally diverse populations. *Nat Rev Genet* [Internet]. 2019;20:520–535. Available from: <https://doi.org/10.1038/s41576-019-0144-0>.
- [18] Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* [Internet]. 2019; Available from: <http://www.nature.com/articles/s41586-019-1310-4>.
- [19] Janssens ACJW. Validity of polygenic risk scores: are we measuring what we think we are? *Hum Mol Genet.* 2019;28:R143–R150.
- [20] Vilhjálmsson BJ, Yang J, Finucane HK, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet.* 2015;97:576–592.
- [21] Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc.*

2020;3:11–13.

- [22] Khera A V, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* [Internet]. 2018;50:1219–1224. Available from: <https://doi.org/10.1038/s41588-018-0183-z>.
- [23] Wainberg M, Magis AT, Earls JC, et al. Multiomic blood correlates of genetic risk identify presymptomatic disease alterations. *Proc Natl Acad Sci* [Internet]. 2020;117:21813 LP – 21820. Available from: <http://www.pnas.org/content/117/35/21813.abstract>.
- [24] Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev*. 1959;3.3:210–229.
- [25] Rosenblatt F. The perceptron, a perceiving and recognizing automaton (Project Para), Cornell Aeronautical Laboratory. 1957.
- [26] Cortes C, Vapnik V. Support-Vector Networks. *Mach Learn*. 1995;20:273–297.
- [27] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press. 2016.
- [28] Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends Genet* [Internet]. 2018;34:301–312. Available from: <http://dx.doi.org/10.1016/j.tig.2017.12.005>.
- [29] Ho DSW, Schierding W, Wake M, et al. Machine Learning SNP Based Prediction for Precision Medicine. *Front Genet* [Internet]. 2019;10:1–10. Available from:

<https://www.frontiersin.org/article/10.3389/fgene.2019.00267/full>.

- [30] Okser S, Pahikkala T, Airola A, et al. Regularized Machine Learning in the Genetic Prediction of Complex Traits. *PLoS Genet* [Internet]. 2014;10:e1004754. Available from: <https://dx.plos.org/10.1371/journal.pgen.1004754>.
- [31] Wise AL, Gyi L, Manolio TA. EXclusion: Toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet* [Internet]. 2013;92:643–647. Available from: <http://dx.doi.org/10.1016/j.ajhg.2013.03.017>.
- [32] Bianchi I, Lleo A, Gershwin ME, et al. The X chromosome and immune associated genes. *J Autoimmun* [Internet]. 2012;38:J187–J192. Available from: <http://dx.doi.org/10.1016/j.jaut.2011.11.012>.
- [33] Gao F, Chang D, Biddanda A, et al. XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome. *J Hered*. 2015;106:666–671.
- [34] Chang D, Gao F, Slavney A, et al. Accounting for eXentricities: Analysis of the X Chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. *PLoS One*. 2014;9:1–31.
- [35] Saetre G-P, Ravinet M. Evolutionary Genetics. Concepts, Analysis, and Practice. 2019.
- [36] Orr HA. Fitness and its role in evolutionary genetics. *Nat Rev Genet*. 2009;10:531–539.
- [37] Darwin C. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the

Struggle for Life. 1859.

- [38] Hoekstra HE, Hoekstra JM, Berrigan D, et al. Strength and tempo of directional selection in the wild. *Proc Natl Acad Sci* [Internet]. 2001;98:9157 LP – 9160. Available from: <http://www.pnas.org/content/98/16/9157.abstract>.
- [39] Masel J. Genetic drift. *Curr Biol* [Internet]. 2011;21:R837–R838. Available from: <https://doi.org/10.1016/j.cub.2011.08.007>.
- [40] KIMURA M. The neutral theory of molecular evolution: A review of recent evidence. *Japanese J Genet*. 1991;66:367–386.
- [41] Morjan CL, Rieseberg LH. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Mol Ecol*. 2004;13:1341–1356.
- [42] Hublin JJ. The origin of Neanderthals. *Proc Natl Acad Sci U S A* [Internet]. 2009;106:16022–16027. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19805257>.
- [43] Green, R.E., Krause, J., Briggs, A., W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., Hansen, N.F., Durand, E., Y., Malaspinas, A., Jensen, J., D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano HA. A Draft Sequence of the Neandertal Genome. *Science* [Internet]. 2010;328:710–722. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20448178>.
- [44] Sankararaman S, Mallick S, Dannemann M, et al. The genomic landscape of Neanderthal ancestry in present-

- day humans. *Nature* [Internet]. 2014;507:354. Available from: <http://dx.doi.org/10.1038/nature12961>.
- [45] Vernot B, Akey JM. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science* (80-) [Internet]. 2014;343:1017–1021. Available from: <http://classic.sciencemag.org/content/343/6174/1017.full.pdf>.
- [46] Reich D, Patterson N, Kircher M, et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet*. 2011;89:516–528.
- [47] Slon V, Mafessoni F, Vernot B, et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* [Internet]. 2018; Available from: <http://www.nature.com/articles/s41586-018-0455-x>.
- [48] Jacobs GS, Hudjashov G, Saag L, et al. Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell* [Internet]. 2019; Available from: <https://doi.org/10.1016/j.cell.2019.02.035>.
- [49] Rentzsch P, Witten D, Cooper GM, et al. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* [Internet]. 2019;47:D886–D894. Available from: <https://doi.org/10.1093/nar/gky1016>.
- [50] Harris K, Nielsen R. The genetic cost of neanderthal introgression. *Genetics*. 2016;203:881–891.
- [51] Silvert M, Quintana-Murci L, Rotival M. Impact and Evolutionary Determinants of Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation. *Am J*

- Hum Genet [Internet]. 2019/05/30. 2019;104:1241–1250.
Available from:
<https://pubmed.ncbi.nlm.nih.gov/31155285>.
- [52] Telis N, Aguilar R, Harris K. Selection against archaic hominin genetic variation in regulatory regions. *Nat Ecol Evol* [Internet]. 2020; Available from: <https://doi.org/10.1038/s41559-020-01284-0>.
- [53] Simonti CN, Vernot B, Bastarache L, et al. The phenotypic legacy of admixture between modern humans and Neandertals. *Science (80-)* [Internet]. 2016;351:737 LP – 741. Available from: <http://science.sciencemag.org/content/351/6274/737.abstract>.
- [54] Consortium TST 2 D, Williams AL, Jacobs SBR, et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* [Internet]. 2013;506:97. Available from: <http://dx.doi.org/10.1038/nature12828>.
- [55] Zeberg H, Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* [Internet]. 2020;587:610–612. Available from: <https://doi.org/10.1038/s41586-020-2818-3>.
- [56] McArthur E, Rinker DC, Capra JA. Quantifying the contribution of Neanderthal introgression to the heritability of complex traits. *Nat Commun* [Internet]. 2021;12:1–14. Available from: <http://dx.doi.org/10.1038/s41467-021-24582-y>.

- [57] Quach H, Rotival M, Pothlichet J, et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell* [Internet]. 2016;167:643-656.e17. Available from: <http://www.sciencedirect.com/science/article/pii/S009286741631306X>.
- [58] Dannemann M, Andrés AM, Kelso J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *Am J Hum Genet.* 2016;98:22–33.
- [59] Enard D, Petrov DA. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* [Internet]. 2018;175:360-371.e13. Available from: <https://doi.org/10.1016/j.cell.2018.08.034>.
- [60] Brinkworth JF, Barreiro LB. The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. *Curr Opin Immunol* [Internet]. 2014;31:66–78. Available from: <http://www.sciencedirect.com/science/article/pii/S0952791514001149>.
- [61] Okin D, Medzhitov R. Evolution of Inflammatory Diseases. *Curr Biol* [Internet]. 2012;22:R733–R740. Available from: <https://doi.org/10.1016/j.cub.2012.07.029>.
- [62] Mikhalevich N, Becknell B, Caligiuri MA, et al. Responsiveness of naive CD4 T cells to polarizing cytokine determines the ratio of Th1 and Th2 cell differentiation. *J Immunol.* 2006;176:1553–1560.

- [63] Sharpe AH, Pauken KE. The diverse functions of the PD1 inhibitory pathway. *Nat Rev Immunol* [Internet]. 2018;18:153–167. Available from: <https://doi.org/10.1038/nri.2017.108>.
- [64] Watanabe N, Gavrieli M, Sedy JR, et al. BTLA is a lymphocyte inhibitory receptor with similarities to CTLA-4 and PD-1. *Nat Immunol*. 2003;4:670–679.
- [65] Lines JL, Pantazi E, Mak J, et al. VISTA is an immune checkpoint molecule for human T cells. *Cancer Res*. 2014;74:1924–1932.
- [66] Du W, Yang M, Turner A, et al. TIM-3 as a Target for Cancer Immunotherapy and Mechanisms of Action. *Int J Mol Sci*. 2017;18.
- [67] Grosso JF, Kelleher CC, Harris TJ, et al. LAG-3 regulates CD8+ T cell accumulation and effector function in murine self- and tumor-tolerance systems. *J Clin Invest*. 2007;117:3383–3392.
- [68] Rackaityte E, Halkias J. Mechanisms of Fetal T Cell Tolerance and Immune Regulation. *Front Immunol*. 2020;11:588.
- [69] Barker CF, Billingham RE. Immunologically privileged sites. *Adv Immunol*. 1977;25:1–54.
- [70] Theofilopoulos AN, Kono DH, Baccala R. The multiple pathways to autoimmunity. *Nat Immunol*. 2017;18:716–724.
- [71] Akiyama T, Shimo Y, Yanai H, et al. The tumor necrosis factor family receptors RANK and CD40 cooperatively

establish the thymic medullary microenvironment and self-tolerance. *Immunity*. 2008;29:423–437.

- [72] Jacobson EM, Huber AK, Akeno N, et al. A CD40 Kozak sequence polymorphism and susceptibility to antibody-mediated autoimmune conditions: the role of CD40 tissue-specific expression. *Genes Immun*. 2007;8:205–214.
- [73] Hu Y-L, Metz DP, Chung J, et al. B7RP-1 blockade ameliorates autoimmunity through regulation of follicular helper T cells. *J Immunol*. 2009;182:1421–1428.
- [74] Zitti B, Bryceson YT. Natural killer cells in inflammation and autoimmunity. *Cytokine Growth Factor Rev*. 2018;42:37–46.
- [75] Chauhan R, Raina V, Nandi SP. Prevalence of Autoimmune Diseases and Its Challenges in Diagnosis. *Crit Rev Immunol*. 2019;39:189–201.
- [76] Rubtsova K, Marrack P, Rubtsov A V. Sexual dimorphism in autoimmunity. *J Clin Invest*. 2015;125:2187–2193.
- [77] Chun S, Casparino A, Patsopoulos NA, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat Genet*. 2017;49:600–605.
- [78] Farh KK-H, Marson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* [Internet]. 2014;518:337. Available from: <https://doi.org/10.1038/nature13835>.
- [79] Zenewicz LA, Abraham C, Flavell RA, et al. Unraveling the genetics of autoimmunity. *Cell* [Internet]. 2010;140:791–

797. Available from:
<https://pubmed.ncbi.nlm.nih.gov/20303870>.
- [80] Khramtsova EA, Davis LK, Stranger BE. The role of sex in the genomics of human complex traits. *Nat Rev Genet* [Internet]. 2019;20:173–190. Available from: <https://doi.org/10.1038/s41576-018-0083-1>.
- [81] Gerussi A, Carbone M. Primary Biliary Cholangitis [Internet]. *Autoimmune Liver Dis.* 2020. p. 123–141. Available from: <https://doi.org/10.1002/9781119532637.ch7>.
- [82] Rodrigues PM, Perugorria MJ, Santos-Laso A, et al. Primary biliary cholangitis: A tale of epigenetically-induced secretory failure? *J Hepatol* [Internet]. 2018; Available from: <http://www.sciencedirect.com/science/article/pii/S0168827818323626>.
- [83] Carbone M, Sharp SJ, Flack S, et al. The UK-PBC risk scores: Derivation and validation of a scoring system for long-term prediction of end-stage liver disease in primary biliary cholangitis. *Hepatology*. 2016;63:930–950.
- [84] Gerussi A, Lucà M, Cristoferi L, et al. New Therapeutic Targets in Autoimmune Cholangiopathies [Internet]. *Front. Med.* 2020. p. 117. Available from: <https://www.frontiersin.org/article/10.3389/fmed.2020.00117>.
- [85] Gerussi A, Carbone M, Asselta R, et al. Genetics of Autoimmune Liver Diseases BT - Liver Immunology :

Principles and Practice. In: Gershwin ME, M. Vierling J, Tanaka A, et al., editors. Cham: Springer International Publishing; 2020. p. 69–85. Available from: https://doi.org/10.1007/978-3-030-51709-0_5.

- [86] Gerussi A, Carbone M, Corpechot C, et al. The genetic architecture of primary biliary cholangitis. *Eur J Med Genet* [Internet]. 2021;64:104292. Available from: <https://doi.org/10.1016/j.ejmg.2021.104292>.
- [87] Gerussi A, Cristoferi L, Carbone M, et al. The immunobiology of female predominance in primary biliary cholangitis. *J Autoimmun* [Internet]. 2018;95:124–132. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0896841118305936>.
- [88] Manno V, Gerussi A, Carbone M, et al. A National Hospital-Based Study of Hospitalized Patients With Primary Biliary Cholangitis. *Hepatol Commun* [Internet]. 2019;3:1250–1257. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/31497745>.
- [89] European Association for the Study of the Liver. EASL Clinical Practice Guidelines: The diagnosis and management of patients with primary biliary cholangitis. *J Hepatol* [Internet]. 2017;145:167–172. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0168827817301861>.
- [90] Węsierska-Gądek J, Klima A, Ranftler C, et al. Characterization of the antibodies to p62 nucleoporin in

- primary biliary cirrhosis using human recombinant antigen. *J Cell Biochem.* 2008;104:27–37.
- [91] Lleo A, Bowlus CL, Yang GX, et al. Biliary apotopes and anti-mitochondrial antibodies activate innate immune responses in primary biliary cirrhosis. *Hepatology.* 2010;52:987–996.
- [92] Kakuda Y, Harada K, Sawada-Kitamura S, et al. Evaluation of a new histologic staging and grading system for primary biliary cirrhosis in comparison with classical systems. *Hum Pathol* [Internet]. 2013;44:1107–1117. Available from: <http://dx.doi.org/10.1016/j.humpath.2012.09.017>.
- [93] Hirschfield GM, Beuers U, Kupcinskas L, et al. A placebo-controlled randomised trial of budesonide for primary biliary cholangitis following an insufficient response to UDCA. *J Hepatol* [Internet]. 2020; Available from: <https://doi.org/10.1016/j.jhep.2020.09.011>.
- [94] Devarajan P, Chen Z. Autoimmune effector memory T cells: the bad and the good. *Immunol Res* [Internet]. 2013;57:12–22. Available from: <https://pubmed.ncbi.nlm.nih.gov/24203440/>.
- [95] Lleo A, Wang G-Q, Gershwin ME, et al. Primary biliary cholangitis. *Lancet* [Internet]. 2020;396:1915–1926. Available from: [https://doi.org/10.1016/S0140-6736\(20\)31607-X](https://doi.org/10.1016/S0140-6736(20)31607-X).
- [96] Beuers U, Hohenester S, de Buy Wenniger LJM, et al. The biliary HCO₃⁻ umbrella: A unifying hypothesis on

- pathogenetic and therapeutic aspects of fibrosing cholangiopathies. *Hepatology* [Internet]. 2010;52:1489–1496. Available from: <https://doi.org/10.1002/hep.23810>.
- [97] Salas JT, Banales JM, Sarvide S, et al. Ae2a, b-deficient mice develop antimitochondrial antibodies and other features resembling primary biliary cirrhosis. *Gastroenterology* [Internet]. 2008;134. Available from: <https://doi.org/10.1053/j.gastro.2008.02.020>.
- [98] Selmi C, Gershwin ME, Lindor KD, et al. Quality of life and everyday activities in patients with primary biliary cirrhosis. *Hepatology*. 2007;
- [99] Dyson JK, Blain A, Foster Shirley MD, et al. Geo-epidemiology and environmental co-variate mapping of primary biliary cholangitis and primary sclerosing cholangitis. *JHEP Reports* [Internet]. 2021;3. Available from: <https://doi.org/10.1016/j.jhepr.2020.100202>.
- [100] Boonstra K, Beuers U, Ponsioen CY. Epidemiology of primary sclerosing cholangitis and primary biliary cirrhosis: A systematic review. *J Hepatol* [Internet]. 2012;56:1181–1188. Available from: <http://dx.doi.org/10.1016/j.jhep.2011.10.025>.
- [101] Corpechot C, Chrétien Y, Chazouillères O, et al. Demographic, lifestyle, medical and familial factors associated with primary biliary cirrhosis. *J Hepatol* [Internet]. 2010;53:162–169. Available from: <http://www.sciencedirect.com/science/article/pii/S0168827810001820>.

- [102] Jones DEJ, Watt FE, Metcalf JV, et al. Familial primary biliary cirrhosis reassessed: a geographically-based population study. *J Hepatol* [Internet]. 1999;30:402–407. Available from: [https://doi.org/10.1016/S0168-8278\(99\)80097-X](https://doi.org/10.1016/S0168-8278(99)80097-X).
- [103] Selmi C, Mayo MJ, Bach N, et al. Primary biliary cirrhosis in monozygotic and dizygotic twins: Genetics, epigenetics, and environment. *Gastroenterology*. 2004;127:485–492.
- [104] Mells GF, Hirschfield GM. Genetics of Primary Biliary Cirrhosis. *eLS*. 2013;
- [105] Cordell HJ, Han Y, Mells GF, et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat Commun* [Internet]. 2015;6:8019. Available from: <http://www.nature.com/ncomms/2015/150922/ncomms9019/full/ncomms9019.html> <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4580981/> &rendertype=abstract
- [106] Liu X, Invernizzi P, Lu Y, et al. Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. *Nat Genet* [Internet]. 2010;42:658–660. Available from: <http://dx.doi.org/10.1038/ng.627>.
- [107] Hirschfield GM, Liu X, Xu C, et al. Primary Biliary Cirrhosis Associated with HLA, IL12A, and IL12RB2 Variants. *N Engl J Med* [Internet]. 2009;360:2544–2555. Available from: <https://doi.org/10.1056/NEJMoa0810440>.

- [108] Group TIPBCGS, Juran BD, Lammert C, et al. Immunochip analyses identify a novel risk locus for primary biliary cirrhosis at 13q14, multiple independent associations at four established risk loci and epistasis between 1p31 and 7q32 risk variants. *Hum Mol Genet* [Internet]. 2012;21:5209–5221. Available from: <https://doi.org/10.1093/hmg/dds359>.
- [109] Liu JZ, Almarri MA, Gaffney DJ, et al. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat Genet* [Internet]. 2012;44:1137. Available from: <http://dx.doi.org/10.1038/ng.2395>.
- [110] Mells GF, Floyd JAB, Morley KI, et al. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat Genet* [Internet]. 2011;43:329. Available from: <http://dx.doi.org/10.1038/ng.789>.
- [111] Nakamura M, Nishida N, Kawashima M, et al. Genome-wide Association Study Identifies TNFSF15 and POU2AF1 as Susceptibility Loci for Primary Biliary Cirrhosis in the Japanese Population. *Am J Hum Genet* [Internet]. 2012;91:721–728. Available from: <https://doi.org/10.1016/j.ajhg.2012.08.010>.
- [112] Qiu F, Tang R, Zuo X, et al. A genome-wide association study identifies six novel risk loci for primary biliary cholangitis. *Nat Commun*. 2017;14828.
- [113] Dendrou CA, Petersen J, Rossjohn J, et al. HLA variation and disease. *Nat Rev Immunol* [Internet]. 2018;18:325–339. Available from:

[http://dx.doi.org/10.1038/nri.2017.143.](http://dx.doi.org/10.1038/nri.2017.143)

- [114] Donaldson PT, Baragiotta A, Heneghan MA, et al. HLA class II alleles, genotypes, haplotypes, and amino acids in primary biliary cirrhosis: A large-scale study. *Hepatology* [Internet]. 2006;44:667–674. Available from: <https://doi.org/10.1002/hep.21316>.
- [115] Invernizzi P, Selmi C, Poli F, et al. Human leukocyte antigen polymorphisms in italian primary biliary cirrhosis: A multicenter study of 664 patients and 1992 healthy controls. *Hepatology* [Internet]. 2008;48:1906–1912. Available from: <https://doi.org/10.1002/hep.22567>.
- [116] Kawashima M, Hitomi Y, Aiba Y, et al. Genome-wide association studies identify PRKCB as a novel genetic susceptibility locus for primary biliary cholangitis in the Japanese population. *Hum Mol Genet* [Internet]. 2017;26:650–659. Available from: <https://doi.org/10.1093/hmg/ddw406>.
- [117] Wang C, Zheng X, Tang R, et al. Fine mapping of the MHC region identifies major independent variants associated with Han Chinese primary biliary cholangitis. *J Autoimmun*. 2020;107.
- [118] Zhao DT, Liao HY, Zhang X, et al. Human leucocyte antigen alleles and haplotypes and their associations with antinuclear antibodies features in Chinese patients with primary biliary cirrhosis. *Liver Int*. 2014;34:220–226.
- [119] Alvarez F, Berg PA, Bianchi FB, et al. International Autoimmune Hepatitis Group Report: Review of criteria for

diagnosis of autoimmune hepatitis. *J Hepatol*. 1999;31:929–938.

- [120] Cordell HJ, Fryett JJ, Ueno K, et al. An international genome-wide meta-analysis of primary biliary cholangitis: novel risk loci and candidate drugs. *J Hepatol* [Internet]. 2021; Available from: <https://www.sciencedirect.com/science/article/pii/S0168827821003342>.
- [121] Bai J, Wu L, Chen X, et al. Suppressor of Cytokine Signaling-1/STAT1 Regulates Renal Inflammation in Mesangial Proliferative Glomerulonephritis Models. *Front Immunol* [Internet]. 2018;9:1982. Available from: <https://www.frontiersin.org/article/10.3389/fimmu.2018.01982>.
- [122] Carbone M, Milani C, Gerussi A, et al. Primary biliary cholangitis: a multifaceted pathogenesis with potential therapeutic targets. *J Hepatol* [Internet]. 2020;73:965–966. Available from: <https://doi.org/10.1016/j.jhep.2020.05.041>.
- [123] Bossini-Castillo L, Martin J-E, Broen J, et al. A GWAS follow-up study reveals the association of the IL12RB2 gene with systemic sclerosis in Caucasian populations. *Hum Mol Genet* [Internet]. 2011/11/10. 2012;21:926–933. Available from: <https://pubmed.ncbi.nlm.nih.gov/22076442/>.
- [124] Morahan G, Huang D, Ymer SI, et al. Linkage disequilibrium of a type 1 diabetes susceptibility locus with a regulatory IL12B allele. *Nat Genet* [Internet].

- 2001;27:218–221. Available from:
<https://doi.org/10.1038/84872>.
- [125] Hsu W, Zhang W, Tsuneyama K, et al. Differential mechanisms in the pathogenesis of autoimmune cholangitis versus inflammatory bowel disease in interleukin-2R α -/- mice. *Hepatology* [Internet]. 2009;49:133–140. Available from:
<https://doi.org/10.1002/hep.22591>.
- [126] Yoshida K, Yang G-X, Zhang W, et al. Deletion of interleukin-12p40 suppresses autoimmune cholangitis in dominant negative transforming growth factor beta receptor type II mice. *Hepatology* [Internet]. 2009;50:1494—1500. Available from:
<https://europepmc.org/articles/PMC2783300>.
- [127] Márquez A, Kerick M, Zhernakova A, et al. Meta-analysis of Immunochip data of four autoimmune diseases reveals novel single-disease and cross-phenotype associations. *Genome Med.* 2018;10:97.
- [128] Tai Y, Wang Q, Korner H, et al. Molecular Mechanisms of T Cells Activation by Dendritic Cells in Autoimmune Diseases [Internet]. *Front. Pharmacol.* . 2018. p. 642. Available from:
<https://www.frontiersin.org/article/10.3389/fphar.2018.00642>.
- [129] Han Y, Jia Q, Jahani PS, et al. Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat Commun.*

2020;11:1776.

- [130] Zhang Y, Yang J, Zhang J, et al. Genome-wide search followed by replication reveals genetic interaction of CD80 and ALOX5AP associated with systemic lupus erythematosus in Asian populations. *Ann Rheum Dis*. 2016;75:891–898.
- [131] Brigl M, Brenner MB. CD1: Antigen Presentation and T Cell Function. *Annu Rev Immunol* [Internet]. 2004;22:817–890. Available from: <https://doi.org/10.1146/annurev.immunol.22.012703.104608>.
- [132] Beecham AH, Patsopoulos NA, Xifara DK, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet*. 2013;45:1353–1360.
- [133] Ellinghaus D, Jostins L, Spain SL, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet*. 2016;48:510–518.
- [134] Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet*. 2015;47:979–986.
- [135] Trynka G, Hunt KA, Bockett NA, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet*. 2011;43:1193–1201.

- [136] Ferrari SM, Fallahi P, Elia G, et al. Autoimmune Endocrine Dysfunctions Associated with Cancer Immunotherapies. *Int J Mol Sci.* 2019;20.
- [137] Cooper JD, Smyth DJ, Smiles AM, et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet.* 2008;40:1399–1401.
- [138] Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature.* 2014;506:376–381.
- [139] Petukhova L, Duvic M, Hordinsky M, et al. Genome-wide association study in alopecia areata implicates both innate and adaptive immunity. *Nature.* 2010;466:113–117.
- [140] Renton AE, Pliner HA, Provenzano C, et al. A genome-wide association study of myasthenia gravis. *JAMA Neurol.* 2015;72:396–404.
- [141] Kuehn HS, Ouyang W, Lo B, et al. Immune dysregulation in human subjects with heterozygous germline mutations in CTLA4. *Science.* 2014;345:1623–1627.
- [142] Andlauer TFM, Buck D, Antony G, et al. Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Sci Adv.* 2016;2:e1501678.
- [143] Olalekan SA, Cao Y, Hamel KM, et al. B cells expressing IFN- γ suppress Treg-cell differentiation and promote autoimmune experimental arthritis. *Eur J Immunol.* 2015;45:988–998.
- [144] Tsuda M, Moritoki Y, Lian ZX, et al. Biochemical and immunologic effects of rituximab in patients with primary

biliary cirrhosis and an incomplete response to ursodeoxycholic acid. *Hepatology*. 2012;55:512–521.

- [145] Trivedi PJ, Adams DH. Mucosal immunity in liver autoimmunity: A comprehensive review. *J Autoimmun* [Internet]. 2013;46:97–111. Available from: <http://dx.doi.org/10.1016/j.jaut.2013.06.013>.
- [146] Dinarello CA. Overview of the IL-1 family in innate inflammation and acquired immunity. *Immunol Rev* [Internet]. 2018;281:8–27. Available from: <https://pubmed.ncbi.nlm.nih.gov/29247995>.
- [147] Barbier L, Ferhat M, Salamé E, et al. Interleukin-1 family cytokines: Keystones in liver inflammatory diseases. *Front Immunol*. 2019;10:1–19.
- [148] Hydes TJ, Blunt MD, Naftel J, et al. Constitutive Activation of Natural Killer Cells in Primary Biliary Cholangitis. *Front Immunol*. 2019;10:1–12.
- [149] Sokol CL, Luster AD. The chemokine system in innate immunity. *Cold Spring Harb Perspect Biol* [Internet]. 2015;7:a016303. Available from: <https://pubmed.ncbi.nlm.nih.gov/25635046>.
- [150] Ronca V, Mancuso C, Milani C, et al. Immune system and cholangiocytes: A puzzling affair in primary biliary cholangitis. *J Leukoc Biol* [Internet]. 2020;n/a. Available from: <https://doi.org/10.1002/JLB.5MR0320-200R>.
- [151] Matta B, Song S, Li D, et al. Interferon regulatory factor signaling in autoimmune disease. *Cytokine* [Internet]. 2017;98:15–26. Available from:

<http://www.sciencedirect.com/science/article/pii/S104346617300406>.

- [152] Bae HR, Leung PSC, Tsuneyama K, et al. Chronic expression of interferon-gamma leads to murine autoimmune cholangitis with a female predominance. *Hepatology*. 2016;64:1189–1201.
- [153] Poupon R, Ping C, Chrétien Y, et al. Genetic factors of susceptibility and of severity in primary biliary cirrhosis. *J Hepatol* [Internet]. 2008;49:1038–1045. Available from: <https://doi.org/10.1016/j.jhep.2008.07.027>.
- [154] Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinforma* [Internet]. 2016;54:1.30.1-1.30.33. Available from: <https://doi.org/10.1002/cpbi.5>.
- [155] Shimoda S, Harada K, Niilo H, et al. CX3CL1 (fractalkine): a signpost for biliary inflammation in primary biliary cirrhosis. *Hepatology* [Internet]. 2010;51:567–575. Available from: <https://pubmed.ncbi.nlm.nih.gov/19908209>.
- [156] Kotake S, Udagawa N, Takahashi N, et al. IL-17 in synovial fluids from patients with rheumatoid arthritis is a potent stimulator of osteoclastogenesis. *J Clin Invest* [Internet]. 1999;103:1345–1352. Available from: <https://doi.org/10.1172/JCI5703>.
- [157] Singh RP, Hasan S, Sharma S, et al. Th17 cells in inflammation and autoimmunity. *Autoimmun Rev* [Internet].

- 2014;13:1174–1181. Available from:
<https://www.sciencedirect.com/science/article/pii/S1568997214001633>.
- [158] Schmidt T, Schwinge D, Rolvien T, et al. Th17 cell frequency is associated with low bone mass in primary sclerosing cholangitis. *J Hepatol* [Internet]. 2019;70:941–953. Available from:
<https://www.sciencedirect.com/science/article/pii/S0168827819300169>.
- [159] Lleo A, Bian Z, Zhang H, et al. Quantitation of the RANK-RANKL axis in primary biliary cholangitis. *PLoS One*. 2016;11:1–15.
- [160] Kukurba KR, Parsana P, Balliu B, et al. Impact of the X chromosome and sex on regulatory variation. *Genome Res*. 2016;26:768–777.
- [161] Deng X, Berletch JB, Nguyen DK, et al. X chromosome regulation: Diverse patterns in development, tissues and disease. *Nat Rev Genet*. 2014;15:367–378.
- [162] Libert C, Dejager L, Pinheiro I. The X chromosome in immune functions: when a chromosome makes the difference. *Nat Rev Immunol* [Internet]. 2010;10:594. Available from: <http://dx.doi.org/10.1038/nri2815>.
- [163] Tukiainen T, Villani A-C, Yen A, et al. Landscape of X chromosome inactivation across human tissues. *Nature* [Internet]. 2017;550:244. Available from: <http://dx.doi.org/10.1038/nature24265>.
- [164] Vicoso B, Charlesworth B. Evolution on the X

chromosome: unusual patterns and processes. *Nat Rev Genet* [Internet]. 2006;7:645–653. Available from: <https://doi.org/10.1038/nrg1914>.

- [165] Tukiainen T, Pirinen M, Sarin AP, et al. Chromosome X-Wide Association Study Identifies Loci for Fasting Insulin and Height and Evidence for Incomplete Dosage Compensation. *PLoS Genet*. 2014;10.
- [166] Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*. 2005;434:400–404.
- [167] Sidorenko J, Kassam I, Kemper KE, et al. The effect of X-linked dosage compensation on complex trait variation. *Nat Commun* [Internet]. 2019;10:1–11. Available from: <http://dx.doi.org/10.1038/s41467-019-10598-y>.
- [168] Sybert VP, McCauley E. Turner's Syndrome. *N Engl J Med*. 2004;1227–1238.
- [169] Invernizzi P, Miozzo M, Battezzati PM, et al. Frequency of monosomy X in women with primary biliary cirrhosis. *Lancet*. 2004;363:533–535.
- [170] Invernizzi P, Miozzo M, Selmi C, et al. X Chromosome Monosomy: A Common Mechanism for Autoimmune Diseases. *J Immunol* [Internet]. 2005;175:575–578. Available from: <http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.175.1.575>.
- [171] Souyris M, Cenac C, Azar P, et al. TLR7 escapes X chromosome inactivation in immune cells. *Sci Immunol*.

2018;3.

- [172] Sandford RN, Neuberger JM, Mells GF, et al. Calcineurin Inhibitors and the IL12A Locus Influence Risk of Recurrent Primary Biliary Cirrhosis After Liver Transplantation. *Am J Transplant*. 2013;13:1110–1111.
- [173] Patel YA, Henson JB, Wilder JM, et al. The impact of human leukocyte antigen donor and recipient serotyping and matching on liver transplant graft failure in primary sclerosing cholangitis, autoimmune hepatitis, and primary biliary cholangitis. *Clin Transplant* [Internet]. 2018;32:e13388. Available from: <https://doi.org/10.1111/ctr.13388>.

CHAPTER 2

An international genome-wide meta-analysis of primary biliary cholangitis: novel risk loci and candidate drugs

Heather J Cordell¹, James J Fryett¹, Kazuko Ueno², Rebecca Darlay¹, Yoshihiro Aiba ³, Yuki Hitomi⁴, Minae Kawashima⁴, Nao Nishida⁴, Seik-Soon Khor², Olivier Gervais ⁵, Yosuke Kawai², Masao Nagasaki⁵, Katsushi Tokunaga², Ruqi Tang⁶, Yongyong Shi⁷, Zhiqiang Li⁷, Brian D Juran⁸, Elizabeth J Atkinson⁹, Alessio Gerussi¹⁰, Marco Carbone¹⁰, Rosanna Asselta¹¹, Angela Cheung⁸, Mariza de Andrade⁹, Aris Baras¹², Julie Horowitz¹², Manuel A R Ferreira¹², Dylan Sun¹², David E Jones¹³, Steven Flack¹⁴, Ann Spicer¹⁴, Victoria L Mulcahy¹⁴, Jinyoung Byan¹⁵, Younghun Han¹⁵, Richard N Sandford¹⁴, Konstantinos N Lazaridis⁸, Christopher I Amos¹⁵, Gideon M Hirschfield¹⁶, Michael F Seldin¹⁷, Pietro Invernizzi¹⁰, Katherine A Siminovitch¹⁸, Xiong Ma⁶, Minoru Nakamura¹⁹, George F Mells²⁰, for the PBC Consortia: Canadian PBC Consortium; Chinese PBC Consortium; Italian PBC Study Group; Japan-PBC-GWAS Consortium; US PBC Consortium; UK-PBC Consortium

¹Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

²Genome Medical Science Project, National Center for Global Health and Medicine (NCGM), Tokyo, Japan

³Clinical Research Center, National Hospital Organization, Nagasaki Medical Center, Omura, Japan

⁴Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

⁵Human Biosciences Unit for the Top Global Course Center for the Promotion of Interdisciplinary Education and Research, Kyoto University, Kyoto, Japan,

⁶Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan

⁷Division of Gastroenterology and Hepatology, Key Laboratory of Gastroenterology and Hepatology, Ministry of Health, State Key Laboratory for Oncogenes and Related Genes, Renji Hospital, School of Medicine, Shanghai JiaoTong University, Shanghai Institute of Digestive Disease, Shanghai, China

⁸Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Center for Brain Science, Shanghai Jiao Tong University, Shanghai, China

⁹Affiliated Hospital of Qingdao University and Biomedical Sciences Institute of Qingdao University (Qingdao Branch of SJTU Bio-X Institutes), Qingdao University, Qingdao, China

¹⁰Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, Minnesota, United States of America

¹¹Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, United States of America

¹²Division of Gastroenterology and Center for Autoimmune Liver Diseases, Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy

¹³European Reference Network on Hepatological Diseases (ERN RARE-LIVER), San Gerardo Hospital, Monza, Italy

¹⁴Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy

¹⁵Humanitas Clinical and Research Center, IRCCS, Rozzano, Milan, Italy

¹⁶Regeneron Genetics Center, Tarrytown, New York, United States of America

¹⁷Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

¹⁸Academic Department of Medical Genetics, University of Cambridge, Cambridge, United Kingdom

¹⁹Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, Texas, United States of America

²⁰Toronto Centre for Liver Disease, Division of Gastroenterology and Hepatology, University of Toronto, Toronto, Ontario, Canada

²¹University of California, Davis, California, United States of America

²²Departments of Medicine, Immunology and Medical Sciences,
University of Toronto, Toronto, Ontario, Canada

²³Mount Sinai Hospital, Lunenfeld-Tanenbaum Research
Institute and Toronto General Research Institute, Toronto,
Ontario, Canada

Published in *J Hepatol.* 2021.
doi:<https://doi.org/10.1016/j.jhep.2021.04.055>

PhD Candidate contribution: interpretation of the results and
critical feedback

Abstract

Background: Primary biliary cholangitis (PBC) is a chronic liver disease in which autoimmune destruction of the small intrahepatic bile ducts eventually leads to cirrhosis. Many patients have inadequate response to licensed medications, motivating the search for novel therapies.

Methods: Previous genome-wide association studies (GWAS) and meta-analyses (GWMA) of PBC have identified numerous risk loci for the condition, providing insight into its aetiology. We undertook the largest GWMA of PBC to date, combining new and existing genotype data for 10,516 cases and 20,772 controls from five European and two East Asian cohorts, aiming to identify additional risk loci, and prioritise candidate genes for *in silico* drug efficacy screening to identify agents potentially suitable for re-purposing to this condition.

Results: We identified 56 genome-wide significant loci (20 novel) including 46 in European, 13 in Asian, and 41 in combined cohorts; and a 57th genome-wide significant locus (also novel) in conditional analysis of the European cohorts. Candidate genes at newly identified loci include *BLIMP1*, *CCR6*, *FCRL3*, *INAVA*, *IRF7*, *CD226*, and *IL12RB1*, each having key roles in immunity. Pathway analysis reiterated the likely importance of pattern recognition receptor and TNF signalling; Jak-STAT signalling; and differentiation of T_H1 and T_H17 cells in the pathogenesis of the disease. Drug efficacy screening identified several

medications predicted to be therapeutic in PBC, some well-established in the treatment of other autoimmune disorders.

Conclusion: This study provides a hierarchy of agents that could be trialled in PBC and emphasises the value of genetic and genomic approaches to drug discovery in complex disorders.

Introduction

Primary biliary cholangitis (PBC) is a chronic liver disease in which autoimmune injury to the small intra-hepatic bile ducts eventually leads to cirrhosis. Only two medications, ursodeoxycholic acid (UDCA) and obeticholic acid (OCA), are licensed for treatment of PBC. Many patients have inadequate response to both agents leaving them at risk of progressive liver disease. Notwithstanding recent advances, novel therapies are needed for this condition.

Delineating the genetic architecture of PBC can provide insight into its aetiology – and more specifically, identify potential drug targets. Therefore, over the past decade, our respective groups have undertaken genome-wide association studies (GWAS) of PBC in Canadian-US¹, Italian², British³, Japanese⁴, and Chinese⁵ cohorts; and in 2015, we undertook a genome-wide meta-analysis (GWMA) of the Canadian-US, Italian, and British discovery panels⁶. These studies have identified genome-wide significant associations at the human leukocyte antigen (HLA) locus and 42 non-HLA loci.

Our GWMA in 2015 did not include the Japanese or Chinese discovery panels. Furthermore, since 2015, our respective groups have undertaken genome-wide genotyping of substantially expanded Canadian, Italian, UK, and US cohorts. Therefore, we present an updated GWMA of PBC that includes these expanded cohorts, as well as the Japanese and Chinese discovery panels. In this study, we aimed to: (1) capitalise on the

increased sample size to discover additional risk loci for PBC; (2) formally explore population-specific genetic heterogeneity at known and newly identified risk loci; (3) integrate GWMA statistics with publicly available gene expression, epigenetic, and proteomic datasets to pinpoint causal variants and prioritise candidate genes; and (4) use these candidate genes for *in silico* drug efficacy screening to identify agents that might be suitable for re-purposing to PBC.

Methods

Participants and genotyping are summarised in **Table 1** and detailed in the **Supplementary Text***. Written informed consent was obtained from each participant. The research conformed to the ethical guidelines of the 1975 Declaration of Helsinki.

Quality control

For the European and Japanese panels, QC checks were performed at Newcastle University, UK, using the software package PLINK⁷. Specific QC thresholds to determine outliers were based on visual inspection and varied by panel. For the European panels, we first removed variants with minor allele frequency (MAF) <0.01; genotype call rate <97% (<95% for the ‘old’ Italian, WTCCC3, and ‘new’ US panels); or significant deviation from Hardy Weinberg Equilibrium (HWE) ($p < 10^{-6}$). We then removed samples with rates of missing data >2% (>4% for the new US panel); whole-genome heterozygosity >3.25 standard deviations from the mean; apparent gender discrepancies (based on X-chromosomal heterozygosity >0.2 for men and <0.2 for women); estimated proportion of identity-by-descent sharing with another sample >0.1 (based on subsets of between 38,000

*The supplementary material, including supplementary figures and tables, listed in this chapter, can be retrieved at [https://www.journal-of-hepatology.eu/article/S0168-8278\(21\)00334-2/fulltext](https://www.journal-of-hepatology.eu/article/S0168-8278(21)00334-2/fulltext)

and 97,000 variants pruned for linkage disequilibrium [LD]); or that did not cluster with the CEU HapMap2 population (based on visual inspection of the first 2 principal components). For the Japanese panel, we used the dataset described in Kawashima et al⁴, except for the additional removal of 4 cases and 10 controls with apparent gender discrepancies.

All samples recruited in China were processed and analysed on Chinese servers to comply with the Regulation of the People's Republic of China on the Administration of Human Genetic Resources. Thus, for the Chinese panel, QC checks were undertaken on a local server in Shanghai, China. Variants were removed with MAF <0.5%, genotype call rate <95%, or deviation from HWE in controls $p < 1 \times 10^{-6}$. Samples were removed with rates of missing data $\geq 5\%$ or pairwise identity-by-state, PI_HAT > 0.25 . Population outliers were identified for exclusion using principal component analysis.

Genome-wide imputation and post-imputation quality control

For the European and Japanese panels, we used the autosomal variants and samples passing QC to carry out genome-wide imputation within each of these panels using the Michigan Imputation Server with Eagle2 phasing⁸, informed by the 1000 Genomes Phase 3 reference panel. Following imputation, we discarded variants with imputation $R^2 < 0.5$; non-unique alleles at the same position; or imputation call rate <90% (based on assigning genotypes according to the most likely genotype call

and setting genotypes to missing if the most likely genotype call had posterior probability <0.9). We also used the resulting common set of imputed variants to check for sample duplicates/relationships across the European panels (based on estimated identity-by-descent sharing using 25,873 variants pruned for LD) and removed 1 person from each of the 137 identified relative pairs.

For the Chinese panel, QC checks were undertaken on a local server in Shanghai, China, using SHAPEIT⁹ and IMPUTE2¹⁰, and the 1000 Genomes Phase 3 reference panel. Following imputation , we discarded variants with call rates <95% (having set genotypes to missing if the most likely genotype call had posterior probability <0.9), MAF <0.01, or HWE p <1×10⁻⁶ in controls. The resulting imputation summary statistics (log odds ratios [lnORs], standard errors, and p values) were submitted were submitted without individual-level data to Newcastle, UK, for meta-analysis with the other panels.

Statistical analysis of European and Japanese cohorts

Within each panel, we carried out association analysis of the genome-wide imputed data using logistic regression of disease phenotype on SNP genotype (coded 0,1,2) in PLINK⁷, with the first 10 principal components (from a pruned set of SNPs with the HLA region removed) included as covariates to help correct for population stratification. (The rationale for removing the HLA region was that inclusion of SNPs in this region would risk generating components that explain variation primarily caused by

strong HLA-disease association, rather than population stratification.) For all but the new Canadian-UK panel, the resulting genomic control (GC) inflation factor λ was modest (<1.026); therefore, we carried out GC correction within each panel by multiplying the standard error (SE) of the estimated InOR for each SNP by . For the new Canadian-UK panel, λ was somewhat inflated at 1.091; therefore, we re-analysed the new Canadian-UK data using a logistic mixed model score test (including the first 10 principal components as covariates) as implemented in the GMMAT package¹¹, resulting in a slightly deflated λ of 0.971. The SE of the estimated InOR for each SNP from PLINK was then (conservatively) adjusted to match that implied by the GMMAT test statistic. Specifically, we multiplied the PLINK-derived SE for each SNP by a SNP-specific factor g , where g was chosen so that the resulting c^2 test statistic ($\text{InOR}/g \text{ SE})^2$ for that SNP had a P value equal to the P value from GMMAT. Genomic control correction was also carried out for the Chinese summary statistics ($\lambda = 1.050$) by multiplying the SE of the estimated InOR for each SNP by .

Meta-analysis of European, Asian, and combined cohorts

We used the software package META¹² to perform fixed effect meta-analysis of the resulting InORs and adjusted SEs from (1) the five European panels; (2) the two Asian panels; and (3) all seven panels, in each case restricting the analysis to variants that (following post-imputation QC) appeared within all contributing panels. Within each meta-analysed set (European,

Asian, and combined), a further GC correction was carried out (to adjust for the inflation factors of $\lambda = 1.041$, $\lambda = 1.033$, and $\lambda = 1.080$, seen within the European, Asian, and combined cohorts, respectively) to produce the final set of genome-wide results reported below. Specifically, as for the individual panels above, the SE of the final InOR for each SNP was multiplied by , and the test statistic and *P*-value re-calculated accordingly. This use of “double” GC might be considered overly conservative, given that part of the observed inflation could be due to polygenicity. We explored this using LD score regression (LDSR)¹³ to compare our original results with those obtained using no GC (or GMMAT-derived) correction at all. We also compared our results from all panels combined with those obtained using trans-ethnic meta-regression analysis as implemented in the software package MR-MEGA¹⁴. See **Supplementary Text** for further details.

Prioritisation of candidate causal variants and candidate genes

We used the FINEMAP¹⁵ package and Conditional and Joint Analysis (COJO)¹⁶ implemented within GCTA¹⁷ to refine and look for independent associations within genome-wide significant risk loci. We used FINEMAP to construct ‘credible sets’ of variants most likely to be causal in PBC; and the ENSEMBL Variant Effect Predictor¹⁸, FUMA GWAS¹⁹ platform, and reference panels from the Avon Longitudinal Study of Parents and Children (ALSPAC)²⁰ and the INTERVAL study²¹ for mapping and functional

annotation of the first set of ‘credible causal variants’ at each risk locus.

Adapting the approach of Barbeira *et al.* (2018)²², we used the MetaXcan package; our European GWMA summary statistics; and reference panels from the Genotype-Tissue Expression (GTEx)²³ project, ALSPAC, and the INTERVAL study to derive genome-wide genetic prediction models of gene expression, DNA methylation, and serum protein levels in cases and controls. We then used these models to correlate predicted gene expression, DNA methylation, and serum protein levels with disease status in transcriptome-wide, methylome-wide, and serum proteome-wide association studies (TWAS, MWAS, and PWAS, respectively).

We used the moloc package²⁴ to look for co-localisation of association signals from our GWMA of the European panels with those derived from mapping of methylation, expression, and protein-quantitative trait loci (mQTLs, eQTLs, pQTLs) in ALSPAC, the GTEx project, and the INTERVAL study, respectively. Finally, we used the DEPICT package²⁵ to systematically prioritise the most likely causal gene at risk loci based on reported gene function.

Enrichment analysis

We used the STRING Database²⁶ to look for enrichment of protein-protein interactions and functional annotations amongst candidate genes; and the DAVID bioinformatics resource²⁷ to

look for enrichment of KEGG pathways by genes with minimum $P_{\text{GWMA}} < 0.01$.

Network-based *in silico* drug efficacy screening

We employed the approach of Guney *et al.* (2016)²⁸ in which known drug targets and candidate genes for a disease are used to estimate a drug-disease proximity measure, z , that quantifies the closeness (or proximity) of the drug and disease gene networks, respectively, correcting for the known biases of the interactome. For our analysis, we used the drug targets listed in DrugBank (accessed June 2019) and candidate genes for PBC prioritised as above. See **Supplementary Text** for further details.

Results

GWMA identifies 23 additional genome-wide significant risk loci for PBC

Following QC, the European panels consisted of 5,186,747 variants across 8,021 cases and 16,489 controls; Asian panels, 5,347,815 variants across 2,495 cases and 4,283 controls; and all panels combined, 2,817,608 variants across 10,516 cases and 20,772 controls (**Table 1**). The substantial reduction in the number of variants in the combined versus the European or Asian panels resulted from lack of overlap between variants passing post-imputation QC, explained by use of different genotyping platforms across cohorts, and different LD patterns in Europeans compared to Asians. Genome-wide meta-analysis of the European panels identified 46 loci at genome-wide significance ($P < 5 \times 10^{-8}$); Asian panels, 13 loci at genome-wide significance; and all panels combined, 41 loci at genome-wide significance (**Suppl. Figure 1**). Altogether, we identified 56 genome-wide significant risk loci in one or other meta-analysis (**Suppl. Table 1, Suppl. Figure 2**). Using COJO, we identified an additional risk locus at 19p13.11 that was genome-wide significant in conditional analysis of the European panels ($P = 4.66 \times 10^{-8}$), having narrowly missed this threshold in the main, unconditional analysis ($P = 6.55 \times 10^{-8}$) (**Suppl. Figure 2.57**). Thus, the total number of genome-wide significant risk loci identified in the current study was 57. Of these, 21 have not been

identified in previous studies; and two, 1q23.1 and 11q24.3, have previously been identified at suggestive rather than genome-wide significance (**Tables 2A&B**)^{4,29}.

At six newly identified or newly confirmed risk loci, we considered evidence of association to be conclusive because: (1) an unequivocal association signal was evident in both the European and Asian panels; and (2) where the lead variant at the locus was different in the European versus the Asian panels, permutation testing confirmed the significance of a signal in the validating dataset, located in proximity to the primary signal in the index dataset ($P_{\text{permutation}} < 0.00217$, corresponding to $P < 0.05$ Bonferroni-corrected for 23 tests; see **Supplementary Text** for further details) (**Table 2A, Suppl. Table 1, Suppl. Figure 2**).

At 17 newly identified or newly confirmed risk loci, we considered evidence of association to be strong but not conclusive because unequivocal association was evident in the European but not the Asian panels (**Table 2B, Suppl. Table 1, Suppl. Figure 2**). We note, however, that most of these 17 loci achieved levels of significance suggestive for validation, including two loci with suggestive permutation P-values (4q24 [2], $P_{\text{permutation}} = 0.0040$, and 5q21.1, $P_{\text{permutation}} = 0.0032$).

We confirmed genome-wide significant association at 34 of 43 previously identified risk loci for PBC – but not at nine previously identified risk loci. Seven of these nine loci nevertheless showed a convincing association signal, albeit at $P > 5 \times 10^{-8}$ (**Suppl. Table 2, Suppl. Figure 3**). We found no evidence of association at the 15q25.1 locus (harbouring *IL16*) that was discovered and

validated in the Chinese GWAS by Qiu *et al.* (2017)⁵; this is explained by the absence of a signal in the Japanese and European panels. Coverage of the 19p13.2 locus was too sparse to test association.

Using FINEMAP and COJO, we found that at most risk loci, the association signal was best explained by a single variant – but at 16 risk loci, it was best explained by two or more independent variants (**Suppl. Table 3**). Notable examples include the 2q32.2 locus harbouring *STAT4*, with three independent variants; 3q25.33 (*IL12A*, three variants); 7q32.1 (*IRF5*, two variants); and 16p13.13 (*CLEC16A*, two variants) – all consistent with previous studies showing two or more independent associations at each of these loci.

We compared our original results to those obtained using no GC (or GMMAT-derived) correction. As expected, with no correction, all loci previously identified as genome-wide significant reached slightly higher levels of significance, while a few loci that did not reach genome-wide significance in our original analysis, now (just) did so (**Suppl. Figure 4, Suppl. Table 4**). We also compared our original results from all panels combined with those obtained using trans-ethnic meta-regression analysis as implemented in MR-MEGA. Results from MR-MEGA were highly concordant with those from our original analysis (**Suppl. Figure 5**), with an independent, genome-wide significant association signal identified at 7q32.1 that exhibited significant heterogeneity in the direction of effects between the Asian and European cohorts (**Suppl. Table 5, Suppl. Figure 6**).

Primary biliary cholangitis shows genetic correlation with other autoimmune conditions

Recognising that most risk loci for PBC are also risk loci for other autoimmune conditions (**Suppl. Table 6**), we used LDSR implemented via LD Hub²⁹ to formally evaluate genetic correlation between PBC (using summary statistics from our European panels) and complex traits with GWAS summary statistics in the LD Hub database. We found significant genetic correlation between PBC and other immune-mediated inflammatory disorders, including systemic lupus erythematosus (SLE, $rg = 0.54$, $P = 2.87 \times 10^{-14}$), rheumatoid arthritis (RA, $rg = 0.26$, $P = 3.77 \times 10^{-5}$), and inflammatory bowel disease (IBD, $rg = 0.23$, $P = 6.97 \times 10^{-5}$) (**Suppl. Table 7**). We were unable to test genetic correlation of PBC with autoimmune thyroid disease, Sjögren syndrome, or systemic sclerosis because GWAS summary statistics for these conditions were not available in LD Hub at the time of interrogation (19.09.2019).

The genetic architecture of PBC is broadly shared across European and Asian populations

To formally evaluate consistency between European and Asian signals, we applied permutation testing where warranted and standard meta-analysis measures of heterogeneity to the lead variants at each of the 56 genome-wide significant risk loci identified or confirmed in the main, unconditional analyses (**Suppl. Table 1**). We found reasonable concordance between risk loci operating in European and Asian populations,

considering (1) the much smaller sample size of the Asian panels; and (2) the interrogation of different variants in the European compared to the Asian panels, for reasons given above. (For a detailed commentary of each risk locus, please see **Suppl. Figure 2**.) With few exceptions, we also found concordance between the InORs seen in the combined Asian and combined European cohorts (**Suppl. Figure 7**).

To investigate overall concordance in the genetic basis of PBC between European and Asian populations, we estimated the proportion of trait variance explained (on the liability scale) in the Japanese cohort (for which individual-level genotype data were available) by sets of variants chosen according to their *P*-values in the European GWMA (see **Supplementary Text**). Regardless of the *P*-value threshold and the assumed trait prevalence, variants showing some level of association in the European GWMA explained more of the trait variance than an equivalent number of randomly chosen variants – in most instances, significantly more – supporting the conclusion that loci influencing risk of PBC in European populations also influence its risk in Asian populations (**Suppl. Table 8**).

Thus, while equivalently powered cohorts, accurately genotyped at the same set of genetic variants, would be required to fully address the question of population-specific genetic heterogeneity, considered as a whole, our results provide preliminary evidence that the genetic architecture of PBC is broadly shared across European and Asian populations.

Co-localisation and DEPICT enable prioritisation of candidate genes

In functional annotation, we found that credible causal variants include missense variants in 21 genes at 14 risk loci; splice variants in eight genes at five risk loci; and stop variants in two genes at two risk loci (**Suppl. Table 9**). Few of these variants are predicted to be deleterious. Conversely, credible causal variants at all genome-wide significant risk loci map to chromatin interacting regions (CIRs), mQTLs, eQTLs, or pQTLs (**Suppl. Tables 10 – 12**); and in the MWAS, TWAS, and PWAS, we predicted differential methylation, transcription, or translation of genes at and beyond GWMA-significant loci (**Suppl. Tables 13 – 15, Suppl. Figure 8**). These observations suggest that the genetic architecture of PBC might confer susceptibility to disease mainly by influencing the regulation of expression of causal genes. Therefore, we sought co-localisation of GWMA with mQTL, eQTL, or pQTL association signals, aiming to pinpoint causal variants and genes across the whole genome. Using moloc, we identified 251 co-localisation models with PPA ≥ 0.80 , implicating variants and genes at 60 loci (**Suppl. Table 16, Suppl. Figure 8C**). Of these 60 loci, 28 correspond to genome-wide significant risk loci, where co-localisation models implicate candidate genes such as *IL12RB2* (1p31.3), *FCRL3* (1q23.1),

and *INAVA* (1q32.1), amongst others. Association at the other 32 loci did not reach genome-wide significance in the GWMA; co-localisation models nevertheless implicate highly plausible candidate genes at some of these loci, such as *CCL21* (9p13.3) and *IL2RB* (22q12.3).

As expected, we found that candidate genes implicated by co-localization were broadly concordant with those implicated by functional annotation of credible causal variants, and by the MWAS, TWAS, and PWAS. As in previous studies, we also observed that candidate genes at disparate risk loci are evidently related in function, e.g., *IL12A* (3q25.33), *IL12B* (5q33.3), *IL12RB1* (19p13.11), and *IL12RB2* (1p31.3). Therefore, we used DEPICT²⁵ to systematically prioritise candidate genes at genome-wide significant risk loci based on their reported functions. In this way, we identified 82 candidate genes with FDR <5% across 48 loci (**Suppl. Table 17**). As expected, genes prioritised by DEPICT have considerable overlap with those prioritised by the other approaches (**Suppl. Table 18**).

We used the information garnered above to finalise a list of top candidate genes at genome-wide significant risk loci (**Suppl. Table 18**). Using STRING²⁶, we found these candidate genes to be highly enriched for protein-protein interactions ($P < 1.0 \times 10^{-16}$, **Suppl. Figure 9**); and enriched at FDR <5% for the KEGG pathways, T_H1 and T_H2 cell differentiation, T_H17 cell differentiation, and toll-like receptor (TLR), RIG-I-like receptor (RLR), TNF, NF κ B, and Jak-STAT signalling pathways, amongst others. For comparison, we undertook enrichment analysis of

1388 genes with minimum $P_{GWMA} < 0.01$ using DAVID²⁷, which identified enrichment at FDR <5% of the KEGG pathways, Antigen processing and presentation, Fc γ R-mediated phagocytosis, NK cell-mediated cytotoxicity, and T cell receptor, B cell receptor, PI3K-AKT, Fc ϵ RI, Jak-STAT, NF κ B, and MAPK signalling pathways, amongst others (**Suppl. Table 19**).

In silico drug efficacy screening identifies agents potentially suitable for re-purposing to PBC

In the approach of Guney *et al.* (2016)²⁸, the more negative the value of z , the closer the proximity of the drug and disease gene networks. A cut-off of $z \leq -0.15$ is taken to show that the drug is proximal to the disease and might thus exert pharmacological effects on it. In our analysis, we identified many agents with $z \leq -0.15$, which are therefore predicted to exert pharmacological effects on PBC (**Table 3**, **Suppl. Table 20**). Top-ranking drugs predicted to potentially ameliorate the disease included several immunomodulators, such as Ustekinumab, an anti-IL-12/23 monoclonal antibody used for psoriasis and Crohn's disease ($z = -4.72$); Abatacept, a CTLA-4 fusion protein used for RA, juvenile idiopathic arthritis (JIA), and psoriatic arthritis ($z = -4.60$); and Belatacept, a CTLA-4 fusion protein used in organ transplantation ($z = -4.49$). Of interest, other top-ranking agents include the retinoids, Etretinate and its metabolite Acitretin, both used for treatment of psoriasis ($z = -3.88$ and $z = -4.55$,

respectively). Unsurprisingly, top-ranking drugs predicted to potentially exacerbate PBC included the pharmacological interferons, such as Interferon alfa-n1 and Interferon alfa-n3 ($z = -3.74$ and $z = -3.65$, respectively). Amongst recognised treatments for PBC, bezafibrate and fenofibrate scored $z = -0.92$ and $z = -0.66$, respectively, and are thus predicted to exert pharmacological effects on PBC. Conversely, UDCA and OCA scored $z = +0.10$ and $z = +1.11$, respectively, which means they are not predicted by this approach to treat the genetically determined component of disease in PBC.

Discussion

We report the largest GWMA of PBC undertaken to date, with a sample size four times greater than that of our previous study. In this better-powered study, we identified 21 additional genome-wide significant risk loci; showed that the genetic architecture of PBC is broadly shared across European and Asian populations; prioritised candidate genes at known and newly identified genome-wide significant risk loci; and used those candidate genes to identify medications predicted to treat the genetically determined component of disease in PBC, which might therefore be suitable for re-purposing to this condition.

Candidate genes at newly identified or newly confirmed risk loci provide additional insights into the pathogenesis of PBC (**Figure 1**). Thus, *INAVA* (1q32.1) amplifies pattern recognition receptor (PRR) signalling; *DNMT3A* (2p23.3), *ZC3HAV1* (7q34), and *TRIM14* (9q22.33) are each involved in RLR signalling; *TET2* (4q24) represses transcription of IL-6; and *PVT1* (8q24.21) regulates inflammation via NF κ B and MAPK pathways. Chemokine receptor 6 (CCR6, 6q27) interacts with CCL20 in the chemotaxis of dendritic cells and lymphocytes to inflamed epithelia; *ST8SIA4* (5q21.1) is required for the interaction of CCR7 with CCL21 in the trafficking of immune cells to secondary lymphatic organs; and *CD226* (18q22.2) participates in lymphocyte and NK cell adhesion and signalling. Fc receptor-like protein 3 (*FCRL3*, 1q23.1), *ID2* (2p25.1), *TET2* (4q24), *RARB* (3p24.2), *NDFIP1* (5q31.3), *ITGB8* (7p21.1), and *CD226*

(18q22.2) are each involved in the differentiation of T_H1 or T_H17 cells, or Tregs. Not unexpectedly, enrichment analysis of candidate genes reiterated the importance of PRR, TNF and NF_kB signalling, and T_H1 and T_H17 cell differentiation in this disease. These findings are consistent with functional data emphasising the importance of innate immune cell hypersensitivity, chemokine signalling and immune cell trafficking, and T_H1/T_H17 cell polarisation in the pathogenesis of PBC, as summarised by Gulamhusein and Hirschfield (2020)²⁹ in their recent review.

There is considerable current interest in the ‘Druggable Genome’, *i.e.*, the use of genome-wide approaches to find targets for drug discovery (for example, see the Open Targets initiative at <https://www.opentargets.org/>). In the current study, having prioritised candidate genes, we used network-based *in silico* drug efficacy screening to identify agents potentially suitable for re-purposing to PBC. Given our other findings – including genetic correlation of PBC with SLE, RA and IBD – it is expected that the top-ranking medications should include immunomodulators already approved for the treatment of RA, JIA, IBD, MS, or psoriasis.

The evidence to support re-purposing of those immunomodulators to PBC is circumstantial yet convincing – but circumspection is required. For example, in the current study, LDSR demonstrated genetic correlation with IBD; enrichment analysis showed association with ‘T_H1 and T_H2 cell differentiation’; and drug efficacy screening suggested that

Ustekinumab, an anti-IL-12/23 monoclonal antibody used for treatment of Crohn's disease, might exert pharmacological effects on PBC. Therefore, it is notable that Ustekinumab showed minimal effect in PBC in the clinical trial of Hirschfield *et al.* (2016)³⁰. Similarly, drug efficacy screening suggested that Abatacept, a CTLA-4 fusion protein used for treatment of RA, might be effective for treatment of PBC – but Abatacept showed no effect in PBC in the clinical trial of Bowlus *et al.* (2019)³¹. A potential explanation for these discrepant observations, also expounded by Bowlus *et al.*³¹, is that the evaluation of immunomodulators in PBC might require a change in clinical trial concept and design. Thus, immunomodulators might require immunological rather than cholestatic endpoints; might be more effective in early disease, before the cholestatic liver injury predominates; and might require combined treatment of both the autoimmune and cholestatic injuries. Re-design of clinical trials in PBC might be contentious; the use of genomic data to prioritise potential agents for PBC is not: new treatments for PBC are needed – and the druggable genome provides a framework to find them.

It is notable that in drug efficacy screening, UDCA – well-established as first-line treatment for PBC – was not predicted to be therapeutic in this condition. One possibility is that UDCA serves primarily to treat a cholestatic liver injury that is critical to disease progression but orthogonal to the genetically determined, autoimmune processes that confer risk of disease. Conversely, OCA (a potent FXR agonist) and the fibrates,

bezafibrate and fenofibrate (PPAR-a/d/c and PPAR-a agonists, respectively), are expected to have immune-modulatory as well as anti-cholestatic effects^{32,33}.

We acknowledge two major limitations of the study. First, the absence of an independent validation cohort meant we were unable to confirm several newly identified risk loci. Other strategies, such as cross-phenotype meta-analysis, may be required for external validation of these loci. And second, the use of different genotyping platforms across cohorts meant that at many risk loci, the lead variant in the European panels was not represented in the Asian panels, or *vice versa*. This, together with marked disparity in the sample size of the European versus the Asian panels, meant that we were unable to fully address the question of population-specific genetic heterogeneity.

Conclusions

In conclusion, our large, trans-ethnic GWMA of PBC has identified additional risk loci; found little evidence for population-specific genetic heterogeneity; and, through functional annotation of credible causal variants and multi-omic analysis, allowed us to prioritise candidate genes, and thereby prioritise drugs potentially suitable for re-purposing to PBC. This study emphasises the value of genomic approaches to provide biological insight and guide the development of novel therapies.

Table Legends

Table 1. Discovery panels included in the current study.

Table 2: Results for the lead variant at newly identified or newly confirmed risk loci in genome-wide meta-analysis of the European, Asian, or combined panels. In **Table 2A**, evidence of association was taken to be conclusive because: (1) an unequivocal association signal at the same locus was observed in both the European and the Asian panels; and (2) where the lead variant at the locus was different in the European versus the Asian panels, permutation testing confirmed the significance of the signal in the validating dataset (see **Main** and **Supplementary Text**, and **Suppl. Table 1**). In **Table 2B**, evidence of association was taken to be strong but not conclusive because association in one dataset was not supported by an unequivocal association signal in the other. Gene, candidate gene at the risk locus (which is not necessarily the mapped gene); Chr, chromosome; BP, base pair position; A1, tested allele; A2, alternative allele; $P_{perm.}$, permutation P-value; OR, odds ratio; CI, confidence interval.

Table 3. Selected, top-ranking agents from *in silico* drug efficacy screening.

Figure Legends

Figure 1. Candidate genes for PBC (**red bold** or **red filled**) emphasise the potential importance of T cell activation, and T_{FH} , T_{H1} , T_{H17} , T_{REG} and B cells (**panel A**); pattern recognition receptor and TNF signalling in antigen presenting cells (**panel B**); and signalling by the IL-12 family of cytokines (**panel C**) in the pathogenesis of PBC.

Table 1

Panel (Ref)	Cases	Controls	Variants*	Platform
European panels				
'Old' Italian (2)	444	901	13,113,694	Illumina Human610-Quad (Cases), Illumina 1M-Duo (Controls)
WTCCC3 (3)	1,816	5,155	12,881,032	Illumina Human-660 W Quad (Cases), Illumina 1M-Duo (Controls)
'New' Canadian-UK	4,615	9,233	8,656,760	Illumina HumanCoreExome
'New' Italian	255	579	9,264,788	Illumina HumanCoreExome
'New' US	891	621	9,964,354	Illumina Infinium Global Screening Array (GSA) v1
European combined	8,021	16,489	5,186,747	-
Asian panels				
Japanese (4)	1,377	1,495	7,308,269	Affymetrix Axiom Genome-Wide ASI 1
Chinese (5)	1,118	2,788	6,934,908	HumanOmniZhongHua-8
Asian combined	2,495	4,283	5,347,815	-
All combined	10,516	20,772	2,817,608	-

* Number of variants following pre- and post-imputation quality control

Table 2A

Locus	Lead variant in the European panels			Lead variant in the Asian panels			Lead variant in the combined panels		
	Variant:A1/A2 (Chr:BP)	P (P _{perm.})	Beta (SE)	Variant:A1/A2 (Chr:BP)	P (P _{perm.})	Beta (SE)	Variant:A1/A2 (Chr:BP)	P (P _{perm.})	Beta (SE)
2p25.1 <i>ID2</i>	rs891058:A/G 2:8,442,547	5.39×10 ⁻⁷	-0.12	rs3111414:C/G 2:8,443,859	1.75×10 ⁻⁴	0.17	rs13416555:G/C 2:8,441,735	2.95×10 ⁻⁸	-0.12
		-	0.02		-0.0017	0.04			0.02
2q21.3 <i>TMEM163</i>	rs859767:G/A 2:135,341,200	1.54×10 ⁻⁹	-0.14	rs842349:T/G 2:135,342,452	1.76×10 ⁻⁹	-0.24	rs859767:G/A 2:135,341,200	8.94×10 ⁻¹⁶	-0.16
		-	0.02		<0.0001	0.04			0.02
6q21 <i>PRDM1</i>	rs58926232:G/C 6:10,6563,612	6.75×10 ⁻⁷	0.14	rs4134466:A/G 6:106,577,368	6.71×10 ⁻⁷	0.20	rs742108:A/G 6:106,582,920	3.16×10 ⁻⁸	0.13
		-	0.03		-0.0001	0.04			0.02
6q27 <i>CCR6</i>	rs3093024:A/G 6:167,532,793	2.37×10 ⁻⁶	0.10	rs4709148:T/C 6:167,521,676	2.18×10 ⁻¹⁰	-0.25	rs968334:T/C 6:167,526,096	3.98×10 ⁻¹⁰	0.12
		-0.0001	0.02		-	0.04			0.02
11q24.3* <i>ETS1</i>	rs10893872:T/C 11:128,325,553	9.07×10 ⁻⁶	0.10	rs11430718:G/GA 11:128,307,445	1.11×10 ⁻⁶	-0.19	rs10893872:T/C 11:128,325,553	9.77×10 ⁻⁹	0.11
		-	0.02		<0.0001	0.04			0.02
14q13.2 <i>FAM177A1</i>	rs712315:A/T 14:35,409,701	5.70×10 ⁻⁷	0.15	rs199892962:AT/A 14:35,646,404	4.36×10 ⁻⁶	0.20	rs799469:G/A 14:35,444,425	1.73×10 ⁻⁹	0.15
		-	0.03		-0.0020	0.04			0.03

*Note that 11q24.3 was previously identified at suggestive level of significance in the study by Kawashima *et al.* (2017)

Table 2B

Locus	Lead variant in the European panels			Lead variant in the Asian panels			Lead variant in the combined panels			
	Gene	Variant (Chr:BP)	P	Beta (SE)	Variant (Chr:BP)	P (P _{perm.})	Beta (SE)	Variant (Chr:BP)	P	Beta (SE)
1q23.1*		rs945635:G/C 1:157,670,290	1.59×10 ⁻⁸	-0.12 -0.02	rs60459521:G/C 1:157,147,588	1.25×10 ⁻³	-0.46 0.14	rs11264790:T/C 1:157,636,074	2.25×10 ⁻⁸	-0.11 0.02
<i>FCRL3</i>										
1q32.1		rs55734382:T/C 1:201,019,059	2.06×10 ⁻⁹	-0.14 0.02	rs117214467:C/T 1:200,436,787	8.55×10 ⁻³	-0.33 0.13	rs12122721:A/G 1:200,984,480	6.95×10 ⁻⁷	-0.11 0.02
<i>INAVA</i>										
2p23.3		rs34655300:T/C 2:25,514,333	5.23×10 ⁻¹⁰	0.14 0.02	rs893589:A/G 2:25,259,442	9.41×10 ⁻⁴	0.15 0.05	rs6711622:A/G 2:25,531,350	3.89×10 ⁻⁸	0.11 0.02
<i>DNMT3A</i>										
3p24.2		rs6550965:A/C 3:25,383,587	3.65×10 ⁻¹⁴	0.16 0.02	rs6807549:T/G 3:24,951,404	1.37×10 ⁻³	0.17 0.05	rs6550965:A/C 3:25,383,587	1.50×10 ⁻¹⁴	0.15 0.02
<i>RARB</i>										
4q24 [2]		rs7663401:C/T 4:106,128,954	2.76×10 ⁻⁸	-0.13 0.02	rs79109654:T/C 4:106,170,514	8.56×10 ⁻⁵	0.37 0.09	rs2007403:T/C 4:106,131,210	6.19×10 ⁻¹⁰	0.13 0.02
<i>TET2</i>										
5q21.1		rs141002831:T/TCA 5:100,202,282	1.47×10 ⁻⁷	0.12 0.02	rs157181:A/C 5:100,103,288	3.94×10 ⁻⁵	0.21 0.05	rs60643069:GA/G 5:100,238,073	2.48×10 ⁻⁹	0.13 0.02
<i>ST8SIA4</i>										
5q31.3		rs10062349:G/A 5:141,509,597	7.36×10 ⁻⁸	-0.12 0.02	rs3761757:A/C 5:141,488,219	7.48×10 ⁻³	-0.14 0.05	rs6874308:C/T 5:141,506,911	4.67×10 ⁻⁸	-0.11 0.02
<i>NDFIP1</i>										
7p21.1		rs7805218:A/G 7:20,378,801	4.12×10 ⁻⁸	0.13 0.02	rs77984571:C/G 7:20,512,650	7.54×10 ⁻³	-0.14 0.05	rs7786537:C/G 7:20,427,776	1.12×10 ⁻⁵	-0.11 0.02
<i>ITGB8</i>										
7q34		rs370193557:GAAT/G 7:138,729,543	1.89×10 ⁻⁸	0.12 0.02	rs12056141:G/A 7:138,797,730	1.05×10 ⁻³	0.18 0.05	rs370193557:G/GAAT 7:138,729,543	9.37×10 ⁻¹⁰	-0.12 0.02
<i>ZC3HAV1</i>										
8q24.21		rs4733851:A/G 8:129,264,420	2.18×10 ⁻⁷	0.11 0.02	rs1902780:C/T 8:129,211,788	5.51×10 ⁻⁴	-0.13 0.04	rs4733851:G/A 8:129,264,420	4.98×10 ⁻⁸	-0.11 0.02
<i>PVT1</i>										
9q22.33		rs11390003:GA/G 9:100,741,912	2.56×10 ⁻⁸	-0.15 0.03	rs10283737:G/T 9:100,780,063	1.24×10 ⁻³	0.15 0.05	rs112500293:T/C 9:100,763,455	7.63×10 ⁻⁹	-0.15 0.03
<i>TRIM14</i>										
10q11.23		rs7097397:A/G 10:50,025,396	2.42×10 ⁻¹⁰	-0.14 0.02	rs76129863:T/C 10:50,437,561	4.83×10 ⁻³	0.56 0.20	rs7922169:T/G 10:50,045,456	5.47×10 ⁻⁸	0.11 0.02
<i>WDFY4</i>										
11p15.5		rs58523027:TAA/T 11:646,986	4.00×10 ⁻⁸	-0.12 0.02	rs3216:C/G 11:214,421	8.17×10 ⁻²	-0.10 0.06	rs9667500:G/A 11:683,761	1.74×10 ⁻⁴	-0.08 0.02
<i>IRF7</i>										
14q32.12		rs72699866:A/G 14:93,114,787	2.89×10 ⁻¹¹	-0.20 0.03	rs76914265:G/C 14:93,219,854	1.16×10 ⁻⁴	-0.30 0.08	rs4904964:C/A 14:93,099,867	2.45×10 ⁻⁸	-0.12 0.02
<i>RIN3</i>										

16q22.1	rs79577483:G/A 16:68,036,939	1.23×10 ⁻¹¹	0.21 0.03	rs698729:G/C 16:68,624,205	1.90×10 ⁻²	-0.12 0.05	rs111644390:TC/T 16:68,046,323	1.18×10 ⁻⁹	0.17 0.03
18q22.2	rs1808094:T/C 18:67,526,026	2.79×10 ⁻⁹	0.13 0.02	rs76486918:T/C 18:67,081,620	2.72×10 ⁻³	-0.91 0.30	rs1808094:T/C 18:67,526,026	1.66×10 ⁻¹⁰	0.12 0.02

*Note that 1q23.1 was previously identified at suggestive level of significance in the study by Kawashima *et al.* (2017)

Table 3

Drug name	z	Description
Ustekinumab	-4.724	Anti-IL-12/23 p40 antibody
Abatacept	-4.596	IgG1 Fc/CTLA-4 fusion protein
Acitretin	-4.548	Oral retinoid
Belatacept	-4.488	IgG1 Fc/CTLA-4 fusion protein
Etretinate	-3.884	Oral retinoid
Denosumab	-3.751	Anti-TNFSF11 antibody
Interferon alfa-n1	-3.744	Glycosylated human interferon alpha
Interferon beta-1a	-3.634	Glycosylated human interferon beta
Fostamatinib	-3.586	Tyrosine kinase inhibitor
Tofacitinib	-3.445	Janus kinase inhibitor
Imatinib	-3.401	Tyrosine kinase inhibitor
Dexchlorpheniramine maleate	-3.382	Antihistamine
Gilteritinib	-3.183	Tyrosine kinase inhibitor
Brigatinib	-3.178	ALK and EGFR inhibitor
Phenformin	-3.006	Biguanide hypoglycemic agent

z, drug-disease proximity measure

Figure 1A

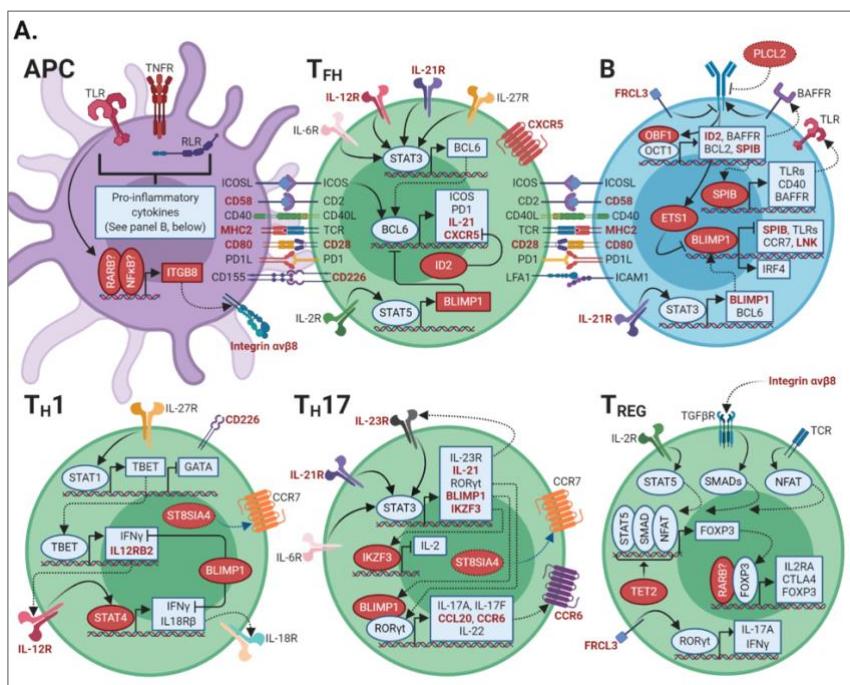


Figure 1B

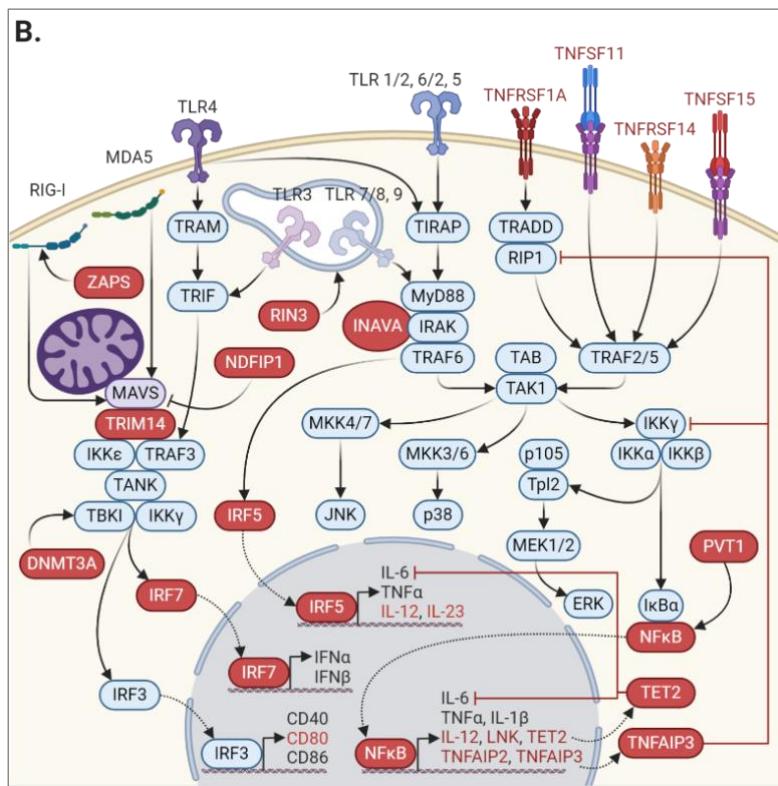
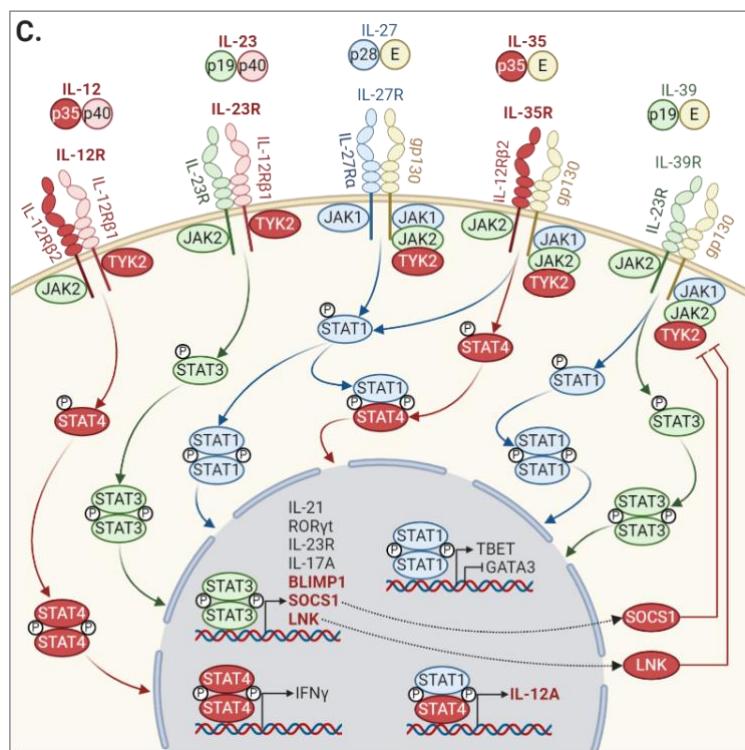


Figure 1C



References

1. Hirschfield, G. M. *et al.* Primary Biliary Cirrhosis Associated with HLA, IL12A, and IL12RB2 Variants. *N. Engl. J. Med.* **360**, 2544–2555 (2009).
2. Liu, X. *et al.* Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. *Nat. Genet.* **42**, 658–660 (2010).
3. Mells, G. F. *et al.* Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **43**, 329 (2011).
4. Kawashima, M. *et al.* Genome-wide association studies identify PRKCB as a novel genetic susceptibility locus for primary biliary cholangitis in the Japanese population. *Hum. Mol. Genet.* **26**, 650–659 (2017).
5. Qiu, F. *et al.* A genome-wide association study identifies six novel risk loci for primary biliary cholangitis. *Nat. Commun.* 14828 (2017). doi:10.1038/ncomms14828
6. Cordell, H. J. *et al.* International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.* **6**, 8019 (2015).
7. Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. ShamShaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. F, and P. C. S.

PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

8. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
9. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
10. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genet.* **5**, e1000529 (2009).
11. Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am. J. Hum. Genet.* **98**, 653–666 (2016).
12. Liu, J. Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet.* **42**, 436–440 (2010).
13. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
14. Mägi, R. *et al.* Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* **26**, 3639–3650 (2017).

15. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
16. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1-3 (2012).
17. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
18. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
19. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
20. Relton, C. L. *et al.* Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int. J. Epidemiol.* **44**, 1181–1190 (2015).
21. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
22. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
23. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

24. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).
25. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
26. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
27. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
28. Guney, E., Menche, J., Vidal, M. & Barabasi, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **7**, 10331 (2016).
29. Tanaka, A. *et al.* Genetic association of Fc receptor-like 3 polymorphisms with susceptibility to primary biliary cirrhosis: ethnic comparative study in Japanese and Italian patients. *Tissue Antigens* **77**, 239–243 (2011).
30. Hirschfield, G. M. *et al.* Ustekinumab for patients with primary biliary cholangitis who have an inadequate response to ursodeoxycholic acid: A proof-of-concept study. *Hepatology* **64**, 189–199 (2016).
31. Bowlus, C. L. *et al.* Therapeutic trials of biologics in primary biliary cholangitis: An open label study of abatacept and

- review of the literature. *J. Autoimmun.* **101**, 26–34 (2019).
- 32. Christofides, A., Konstantinidou, E., Jani, C. & Boussiotis, V. A. The role of peroxisome proliferator-activated receptors (PPAR) in immune responses. *Metabolism* **114**, 154338 (2021).
 - 33. Chen, M. L., Takeda, K. & Sundrud, M. S. Emerging roles of bile acids in mucosal immunity and inflammation. *Mucosal Immunol.* **12**, 851–861 (2019).

CHAPTER 3

X chromosome contribution to the genetic architecture of primary biliary cholangitis

Rosanna Asselta^{1,2}, Elvezia Maria Paraboschi^{1,2}, Alessio Gerussi^{3,4}, Heather J. Cordell⁵, George F. Mells⁶, Richard N. Sandford⁶, David E. Jones⁷, Minoru Nakamura⁸, Kazuko Ueno⁹, Yuki Hitomi¹⁰, Minae Kawashima¹⁰, Nao Nishida¹⁰, Katsushi Tokunaga^{9,10}, Masao Nagasaki¹¹, Atsushi Tanaka¹², Ruqi Tang¹³, Zhiqiang Li^{14,15}, Yongyong Shi^{14,15}, Xiangdong Liu¹⁶, Ma Xiong¹³, Gideon Hirschfield^{17,18}, Katherine A. Siminovitch^{19,20,21,22}, Canadian-US PBC Consortium#, Italian PBC Genetics Study Group#, UK-PBC Consortium#, Japan PBC-GWAS Consortium#, Marco Carbone^{3,4}, Giulia Cardamone^{1,2}, Stefano Duga^{1,2}, M. Eric Gershwin²³, Michael F. Seldin²³, Pietro Invernizzi^{3,4}.

¹Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, 20090 Pieve Emanuele, Milan, Italy

²Humanitas Clinical and Research Center, IRCCS, Via Manzoni 56, 20089 Rozzano, Milan, Italy

³Division of Gastroenterology and Center for Autoimmune Liver Diseases, Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy

⁴European Reference Network on Hepatological Diseases (ERN RARE-LIVER), San Gerardo Hospital, Monza, Italy;

⁵Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, NE1 3BZ, UK

⁶Academic Department of Medical Genetics, Cambridge University, Cambridge, UK, CB2 0QQ

⁷Faculty of Medical Sciences, Newcastle University, Newcastle, UK

⁸Clinical Research Center, National Hospital Organization (NHO), Nagasaki Medical Center, and Department of Hepatology, Nagasaki University Graduate School of Biomedical Sciences, Omura, Nagasaki, Japan

⁹Genome Medical Science Project, National Center for Global Health and Medicine (NCGM), Tokyo, Japan

¹⁰Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

¹¹Human Biosciences Unit for the Top Global Course Center for the Promotion of Interdisciplinary Education and Research, Kyoto University, Kyoto, Japan, and Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan

¹²Department of Medicine, Teikyo University School of Medicine, Tokyo 173-8605, Japan

¹³Division of Gastroenterology and Hepatology, Key Laboratory of Gastroenterology and Hepatology, Ministry of Health, State Key Laboratory for Oncogenes and Related Genes, Renji Hospital, School of Medicine, Shanghai Jiao Tong University,

Shanghai Institute of Digestive Disease, 145 Middle Shandong Road, Shanghai 200001, China

¹⁴Affiliated Hospital of Qingdao University and Biomedical Sciences Institute of Qingdao University (Qingdao Branch of SJTU Bio-X Institutes), Qingdao University, Qingdao, China;

¹⁵Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Center for Brain Science, Shanghai Jiao Tong University, Shanghai, China

¹⁶Key Laboratory of Developmental Genes and Human Diseases, Institute of Life Sciences, Southeast University, Nanjing, Jiangsu 210096, China

¹⁷Toronto General Hospital Research Institute, Toronto, Ontario Canada

¹⁸Department of Medicine, University of Toronto, Toronto, M5S 3H2 Ontario, Canada

¹⁹Mount Sinai Hospital, Lunenfeld Tanenbaum Research Institute and Toronto General Research Institute, Toronto, Canada, M5G 1X5

²⁰Department of Medicine, University of Toronto, Toronto, Ontario, Canada, M5S 3H2

²¹Department of Immunology, University of Toronto, Toronto, Ontario, Canada, M5S 3H2

²²Institute of Medical Sciences, University of Toronto, Toronto, Ontario M5S 3H2

²³University of California – Davis, Davis, California, USA, 95616

A full list of consortium members appears in the supplementary material

Published in *Gastroenterology*. 2021;2483–2495.
doi:10.1053/j.gastro.2021.02.061

PhD Candidate contribution: support in the analysis, interpretation of the results and critical feedback, support in the writing process (review and editing).

Abstract

Background & aims: Genome-wide association studies (GWAS) in primary biliary cholangitis (PBC) have failed to find X chromosome (chrX) variants associated with the disease. Here, we specifically explore the chrX contribution to PBC, a sexually-dimorphic complex autoimmune disease.

Methods: We performed a chrX-wide association study (XWAS), including genotype data from five GWAS (from Italy, UK, Canada, China, Japan; 5,244 cases, 11,875 controls).

Results: Single-marker association analyses found ~100 loci displaying $P < 5 \times 10^{-4}$; enrichment analyses, performed on these genes, revealed that they share features, such as a differential methylation in PBC ($P=0.013$), an increased production of circular RNAs ($P=0.010$), a higher density of super-enhancers ($P=0.030$), and an enrichment in binding sites for immunologically relevant transcription factors (NFATs, FOXO4; $P < 1.5 \times 10^{-8}$). While the transethnic meta-analysis evidenced only a suggestive signal (rs2239452, mapping within the PIM2 gene; OR=1.17, 95%CI=1.09-1.26; $P=9.93 \times 10^{-8}$), the population-specific meta-analysis showed a genome-wide significant locus in East Asians pointing to the same region (rs7059064, mapping within the GRIPAP1 gene; $P=6.2 \times 10^{-9}$, OR=1.33, CI=1.21-1.46). Indeed, rs7059064 tags a unique LD block including seven genes: TIMM17B, PQBP1, PIM2, SLC35A2, OTUD5, KCND1, and GRIPAP1, as well as a super-enhancer (GH0XJ048933 within OTUD5) targeting all these genes. GH0XJ048933 is

predicted to target also *FOXP3*, the main T regulatory cell lineage-specification factor. Consistently, *OTUD5* and *FOXP3*RNA levels were upregulated in PBC cases (1.75- and 1.64-fold, respectively).

Conclusion: This work represents the first comprehensive study of the chrX contribution to the genetics of an autoimmune liver disease and revealed a novel PBC-related genome-wide significant locus.

Introduction

Primary biliary cholangitis (PBC) is a complex disease in which an inappropriately activated immune response, characterized by high-titer serum antimitochondrial autoantibodies (AMA) as well as disease-specific antinuclear autoantibodies, leads to a progressive damage of intrahepatic bile ducts that may eventually cause liver failure^{1,2}. The disease is characterized by a striking female predominance (female:male prevalence ratio up to 8:1), with evidence of a significant contribution of X chromosome (chrX) defects to PBC pathogenesis: in fact, women with PBC show a significantly higher frequency of X monosomy in peripheral leukocytes compared to age-matched healthy women^{3,4}. However, there is a substantial lack of explanation for female predominance that is also emphasized by the absence of risk loci mapping on chrX⁵.

PBC is characterized by a strong genetic predisposition, with the major histocompatibility complex (MHC) class-II haplotypes (primarily *HLA-DRB1*, *DQB1*, and *DPB1*) showing the strongest association with the disease^{6–9}. In addition, genome-wide association studies (GWAS) have identified more than 40 non-MHC loci contributing to the disease risk. Most of these non-MHC loci implicate genes that contribute to cell-mediated immune mechanisms^{10–17}. These GWAS studies show an overlap in susceptibility loci between European and East-Asian populations, albeit with some degree of locus heterogeneity^{10–17}.

Notwithstanding these efforts, only a modest fraction of PBC heritability (~15%) has been explained¹⁸. Of note, the role of chrX in PBC still remains largely unknown, with no association signal reported at a genome-wide threshold of significance. This could also be explained by the fact that chrX polymorphisms have not been included in GWAS analysis, and, especially in the past, also by the lack of chrX-specific bioinformatics pipelines to be used in the analytic steps¹⁹. These limitations have indeed a more general impact on genetics of complex diseases: chrX constitutes 5% of the nuclear genome and mutations in genes mapping on this chromosome account for ~10% of Mendelian disorders²⁰; nevertheless, only 114 chrX susceptibility loci (0.8%) at $P \leq 5 \times 10^{-8}$ have been described, on a total of ~15,000 signals identified by GWAS studies for more than 300 traits²¹.

Here, we examine the chrX contribution to the genetic architecture of PBC by applying an analysis pipeline accounting for X-specific quality check (QC), imputation, and association tests^{22,23}.

Methods

Study design and participants

This study included genotype data on chrX principally derived from five previously performed GWAS (Supplementary Table 1[†])^{7,11–13,15–17}. All participants gave written informed consent for genetic studies. Local Institutional Review Boards approved the respective study protocols.

All cases met internationally accepted criteria for PBC²⁴. Most individuals were positive for serum AMA. Nevertheless, AMA positivity was not used as an inclusion criterion, considering previous data suggesting no effect of AMA status on the profile of disease-associated loci⁷.

QC of genotype data

QC steps were applied with a stepwise procedure separately for each dataset. First, we removed individuals: i) showing cryptic relatedness based on identity-by-state status ($\text{PI_Hat} > 0.10$), ii) having >10% missing genotypes, iii) with reported sex not matching the heterozygosity rates observed on chrX; and iv) with significant differences in call rate between cases and controls²⁵. Next, we excluded nucleotide polymorphisms (SNPs) having: i) >10% missingness throughout the dataset, ii) a minor allele

[†]The supplementary material, including supplementary figures and tables, listed in this chapter, can be retrieved at <https://www.sciencedirect.com/science/article/abs/pii/S0016508521004662?via%3Dihub>

frequency (MAF) <0.005, iii) a departure from the Hardy-Weinberg equilibrium in control females ($P<1*10^{-4}$), iv) significant differences in MAF between males and females in control individuals ($P<0.05/\text{number of SNPs}$), and v) a location in pseudoautosomal regions. Finally, we also removed SNPs exhibiting differential missingness between males and females ($P<1*10^{-4}$).

Correction for population stratification

We corrected for possible population stratification using chrX-derived principal components (PCs), which have been demonstrated to provide a more accurate population stratification correction for XWAS in admixed populations²². This procedure was performed using the principal component analysis method implemented in the EIGENSOFT program (https://genetics.med.harvard.edu/reich/Reich_Lab/Software.html)^{26,27}, after pruning for linkage disequilibrium (LD) and removing large LD blocks²⁸. For assessment and correction for population stratification we used the first 10 PCs of each dataset²⁷, and excluded all individuals inferred to be of an ancestry different from that of the specific dataset.

Imputation

Prephasing was performed using the SHAPEIT software, v2.17 (<https://mathgen.stats.ox.ac.uk/shapeit>)²⁹, using the parameters suggested for chrX. Datasets were imputed using the IMPUTE2 software,

v2.3.2

(https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference_5)³⁰, based on 1000 Genomes Project whole-genome and whole-exome haplotype data (reference panel: 1000Genome Phase3)³¹. IMPUTE2 has improved the imputation accuracy on chrX by taking into account the reduced effective population size available for this chromosome, by assuming that it is 25% less than that of the autosomes. As recommended by IMPUTE2 authors³⁰, the effective population size was set to 20,000, and k value to 1,000. Variants with MAF <0.005 or with informativeness <0.7 were considered of low confidence, and hence not considered in further analyses. Imputed datasets were finally submitted to QC steps, using the PLINK-XWAS v1.1 software (<http://keinanlab.cb.bscb.cornell.edu/content/xwas>)²³, using the above-described criteria.

Single-SNP association analyses

Single-SNP association tests were performed using PLINK-XWAS v1.1²³.

We assumed uniform and complete X-inactivation in females and a similar effect size between males and females. Hence, females are considered to have 0, 1, or 2 copies of an allele (like in autosomal analyses), whereas males are considered to have 0 or 2 copies of the same allele (i.e. male hemizygotes are considered equivalent to female homozygotes). This test is implemented in PLINK under the Model-2 option.

We then performed a second test by analyzing separately each sex (cases vs controls), with males coded as either having 0 or 2

copies of an allele as above. The female-only and male-only P values were then combined using the weighted Stouffer's method³², which allows combining P values not only accounting for potential effect size and direction between males and females, but also weighting the two test statistics (by using the square-root of the male/female sample size).

All the samples recruited in China were processed and analyzed as described above in a Chinese server to comply with the Regulation of the People's Republic of China on the Administration of Human Genetic Resources. The summary statistics, with no individual-level data were used for all subsequent analyses (e.g., meta-analysis with other panels).

Quantile-quantile (QQ) plots, genomic inflation factors (λ) calculations, and Manhattan plots were obtained using the R program (<https://www.r-project.org/>)³³. Single-SNP association results were clumped by the PLINK 1.9 software (<https://www.cog-genomics.org/plink/1.9/>), adopting $P<0.001$, $r^2>0.5$, and 250kb as parameters.

Meta-analysis

We filtered the SNP lists to include only those polymorphisms for which the association result was available from all cohorts (110,370 SNPs). Meta-analysis was performed both by combining data of all analyzed populations (transethnic meta-analysis) and by separately considering Caucasian and East-Asians populations.

The transethnic meta-analysis was carried out by using the MR-MEGA (Meta-Regression of Multi-Ethnic Genetic Association) software, which models allelic effects of a variant across datasets, weighted by their corresponding standard errors, in a linear regression framework, including the axes of genetic variation as covariates³⁴.

The Caucasian- and East-Asian-specific meta-analyses were performed using the Stouffer's method, taking into account weights and effect directions, as implemented in the Metainter software³⁵. This software uses a modified version of the meta-analytic approach based on multivariate generalized least squares estimation suggested by Becker and Wu³⁶, and is equivalent to the fixed effect model. Meta-analysis results were clumped together using the SECA software³⁷, extracting subsets of independent SNPs via LD. The procedure was "P-value informed", using $r^2 > 0.1$ and 1Mb (in LD with the index SNP) as parameters.

Finally, the genome-wide associated PBC risk locus in Asians was closely examined, by considering SNPs in the region surrounding the top hit (i.e., rs7059064; $\pm 200\text{kb}$). Pairwise LD among the SNPs was calculated to detect potential independent signals. SNPs showing $P_{\text{meta}} < 0.01$ and low LD with the rs7059064 SNP ($r^2 < 0.5$) were selected for conditional analysis.

In all our analyses, we considered loci with $P < 5 \times 10^{-8}$ (genome-wide level) as significant, and loci with $P < 5 \times 10^{-5}$ as suggestive of association. Although, $P < 1 \times 10^{-5}$ is the threshold at which, under the null hypothesis, one false positive result is expected per X-

wide scan of ~100K SNPs, we chose the less stringent threshold of $P < 5 \times 10^{-5}$, based on the high level of LD characterizing chrX³⁸.

Measurements of mRNA levels

Peripheral blood mononuclear cells (PBMCs) were isolated by centrifugation on a Lympholyte Cell separation medium (Cederlane Laboratories Limited, Hornby, Canada) gradient. Total RNA was isolated using the EuroGold Trifast kit (Euroclone, Wetherby, UK).

Random examers (Promega, Madison, USA) and the Superscript-III Reverse Transcriptase (Thermo Fisher Scientific, Waltham, USA) were used to perform first-strand cDNA synthesis, following the manufacturer's instructions. Semi-quantitative real-time reverse transcriptase-polymerase chain (RT-PCR) reactions were accomplished by using 1 μ l of the RT reaction, the SYBR Premix Ex Taq II (TaKaRa, Japan), and a touchdown thermal protocol on a LightCycler 480 (Roche, Basel, Switzerland). *HMBS* (hydroxymethylbilane synthase) was used as housekeeping gene. Reactions were performed in triplicate, and expression data analyzed using the GeNorm software³⁹. Primer sequences will be provided upon request.

Results

For evaluating the contribution of chrX to the genetic architecture of PBC, we extended the chrX marker sets from five GWAS cohorts by imputing non-pseudoautosomal regions in a total of ~17,000 individuals. For the analyses, we obtained up to 240,385 high-quality SNPs (**Supplementary Table 1**).

Single-Nucleotide Polymorphism association analysis within individual cohorts

We performed two different tests: within each cohort, the associations were studied considering males and females together (Test 1), or separately. For the separate analysis males and females were combined using the Stouffer method (Test 2). QQ plots for each test, along with the corresponding I calculations, showed well-calibrated test statistic distributions (**Supplementary Figure 1**).

Association analyses did not reveal any genome-wide significant signal (**Figure 1A**), with the most significant being a signal within the *OTUD5* gene (rs3027490, $P=4.80 \times 10^{-6}$; OR=1.39 CI=1.028-1.88; Japanese cohort, Test 1; **Supplementary Table 2**). The association signals were consistent between the two used association methods within each population (**Supplementary Tables 2, 3**). In particular, a total of 115 and 104 SNPs in the five cohorts displayed a nominal $P<0.0005$ for Test 1 and Test 2 analysis, respectively, with 79 overlapping signals; >40% of signals were within “gene-desert” regions (**Supplementary**

Table 4). Genes pinpointed by these signals showed few overlaps among populations (**Figure 1B**).

Transethnic meta-analysis

Based on single-SNP association results, we performed transethnic meta-analyses including all the five cohorts by using two approaches: 1) results from the Test-1 analysis were directly combined; 2) results from the Test-2 analysis were used in a sex-differentiated meta-analysis. The genomic inflation factors for these meta-analyses were between 9.979 and 1.114 (**Supplementary Figure 2**), indicating only a minimal residual bias.

Adopting the genome-wide significance threshold, the transethnic meta-analysis revealed the presence of only one interesting signal: the region tagged by the rs2239452 variant, which maps in the *PIM2* gene (suggestive $P_{meta}=9.93*10^{-8}$) (**Figure 2A; Table 1**). This signal was found considering all the five cohorts together, and seems to be sustained by the female component of the cohorts (suggestive $P_{meta-females}=1.34*10^{-5}$) (**Figure 2B; Table 1**).

Population-specific meta-analysis evidenced a novel PBC locus

Because of the evidence for locus heterogeneity in PBC susceptibility among different ethnicities^{10–17}, we also performed separate European and East Asian-specific meta-analyses, as well sex-specific meta-analyses, using the same strategy

described above; genomic inflation factors for these meta-analyses were well calibrated (**Supplementary Figure 3**).

The population-specific meta-analysis evidenced one locus with an association signal at genome-wide significance, i.e. the region tagged by the rs7059064 polymorphism, which maps within the *GRIPAP1* gene ($P_{meta}=6.17*10^{-9}$; OR=1.33, 95%CI=1.21-1.46) (**Figures 2, 3; Table 1; Supplementary Table 8**). This signal was found in East Asians, and corresponds to the top region evidenced by the transethnic meta-analysis (the *GRIPAP1* and the *PIM2* genes are only 53 kb apart).

In Europeans, there were only suggestive associations, one in an intergenic region (rs62604490; $P_{meta}=2.98*10^{-6}$) (**Table 1; Supplementary Figure 4**), and a second mapping within the *FGF13* gene (rs73241097; $P_{meta}=6.77*10^{-6}$) (**Table 1; Supplementary Figure 5**).

The sex-stratified analysis found a novel suggestive signal among East-Asian males (i.e., rs113885580 mapping within the *OTC* gene, $P_{meta-males}=1.06*10^{-5}$) (**Figure 2; Table 1; Supplementary Figure 6**), and evidenced that both the signal in the *GRIPAP1* region and the intergenic region pinpointed by the rs62604490 SNP are sustained by the female component ($P_{meta-females}=4.64*10^{-8}$ and $P_{meta-females}=4.24*10^{-6}$, respectively). The strongest signal among East-Asian females corresponded to rs2283734, a SNP mapping in the *PIM2* gene (**Figure 2**), the same gene highlighted by the transethnic meta-analysis.

Dissecting the genome-wide significant *GRIPAP1/PIM2* locus

The strongest signal evidenced by the meta-analysis was further investigated. **Table 2** shows the association summary statistics for the lead SNP (rs7059064) in each of the analyzed cohorts, including the European ones: there are indeed differences between the frequencies of the rs7059064-G minor allele in European cases (from 9.7 to 11.4%) vs those observed in East Asian patients (from 16 to 16.2%), thus possibly explaining the lack of association observed among Europeans. The effect of the rs7059064-G allele among East Asians was comparable between males and females (OR=1.50, 95%CI=1.24-1.81 for females; OR=1.53, 95%CI=0.99-2.38 for males), thus indicating that the apparent major contribution of females to the association signal simply stems on the higher number of analyzed female patients (**Supplementary Table 1**). Also, we observed a certain degree of heterogeneity among European populations ($P=0.15$), considering that, though not significantly, the rs7059064-G allele seems to exert a protective effect in British patients [OR=0.89, 95%CI=0.78-1.02], has no effect among Canadians [OR=0.99, 95%CI=0.77-1.25], whereas appears to confer predisposition towards PBC in Italian patients [OR=1.21, 95%CI=0.92-1.59]. Within a $\pm 200\text{kb}$ window centered on the rs7059064 polymorphism, there were 25 SNPs with association signals at $P_{\text{meta}}<0.01$. However, after conditional analysis, none of them remained significant, indicating that rs7059064 tagged a single haplotype that could account for the association signal in this

region. Indeed, this region is characterized by a unique LD block including seven genes (**Figure 3A, B**^{40,41}): *TIMM17B*, *PQBP1*, *PIM2*, *SLC35A2*, *OTUD5*, *KCND1*, and *GRIPAP1*. Among these genes, only *PQBP1* was previously associated at the genome-wide level with a phenotype (i.e., type II diabetes mellitus; <https://www.ebi.ac.uk/gwas/>).

With the exception of *KCND1*, each of these genes show expression in most tissues, including liver and whole blood (**Figure 3C**). While this region does not contain significant expression quantitative-trait loci (eQTLs; GTEx portal and eQTL Catalogue at EBI), it is characterized by the presence of a strong epigenetic signature (the activating H3K27Ac histone mark) within *OTUD5* intron 2, associated with the presence of a SE (element ID: GH0XJ048933; **Figure 3B, 4A**). Two SNPs, in perfect LD with the top-hit rs7059064, fall within this SE (**Supplementary Table 8**). GH0XJ048933 is known to target 13 genes, including six out of seven mapping in the PBC-associated region (*KCND1* is the only one not targeted by the enhancer; data from FANTOM5 Human Enhancers project⁴²; **Figure 4B**^{43,44}). Among these 13 genes, we found the immunologically relevant transcription factor forkhead box P3 (*FOXP3*). Hence, to further study the potential impact of the identified haplotype, we evaluated the expression levels of both *OTUD5* and *FOXP3* by semi-quantitative real-time RT-PCRs comparing PBMCs from 16 female PBC patients and 18 healthy female controls. Only females were examined due to the possibility of confounding sex effects (especially for *FOXP3*; see Discussion). We found a

significant 1.75- and 1.64-fold upregulation in PBC patients of *OTUD5* ($P=0.0013$) and *FOXP3* ($P=0.046$), respectively (**Figure 5**). For the other top loci (the intergenic rs62604490 polymorphism, the *FGF13* locus, and the *OTC* gene), their main features are illustrated in **Supplementary Figures 4-6**.

Discussion

GWAS have been a fruitful method to disclose genes/regions involved in the predisposition to complex diseases, however, chrX is notable for the paucity of associated loci¹⁹. For example, the most recent meta-analysis on multiple sclerosis identified 233 loci associated with the disease at genome-wide level, but just one locus was reported on chrX⁴⁵. In our study, we adopted an analysis pipeline specifically designed for chrX, to search for novel potential contributors to PBC heritability, and, possibly, to its female preponderance.

Indeed, the best association signal observed both in the transethnic and in the population-specific meta-analyses points to a unique LD region characterized by the presence of 7 genes (*TIMM17B*, *PQBP1*, *PIM2*, *SLC35A2*, *OTUD5*, *KCND1*, and *GRIPAP1*) and a SE, GH0XJ048933 (within *OTUD5*), which presents features with a potential impact on PBC pathogenesis. First, the enhancer is site of active transcription of an enhancer RNA (eRNA), which has been described as significantly expressed in blood, thymus, and spleen, as well as in blood cells such as neutrophils, natural killer, T, and B cells (Figure 4A, C; FANTOM5 data). This type of non-coding RNAs usually contributes to the enhancer activity and to the in-cis regulation of nearby genes⁴⁶. Second, the enhancer is enriched in binding sites for immune-related NFAT transcription factors (particularly, NFATC1 and NFATC3), thus stressing its possible involvement in an immune-mediated regulation of target genes. Third, the

enhancer targets 13 genes that, by integrating gene expression, protein expression, and methylation data, seem to be strongly co-regulated (**Figure 4C**), as it could be predicted for an enhancer having its cognate promoters located in the same topologically associating domain (TAD) (**Figure 4D**). Fourth, GH0XJ048933 targets also the *FOXP3* gene. *FOXP3* is a specific marker of T regulatory cells (Tregs), which are critical for the correct maintenance of immune tolerance (especially self-tolerance) and have been implicated in the pathogenesis of many autoimmune diseases⁴⁶. Fifth, *FOXP3* interacts with important determinants of the immune response (**Figure 4E**), and the transcript is among the few ones mapping on chrX to show a significant differential expression between males and females (in blood, P=0.0082; **Figure 4F**)⁴⁷. Last, but not least, different *Foxp3* transgenic mouse models have been developed^{48–50}; particularly interesting are: i) the *Foxp3*^{-/-} knockout mice, which developed an intense multiorgan inflammatory response and loss of CD4+ CD25+ Treg cells⁴⁹; ii) the *Foxp3* conditional-knockout mice (*Foxp3*^{flx}R26Cre^{ERT2}), which showed increased levels of IgE and autoantibodies⁴⁸; and, more importantly, iii) the so called Scurfy mice (*Foxp3*stmutant), i.e. animals that have a mutation in *Foxp3* that results in the complete abolition of *Foxp3*⁺ Tregs, which are all characterized, at 3-4 weeks of age, by the presence of high-titer serum AMA of all isotypes, by moderate to severe lymphocytic infiltrates surrounding portal areas and by evidence of biliary duct damage⁵⁰.

Together with *FOXP3*, at least three additional genes with potential implications in PBC - *PIM2*, *OTUD5*, and *GRIPAP1* - could be regulated by the GH0XJ048933 SE (**Figures 3, 4**). The proviral integration site for Moloney murine leukemia virus 2 (*PIM2*) is a serine/threonine kinase belonging to the PIM family, playing fundamental roles in proliferation/differentiation processes, and with known implications in cancer⁵¹. A growing number of studies have also implicated PIM2 in regulating the immune response, in particular with the description of a circuit linking the PIM2 protein with FOXP3: PIM2, induced by FOXP3, was demonstrated to be essential for the expansion of Tregs and, contrariwise, PIM2 was also described as being able to inhibit the suppressive function of Tregs by phosphorylating FOXP3⁵². Concerning the *OTUD5* gene, it codes for a member of the OTU (ovarian tumor) domain-containing cysteine protease superfamily. Also known as DUBA (deubiquitinating enzyme A), the *OTUD5* protein was shown to suppress type-I interferon (IFN-I) dependent innate immune response, by cleaving the poly-ubiquitin chain from the IFN-I adaptor protein, thus causing the disassociation of the adaptor from the downstream signaling complex, and ultimately the interruption of the IFN-I signaling cascade⁵³. As for *GRIPAP1* (GRIP1-associated protein 1), this gene codes for a guanine nucleotide exchange factor for the Ras family of small G proteins⁵³. Indeed, in a study aimed at identifying autoantibodies in PBC directed against GWBs (glycine-tryptophan-containing bodies, i.e. cytoplasmic domains that are involved in mRNA processing), Stinton and colleagues⁵⁴

were able to demonstrate that GRIPAP1 is one of the most common GWB autoantigen targets, being present in 17% of analyzed patients.

Although we demonstrated that *OTUD5* and *FOXP3* are differentially expressed in PBC patients, a major limitation of our study is the lack of functional studies, from one hand unraveling the molecular mechanisms linking SE GH0XJ048933 and its molecular targets, from the other explaining how genetic variants in this region could influence these mechanisms.

Conclusions

In conclusion, from the extensive analysis of chrX it emerges a number of genes possibly contributing, each with a modest effect, to PBC. This is not trivial, especially considering that chrX can be regarded as an “immunologic” chromosome (it contains the largest number of immune-related genes compared to other chromosomes)⁵⁵. Our major finding is however the identification of a genome-wide significantly associated locus, i.e. the one tagged by the rs7059064 polymorphism. This locus is characterized by presence of different genes and of a superenhancer possibly involved in their co-regulation, as well as in the regulation of *FOXP3* (which located in the same TAD). Future studies are mandatory for explaining the role of SE GH0XJ048933 and its targets in PBC.

Table Legends

Table 1. Meta-analysis results: list of top independent suggestive signals ($P < 5 \times 10^{-5}$).

Table 2. Association data for the lead rs7059064 polymorphism in all analyzed populations.

Figure Legends

Figure 1. Single-SNP association analysis results.

- A)** Manhattan plots showing the associations of chrX SNPs with PBC in the analyzed cohorts (CA, Canadians; ITA, Italians; UK, British; JP, Japanese; CH, Chinese) for the Stouffer analysis (Test 2). The blue line represents the $P=1*10^{-5}$ significance level. SNPs showing lowest P values are indicated by an arrow.
- B)** Venn diagrams show the number of genes mapping in correspondence/proximity of SNPs at $P<0.0005$ for each population. Chinese and Japanese show the major number of overlapping signals (genes are listed); the only gene shared by three populations is also highlighted.
- C,D)** The tables show the results of the enrichment analyses performed with the GSEA tool. We included either GO enrichment terms and all available immunologic signatures (C), or the Transcription-Factor Targets database terms (D). The first column describes the identified enrichments with regard to GO terms, dataset GSE reference numbers (as deposited in the GEO repository, <https://www.ncbi.nlm.nih.gov/gds/>), or TRANSFAC v7.4 identification codes and sequence of the transcription-factor binding sites (<http://gene-regulation.com/pub/databases.html>). The P value refers to the hypergeometric distribution for the number of genes in the intersection of the query set with a set from the used database. The FDR q-value is the false discovery rate analog of hypergeometric P-value after correction for

multiple testing (Benjamini/Hochberg). In D, the list of genes with overlapping conserved motifs (indicated in blue) is also reported. **E)** PBC-associated loci on chrX are enriched in differentially methylated genes. On the left, the Venn diagram shows the number of genes: i) mapping in chrX (subdivided in 668 protein coding and 1,628 non-coding genes); ii) tagged by SNPs associated with PBC at a nominal $P<0.0005$; and iii) previously reported as differentially methylated in PBC (see Supplementary Material for details). A significant fraction of PBC-associated genes (nine, listed) overlaps with differentially methylated genes in PBC. On the right, the table shows enrichment-analysis results: for the random sets analysis, the average values calculated on 1,000 iterations are indicated. SD is shown in parenthesis. The % refers to the percentage of times in which the same or a larger number of differentially methylated genes was obtained in the 1,000 iterations as compared to the PBC dataset.

Figure 2. Manhattan plots of meta-analyses.

Manhattan plots summarizing the results of transethnic (**A-C**), population-specific (**D-I**), and sex-stratified meta-analyses (**B,C,E,F,H,I**). The horizontal lines represent the suggestive $P=5*10^{-5}$ and the genome-wide Bonferroni-corrected $P=5*10^{-8}$ significance levels. SNPs showing lowest P values are indicated by an arrow (if intragenic, the relevant gene is also indicated); those reported in red, survive to the Bonferroni correction for multiple testing.

Figure 3. The *GRIPAP1/PIM2* locus.

- A)** Plot of the regional association signals surrounding the rs7059064 top hit in East Asians. The plot was built using the LocusTrack site (<https://gump.qimr.edu.au/general/gabrieC/LocusTrack/>).⁴⁶
- B)** Screenshot from the UCSC Genome browser (<http://genome.ucsc.edu/>; GRCh37/hg19) highlighting the PBC-associated LD region (coordinates chrX:48,750,000-48,865,000). The panel shows the following tracks: i) the ruler with the scale at the genomic level; ii) chrX nucleotide numbering; iii) the UCSC RefSeq track; iv) ENCODE data (<https://www.encodeproject.org/>) for H3K4Me1, H3K4Me3, H3K27Ac histone marks, derived from seven cell lines; v) enhancers (grey bars) and promoters (red bars) from GeneHancer⁴¹ with the GH0XJ048933 enhancer targets; vi) interactions (curved lines) connecting GeneHancer regulatory elements/genes; vii) basewise conservation track.
- C)** Expression panel across tissues of the genes depicted in panel A (GTEx data; <https://gtexportal.org/home/>).

Figure 4. The GH0XJ048933 SE codes for an eRNA and co-regulates the genes of the *GRIPAP1/PIM2* locus.

- A)** Screenshot from the UCSC Genome browser showing the GH0XJ048933 SE region (chrX:48,791,000-48,802,000). Listed tracks are: i) the ruler with the scale at the genomic level; ii) chrX nucleotide numbering; iii) the track for eRNAs from the FANTOM5 Human Enhancers project

(http://slidebase.binf.ku.dk/human_enancers/); iv) the UCSC RefSeq track; v) ENCODE data for H3K4Me1, H3K4Me3, H3K27Ac histone marks, derived from seven cell lines; vi) the enhancers/promoters track from GeneHancer;⁴⁷ the grey bar indicates the GH0XJ048933 enhancer; vii) interactions connecting GeneHancer regulatory elements and genes (interactions with *OTUD5*, *PIM2*, *PQBP1*, *FOXP3* are depicted); viii) basewise conservation track; ix) the dbSNP(151) track for common polymorphisms.

B) Integration of gene expression (GE), protein expression (PE), copy number (CN) and methylation (ME) relative to the 13 genes regulated by the GH0XJ048933 SE. Data come from the TCGA portal (<https://tcga-data.nci.nih.gov/docs/publications/tcga/?>). Circle plot was built by using the Zodiac tool (<http://www.compgenome.org/zodiac/>)⁴³. Only significant intergenic interactions are shown (FDR≤0.1). Green lines indicate positive interactions.

C) The tables show expression data (>1%) in organs/cells for the eRNA gene mapping within GH0XJ048933. Red bars indicate a significant over-representation of the transcript (FANTOM5 data).

D) TAD structure of the chrX:47,480,000-50,440,000 region. The central TAD contains all genes of the PBC-associated region tagged by rs7059064. The panel was produced through the 3D-Genome Browser (<http://3dgenome.org>)⁴⁴, using Hi-C data produced in HepG2 cells (hepatocytes) and generated by the Dekker Laboratory (resolution: 40kb).

E) FOXP3 interactome. The best 10 interactions are shown (highest confidence=90%). Evidence are based on text-mining, experiments, databases, co-expression data, gene fusions, co-occurrences. The panel was produced using the STRING tool (<https://string-db.org/>).

F) Violin plots show *FOXP3* RNA expression levels in whole blood and liver, obtained through the GTEx portal, stratified on sex (265 males, 142 females).

Figure 5. *OTUD5* and *FOXP3* are overexpressed in PBC.

Boxplots show expression levels of *OTUD5* (**A**) and *FOXP3* (**B**) measured by semi-quantitative real-time RT-PCR in PBMCs of a PBC case-control cohort. Boxes define the interquartile range; thick lines refer to the median. Results were normalized to expression levels of the *HMBS* housekeeping gene and are presented as rescaled values. The number of subjects is indicated (N). Significance levels of t-tests: *: $P<0.05$; **: $P<0.005$.

Table 1

Populations	Software	SNP	ChrX Position *	A1/A2	P_JP	P_CH	P_CAN	P_ITA	P_UK	P_meta	OR [95% CI]	Locus
ALL COHORTS	MM	rs2239452	48775572	G/C	7.51e-06	5.41e-04	0.97	0.59	0.012	9.9e-08	1.17 [1.09-1.26]	PIM2
FEMALES-ALL COHORTS	MM	rs2239452	48775572	G/C	4.35e-05	4.83e-4	0.48	0.36	5.11e-3	1.3e-05	1.11 [1.02-1.21]	PIM2
MALES-ALL COHORTS	MM	rs201130692	132978723	-/A	0.015	0.88	5e-04	0.045	0.25	3.1e-05	3.16 [1.8-5.42]	GPC3
					P_JP	P_CH	P_meta	OR [95% CI]	Locus			
EAST ASIANS	MI	rs7059064	48837087	G/A	8.1e-06	1.75e-04	6.2e-09	1.33 [1.21-1.46]	GRIPAP1			
EAST-ASIAN FEMALES	MI	rs2283734	8773556	A/G	4.15e-05	2.91e-04	4.64e-08	1.38 [1.23-1.56]	PIM2			
EAST-ASIAN MALES	MI	rs113885580	38236645	G/A	0.0075	3.84e-04	1.06e-05	2.36 [1.61-3.46]	OTC			
					P_CAN	P_ITA	P_UK	P_meta	OR [95% CI]	Locus		
CAUCASIANS	MI	rs62604490	116104694	G/A	0.36	0.0058	6.90e-05	2.98e-06	0.75 [0.66-0.85]	Intergenic		
CAUCASIAN FEMALES	MI	rs62604490	116104694	G/A	0.47	7.59e-04	1.09e-04	4.24e-06	0.73 [0.63-0.83]	Intergenic		

Only top signals of each suggestively/genome-wide associated region is reported (see also Figure 2).

For all SNPs presented in this table, directions among cohorts were always consistent, except for rs2239452 (all cohort analysis). A1: tested allele (MAF allele). JP, Japanese; CH, Chinese; CAN, Canadians; ITA, Italians; UK, British; OR, odds ratio; CI, confidence interval, MI, METAINTER; MM, MR-MEG. * According to human genome release Feb. 2009, GRCh37/hg19.

Table 2

	<i>Position</i> *	<i>Minor allele/ Major allele</i>	<i>MAF cases</i>	<i>MAF controls</i>	<i>OR [95% CI]</i>	<i>P value</i>	<i>Population</i>	<i>P_{meta}</i>	<i>Transethnic P_{meta}</i>	
rs7059064	48837087	G/A	0.114	0.119	0.99 [0.77- 1.25]	0.908	Canadian	0.350	9.93e-08	
			0.111	0.0935	1.21 [0.92- 1.59]	0.174	Italian			
			0.0973	0.116	0.89 [0.78- 1.02]	0.0942	British			
			0.162	0.119	1.38 [1.20- 1.59]	8.14e-06	Japanese	6.2e-09		
			0.160	0.123	1.28 [1.13- 1.46]	1.75e-04	Chinese			

Minor allele frequencies (MAF) and P values of association tests are given for all populations (Model-2 analysis). P values are presented for both the population-specific and transethnic meta-analyses.

* According to human genome release Feb. 2009, GRCh37/hg19

Figure 1

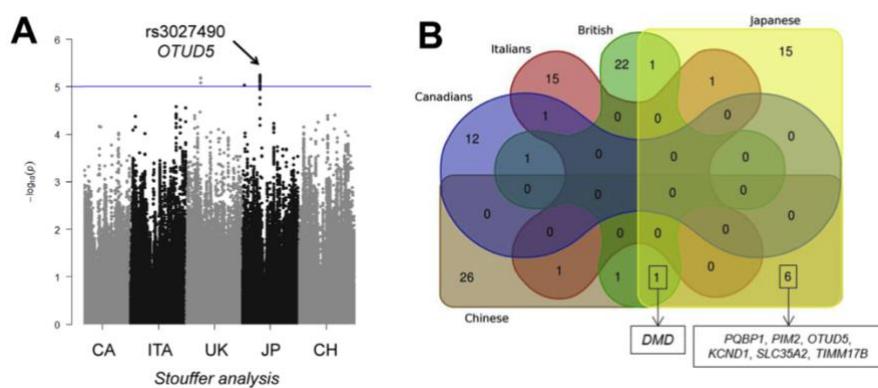


Figure 2

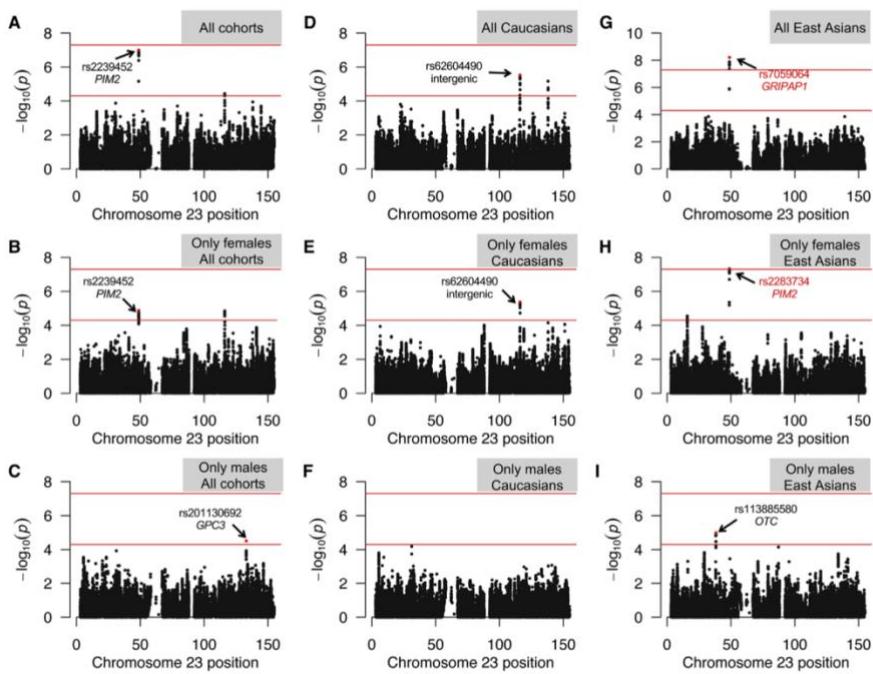


Figure 3

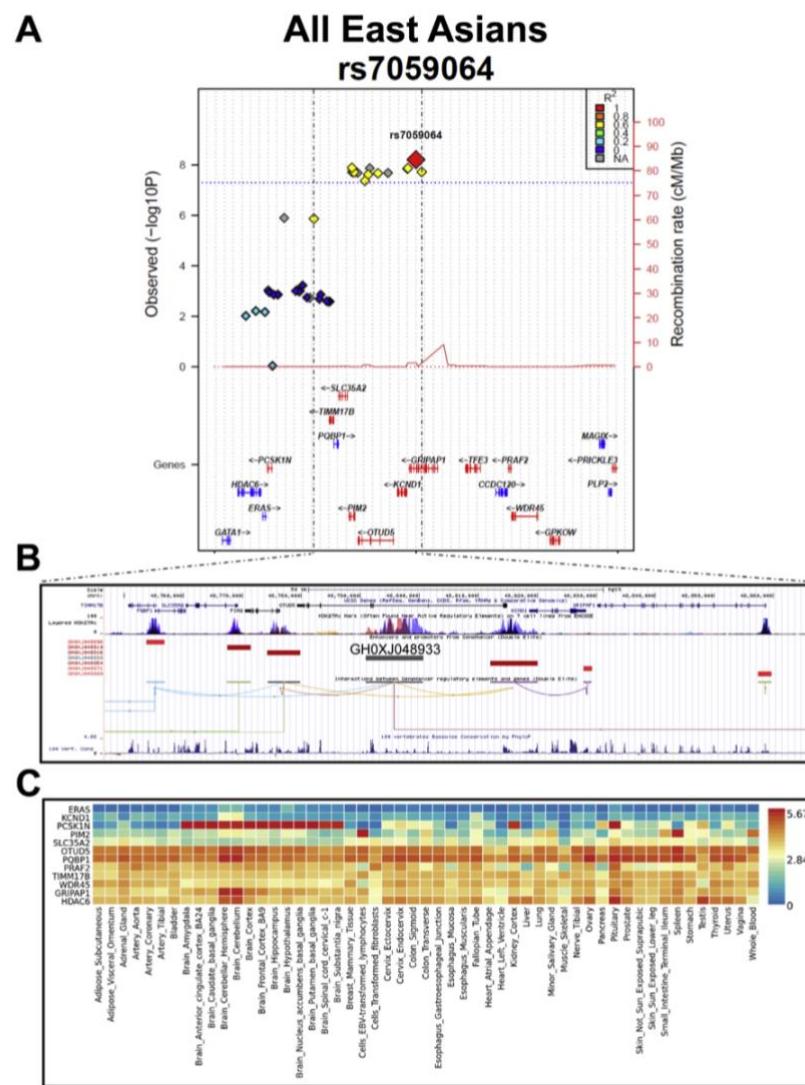


Figure 4

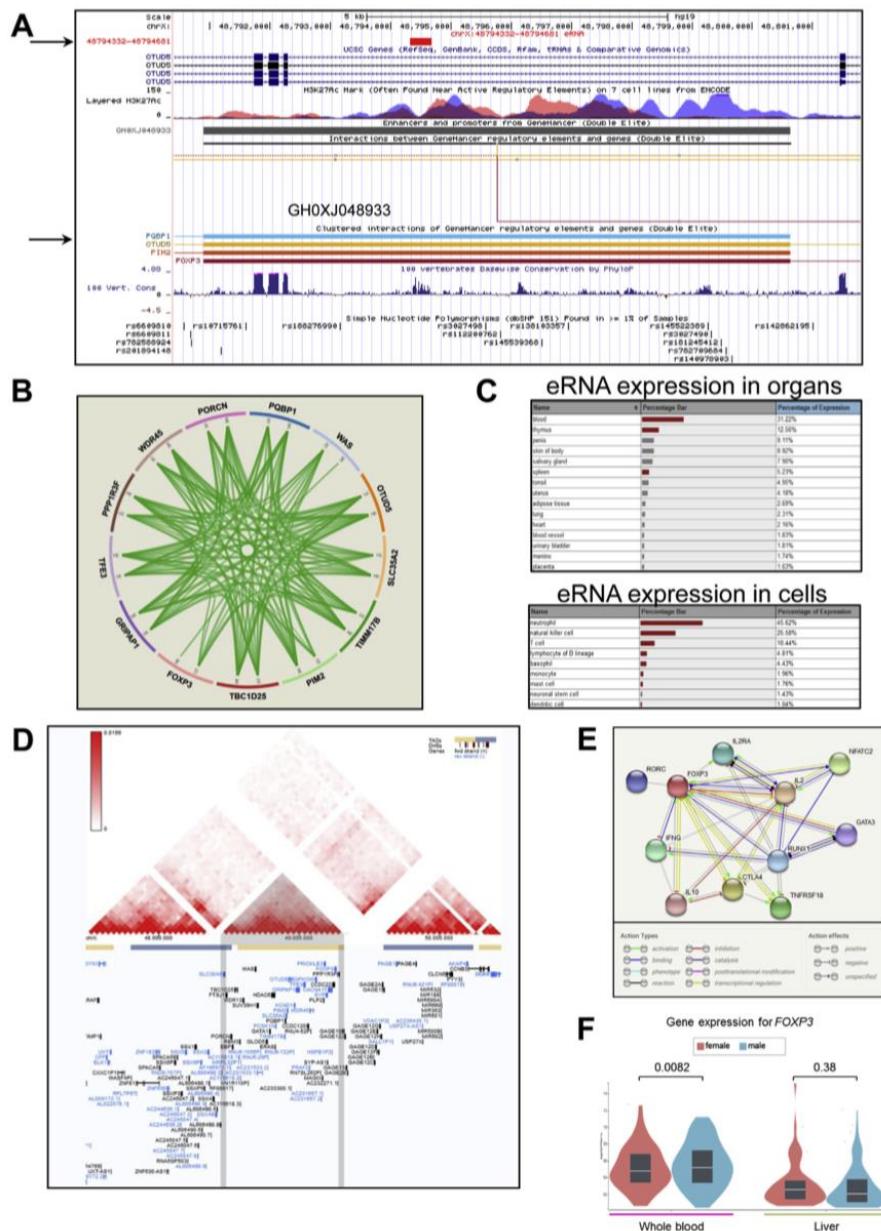
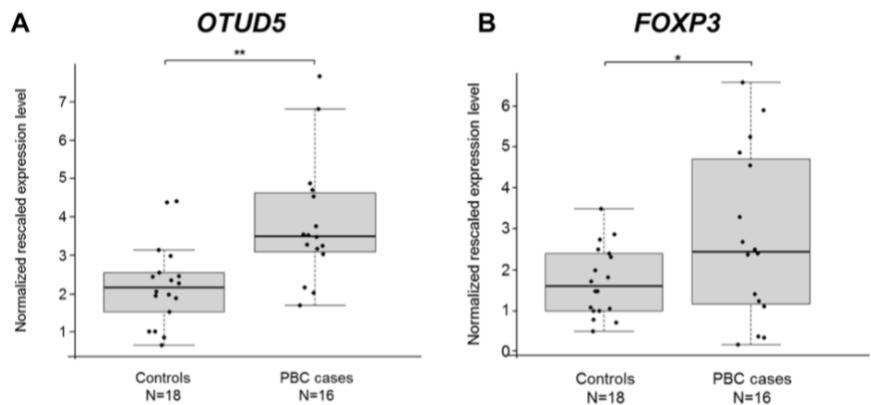


Figure 5



References

1. Marshall M. Kaplan, M.D., and M. Eric Gershwin, M. . Primary Biliary Cirrhosis. *N. Engl. J. Med.* **353**, 1261–73 (2005).
2. Invernizzi, P., Selmi, C. & Gershwin, M. E. Update on primary biliary cirrhosis. *Dig. Liver Dis.* **42**, 401–408 (2010).
3. Gerussi, A., Cristoferi, L., Carbone, M., Asselta, R. & Invernizzi, P. The immunobiology of female predominance in primary biliary cholangitis. *J. Autoimmun.* **95**, 124–132 (2018).
4. Invernizzi, P. *et al.* Frequency of monosomy X in women with primary biliary cirrhosis. *Lancet* **363**, 533–535 (2004).
5. Webb, G. J., Siminovitch, K. A. & Hirschfield, G. M. The immunogenetics of primary biliary cirrhosis: A comprehensive review. *J. Autoimmun.* **64**, 42–52 (2015).
6. Invernizzi, P. *et al.* Human leukocyte antigen polymorphisms in italian primary biliary cirrhosis: A multicenter study of 664 patients and 1992 healthy controls. *Hepatology* **48**, 1906–1912 (2008).
7. Hirschfield, G. M. *et al.* Primary Biliary Cirrhosis Associated with HLA, IL12A, and IL12RB2 Variants. *N. Engl. J. Med.* **360**, 2544–2555 (2009).
8. Invernizzi, P. Human leukocyte antigen in primary biliary cirrhosis: An old story now reviving. *Hepatology* **54**, 714–723 (2011).

9. Invernizzi, P. *et al.* Classical HLA-DRB1 and DPB1 alleles account for HLA associations with primary biliary cirrhosis. *Genes Immun.* **13**, 461–468 (2012).
10. Joshita, S., Umemura, T., Tanaka, E. & Ota, M. Genetics and epigenetics in the pathogenesis of primary biliary cholangitis. *Clin. J. Gastroenterol.* **11**, 11–18 (2018).
11. Liu, X. *et al.* Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. *Nat. Genet.* **42**, 658–660 (2010).
12. Mells, G. F. *et al.* Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **43**, 329 (2011).
13. Liu, J. Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **44**, 1137 (2012).
14. Juran, B. D. *et al.* Immunochip analyses identify a novel risk locus for primary biliary cirrhosis at 13q14, multiple independent associations at four established risk loci and epistasis between 1p31 and 7q32 risk Variants. *Hum. Mol. Genet.* **21**, 5209–5221 (2012).
15. Nakamura, M. *et al.* Genome-wide Association Study Identifies TNFSF15 and POU2AF1 as Susceptibility Loci for Primary Biliary Cirrhosis in the Japanese Population. *Am. J. Hum. Genet.* **91**, 721–728 (2012).
16. Qiu, F. *et al.* A genome-wide association study identifies six novel risk loci for primary biliary cholangitis. *Nat. Commun.* 14828 (2017). doi:10.1038/ncomms14828

17. Kawashima, M. *et al.* Genome-wide association studies identify PRKCB as a novel genetic susceptibility locus for primary biliary cholangitis in the Japanese population. *Hum. Mol. Genet.* **26**, 650–659 (2017).
18. Gulamhusein, A. F., Juran, B. D. & Lazaridis, K. N. Genome-Wide Association Studies in Primary Biliary Cirrhosis. *Semin Liver Dis* **35**, 392–401 (2015).
19. Wise, A. L., Gyi, L. & Manolio, T. A. EXclusion: Toward integrating the X chromosome in genome-wide association analyses. *Am. J. Hum. Genet.* **92**, 643–647 (2013).
20. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Res.* **37**, D793–D796 (2009).
21. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2016).
22. Chang, D. *et al.* Accounting for eXentricities: Analysis of the X Chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. *PLoS One* **9**, 1–31 (2014).
23. Keinan, A. *et al.* XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome. *J. Hered.* **106**, 666–671 (2015).
24. Lindor, K. D. *et al.* Primary biliary cirrhosis. *Hepatology* **50**, 291–308 (2009).
25. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies.

Genet. Epidemiol. **34**, 591–602 (2010).

26. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
27. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904 (2006).
28. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
29. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
30. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genet.* **5**, e1000529 (2009).
31. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
32. Stouffer, S. A., Suchman, E. A., DeVinney, L. C. & Shirley, A. Star, and Robin M. Williams. 1949. The American Soldier: Adjustment During Army Life. *Stud. Soc. Psychol. World War II* **1**,
33. R Core Team, Rf. R: A language and environment for statistical computing. (2013).
34. Mägi, R. *et al.* Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* **26**, 3639–3650 (2017).

35. Vaitsiakhovich, T., Drichel, D., Herold, C., Lacour, A. & Becker, T. METainter: meta-analysis of multiple regression models in genome-wide association studies. *Bioinformatics* **31**, 151–157 (2015).
36. Becker, B. J. & Wu, M.-J. The Synthesis of Regression Slopes in Meta-Analysis. *Stat. Sci.* **22**, 414–429 (2007).
37. Nyholt, D. R. SECA: SNP effect concordance analysis using genome-wide association summary results. *Bioinformatics* **30**, 2086–2088 (2014).
38. Schaffner, S. F. The X chromosome in population genetics. *Nat. Rev. Genet.* **5**, 43–51 (2004).
39. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, RESEARCH0034 (2002).
40. Cuellar-Partida, G., Renteria, M. E. & MacGregor, S. LocusTrack: Integrated visualization of GWAS results and genomic annotation. *Source Code Biol. Med.* **10**, 1 (2015).
41. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*. **2017**, (2017).
42. Lizio, M. *et al.* Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. *Nucleic Acids Res.* **45**, D737–D743 (2017).
43. Zhu, Y. *et al.* Zodiac: A Comprehensive Depiction of Genetic Interactions in Cancer by Integrating TCGA Data. *J. Natl. Cancer Inst.* **107**, (2015).

44. Wang, Y. *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* **19**, 151 (2018).
45. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, (2019).
46. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017).
47. Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244 (2017).
48. Tai, Y., Sakamoto, K., Takano, A., Haga, K. & Harada, Y. Dysregulation of humoral immunity in Foxp3 conditional-knockout mice. *Biochem. Biophys. Res. Commun.* **513**, 787–793 (2019).
49. Lin, W. *et al.* Allergic dysregulation and hyperimmunoglobulinemia E in Foxp3 mutant mice. *J. Allergy Clin. Immunol.* **116**, 1106–1115 (2005).
50. Zhang, W. *et al.* Deficiency in regulatory T cells results in development of antimitochondrial antibodies and autoimmune cholangitis. *Hepatology* **49**, 545–552 (2009).
51. Narlik-Grassow, M., Blanco-Aparicio, C. & Carnero, A. The PIM family of serine/threonine kinases in cancer. *Med. Res. Rev.* **34**, 136–159 (2014).
52. Deng, G. *et al.* Pim-2 Kinase Influences Regulatory T Cell Function and Stability by Mediating Foxp3 Protein N-

- terminal Phosphorylation*. *J. Biol. Chem.* **290**, 20211–20220 (2015).
53. Kayagaki, N. *et al.* DUBA: a deubiquitinase that regulates type I interferon production. *Science* **318**, 1628–1632 (2007).
 54. Stinton, L. M., Swain, M., Myers, R. P., Shaheen, A. A. & Fritzler, M. J. Autoantibodies to GW bodies and other autoantigens in primary biliary cirrhosis. *Clin. Exp. Immunol.* **163**, 147–156 (2011).
 55. Bianchi, I., Lleo, A., Gershwin, M. E. & Invernizzi, P. The X chromosome and immune associated genes. *J. Autoimmun.* **38**, J187–J192 (2012).

CHAPTER 4

A novel Polygenic Risk Score in Primary Biliary Cholangitis

Alessio Gerussi^{1,2}, Claudio Cappadona^{4,5}, Davide Paolo Bernasconi³, Laura Cristoferi^{1,2,3}, Maria Grazia Valsecchi³, Marco Carbone^{1,2}, Rosanna Asselta^{4,5}, Pietro Invernizzi^{1,2}

¹Division of Gastroenterology, Center for Autoimmune Liver Diseases, Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy

²European Reference Network on Hepatological Diseases (ERN RARE-LIVER), San Gerardo Hospital, Monza, Italy

³Bicocca Bioinformatics Biostatistics and Bioimaging Centre - B4, School of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy.

⁴Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, 20072 Pieve Emanuele, Milan, Italy

⁵IRCCS Humanitas Clinical and Research Center, Via Manzoni 56, 20089 Rozzano, Milan, Italy

Manuscript in preparation

PhD Candidate contribution: leader in the conceptualization of the study, data curation, analytic process, interpretation of the results and writing the manuscript.

Abstract

Background and Aims: Several genetic variants modulate the genetic risk of primary biliary cholangitis (PBC). We aimed to set up and investigate the accuracy of a polygenic risk score (PRS) based on the most updated list of variants associated with the disease.

Methods: Individual data from two Italian cohorts (“OldIT”, 444 cases and 901 controls, and “NewIT”, 255 cases and 579 controls) were used for analyses. The summary statistics from the most comprehensive international meta-analysis were used as reference data to obtain the effect size estimates. The PRS was calculated following the classic “clumping + thresholding” method. PRS scores were expressed as median and Interquartile Range (IQR). HLA and sex were used as covariates. PLINK 1.9 and R were used for PRS set up and for statistical analyses.

Results: Starting from a pool of 46 genes, a total of 22 variants was selected. We found that PBC patients in the OldIT cohort had a significantly greater risk score than healthy controls: -0.014 (IQR -0.023, 0.005) versus -0.022 (IQR -0.030, -0.013) ($P < 2.2 \times 10^{-16}$). The area under the curve (AUC) value for the model to predict case vs control based only on genetic information was 0.72, while the inclusion of sex to the model increased the AUC to 0.82. We validated our findings in the NewIT cohort, confirming the high accuracy of the model (0.71 without sex and 0.81 with sex included) and its good calibration. In both cohorts, individuals

in the highest stratum had around 14 times higher risk of having PBC than those in the reference group ($P < 10^{-6}$). The PRS explained around 5% of the genetic variance, highlighting the presence of a missing heritability still to be elucidated.

Conclusion: A PRS based on 22 variants together with sex and HLA were strongly associated with the susceptibility to PBC and displayed excellent ability to discriminate between cases and controls in two independent samples of individuals of European ancestry. The PRS identified a subgroup of subjects at high risk that should be the target of tailored follow-up.

Introduction

Primary biliary cholangitis (PBC) is an autoimmune liver disease with a complex genetic background. Heritability of PBC is particularly high, with a sibling relative risk of around 10¹. Several genome-wide association studies (GWAS) and meta-analyses have been performed to dissect its genetic architecture, identifying >50 loci modulating the risk of developing PBC^{1,2}.

Most, if not all, loci bring little additional risk by themselves. A genetic risk score, encompassing the combined effect of different loci, represents a better tool to predict disease risk than single loci³. A genetic risk score for PBC was created in 2015, but it included as reference dataset for summary statistics only the GWAS performed in the UK earlier⁴. In the meantime, two meta-analyses have been performed, increasing the number of loci to be included^{2,5}.

In this study, we aimed to create a polygenic risk score (PRS) to assess its capacity both to discriminate between cases and controls and to evaluate the role of non-genetic covariates such as sex.

Methods

The following methods follow the guidelines for reporting PRS⁶.

Background - Study type and outcome

The study aimed to develop and validate a PRS for PBC. The predicted outcome of the study was disease status (diagnosis of PBC versus healthy control). Therefore, the PRS was generated for risk prediction.

Study population and data

Study design and recruitment

Individual data from the two Italian cohorts (“OldIT” and “NewIT”) included in the meta-analysis by Cordell *et al*⁷ presented in Chapter 2 were used for analysis (secondary data).

Participant demographics, clinical characteristics and ancestry

The OldIT Italian cohort included 444 cases and 901 healthy controls from the general population of Italian ancestry (post quality-checked cohort); 515 were males and 830 females. The NewIT Italian cohort included 834 individuals of Italian ancestry, 255 cases and 579 controls; 335 were males and 499 females. All cases met internationally accepted criteria for PBC diagnosis⁷.

Genetic data and non-genetic variables

A total of 46 genes were selected, based on genome-wide significant threshold of p-value in the meta-analysis by Cordell et al, based on summary statistics in individuals on European ancestry (**Table 1**). For each gene, we considered a genomic region comprising 250 kb upstream and 250 kb downstream of the transcriptional unit, and extracted from the dataset all the SNPs by using PLINK v.1.9⁸.

The OldIT and NewIT cohorts included 105,150 and 74,484 genotyped/imputed SNPs, respectively, mapping within or in proximity (± 250 kb) of 46 genes associated with known predisposition to PBC (**Table 1**).

To account for the role of HLA, we selected and extracted the best associated SNP within each cohort (--assoc function in PLINK). For OldIT the top variant was chr6:32653792:A:G, while for NewIT the top variant was chr6:32429303:A:G. The HLA SNP was recoded as an ordinal variable as follows: homozygous (AA) = 1, heterozygous (AG, GA) = 2, homozygous (GG) = 3.

For quality check and imputation steps we refer to the Methods section of chapter 2 and Cordell *et al*. The only non-genetic variable that was used for the model was sex.

Risk model development and application

Polygenic Risk Score construction and estimation

The PRS was calculated following the classic “clumping + thresholding” (C + T) method, where SNP clumping and a GWAS

p-value thresholding are performed to control for linkage disequilibrium (LD) and adjust GWAS estimated effect sizes, respectively³. The sum of risk alleles of an individual was calculated using the risk allele effect size estimates from the summary statistics from Cordell *et al*⁷ presented in Chapter 2.

The PRS was calculated by using the top variant for each of the 46 candidate genes. Several PRS were calculated at a range of P-value thresholds and then the PRS that explained the highest phenotypic variance was selected (threshold of association of $p < 0.0000001$). With this set, a total of 22 variants were included, of which 9 (41% were new variants not included in the previous PRS; 24 variants were not found in the OldIT and NewIT datasets under study. The independence of the variants was confirmed through clumping using PLINK 1.9.

The PRS was then calculated using PLINK 1.9 --score command and --sum option, assuming an additive model for independent variants. The PRS explaining the highest phenotypic variance was identified by using the R program⁹, by testing the association between the PRS and PBC phenotype in our dataset using a logistic regression approach.

Integrated risk model description and fitting

Different combinations of non-genetic variants (Sex), HLA variant (treated as factor) and 10 principal components of ancestry were included in the model.

Risk model evaluation

PRS distribution

PRS scores were expressed as median and Interquartile Range (IQR).

Risk model predictive ability

Comparison between median PRS scores between cases and controls were performed by Mann-Whitney U test.

The PRS was divided into four strata dividing the entire range of PRS values into 4 evenly spaced intervals of equal distance. The relative proportion of cases and controls in each strata was calculated. The association between PRS strata and risk of PBC was expressed as OR \pm 95%CI. ORs were calculated comparing the lowest stratum as reference against all the others. p-values from Chi-square test are reported for all strata against the reference stratum.

Heritability

To identify the PRS with the best fit, we first performed a logistic regression between our phenotype and PRSs calculated at different p-values thresholds. The 'glm' function in R was used for this step^{9,10}. Then, for each p-value threshold, we estimated the SNP-heritability, that is, the phenotypic variance explained by the SNPs used to calculate the PRS. This was done by calculating the McFadden's R squared, which compares our PRS model to a null model^{11,12}. The PRS model used assumes an

independent, additive effect of all the common SNPs which satisfy the p-value threshold. Finally, the PRS with the highest McFadden's R squared value, which explains the highest phenotypic variance, was selected.

Risk model discrimination

Diagnostic accuracy was evaluated using receiver operating characteristic (ROC) curves. The performance of the PRS and integrated risk models was evaluated by Area under the ROC curve (AUC), which was reported together with its 95% CI. Nonparametric stratified bootstrapping was used to compute confidence bands for ROC curves.

Negative and positive predictive values (NPV, PPV), specificity and sensitivity, and accuracy are reported. The optimal threshold of the PRS was chosen maximizing the Youden index.

Risk model calibration

Calibration was assessed after calculating risk predictions according to a logistic regression model which included continuous PRS.

Individual predicted risks were then divided into ten equally sized categories (i.e. according to deciles). A calibration plot was then produced by comparing the mean predicted risk in each decile (displayed in the x axis) with the observed risk, calculated as the proportion of PBC cases within each decile (displayed in the y axis). Brier score, corresponding to the mean squared error of the prediction, was also calculated together with its 95% CI.

Results

Set up of a Polygenic Risk Score

An Italian cohort of 444 cases and 901 healthy controls from the general population (post quality-checked cohort) was the cohort under study. It included 105,150 genotyped/imputed SNPs mapping within or in proximity (± 250 kb) of 46 genes associated with known predisposition to PBC².

After clumping, a PRS utilizing 22 disease susceptibility loci was created (**Table 2**). For each individual, a PRS was calculated to evaluate the risk of the disease.

Risk model predictive ability

We compared the distribution of PRS between PBC cases and controls and found that PBC patients had a significantly greater risk score than healthy controls (median value in cases -0.014 (IQR -0.023, 0.005) versus median value in controls -0.022 (IQR -0.030, -0.013), Wilcoxon test $p < 2.2 \times 10^{-16}$) (**Figure 1A** and **1B**); differences were maintained in subgroup analysis by sex (**Supplementary Figure 1**), as well as sex and HLA status (**Supplementary Figure 2**).

To explore the effect of PRS in more detail, we divided the PRS into four strata of equal dimension. The relative proportion of cases increased consistently across different strata for increasing intervals of the PRS (**Figure 1C**); this effect was conserved when strata were stratified by sex (**Supplementary Figure 1**), and by sex and HLA status (**Supplementary Figure**

2). The OR increased with the increasing PRS groups, using the first stratum as reference. Individuals in the highest strata have around 14 times higher risk than those in the reference group (95% CI: 1.291-1.482) (**Figure 1D** and **Supplementary Table 1**). The ORs when also sex and HLA were added to the model are reported in **Table 3**.

Risk model discrimination

To assess the ability of PRS to discriminate correctly between cases and controls, we used ROC curves and calculated AUC. The PRS composed of 22 SNPs, the HLA SNP, and sex showed good ability to identify individuals who are at the increased risk for developing PBC (AUC: 0.8273, 95% CI: 0.8049-0.8497). The maximum value of Youden's index was 1.37. Sensitivity was 0.86, specificity 0.65, positive predictive value was 0.55, negative predictive value was 0.90, accuracy was 0.72.

Comparing the PRS model including all 22 SNPs and the model without the HLA tag SNP chr6:32653792:A:G, we found that removal of the HLA SNP from the PRS only slightly decreased the AUC to 0.81 (95% CI: 0.7907-0.8368). The AUC for the tag SNP chr6:32653792:A:G alone decreased to 0.63 (95% CI: 0.5989-0.6612) (**Table 4** and **Figure 2**).

Risk model calibration

Logistic regression models including PRS (**Figure 3A**) and PRS with sex and HLA as covariates (**Figure 3B**) produced individual predicted risks that were in strong agreement with the observed

risks, indicating a good level of calibration for both, even though the integrative model was slightly better as indicated by the lower Brier score.

Replication of results in an independent cohort

To replicate the genetic risk model in an independent population, we next examined the PRS in a second Italian cohort (“NewIT”). The validation set (“NewIT”) included 834 individuals of Italian ancestry, 255 cases and 579 controls; 335 were males and 499 females. In the validation cohort, 74,484 variants were included. The PRS was generated by weighting 22 SNPs (**Table 5**). In this replication data set, the PRS index was also significantly higher in patients with PBC, with a median score of -0.013 (-0.024,-0.003) compared to controls with a median score of -0.02 (-0.029,-0.013) (Mann-Whitney test $P = 7.40 \times 10^{-13}$) (**Figure 4A** and **4B**). Differences were maintained in subgroup analyses by sex (**Supplementary Figure 3**) and by sex and HLA (**Supplementary Figure 4**).

The PRS were divided into four strata of equal dimension. The relative proportion of cases increased consistently across different strata for increasing intervals of the PRS (**Figure 4C**). This effect was conserved when strata were stratified by sex (**Supplementary Figure 3**) and by sex and HLA (**Supplementary Figure 4**). The individuals in the fourth stratum had significantly increased risk compared with the individuals in the reference group (OR: 13.95, 95% CI: 4.30-45.22, **Figure 3D**

and **Supplementary Table 2**). The ORs when also sex and HLA was added to the model are reported in **Table 6**.

The PRS composed of 22 SNPs, the HLA SNP and sex showed good ability to identify individuals who are at the increased risk for developing PBC (AUC: 0.8138, 95% CI: 0.7832-0.8445). Maximum value of Youden's index was 0.29. Sensitivity was 0.81, specificity 0.67, positive predictive value was 0.52, negative predictive value was 0.88, accuracy was 0.71.

Comparing the PRS model including all 22 SNPs and the model without the HLA tag SNP chr6:32653792:A:G, we found that removal of the HLA SNP from the PRS only slightly decreased the AUC to 0.79 (95% CI: 0.7667-0.8291). The AUC for tag SNP chr6:32653792:A:G alone decreased to 0.6523 (95% CI: 0.6108-0.6939).

There was little difference between the PRS model containing all the SNPs and the model with only non-HLA SNPs (AUC: 0.71 versus 0.68). The AUC including sex showed better discriminatory ability than the one with genetic information alone (AUC 0.79 versus 0.71, **Table 7** and **Figure 5**).

The level of calibration in the NewIT cohort was good and similar to OldIT both for the PRS and the integrative risk score, as also suggested by the almost equal Brier scores (**Figure 6**).

Measurement of the genetic variance

We estimated the cumulative fraction of genetic variance explained by the SNPs included in the PRS; together the 22 SNPs explain 5.6% of genetic susceptibility to PBC in the OldIT

cohort. The NewIT cohort showed similar results with the same set of SNPs, explaining 5.3% of genetic susceptibility.

Discussion

Here, we present a novel PRS together with an integrated risk model of the PRS with sex and HLA to estimate genetic predisposition to PBC. The PRS has been derived and validated in two independent Italian cohorts, using recently established predisposing genetic variants together with their correlated effect sizes, derived from the most recent international meta-analysis. We found that the set-up 22-locus PRS was significantly associated with the increased risk of PBC, and in both cohorts: specifically, the OR of the highest risk group was around 14 when compared with the lowest risk group, which is a magnitude comparable to what found for rare variants in monogenic disorders¹³. The PRS displayed in both cohorts good discrimination between individuals with PBC and healthy controls, which significantly increased when sex and HLA were also included in an integrated risk model⁶.

The associations between alleles and a trait of interest can be combined to quantify heritability and to guide risk stratification⁶. The capacity to identify subjects at significantly higher genetic risk of developing PBC represents at the same time an opportunity and a challenge for the field. No data is available on cost-effectiveness of testing first-degree relatives of PBC patients for PBC-specific autoantibodies and cholestatic liver enzymes. In addition, no specific recommendation on their follow-up has been made in most recent international PBC guidelines¹⁴.

Since strategies to manage first-degree relatives are not established, the integrated risk model could be used to identify those individuals at high risk and, consequently, set up prospective studies tailored to study their natural history. These studies could represent an advancement in the field by identifying early markers of disease onset¹³. Together with lack of data to guide clinical management in these scenarios, the progressive reduction of costs for genome sequencing might also encourage the design and implementation of prospective studies to evaluate the net benefit of such strategies.

Once these strategies are established, they could also be modulated in order to allocate attention and resources across individuals with different levels of genetic risk¹⁵. Although PRSSs typically involve several variants pointing to multiple disease pathways, in terms of prevention and detection strategies they can be useful regardless of the explanation of the underlying mechanism, like occurred for statin therapy for cardiovascular disease or mammography screening for breast cancer¹³.

Despite the progressive increase in the number of genetic variants associated with PBC², the genetic variance of the disease explained by the risk loci identified so far is only ~5%⁴, and this estimate is confirmed also in our study. The “missing” heritability may be due to several factors, including the need for larger cohorts to identify other common variants¹⁶, the lack of whole-genome sequencing studies to pinpoint rare and ultra-rare variants (both in the coding portion of the genome, or mapping in “non canonical” non-coding regions)^{17,18}, the historical neglect of

X chromosome in genome-wide studies despite being rich in immune-related genes^{1,19–21}, the complexity in taking into account gene-gene and gene-environment interactions²², and the established role of epigenetics in the onset and progression of the disease^{23,24}. The performance of the model would likely improve if these factors were addressed and taken into account. Several limitations need to be considered. Around half of the variants associated with PBC in the meta-analysis were not found in our available cohorts: their inclusion would likely improve the performance of the model and the fraction of explained heritability. The addition of sex in our model enhanced the discriminatory power, so that we cannot exclude that other non-genetic factors such as age, age at disease onset, family risk and positivity to anti-mitochondrial antibodies may affect the risk of PBC. However, these clinical data were not available in our study. In addition, our analysis was restricted to individuals of European ancestry. Based on data derived from the recent meta-analysis and the well-known role of ancestry in shaping the genetic architecture of complex traits²⁵, we suggest against the use of this PRS in individuals of non-European ancestry. The available data sets contributed to the published meta-analysis results for PBC, with a risk of bias when evaluating the classification performance; however, their contribution in terms of sample size was quite limited (the Italian subjects being ~8% of individuals of European ancestry and ~7% of all individuals included in the meta-analysis)².

Conclusions

We demonstrated that a PRS comprising 22 PBC risk SNPs together with sex and HLA was strongly associated with the susceptibility to PBC and displayed excellent ability to discriminate between cases and controls in two independent cohorts of individuals of European ancestry. The PRS identified a subgroup of subjects at high risk that should be the target of tailored follow-up. We foresee the application of PRS to first-degree relatives of patients with PBC; prospective studies are needed to evaluate the potential role of PRS in these individuals.

Table Legends

Table 1. List of genes associated with PBC.

Chr, Chromosome

Table 2. List of variants included in the PRS (OldIT).

Abbreviations: Chr, Chromosome; Snp, Single Nucleotide Polymorphism; Bp, Base Pair.

Table 3. Odds ratios for integrative risk model (OldIT).

Risk of PBC, expressed as OR \pm 95% CI by PRS strata, HLA status and sex (integrative risk model). ORs for strata and HLA were calculated comparing the lowest stratum as reference against all the others, while for sex male sex was used as reference; p-values from the glm model are reported.

Abbreviations: HLA, human leukocyte antigen, OR, odds ratio.

Table 4. The association of each PRS model with PBC (OldIT).

wGRS_sex_HLA_10PC includes 23 SNPs, sex and the 10 principal components (PCs); wGRS_sex_10PC includes 22 non-HLA SNPs, sex and the 10 PCs; wGRS_HLA_10PC includes the 23 SNPs and the 10 PCs; wGRS_10PC includes 22 non-HLA SNPs and the 10 PCs; HLA_10PC includes the tag HLA SNP and the 10 PCs in the model.

Abbreviations: AUC, area under curve; CI, confidence interval; HLA, human leukocyte antigen; SNP, single-nucleotide polymorphisms, PC, principal components.

Table 5. List of variants included in the PRS (NewIT).

Abbreviations: Chr, Chromosome; Snp, Single Nucleotide Polymorphism; Bp, Base Pair.

Table 6. Odds ratios for integrative risk model (NewIT).

Risk of PBC, expressed as OR \pm 95%CI by PRS strata, HLA status and sex (integrative risk model). ORs for strata and HLA were calculated comparing the lowest stratum as reference against all the others, while for sex male sex was used as reference; p-values from the glm model are reported.

Abbreviations: HLA, human leukocyte antigen, OR, odds ratio.

Table 7. The association of each PRS model with PBC (NewIT).

wGRS_sex_HLA_10PC includes 23 SNPs, sex and the 10 principal components (PCs); wGRS_sex_10PC includes 22 non-HLA SNPs, sex and the 10 PCs; wGRS_HLA_10PC includes the 23 SNPs and the 10 PCs; wGRS_10PC includes 22 non-HLA SNPs and the 10 PCs; HLA_10PC includes the tag HLA SNP and the 10 PCs in the model.

Abbreviations AUC, area under curve; CI, confidence interval; HLA, human leukocyte antigen; SNP, single-nucleotide polymorphisms, PC, principal components.

Supplementary Table Legends

Supplementary Table 1. Odds ratios for PRS (OldIT)

Risk of PBC, expressed as OR \pm 95%CI by PRS strata. ORs for strata were calculated comparing the lowest stratum as reference against all the others; p-values from the glm model are reported.

Abbreviations: OR, odds ratio.

Supplementary Table 2. Odds ratios for PRS (NewIT)

Risk of PBC, expressed as OR \pm 95%CI by strata. ORs were calculated comparing the lowest stratum as reference against all the others. p-values from the glm model are reported.

Abbreviations: OR, odds ratio.

Figure Legends

Figure 1. Distribution of the PRS in cases and controls and PBC risk in the OldIT cohort.

- A)** Comparison of PRS scores between cases and controls. In the boxplot, boxes define the interquartile range; thick lines refer to the median. Samples were divided into 4 strata (dividing the entire range of PRS values into 4 evenly spaced intervals). The p value was calculated using the Wilcoxon rank sum test.
- B)** Density plot of PRS scores for cases and controls.
- C)** Proportion of cases and controls in each PRS stratum (setting as 100% the overall number of individuals for each stratum).
- D)** Risk of PBC, expressed as OR \pm 95%CI by strata. ORs were calculated comparing the lowest stratum as reference against all the others.

Abbreviations: PBC, Primary Biliary Cholangitis, PRS, Polygenic Risk Score.

Figure 2. ROC curves comparing wGRS in different models in OldIT cohort.

wGRS_sex_HLA_10PC includes 23 SNPs, sex and the 10 principal components (PCs); wGRS_sex_10PC includes 22 non-HLA SNPs, sex and the 10 PCs; wGRS_HLA_10PC includes the 23 SNPs and the 10 PCs; wGRS_10PC includes 22 non-HLA SNPs and the 10 PCs; HLA_10PC includes the tag HLA SNP and the 10 PCs in the model.

Abbreviations: HLA, human leukocyte antigen; PC, principal components, ROC, Receiver Operating Characteristic

Figure 3. Calibration plots for the PRS (A) and the integrative risk model (B) in the OldIT cohort.

Abbreviations: PRS, Polygenic Risk Score.

Figure 4. Distribution of the PRS in cases and controls and PBC risk in the NewIT cohort.

A) Comparison of PRS scores between cases and controls. In the boxplot, boxes define the interquartile range; thick lines refer to the median. Samples were divided into 4 strata (dividing the entire range of PRS values into 4 evenly spaced intervals). The p value was calculated using the Wilcoxon rank sum test.

B) Density plot of PRS scores for cases and controls.

C) Proportion of cases and controls in each PRS stratum (setting as 100% the overall number of individuals for each stratum).

D) Risk of PBC, expressed as OR \pm 95%CI by strata. ORs were calculated comparing the lowest stratum as reference against all the others.

Abbreviations: PBC, Primary Biliary Cholangitis, PRS, Polygenic Risk Score.

Figure 5. ROC curves comparing PRS in different models in the NewIT cohort. wGRS_sex_HLA_10PC includes 23 SNPs, sex and the first 10 principal components (PCs); wGRS_sex_10PC includes 22 non-HLA SNPs, sex and the 10 PCs; wGRS_HLA_10PC includes the 23 SNPs and the 10 PCs; wGRS_10PC includes 22 non-HLA SNPs and the 10 PCs;

HLA_10PC includes the tag HLA SNP and the 10 PCs in the model.

Abbreviations: HLA, human leukocyte antigen; PC, principal components, ROC, Receiver Operating Characteristic

Figure 6. Calibration plots for the PRS (A) and the integrative risk model (B) in the NewIT cohort.

Abbreviations: PRS, Polygenic Risk Score.

Supplementary Figure Legend

Supplementary Figure 1. Distribution of the PRS in cases and controls stratified by sex in the OldIT cohort.

A) Comparison of PRS scores between cases and controls stratified by sex. In the boxplot, boxes define the interquartile range; thick lines refer to the median. Samples were divided into 4 strata (dividing the entire range of PRS values into 4 evenly spaced intervals) and stratified by sex. The p value was calculated using the Wilcoxon rank sum test.

B) Density plot of PRS scores for cases and controls stratified by sex.

C) Proportion of cases and controls in each PRS stratum stratified by sex (setting as 100% the overall number of individuals for each stratum).

Abbreviations: PBC, Primary Biliary Cholangitis, PRS, Polygenic Risk Score.

Supplementary Figure 2. Distribution of the PRS in cases and controls stratified by sex and HLA in the OldIT cohort.

A) Comparison of PRS scores between cases and controls stratified by sex and HLA. In the boxplot, boxes define the interquartile range; thick lines refer to the median. Samples were divided into 4 strata (dividing the entire range of PRS values into 4 evenly spaced intervals) and stratified by sex and HLA. The p value was calculated using the Wilcoxon rank sum test.

B) Density plot of PRS scores for cases and controls stratified by sex and HLA.

C) Proportion of cases and controls in each PRS stratum stratified by sex and HLA (setting as 100% the overall number of individuals for each stratum).

Abbreviations: HLA, human leukocyte antigen; PBC, Primary Biliary Cholangitis, PRS, Polygenic Risk Score.

Supplementary Figure 3. Distribution of the PRS in cases and controls stratified by sex in the NewT cohort.

A) Comparison of PRS scores between cases and controls stratified by sex. In the boxplot, boxes define the interquartile range; thick lines refer to the median. Samples were divided into 4 strata (dividing the entire range of PRS values into 4 evenly spaced intervals) and stratified by sex. The p value was calculated using the Wilcoxon rank sum test.

B) Density plot of PRS scores for cases and controls stratified by sex.

C) Proportion of cases and controls in each PRS stratum stratified by sex (setting as 100% the overall number of individuals for each stratum).

Abbreviations: PBC, Primary Biliary Cholangitis, PRS, Polygenic Risk Score.

Supplementary Figure 4. Distribution of the PRS in cases and controls stratified by sex and HLA in the NewIT cohort.

A) Comparison of PRS scores between cases and controls stratified by sex and HLA. In the boxplot, boxes define the interquartile range; thick lines refer to the median. Samples were divided into 4 strata (dividing the entire range of PRS values into 4 evenly spaced intervals) and stratified by sex and HLA. The p value was calculated using the Wilcoxon rank sum test.

B) Density plot of PRS scores for cases and controls stratified by sex and HLA.

C) Proportion of cases and controls in each PRS stratum stratified by sex and HLA (setting as 100% the overall number of individuals for each stratum).

Abbreviations: HLA, human leukocyte antigen; PBC, Primary Biliary Cholangitis, PRS, Polygenic Risk Score.

Table 1

Chr	Gene	Coordinates of the analyzed region (hg38) - start	Coordinates of the analyzed region (hg38) - end
1	<i>MMEL1</i>	2273723	2773723
1	<i>IL12RB2</i>	67570194	68070194
1	<i>CD58</i>	116815083	117315083
1	<i>FCRL3</i>	157420290	157920290
1	<i>DENND1B</i>	197530966	198030966
1	<i>CACNA1S</i>	200769059	201269059
2	<i>DNMT3A</i>	25264333	25764333
2	<i>TMEM163</i>	135091200	135591200
2	<i>STAT4</i>	191693742	192193742
3	<i>PLCL2</i>	16711265	17211265
3	<i>RARB</i>	25133587	25633587
3	<i>TIMMDC1</i>	118969934	119469934
3	<i>IL12A-AS1</i>	159410283	159910283
4	<i>NFKB1</i>	103290780	103790780

4	<i>TET2</i>	105878954	106378954
5	<i>IL7R</i>	35631130	36131130
5	<i>LOC285626</i>	158509900	159009900
6	<i>OLIG3</i>	137723068	138223068
7	<i>ITGB8</i>	20128801	20628801
7	<i>ELMO1</i>	37132465	37632465
7	<i>TNPO3</i>	128367466	128867466
7	<i>ZC3HAV1</i>	138479543	138979543
9	<i>HEMGN</i>	100491912	100991912
10	<i>WDFY4</i>	49775396	50275396
11	<i>DEAF1</i>	396986	896986
11	<i>CCDC88B,</i>	63860422	64360422
11	<i>POU2AF1</i>	110989365	111489365
11	<i>DDX6</i>	118490104	118990104
12	<i>TNFRSF1A</i>	6190009	6690009
12	<i>ATXN2</i>	111657431	112157431
13	<i>LINC02341</i>	42805002	43305002
13	<i>DLEU1</i>	50561220	51061220
14	<i>RAD51B</i>	68499927	68999927
14	<i>RIN3</i>	92864787	93364787

14	<i>EXOC3L4</i>	103314807	103814807
16	<i>CLEC16A</i>	10924365	11424365
16	<i>IL4R</i>	27153469	27653469
16	<i>DPEP2</i>	67786939	68286939
16	<i>LOC105371388</i>	85769271	86269271
17	<i>ZPBP2</i>	37794893	38294893
17	<i>KANSL1</i>	43899348	44399348
18	<i>CD226</i>	67276026	67776026
19	<i>TYK2</i>	10225652	10725652
19	<i>MAST3</i>	17985882	18485882
19	<i>SPIB</i>	50676742	51176742
22	<i>RPL3</i>	39490078	39990078
Total	46		

Table 2

Chr	Snp	Bp
1	1:2523723:C:T	2523723
1	1:157670290:C:G	157670290
1	1:201019059:C:T	201019059
2	2:25514333:C:T	25514333
2	2:135341200:A:G	135341200
3	3:16961265:A:G	16961265
3	3:159733527:G:T	159733527
5	5:35881130:G:GT	35881130
5	5:158759900:A:G	158759900
6	6:137973068:A:G	137973068
7	7:128617466:A:G	128617466
7	7:138729543:G:GAAT	138729543
9	9:100741912:G:GA	100741912
11	11:646986:T:TAA	646986
11	11:64110422:A:T	64110422
11	11:111239365:C:G	111239365
11	11:118740104:A:AT	118740104
12	12:111907431:A:AC	111907431
14	14:103564807:A:C	103564807
16	16:68036939:A:G	68036939
17	17:38044893:C:CTT	38044893
18	18:67526026:C:T	67526026

Table 3

	OR	lower	upper	p-value
PRS.q4(-0.0693,-0.0459)	reference			
PRS.q4(-0.0459,-0.0226)	3.80	1.09	13.23	0.035
PRS.q4(-0.0226,0.00071)	8.60	4.49	29.72	6.7×10^{-4}
PRS.q4(0.00071,0.0241)	16.56	4.50	60.96	2.44×10^{-5}
Sex (male)	0.078	0.054	0.114	2.0×10^{-42}
HLA AA	reference			
HLA AG/GA	2.43	1.75	3.39	1.5×10^{-7}
HLA GG	3.46	1.31	9.13	0.012

Table 4

OldIT	non-HLA SNP	Tag HLA Snp	Sex	AUC (95% CI)
PRS + sex + HLA + 10 PC	x	x	x	0.8273 (0.8049-0.8497)
PRS + sex + 10 PC	x		x	0.8138 (0.7907-0.8368)
PRS + HLA + 10 PC	x	x		0.7223 (0.6941-0.7505)
PRS + 10 PC	x			0.6909 (0.6619-0.7199)
HLA + 10 PC		x		0.6302 (0.599-0.6613)

Table 5

Chr	Snp	Bp
1	1:201019059:C:T	2523723
1	1:2523723:C:T	157670290
1	1:157670290:C:G	201019059
2	2:135341200:A:G	25514333
2	2:25514333:C:T	135341200
3	3:16961265:A:G	16961265
3	3:159733527:G:T	159733527
5	5:158759900:A:G	35881130
5	5:35881130:G:GT	158759900
6	6:137973068:A:G	137973068
7	7:128617466:A:G	128617466
7	7:138729543:G:GAAT	138729543
9	9:100741912:G:GA	100741912
11	11:646986:T:TAA	646986
11	11:118740104:A:AT	64110422
11	11:64110422:A:T	111239365
11	11:111239365:C:G	118740104
12	12:111907431:A:AC	111907431
14	14:103564807:A:C	103564807
16	16:68036939:A:G	68036939
17	17:38044893:C:CTT	38044893
18	18:67526026:C:T	67526026

Table 6

	OR	lower	upper	p-value
PRS.q4(-0.0628,-0.0414)	reference			
PRS.q4(-0.0414,-0.02)	2.83	0.94	8.53	0.06
PRS.q4(-0.02,0.0013)	4.92	1.65	14.70	5.28 x 10 ⁻³
PRS.q4(0.0013,0.0227)	17.22	4.99	59.38	1.57 x 10 ⁻⁵
Sex (male)	0.11	0.07	0.17	5.0 x 10 ⁻²³
HLA AA	reference			
HLA AG/GA	0.40	0.25	0.65	1.6 x 10 ⁻⁴
HLA GG	0.24	0.14	0.40	3.7 x 10 ⁻⁸

Table 7

NewIT	non-HLA SNP	Tag HLA SNP	Sex	AUC (95% CI)
PRS + sex + HLA + 10 PC	x	x	x	0.8138 (0.7832-0.8445)
PRS + sex + 10 PC	x		x	0.7979 (0.7667-0.8291)
PRS + HLA + 10 PC	x	x		0.7091 (0.6706-0.7476)
PRS + 10 PC	x			0.6842 (0.6446-0.7238)
HLA + 10 PC		x		0.6523 (0.6108-0.6939)

Figure 1

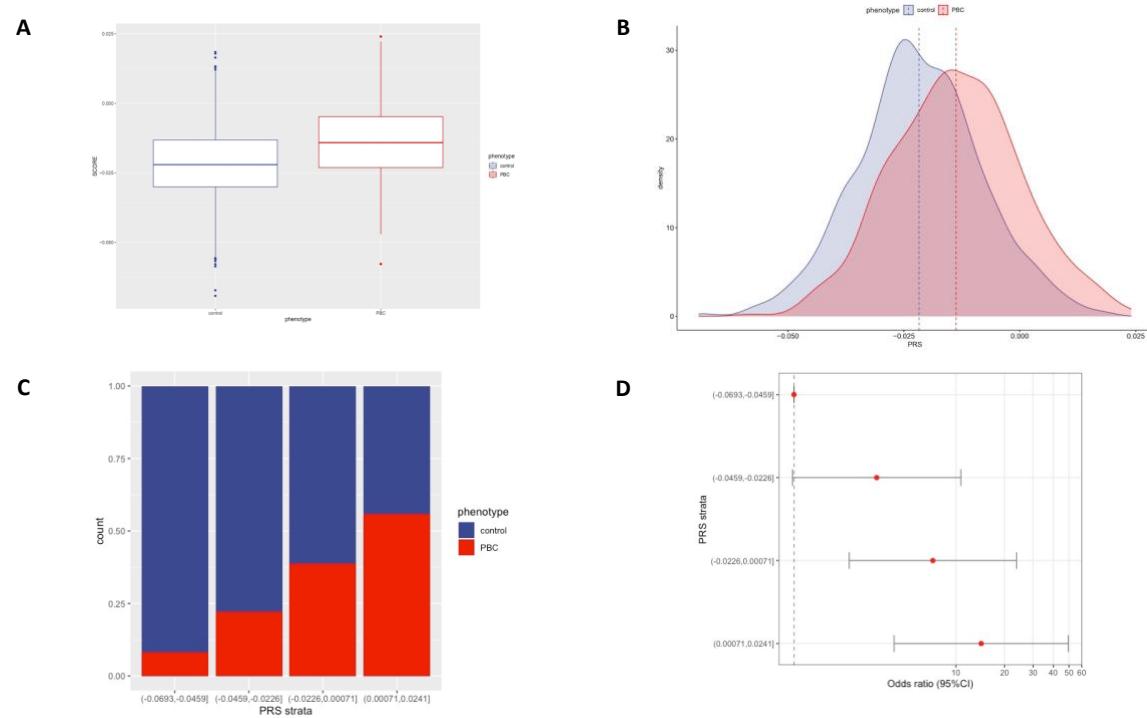


Figure 2

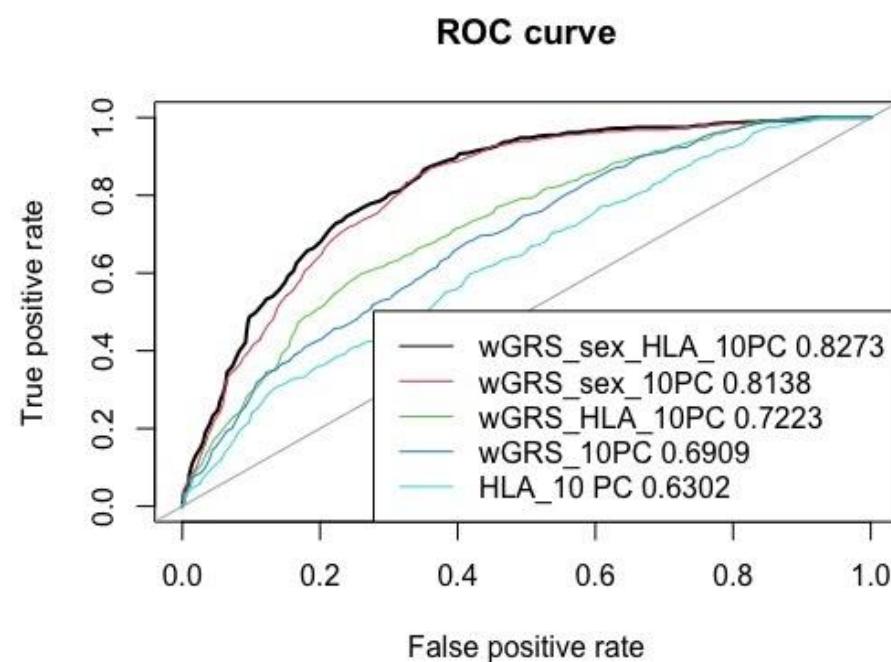


Figure 3

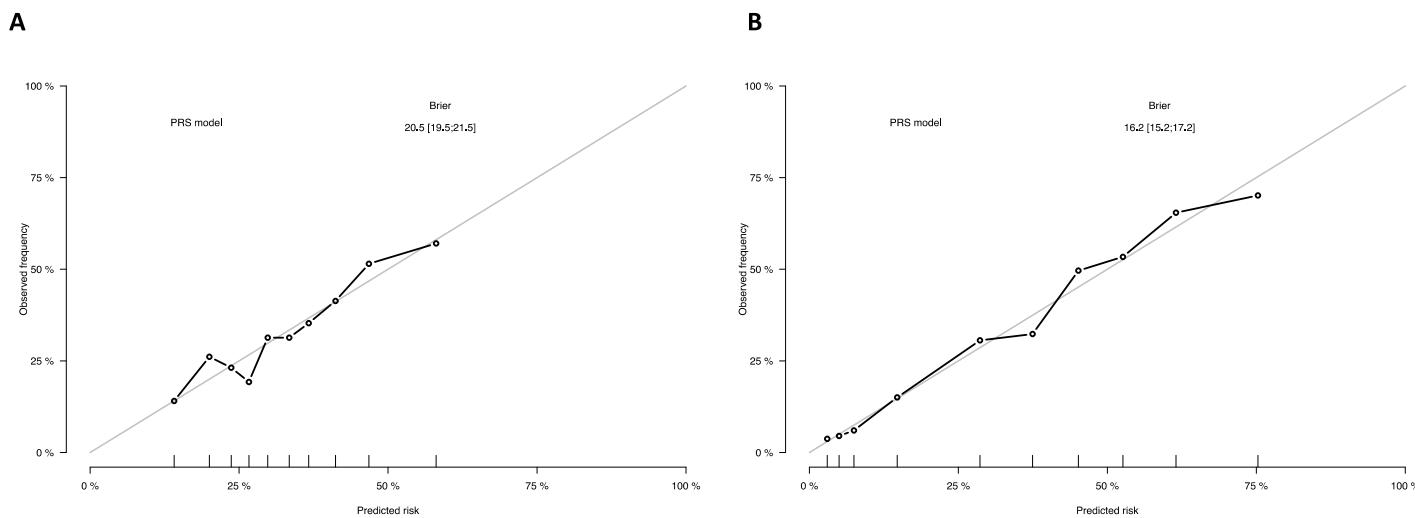


Figure 4

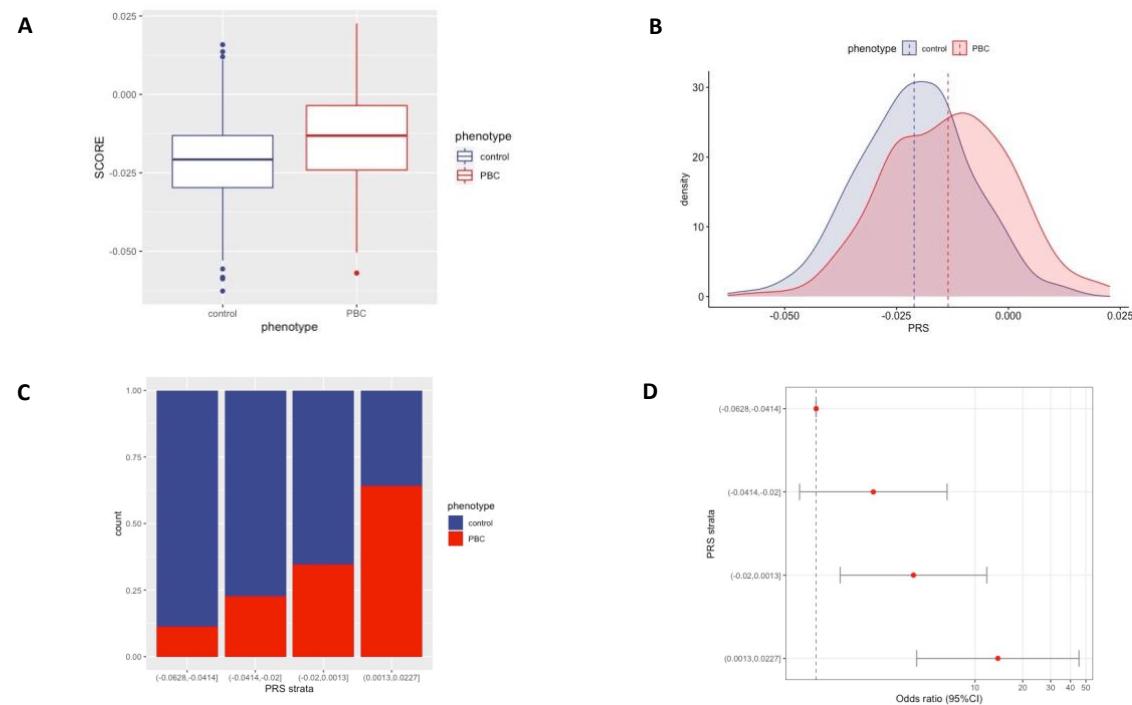


Figure 5

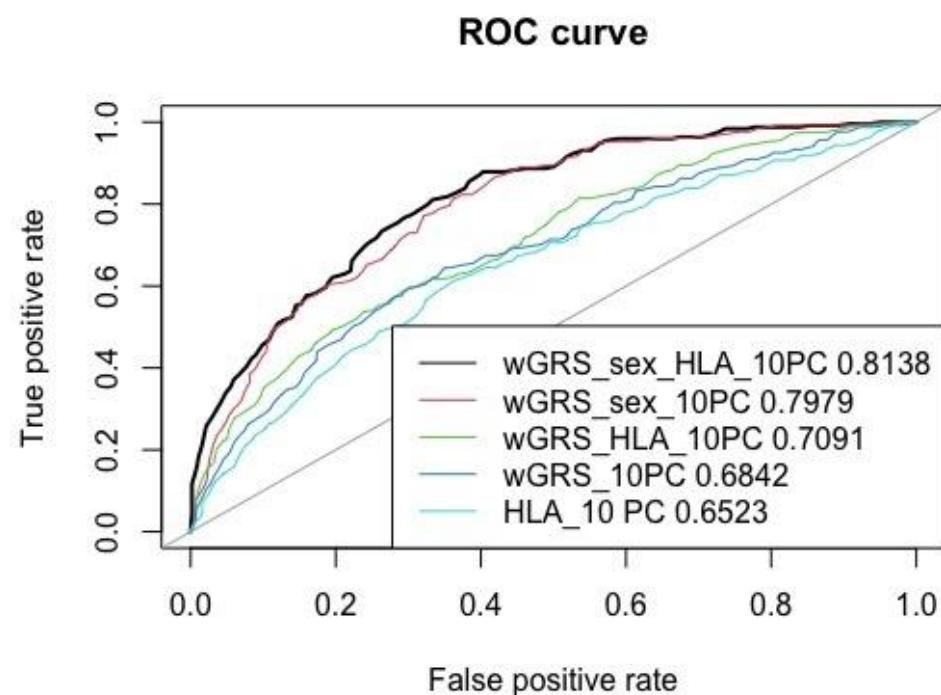
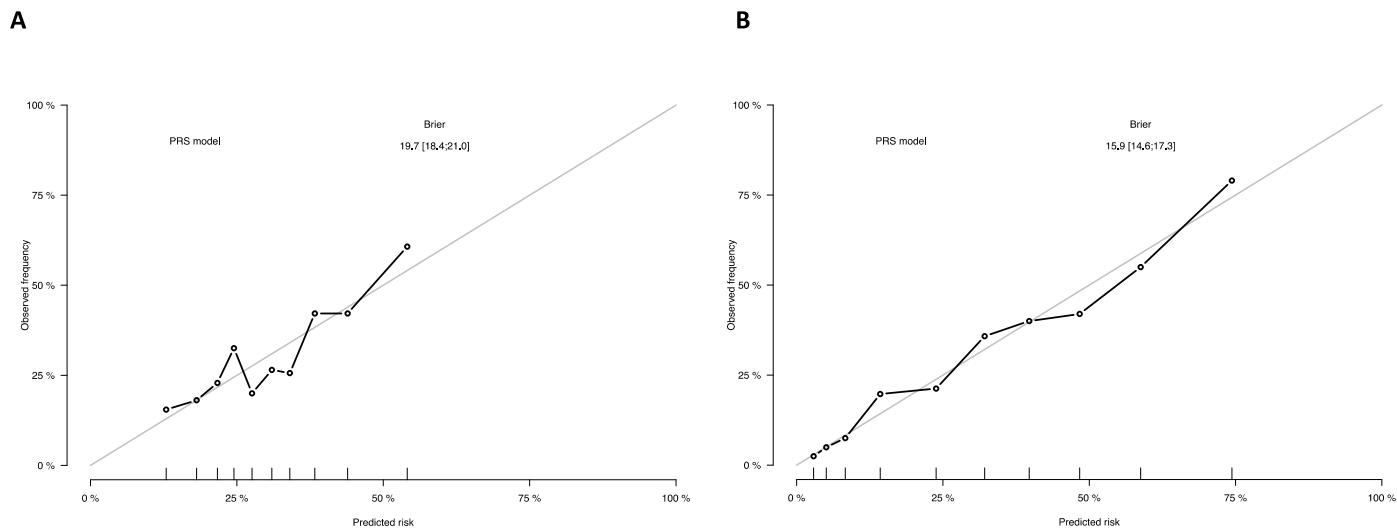


Figure 6



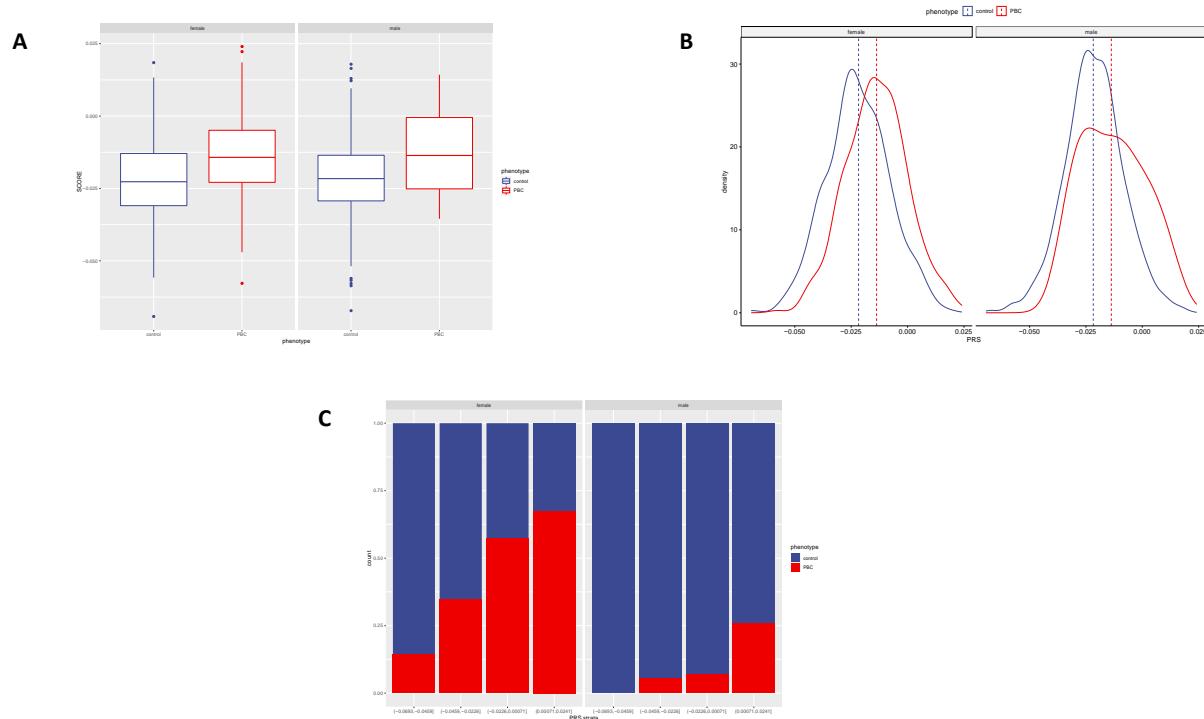
Supplementary Table 1

	OR	lower	upper	p-value
PRS.q4(-0.0693,-0.0459)		reference		
PRS.q4(-0.0459,-0.0226)	3.242	0.979	10.742	0.068
PRS.q4(-0.0226,0.00071)	7.210	2.194	23.696	3.24×10^{-4}
PRS.q4(0.00071,0.0241)	14.340	4.159	49.450	1.08×10^{-6}

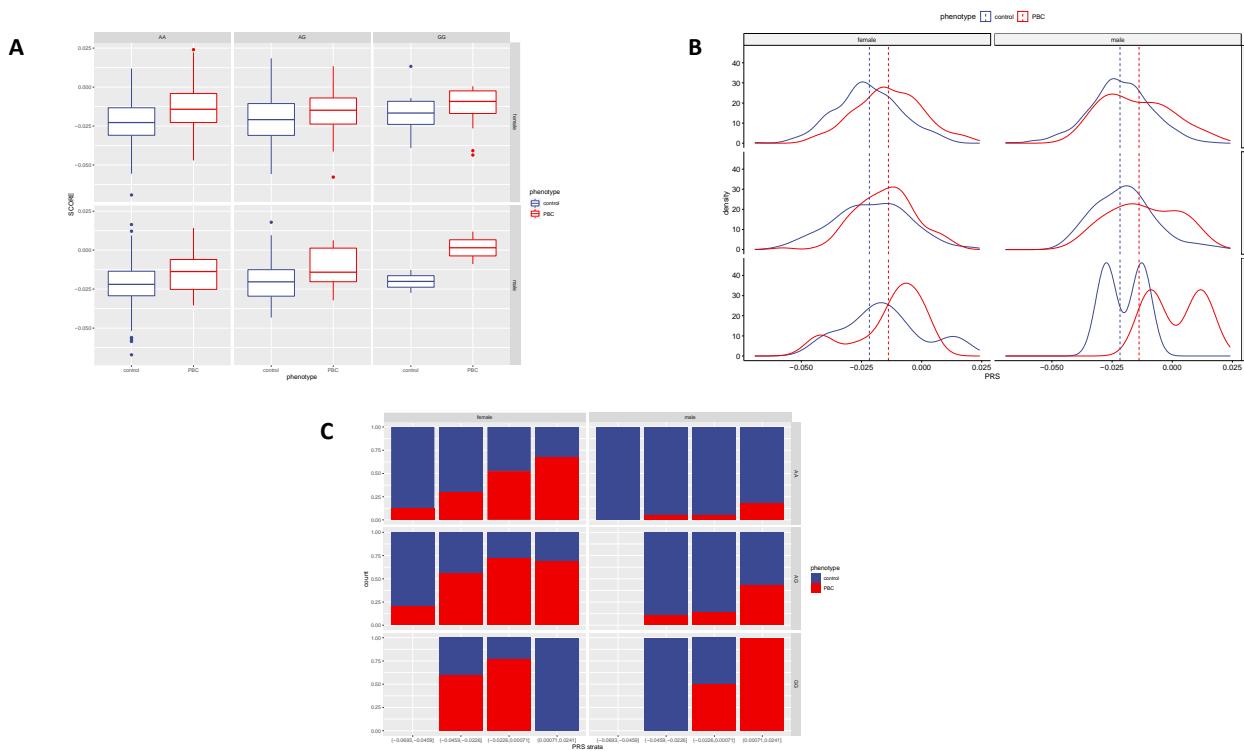
Supplementary Table 2

	OR	lower	upper	p-value
PRS.q4(-0.0628,-0.0414)	reference			
PRS.q4(-0.0414,-0.02)	2.291	0.786	6.682	0.18
PRS.q4(-0.02,0.0013)	4.105	1.419	11.873	8.97 x 10 ⁻³
PRS.q4(0.0013,0.0227)	13.950	4.304	45.218	2.30 x 10 ⁻⁶

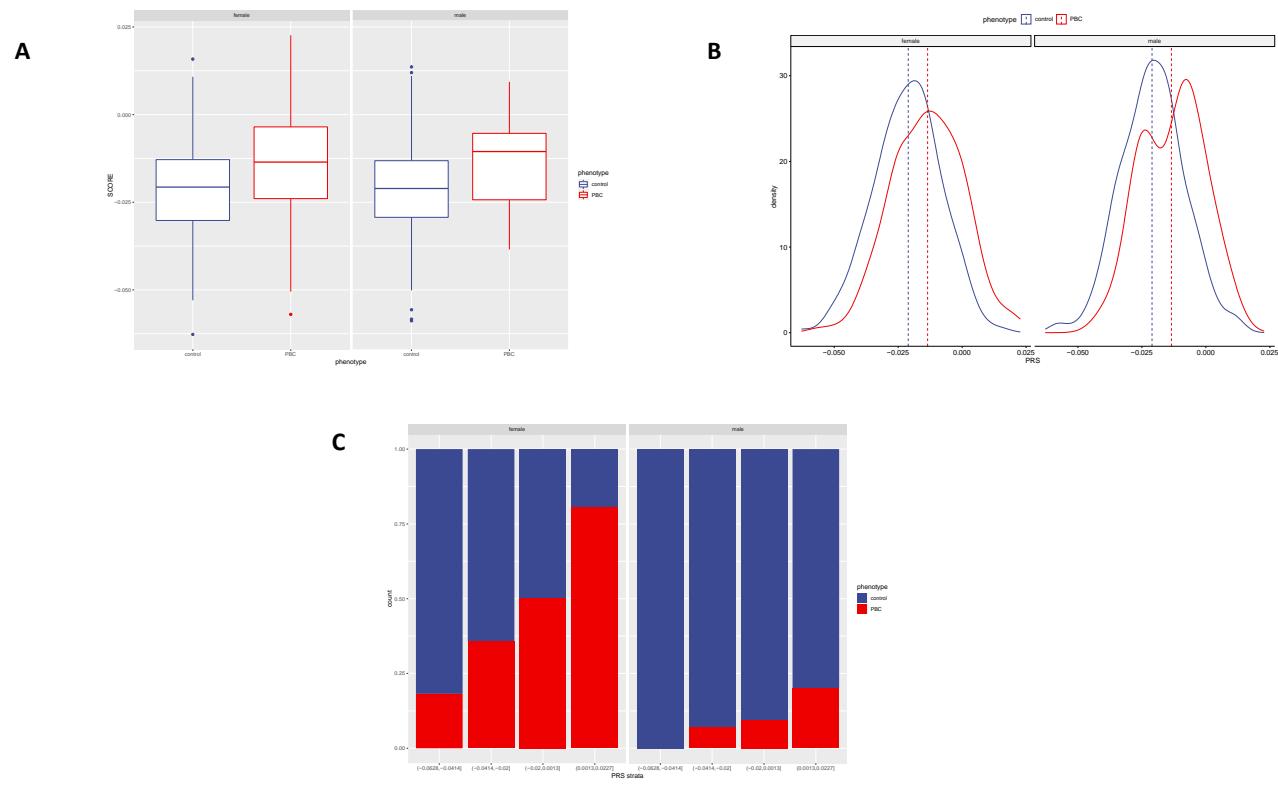
Supplementary Figure 1



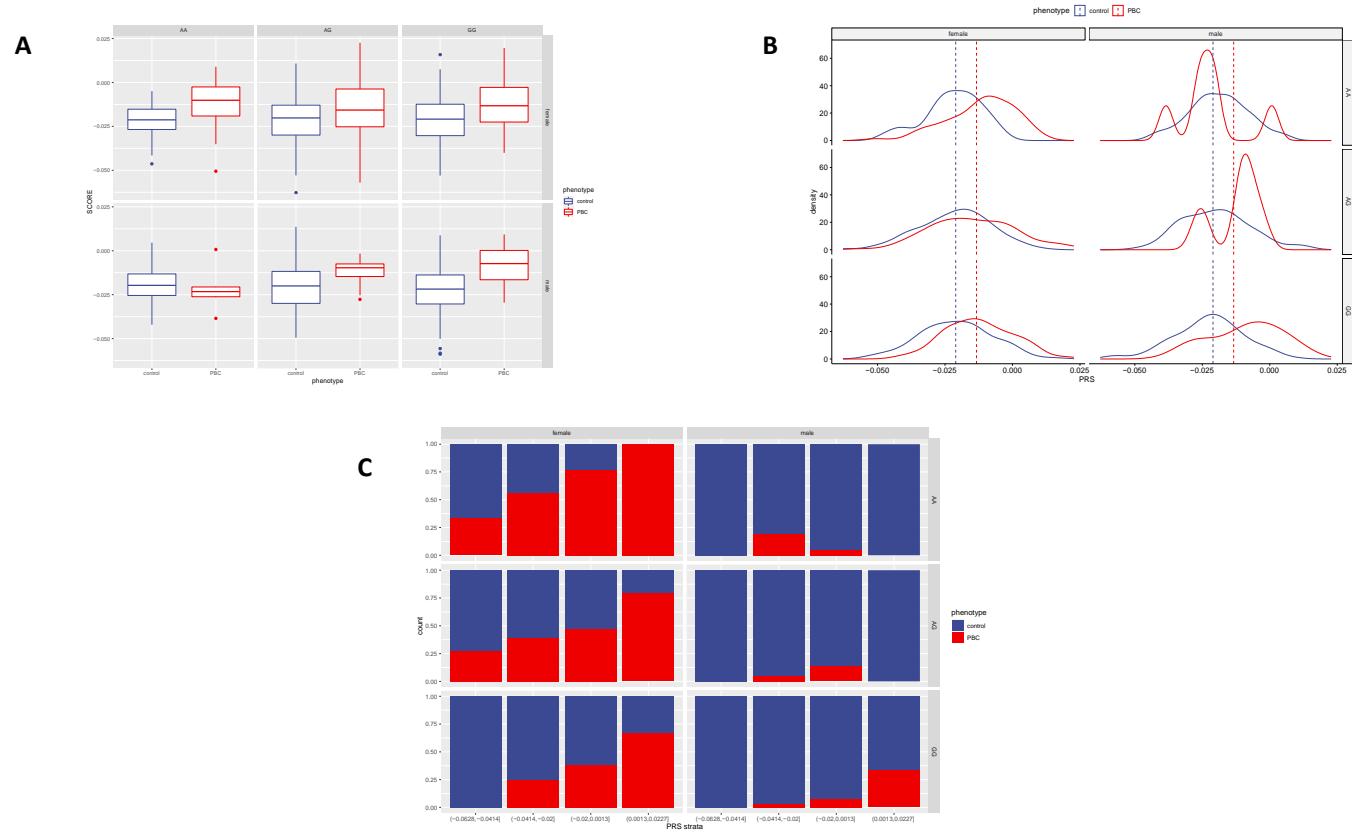
Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 4



References

1. Gerussi, A., Carbone, M., Corpechot, C. & Schramm, C. The genetic architecture of primary biliary cholangitis. *Eur. J. Med. Genet.* **64**, 104292 (2021).
2. Cordell, H. J. *et al.* An international genome-wide meta-analysis of primary biliary cholangitis: novel risk loci and candidate drugs. *J. Hepatol.* (2021). doi:<https://doi.org/10.1016/j.jhep.2021.04.055>
3. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **3**, 11–13 (2020).
4. Tang, R. *et al.* The cumulative effects of known susceptibility variants to predict primary biliary cirrhosis risk. *Genes Immun.* **16**, 193–198 (2015).
5. Cordell, H. J. *et al.* International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.* **6**, 8019 (2015).
6. Wand, H., Lambert, S. A., Tamburro, C. & Iacocca, M. A. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**, (2021).
7. Lindor, K. D. *et al.* Primary biliary cirrhosis. *Hepatology* **50**, 291–308 (2009).
8. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7

(2015).

9. Team, R. C. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* (2016).
10. McCullagh, P. & Nelder, J. A. *Generalized linear models*. (Routledge, 2019).
11. Mittlböck, M. & Schemper, M. Explained variation for logistic regression. *Stat. Med.* **15**, 1987–1997 (1996).
12. McFadden, D. The measurement of urban travel demand. *J. Public Econ.* **3**, 303–328 (1974).
13. Khera, A. V *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
14. European Association for the Study of the Liver. EASL Clinical Practice Guidelines: The diagnosis and management of patients with primary biliary cholangitis. *J. Hepatol.* **145**, 167–172 (2017).
15. Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat. Med.* **27**, (2021).
16. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
17. 100, 000 Genomes Project Pilot Investigators *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N. Engl. J. Med.* **385**,

- 1868–1880 (2021).
- 18. Jiang, X. *et al.* A heterozygous germline CD100 mutation in a family with primary sclerosing cholangitis. *Sci. Transl. Med.* **13**, eabb0036 (2021).
 - 19. Gerussi, A., Cristoferi, L., Carbone, M., Asselta, R. & Invernizzi, P. The immunobiology of female predominance in primary biliary cholangitis. *J. Autoimmun.* **95**, 124–132 (2018).
 - 20. Asselta, R. *et al.* X Chromosome Contribution to the Genetic Architecture of Primary Biliary Cholangitis. *Gastroenterology* 2483–2495 (2021). doi:10.1053/j.gastro.2021.02.061
 - 21. Bianchi, I., Lleo, A., Gershwin, M. E. & Invernizzi, P. The X chromosome and immune associated genes. *J. Autoimmun.* **38**, J187–J192 (2012).
 - 22. Moore, J. H. & Williams, S. M. Epistasis and Its Implications for Personal Genetics. *Am. J. Hum. Genet.* **85**, 309–320 (2009).
 - 23. Zhang, P. & Lu, Q. Genetic and epigenetic influences on the loss of tolerance in autoimmunity. *Cell. Mol. Immunol.* 1–11 (2018). doi:10.1038/cmi.2017.137
 - 24. Marzorati, S., Lleo, A., Carbone, M., Gershwin, M. E. & Invernizzi, P. The epigenetics of PBC: The link between genetic susceptibility and environment. *Clin. Res. Hepatol. Gastroenterol.* **40**, 650–659 (2016).
 - 25. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations.

Nat. Rev. Genet. **20**, 520–535 (2019).

CHAPTER 5

Machine-learning SNP-based prediction for Primary Biliary Cholangitis: a proof-of-concept study

Alessio Gerussi^{1,2}, Damiano Verda³, Claudio Cappadona^{4,5},
Laura Cristoferi^{1,2,6}, Davide Paolo Bernasconi⁶, Marco
Carbone^{1,2}, Marco Muselli³, Pietro Invernizzi^{1,2}, Rosanna
Asselta^{4,5}

¹Division of Gastroenterology, Center for Autoimmune Liver Diseases, Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy

²European Reference Network on Hepatological Diseases (ERN RARE-LIVER), San Gerardo Hospital, Monza, Italy

³Rulex Inc., Newton, MA, USA – www.rulex.ai

⁴Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, 20072 Pieve Emanuele, Milan, Italy

⁵IRCCS Humanitas Clinical and Research Center, Via Manzoni 56, 20089 Rozzano, Milan, Italy

⁶Bicocca Bioinformatics Biostatistics and Bioimaging Centre - B4, School of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy.

Manuscript in preparation

PhD Candidate contribution: leader in the conceptualization of the study, data curation, analytic process, interpretation of the results and writing the manuscript.

Abstract

Background and Aims: The application of Machine Learning (ML) to genetic individual-level data represents a foreseeable advancement for the field. Here, we aimed to evaluate the feasibility and accuracy of a ML-based model for disease risk prediction applied to Primary Biliary Cholangitis.

Methods: Genome-wide significant variants identified in subjects of European ancestry in the recently released 2nd international meta-analysis of GWAS in PBC were used as input data. Quality-checked, individual genomic data from the two Italian cohorts were used. The main analytic steps were the following: import of genotype and phenotype data, feature selection, supervised classification of phenotype by genotype, generation of “if-then” rules for disease prediction by logic learning machine (LLM), and model validation in a different dataset. Ten-fold cross-validation was performed for validation. PLINK 1.9 was used for extraction of variants of interest and preliminary QC steps, while the Rulex software was used to build the ML model.

Results: The training set included 1,345 individuals, 444 were PBC cases and 901 healthy controls. After preprocessing, 41,899 variants entered the analysis. Several configurations of parameters related to feature selection were simulated. The final model had Accuracy of 71.7%, a Matthews value of 0.29, a Youden’s value of 0.21, a sensitivity of 0.28, a specificity of 0.93, a positive predictive value of 0.66 and a negative predictive value of 0.72. Five rules were generated with a covering > 20%. The

rule with the highest covering (22.5) included the following features: female sex AND chr2:191943742 = T AND chr3:119116150 = G AND chr3:11925780 = A AND chr16:86073712 = TTG AND chr17:38057189 = A. The genes involved in the best rule were: *STAT4*, *TIMMDC1*, *LOC105371388*, and *ZPBP2*.

The validation cohort included 834 individuals, 255 cases and 579 controls. By applying the ruleset derived in the training cohort, the Area under the Curve of the model was 0.73.

Conclusion: This study represents the first illustration of a successful analysis of genome-wide association analysis data with ML to study genetic liability of PBC. ML is computationally feasible, generates accurate information that incorporates gene-gene interactions, and can be used for disease prediction in at-risk individuals.

Introduction

Precision medicine aims to tailor diagnosis, follow-up and management of individuals based on their genetic and environmental conditions¹. To accurately model the genetic risk to develop a disease is challenging for complex traits like Primary Biliary Cholangitis (PBC), a rare autoimmune disease of the liver^{2–4}. PBC is characterized by an autoimmune pathogenesis and strong genetic predisposition, with major histocompatibility complex (MHC) class-II haplotypes and non-MHC loci contributing to the genetic risk^{4–6}.

Several Genome-wide association studies (GWAS) have been published associating more than 60 variants with PBC^{4,7}. The polygenic architecture of the disease, together with high levels of heritability⁴ represent a solid rationale to develop polygenic risk scores (PRSs). Yet, PRSs assume that each variant has a linear additive effect on disease⁸; many authors suggested that the limited success in complex disease predictions of PRS is due to their dependency on linear regression^{9,10}. In addition, PRSs typically provide a relative measure of risk evaluated at the level of a group of people but not at individual level^{10,11}.

Machine learning (ML) algorithms can be trained to model the genetic risk of a complex trait, with theoretical advantages related to their ability to handle high-dimensional data by nonlinear effects and the use of individual data^{10,12}. ML algorithms employ minimal a priori assumptions about the nature of the genetic

effects being modeled, potentially taking into account also gene-gene interactions with non-additive effects on disease risk^{10,13}. Our aim was to evaluate the feasibility and accuracy of a ML-based model for genetic risk prediction of PBC.

Methods

Study Design and Participants

All cases met internationally accepted criteria for PBC¹⁴. Training set (“OldIT”) was composed of 1,345 individuals of Italian ancestry, 444 PBC cases and 901 healthy controls; 515 were males and 830 females⁷. Validation set (“NewIT”) was made of 834 individuals of Italian ancestry, 255 cases and 579 controls; 335 were males and 499 females⁷ (see Chapter 2 and 4).

This study included quality-checked, imputed genotype data derived from the recently published international meta-analysis⁷. For a detailed description of the cohorts under analysis we refer to the methods sections of the meta-analysis by Cordell *et al*⁷ in chapter 2, where the time span of data collection, the collection site and setting, relevant population characteristics and any inclusion or exclusion criteria used in original studies can be retrieved.

Patient privacy and ethical issues

All participants gave written informed consent for genetic studies. The research conformed to the ethical guidelines of the 1975 Declaration of Helsinki. The protocol was approved by each participating centre in accordance with local regulations. As far as the application of ML to clinical data is concerned, the Rulex proprietary software used in this work is compliant with the

strictest data privacy regulations, such as the European Union's General Data Protection Regulation (GDPR). GDPR allows automated ML predictions only if the clear explanation of the logic used to make each decision is provided, which is difficult with black box models.

This section of methods adheres to the guidelines and quality criteria for artificial intelligence-based prediction models in healthcare¹⁵.

Data preprocessing

Selection of genetic variants

The recent international meta-analysis has identified 57 loci⁷ associated with PBC at genome-wide level of significance in patients with European and East Asian ancestry. For the current study, we selected those at genome-wide significant threshold in Europeans, eventually including 46 loci (**Table 1**).

To take into account the linkage disequilibrium (LD) structure of the regions harboring the selected loci (hence avoiding to miss genetic information), for each genome-wide significantly associated locus, a region spanning $\pm 250\text{kb}$ upstream and downstream of the corresponding coding sequence was considered. We hence extracted SNP genotype data from the imputed set of data using PLINK 1.9¹⁶.

Import of PLINK files in the Rulex environment

Rulex is a novel ML software able to make intelligible predictive models (www.rulex.ai). The Logic Learning Machine (LLM) is the core machine learning algorithm of Rulex and represents a method of supervised data mining based on an efficient implementation of the Switching Neural Network model. Rather than producing a math function, the LLM produces conditional logic rules, fulfilling the definition of explainable Artificial Intelligence (AI)¹⁷, as opposed to deep learning and other “black-box” AI algorithms^{18,19}. A list of published works using Rulex in the field of biology and medicine can be found here^{20–27}. Further technical material is available upon request.

PLINK files (.map and .ped) were imported into the Rulex environment by the Import command from text operator and parsed accordingly (**Figure 1 - Import and Parsing**). Sex and a pre-selected top HLA variant (see also chapter 4) were used as additional features.

Model development

Feature selection

The output of the study was the variable Status identifying whether a subject of the study was a PBC patient (case) or a control. Features with a number of mode values above 95% were removed by the Rulex Fill/Clean operator ((**Figure 1 - Import and Parsing**). To avoid redundancy in input features, where

redundancy stands for high correlation, Rulex was used to rank all available features versus the disease status by univariate association based on Cramer's V. After this univariate ranking, all the features entered a greedy forward selection process (**Figure 1 - Feature Selection**).

Greedy forward selection is a popular technique for feature subset selection²⁸. The main advantages of this approach are its simplicity and its computational scalability, which makes it applicable to many practical problems, including the most complex ones. The algorithm starts with an empty set of features; then, the additional feature is added iteratively to the set, provided that it meets a predefined performance measure.

Since our aim was to avoid preliminary multivariate steps before LLM, we avoided the use of indicators such as the Akaike Information Criterion (AIC)²⁸, since it would implicitly introduce an auxiliary multivariate model by selecting features according to their performance and number. In this way, the landscape of input features was filtered only based on direct correlations (between the input features with the output and among input features with each other). At each step, the selected feature was the one with the highest correlation with the output, provided that its correlation with any of the already selected inputs did not exceed a threshold value t .

The second parameter that was used for tuning was the maximum number of features n to be selected. The procedure terminated and no additional features were considered for

inclusion if, after a given iteration of the described greedy procedure, n features were included.

Internal validation

The final list of variants at the end of the feature selection process was used as input for the Logic Learning Machine operator, to build intelligible rules to predict the disease status. The Rulex “Logic Learning Machine” (LLM) operator was nested within a process working in cross-validation classifying features associated with the output of interest (**Figure 1 - Process working in cross-validation**). For internal validation, LLM operated under a 10-fold cross-validation approach. Cross-validation involves partitioning a sample of data into complementary subsets, performing analyses on one subset (the training set) and validating the analyses on the other subset (the test set). The time for a complete run (from import to rules generation) was on average around 2 hours and half on a 64gb working station.

The output of each run of the model was a list of rules (called ruleset). A metrics section of the pipeline was dedicated to the evaluation of the performance of the rulesets in training and test sets (**Figure 1 - Metrics**).

Hyperparameter Tuning

Several sets of parameters were evaluated before choosing the final optimal model (**Figure 1 - Hyperparameters**). The parameters undergoing tuning were:

- correlation among features*: three thresholds t based on Cramer's V value were evaluated (0.7, 0.8 and 0.9);
- number of pre-selected and selected features*: several fixed combinations of thresholds for pre-selected features and selected features were evaluated (**Supplementary Table 1**);
- max error (errmax)*: $errmax$ represents the maximum level of error for each rule included in the ruleset. In other words, this corresponded to the maximum percentage of cases belonging to output classes different from the predicted one which verified the rule.

Changes in the performance of the model after modifications of each parameter was evaluated based on the following metrics: *Sensitivity*, *Specificity*, *Positive Predictive Value*, *Negative Predictive Value*, *Accuracy*, *Matthews*, *Youden's index* (**Figure 1 - Metrics**).

Sensitivity was defined as:

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Specificity was defined as:

$$\frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Positive Predictive Value was defined as:

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Negative Predictive Value was defined as:

$$\frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}}$$

Accuracy was defined as:

$$\frac{\text{True Positives} + \text{True Negative}}{\text{True Positives} + \text{False Negative} + \text{False Positives} + \text{True Negative}} \times 100$$

The *Matthews correlation coefficient* was defined as:

$$\frac{\text{True Positives} \times \text{True Negatives} - \text{False Positives} \times \text{False Negatives}}{\sqrt{(\text{True Positives} + \text{False Positives})(\text{True Positives} + \text{False Negatives})(\text{True Negatives} + \text{False Positives})(\text{True Negatives} + \text{False Negatives})}}$$

The *Youden's Index* was defined as:

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negative}} + \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} - 1$$

Final model selection and rules generation

After 814 runs of the Rulex process, a full list of metrics including all hyperparameters and accuracy metrics was available to choose the best model. After choosing the combination that maximized the accuracy and specificity of the model, the model was re-run with this parameters set and the final ruleset was generated ((**Figure 1 - LLM Final Model**)).

The quality of each rule was then evaluated based on some metrics specific to the Rulex software. “Covering” is the percentage of samples belonging to the class described by the rule, fulfilling that specific rule; “Error” is the percentage of samples belonging to the other classes fulfilling that specific rule. A Feature ranking plot was also generated to help discriminating the most relevant features.

Validation of the model

To further improve the robustness of our conclusions and reduce the risk of overfitting, we also performed an external validation on a dataset (“NewIT”) that was used only for this purpose (**Figure 2 - Import the validation set**). This dataset will be referred to as the forecast or validation set.

Since the training and the forecast datasets did not include exactly the same features, not all the rules of the final model could be applied into the forecast dataset. To minimize the reduction in prediction power of the model due to this heterogeneity, we adopted a multi-step process. We decided not

to extract a model in the training set based only on the shared features, in order to derive a single model that can be adapted to different forecasting datasets, enhancing model plasticity and generalization.

To do so, a count of the conditions present in the final ruleset not verified in the forecast set due to missing data was calculated. Then, a frequent pattern mining auxiliary layer was built, leveraging the notion that many input features are highly correlated with each other based on LD, so that if a variant is missing another variant in LD could ideally be retrieved²⁸. More specifically, for each condition c , the frequent pattern mining branch identified which conditions, among the shared ones, was the most correlated to c . For biological plausibility, only conditions located on the same chromosome as c and not more than 500000 base pairs were considered as candidate conditions for replacement. Let us refer to the condition (among the candidate ones) which is more correlated to c as c' . The all-confidence score was used as a ranking metric for identifying the most correlated condition²⁹. Rules constituting the model were adapted by substituting each condition c with c' provided that the all-confidence score measuring their correlation met a minimum threshold of 0.9 (**Figure 2 - Identify missing conditions in the validation set - change with correlated variants**).

Finally, the adapted ruleset was applied to the forecasting dataset. The application of this ruleset produced a forecast score, ranging from 0 to 1, for each of the considered individuals. For each subject, the forecast score was initialized to 0.5 and it

increased or decreased according to the rules included in the model that the subject meets. For instance, if a patient verified all the rules that predicted to be a case and no rule that predicted to be a control, its forecast score would be 1. Conversely, if the patient verified all the rules that predicted to be a control and no rules that predicted to be a case, its forecast score would be 0. Comparison between median scores between cases and controls were performed by Wilcoxon signed-rank test. Diagnostic accuracy was evaluated using receiving operator characteristic (ROC) curves. Area under the ROC curve (AUC) is reported together with its 95% CI. Calibration was assessed after calculating risk predictions according to a logistic regression model, which included the continuous forecast score. Individual predicted risks were then divided into ten equally sized categories (i.e. according to deciles). A calibration plot was then produced by comparing the mean predicted risk in each decile (displayed in the x axis) with the observed risk, calculated as the proportion of PBC cases within each decile (displayed in the y axis). Brier score, corresponding to the mean squared error of the prediction, was also calculated together with its 95% CI (**Figure 2 - Run new LLM model in the validation set**).

Results

Description of the training cohort

The training set (“OldIT”) was composed of 1,345 individuals of Italian ancestry, 444 PBC cases and 901 healthy controls. 515 were males and 830 females. The total number of SNPs in the training cohort was 105,150 variants.

Feature Selection

After removal of variants with > 95% of constant values, 41,899 variants were brought forward into the feature selection process. Univariate association between each of the 41,899 variants and the output (case vs control) was performed and features were ranked according to Cramer’s v (**Supplementary Table 2**).

After univariate analysis, an iterative greedy procedure was performed to avoid the inclusion of input features strongly correlated to each other. Different values of hyperparameters (npresel, nsel, errmax, correlation) were evaluated in training and test sets (where the test represented the one-tenth randomly selected portion of the training dataset under cross validation). A total of 814 runs of the workflow were performed for hyperparameter optimization (**Supplementary Table 3**).

The evaluation of the performance of the model was done on the test set, to avoid overfitting the model by choosing the best model on the training set. The configuration that maximized specificity with a good balance in terms of accuracy and precision in the test

set had npresel = 872, nsel = 266, errmax = 0.05 and correlation = 0.8. The full summary of the 814 runs with metrics is presented in **Supplementary Table 4**. The list of selected variants that entered the classification model as input features is shown in **Table 2**.

Final Model

The final LLM model generated 38 rules to classify disease status. The LLM model reached an Accuracy of 71.7%, a Matthews correlation coefficient of 0.29, a Youden's value of 0.21, a Sensitivity of 0.28, a Specificity of 0.93, a Positive Predictive Value of 0.66 and a Negative Predictive Value of 0.72. Covering of rules ranged from 0.45 to 43.28, with a median value of 5.30 (IQR 2.59, 12.61); error ranged from 0.00 to 5.86, with a median value of 3.41 (IQR 1.50, 4.76)

The rule with the highest covering (19.14) predicting PBC was rule 4, with error equal to 3.88. Rule 4 included the following 13 features: female sex AND chr14:92932650 = C AND chr17:43906828 = G AND chr17:43912635 = A AND chr17:44038536 = CA AND chr17:44040823 = C AND chr17:44065263 = T AND chr17:44183317 = C AND chr17:44185431 = T AND chr17:44222335 = G AND chr17:44283022 = A AND chr3:119111870 = T AND chr7:128705730 = T (**Table 3**). The genes involved in the best rule 4 were the following: *RIN3*, *KANSL1*, *TIMMDC1*,

TNPO3. The covering and error of each condition is reported in **Supplementary Table 5**.

The most informative rule that did not include non-genetic information was rule 11, including seven conditions: chr17:38020058 = AC AND chr17:38049589 = T AND chr17:38070071 = C AND chr17:43933579 = C AND chr2:135188248 = A AND chr2:25332696 = C AND chr3:159726324 = C (**Table 4**). The genes involved in rule 11 were the following: *TNPO3*, *KANSL1*, *TMEM163*, *RARB* and *IL12A-AS1*. The covering and error of each condition is reported in **Supplementary Table 6**.

Feature ranking outlines the most relevant variants that have been used by the model to classify the output (**Figure 3**). As expected by the known female predominance of the disease, sex ranked first and its relevance was 0.81, 11.5 times more relevant than the second feature (the HLA SNP) and 13.5 times more relevant than the third one (the first non-HLA SNP).

Among genetic variants, the HLA SNP ranked second, with a relevance of 0.07. Among non-HLA variants the best ones were chr19:50924093 (*SPIB*) (0.06), chr3:119111870 (*TIMMD1*) (0.06) and chr2:191943742 (*STAT4*) (0.05).

Forecast in the validation cohort

The validation set (“NewIT”) included 834 individuals of Italian ancestry, 255 cases and 579 controls; 335 were males and 499 females. In the validation cohort, 74,484 variants were included.

The number of variants shared between the training and the validation cohorts was 64,918. Since conditions in rules are connected by a boolean AND, if a condition was not found within the validation set the whole rule could not be applied. Out of 139 different unique rule conditions, corresponding to 139 different features, 16 (11.5%) were not verified in the validation dataset. Therefore, 28/38 (74%) rules of the original ruleset were effectively applied in the validation cohort.

The median score in cases was significantly higher than controls (0.52 (IQR 0.50-0.56) vs 0.43 (IQR 0.28,0.50) ($p < 0.001$) (**Figure 4A**); for higher scores the number of cases increased consistently (**Figure 4B**). By applying the ruleset derived in the training cohort, the AUC of the model was 0.73 (95% CI 0.69-0.76); the ROC curve in the validation set is presented in **Figure 4C**. The LLM model reached in the validation set an accuracy of 68.5%, a Matthews correlation coefficient of 0.28, a sensitivity of 0.55, a specificity of 0.74, a positive predictive value of 0.48 and a negative predictive value of 0.79. The LLM model produced individual predicted risks that were in strong agreement with the observed risks, indicating a good level of calibration (**Figure 4D**).

Discussion

Our study shows that a ML-based model generated with Rulex to predict genetic susceptibility to PBC is: 1) computationally feasible; 2) methodologically innovative; 3) explainable; and 4) accurate and well calibrated.

More specifically, the Rulex workflow operated within three hours, generated intelligible if-then rules for prediction, considered linked groups of variants instead of single variants alone, and achieved high accuracy in discriminating between PBC cases and controls in a new validation cohort.

As regards methodology employed, the pipeline presented in this study utilized individual data and not summary statistics. It did not evaluate a statistical imbalance of the allele frequency of a variant between cases and controls, but rather how the concomitant presence of genetic variants having a specific DNA base would associate to discriminate between cases and controls. Multivariate computational approaches, such as Rulex, may be able to capture the complex relationships among risk variants for complex traits, which represents a possible innovation for the field¹⁰. Based on the non-linearity of switching neural networks at the core of Rulex LLM³⁰, we can infer that Rulex did approach genetic information in a different manner than standard statistical genetic methods do. For instance, the rule with the highest covering included only one SNP among the most significant ones, based on effect sizes and p-values derived from previous studies; in other words, univariate statistical

significance was not the only parameter to consider for the ML model to predict disease status. Groups instead of single attributes were linked to case/control prediction by Rulex LLM, indirectly inferring also complementary relations among inputs. Although our analysis was not aimed to study epistasis specifically, the boolean AND that links conditions within the same rule might represent a proxy for statistical epistatic interaction. Rulex LLM could represent a novel method to improve the way gene-gene interactions are taken into account³¹. Further studies are needed to assess the applicability of Rulex LLM to study epistatic interactions.

As regards model explainability, the LLM generates if-then rules that are easy to understand. There is growing awareness in the scientific community about the importance of model explainability when ML algorithms are applied in the biomedical field^{17,18}. There is also increasing concern related to the possible risk of gender and racial discrimination enhanced by ML algorithms, and this could be more problematic when the user cannot recognize how the algorithm is generating the output³². The explainability of the LLM rules makes the algorithm particularly interesting for future applications in risk stratification, as compared with black-box models like deep learning algorithms characterized by excellent accuracy counterbalanced by low levels of explainability^{18,33}. The high expectations behind ML should not conceal the recognition that ML algorithms have both benefits and risks; both the architecture of the algorithm and its use should be judiciously balanced³⁴.

In terms of clinical translation, the low prevalence of PBC in the population makes mass screening not feasible: the target population for SNP genotyping and applications of rules for prediction would be at-risk individuals, such first-degree relatives of PBC patients and AMA-isolated positive subjects, who represent a pre-clinical stage of the disease. A recently presented congress communication³⁵ reported that, in a prospective cohort of first-degree relatives of patients with PBC, the prevalence of PBC was 5 cases out of 231 individuals assessed (0.02). The prevalence of PBC-specific autoantibodies in first-degree relatives was 28 out of 231 (0.12), with a higher prevalence in sisters (0.23). If these data are incorporated in our analysis, the final model would have a Likelihood ratio of 4.0, meaning that if a first-degree relative tested positive the risk of PBC would increase from 2% to 27% (2% + 25%)³⁶. As outlined in Chapter 4, clinical strategies to deal with these individuals (first-degree relatives and AMA-isolated cases) are still preliminary. Our LLM model could represent a valuable tool to allocate attention and resources across individuals with higher levels of genetic risk³⁷.

Our work has some limitations. We did not perform an analysis comparable to a GWAS, because we pre-selected top variants from the meta-analysis shown in Chapter 2⁷, both for computational reasons (a smaller collection of variants) and to follow what has been historically performed for polygenic risk scores (pre-selection based on p-value). This approach could have been over-conservative; the next step will be to expand the

analysis to the whole genome (including sexual chromosomes) to assess scalability of our pipeline and its capacity to discover new variants. We might anticipate that LLM rules would employ as conditions some variants that would not reach the established genome-wide threshold of significance by standard methods, since they might be important for classification based on the non-linear association method behind LLM.

As regards accuracy, the model achieved 73% of accuracy in the test set and AUC of 0.73 in the validation set; in addition, the model showed good calibration, with a Brier score of 18.4. As compared to the PRS shown in Chapter 4, the accuracy of the ML model was lower. The two Italian datasets under study were not completely overlapping in terms of genotyped/imputed variants; despite our efforts to confidently swap genetic variants that were highly correlated to those missing in the validation set did recover most of the rules generated on the training set, we were not able to evaluate the full potential of Rulex LLM. In addition, the pre-selection of top variants may have halted the capability of Rulex to leverage variants with lower effect size to make the model more robust. Bigger samples may also be helpful to improve accuracy, since non-linear interactions are typically data-hungry³¹. Moreover, the greedy feature selection process may have been quicker than other strategies such as exhaustive grid search processes at the expense of robustness. Iterative strategies such as gradient evolution algorithms should also be tested to understand whether accuracy can be improved

keeping complexity, and consequently the computational time, controlled³⁸.

Advocates of PRSs do affirm that they will be more diffuse than ML-models based on individual data⁸ (Choi et al Nat Protocols); the main reason behind this statement is that PRSs use summary statistics and do not need individual data, overcoming ethical and logistic limitations related to genetic data sharing. We acknowledge this limitation, underlining that the two approaches should be complementary; further interdisciplinary research is crucial to better understand the role of ML-models in precision genomics.

Conclusions

This study represents the first illustration of a successful analysis of GWAS data with ML to study genetic liability of PBC. ML is computationally feasible and generates accurate information that can be leveraged for disease prediction.

Our work paves the way for future prospective studies targeting relatives of patients with PBC or isolated AMA-positive subjects and aiming at more intensive follow-up for early identification and timely treatment of new PBC cases.

Table Legends

Table 1. Selected loci at genome-wide significant threshold in Europeans.

Second and third columns report coordinates of the analyzed genomic region (hg38, start and end, respectively).

Abbreviations: Chr, Chromosome.

Table 2. List of selected variants entering Logic Learning Machine (LLM) model.

Table 3. Best rule for prediction including sex as condition.

“Covering” is the percentage of samples belonging to the class described by the rule, fulfilling that specific rule; “Error” is the percentage of samples belonging to the other classes fulfilling that specific rule.

Table 4. Best rule for prediction using only genetic information.

“Covering” is the percentage of samples belonging to the class described by the rule, fulfilling that specific rule; “Error” is the percentage of samples belonging to the other classes fulfilling that specific rule.

Figure Legends

Figure 1. Description of the classification pipeline.

This figure summarizes the classification pipeline in the Rulex environment. Information is imported and parsed and then entered into a feature selection branch working in cross-validation. Several hyperparameters are evaluated, and their value can be tuned by modifications of the associated excel file. After completing the pre-defined number of runs, the process generates a number of metrics, including accuracy measures and feature ranking of attributes. At the top of the figure, the LLM operator can be set with the optimal set of parameters and generates the final ruleset.

Abbreviations: LLM, Logic Learning Machine.

Figure 2. Description of the forecast pipeline.

This figure summarizes the forecast pipeline in the Rulex environment. Information is imported and parsed both for the training (via a link to the previous classification workflow) and the validation sets. The LLM model is applied to the validation set. The conditions that are missing in the validation set are analyzed and a dedicated pipeline operates to find in the training set new features that are present in the validation set and are correlated (at a predefined threshold) to the missing conditions. The LLM model is then re-run in the training set to evaluate the change of performance of the model after these changes. Finally, the LLM

model is run in the validation set and metrics of accuracy are calculated.

Abbreviations: LLM, Logic Learning Machine.

Figure 3. Feature Ranking.

The Feature Ranking task computes a set of measures to assess the relevance/usefulness of the input attributes based on the input model. Absolute relevancy gives an aggregate measure of how ‘strong’ is the correlation between a given input attribute and the output.

Figure 4. Application of the LLM model in the validation set.

A) Score distribution between cases and controls in the validation cohort. In the box plot, boxes define the interquartile range; thick lines refer to the median. The p-value was calculated using the Wilcoxon rank sum test. Legend: 1 = healthy control (green color), 2 = PBC case (red color).

B) Proportion of cases and controls by deciles of score in the validation cohort. Legend: 1 = healthy control (green color), 2 = case (red color).

C) ROC curve for case-control discrimination in the validation cohort.

D) Calibration plot with Brier score.

Abbreviations: AUC, Area Under the Curve; PBC, Primary Biliary Cholangitis, ROC, Receiver Operating Characteristics.

Supplementary Table Legend

Supplementary Table 1 Values of npresel and presel features that were evaluated in combination with other hyperparameters for model selection.

Abbreviations: npresel, number of pre-selected features; sel, number of selected features.

Supplementary Table 2 Univariate ranking of variants. Only the first 30 variants are shown as example due to space constraints.

Supplementary Table 3 Sets of hyperparameters evaluated before choosing the final model.

Abbreviations: corr, correlation value; npresel, number of pre-selected features; sel, number of selected features.

Supplementary Table 4 Summary Table of performance based on different sets of hyperparameters.

“errmax” represents the maximum level of error for each rule included in the ruleset. In other words, this corresponded to the maximum percentage of cases belonging to output classes different from the predicted one which verified the rule.

Abbreviations: errmax, maximum level of error for each rule; PPV, Positive Predictive Value; NPV, Negative Predictive Value.

Supplementary Table 5. Covering and error of each condition within the best rule including sex.

“Covering” is the percentage of samples belonging to the class described by the rule, fulfilling that specific rule; “Error” is the percentage of samples belonging to the other classes fulfilling that specific rule.

Supplementary Table 6. Covering and error of each condition within the best rule without sex.

“Covering” is the percentage of samples belonging to the class described by the rule, fulfilling that specific rule; “Error” is the percentage of samples belonging to the other classes fulfilling that specific rule.

Table 1

Chr	Gene	start	end
1	<i>MMEL1</i>	2273723	2773723
1	<i>IL12RB2</i>	67570194	68070194
1	<i>CD58</i>	116815083	117315083
1	<i>FCRL3</i>	157420290	157920290
1	<i>DENND1B</i>	197530966	198030966
1	<i>CACNA1S</i>	200769059	201269059
2	<i>DNMT3A</i>	25264333	25764333
2	<i>TMEM163</i>	135091200	135591200
2	<i>STAT4</i>	191693742	192193742
3	<i>PLCL2</i>	16711265	17211265
3	<i>RARB</i>	25133587	25633587
3	<i>TIMMDC1</i>	118969934	119469934
3	<i>IL12A-AS1</i>	159410283	159910283
4	<i>NFKB1</i>	103290780	103790780
4	<i>TET2</i>	105878954	106378954
5	<i>IL7R</i>	35631130	36131130
5	<i>LOC285626</i>	158509900	159009900
6	<i>OLIG3</i>	137723068	138223068
7	<i>ITGB8</i>	20128801	20628801
7	<i>ELMO1</i>	37132465	37632465
7	<i>TNPO3</i>	128367466	128867466
7	<i>ZC3HAV1</i>	138479543	138979543
9	<i>HEMGN</i>	100491912	100991912
10	<i>WDFY4</i>	49775396	50275396
11	<i>DEAF1</i>	396986	896986
11	<i>CCDC88B,</i>	63860422	64360422
11	<i>POU2AF1</i>	110989365	111489365

11	<i>DDX6</i>	118490104	118990104
12	<i>TNFRSF1A</i>	6190009	6690009
12	<i>ATXN2</i>	111657431	112157431
13	<i>LINC02341</i>	42805002	43305002
13	<i>DLEU1</i>	50561220	51061220
14	<i>RAD51B</i>	68499927	68999927
14	<i>RIN3</i>	92864787	93364787
14	<i>EXOC3L4</i>	103314807	103814807
16	<i>CLEC16A</i>	10924365	11424365
16	<i>IL4R</i>	27153469	27653469
16	<i>DPEP2</i>	67786939	68286939
16	<i>LOC105371388</i>	85769271	86269271
17	<i>ZPBP2</i>	37794893	38294893
17	<i>KANSL1</i>	43899348	44399348
18	<i>CD226</i>	67276026	67776026
19	<i>TYK2</i>	10225652	10725652
19	<i>MAST3</i>	17985882	18485882
19	<i>SPIB</i>	50676742	51176742
22	<i>RPL3</i>	39490078	39990078
Total	46		

Table 2

6:32653792:A:G	11:64017417_1	14:103563421_2
Sex	11:64021605_1	14:103563547_2
7:128588434_1	11:64031798_1	16:11033402_1
7:128590801_1	11:64031798_2	16:11035888_1
7:128669912_1	11:64053157_1	16:11045340_1
7:128681062_1	11:64102948_1	16:11058753_1
7:128705730_1	11:64158950_2	16:11058757_1
7:128705730_2	12:6212681_1	16:11082692_1
7:128714746_1	12:6495275_2	16:11085646_1
7:128718178_1	13:42969049_1	16:11090358_1
7:128720295_1	13:42970446_1	16:11189256_1
7:128723194_1	13:42970446_2	16:11191219_1
7:128723943_1	14:68976059_1	16:11193553_1
10:49999680_1	14:68976059_2	16:11195948_1
10:50003921_1	14:92928693_2	16:11221584_1
11:63908660_2	14:92932650_2	16:11233179_1
11:64011614_2	14:103563195_2	16:11239689_1
11:64011854_2		

16:27165173_1
16:86073712_1
16:86076497_1
16:86079249_1
16:86079388_1
16:86212152_1
16:86213219_1
16:86214690_1
17:37952989_1
17:38020058_2
17:38023745_2
17:38024626_2
17:38026169_2
17:38031030_2
17:38031714_2
17:38031857_2
17:38038389_2
17:38044893_2
17:38045725_2
17:38049589_2

17:38055921_1
17:38055921_2
17:38057189_1
17:38057189_2
17:38066267_2
17:38067533_2
17:38070071_2
17:43899417_1
17:43900215_1
17:43902861_1
17:43903298_1
17:43906828_1
17:43908989_1
17:43910183_1
17:43912635_1
17:43919070_1
17:43919096_1
17:43925966_1
17:43928614_1
17:43932129_1

17:43932173_1
17:43933190_1
17:43933579_1
17:43933673_1
17:43945745_1
17:43945938_1
17:43946223_1
17:43946318_1
17:43946423_1
17:43946875_1
17:43948977_1
17:43949342_1
17:43949448_1
17:43952944_1
17:43954416_1
17:43956139_1
17:43958362_1
17:43960341_1
17:43977049_1
17:43978534_1

17:43991515_1
17:43994252_1
17:43994648_1
17:43996430_1
17:44003397_1
17:44006453_1
17:44010452_1
17:44010463_1
17:44012463_1
17:44017666_1
17:44018399_1
17:44018488_1
17:44019107_1
17:44023087_1
17:44023828_1
17:44025359_1
17:44033097_1
17:44036408_1
17:44038536_1
17:44038785_1

17:44039008_1
17:44040288_1
17:44040823_1
17:44043092_1
17:44047449_1
17:44061608_1
17:44065263_1
17:44078816_1
17:44080465_1
17:44081064_1
17:44089727_1
17:44094471_1
17:44096553_1
17:44103825_1
17:44125288_1
17:44126575_1
17:44133818_1
17:44137009_1
17:44137386_1
17:44147574_1

17:44147721_1
17:44150161_1
17:44156180_1
17:44157676_1
17:44163547_1
17:44181933_1
17:44183317_1
17:44184819_1
17:44185431_1
17:44188755_1
17:44192590_1
17:44196447_1
17:44200015_1
17:44201109_1
17:44205690_1
17:44207887_1
17:44222335_1
17:44228529_1
17:44257473_1
17:44258422_1

17:44272679_1
17:44283022_1
19:10645576_1
19:18236725_1
19:18244560_1
19:50924093_1
19:50925395_1
19:50926265_1
19:50926742_1
19:50927358_1
2:25332696_2
2:135188248_1
2:191917317_1
2:191943742_1
2:191992237_1
2:191992611_1
3:17131082_2
3:119103580_1

3:119111870_1
3:119116150_1
3:119128398_1
3:119174383_1
3:119209027_1
3:119219934_1
3:119222456_1
3:119228508_1
3:119244593_1
1:67875102_1
3:119257802_1
3:119262734_1
3:119262734_2
3:119286713_1
3:119292618_1
3:119297389_1
3:159497430_1
3:159520115_1

3:159556462_1
3:159561337_1
3:159568546_1
3:159726324_1
4:103444533_1
4:103446115_1
4:103476166_1
4:103511747_1
4:103531112_1
4:103538911_1
4:103540780_1
4:103554350_1
4:103555611_1
4:103559876_1
4:103622568_1
4:103622568_2
7:37176353_1
7:37176353_2]

Table 3

Id rule	4
Number of conditions	13
Output attribute	Affection
Output value	Case
Covering %	19,144144
Error %	3,884573
Condition 1	14:92932650_2 = "C"
Condition 2	17:43906828_1 = "G"
Condition 3	17:43912635_1 = "A"
Condition 4	17:44038536_1 = "CA"
Condition 5	17:44040823_1 = "C"
Condition 6	17:44065263_1 = "T"
Condition 7	17:44183317_1 = "C"
Condition 8	17:44185431_1 = "T"
Condition 9	17:44222335_1 = "G"
Condition 10	17:44283022_1 = "A"
Condition 11	3:119111870_1 = "T"
Condition 12	7:128705730_1 = "T"
Condition 13	Sex = F

Table 4

Id rule	11
Number of conditions	7
Output attribute	Affection
Output value	Case
Covering %	12,162162
Error %	4,661487
Condition 1	17:38020058_2 = "AC"
Condition 2	17:38049589_2 = "T"
Condition 3	17:38070071_2 = "C"
Condition 4	17:43933579_1 = "C"
Condition 5	2:135188248_1 = "A"
Condition 6	2:25332696_2 = "C"
Condition 7	3:159726324_1 = "A"

Figure 1

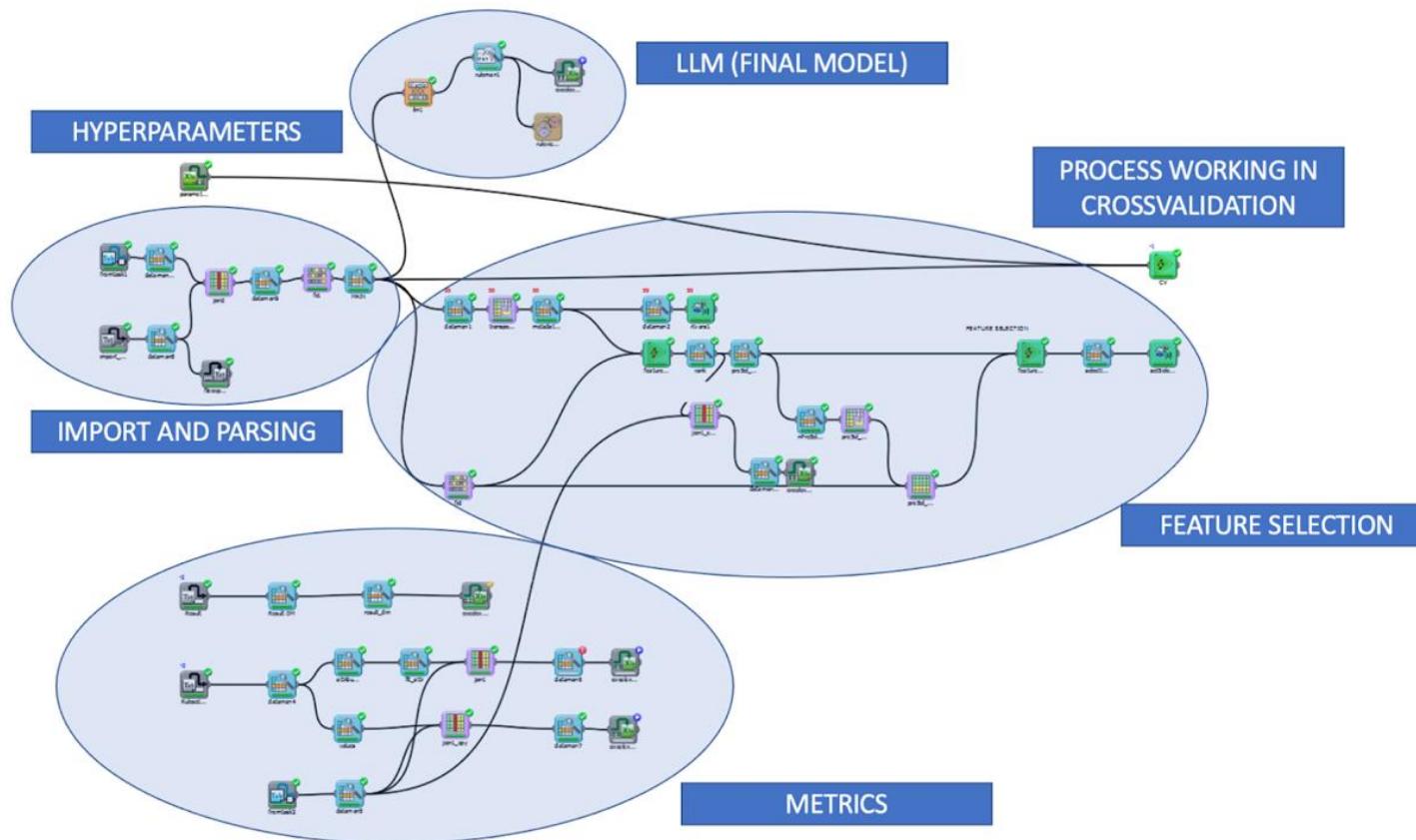


Figure 2

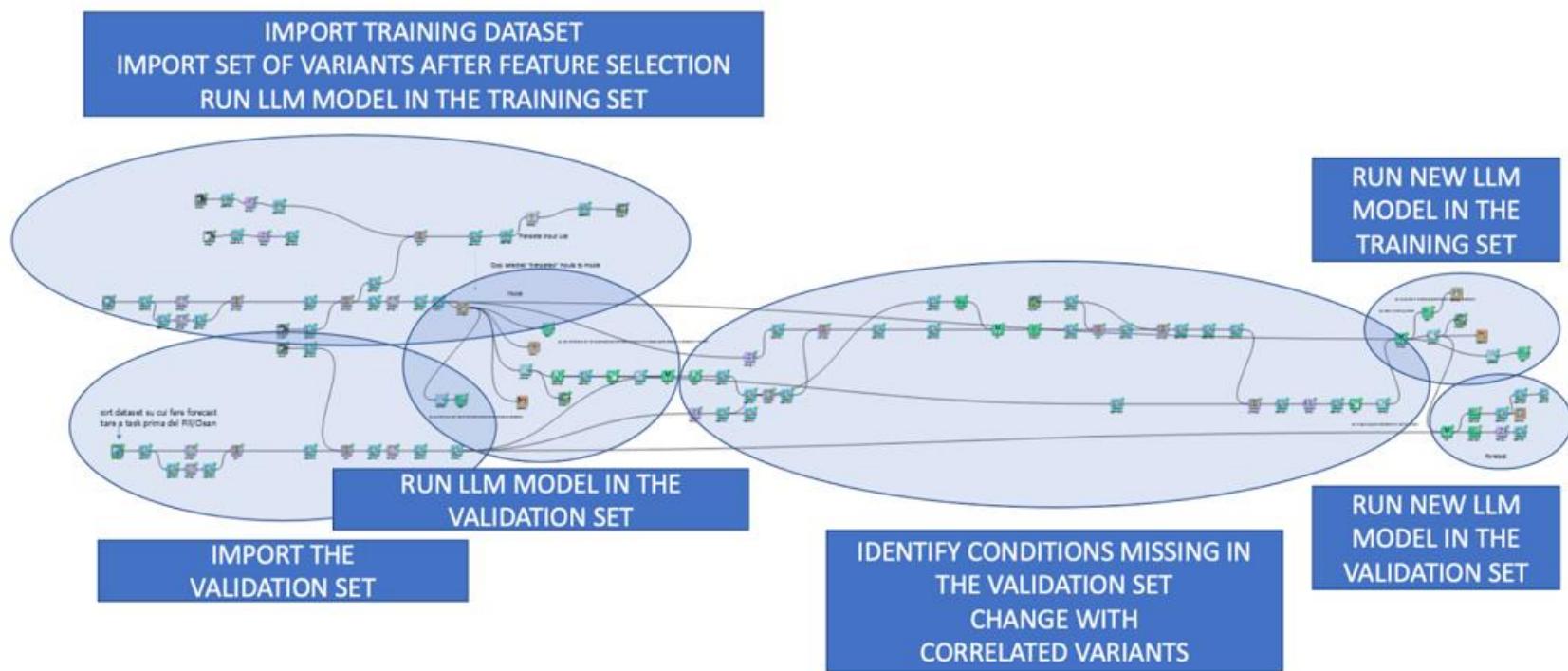


Figure 3

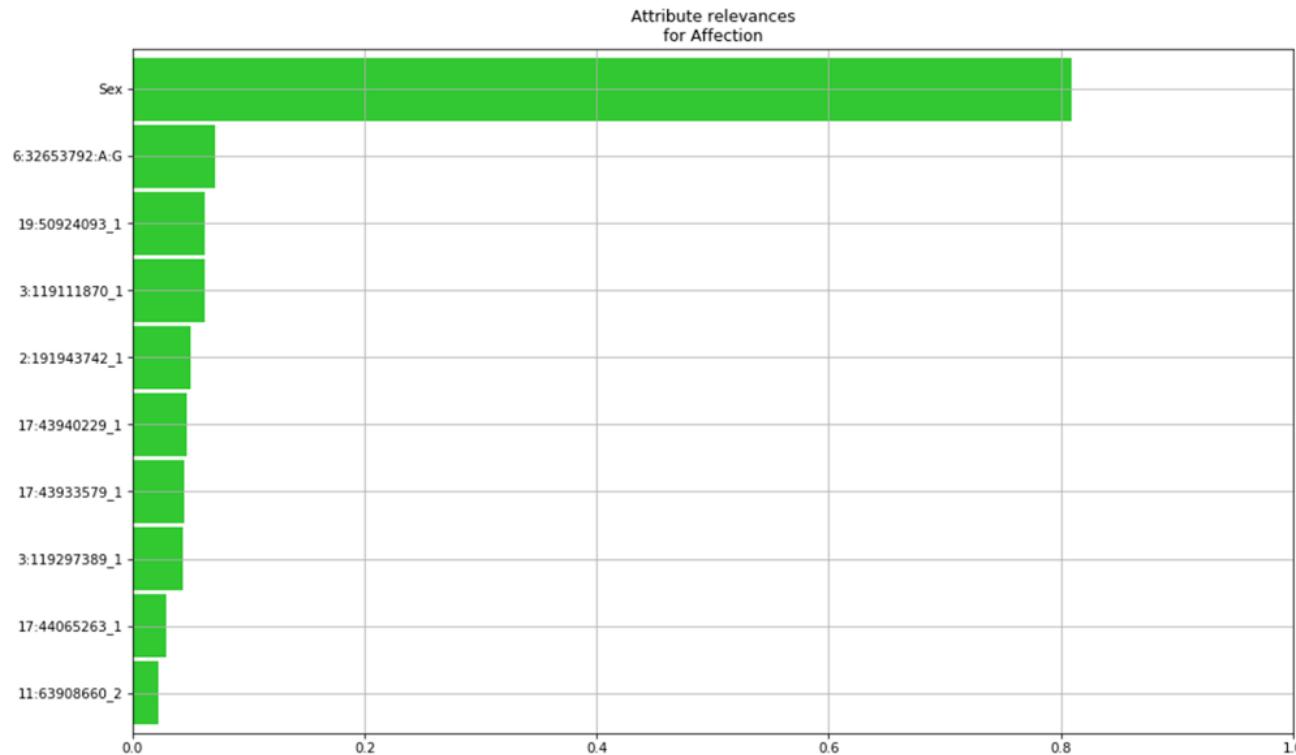
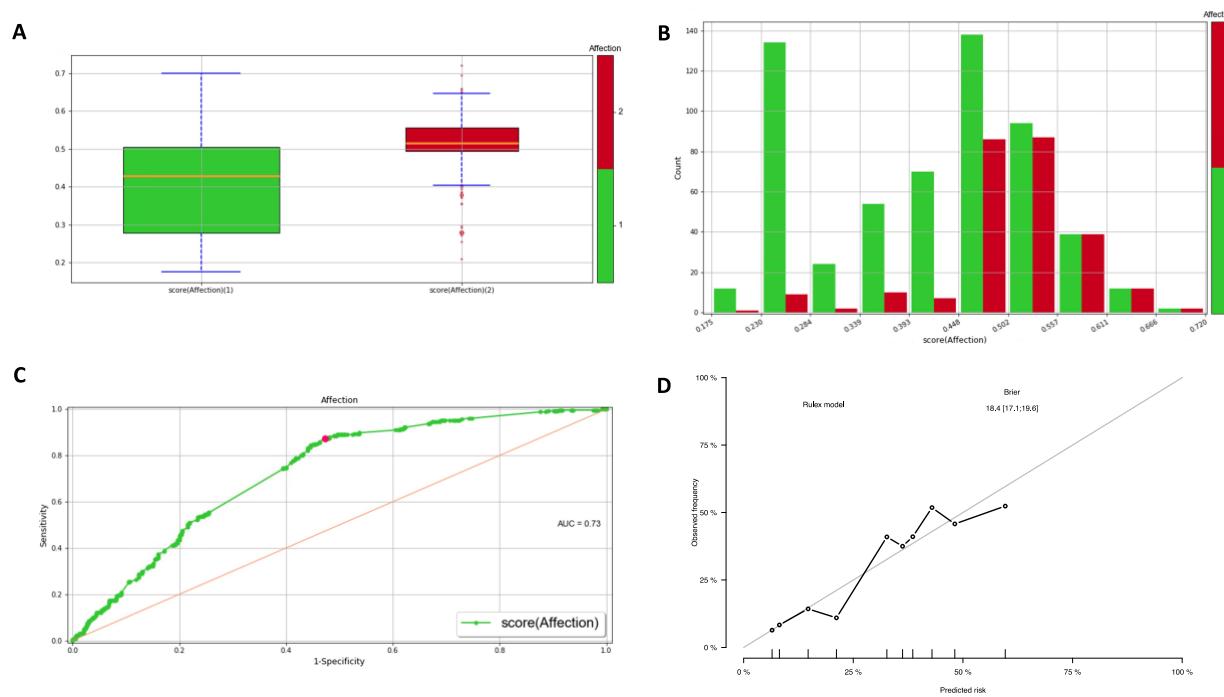


Figure 4



Supplementary Table 1

	Npresel	Sel					
1	200	50	98				
2	248	62	110				
3	296	74	122				
4	344	86	134				
5	392	98	146	194			
6	440	110	158	206			
7	488	122	170	218			
8	536	134	182	230			
9	584	146	194	242	290		
10	632	158	206	254	302		
11	680	170	218	266	314		
12	728	182	230	278	326		
13	776	194	242	290	338	386	
14	824	206	254	302	350	398	
15	872	218	266	314	362	410	
16	920	230	278	326	374	422	
17	968	242	290	338	386	434	482
18	1016	254	302	350	398	446	494

Supplementary Table 2

Attribute	Score	rank
Sex	0.187156	1
6:32653792:A:G	0.034963	2
3:119262734:C:CT	0.018558	3
3:119111870:C:T	0.016562	4
7:37176353:C:CA	0.016303	5
7:128588434:T:TG	0.015746	6
7:128589000:C:T	0.015746	7
3:119103580:G:T	0.015431	8
3:119116150:A:G	0.015268	9
7:37176353:C:CA	0.015237	10
7:128714746:A:G	0.015067	11
7:128714843:C:T	0.015067	12
7:128715299:A:T	0.015067	13
17:43919070:C:T	0.015019	14
17:43919068:G:T	0.015019	15

Attribute	Score	rank
17:43919073:G:T	0.015019	16
19:50927358:A:G	0.01479	17
7:128717234:A:AAT	0.014777	18
7:128717305:A:G	0.014777	19
3:119128398:A:G	0.014731	20
3:119130141:A:G	0.014731	21
4:103446115:A:G	0.014575	22
16:11082692:C:T	0.014553	23
16:11058753:A:C	0.014541	24
7:128713630:A:G	0.014506	25
11:64031798:C:G	0.014402	26
11:64031798:C:G	0.014381	27
16:11195948:A:G	0.014345	28
7:128716007:G:T	0.014325	29
1:67875102:A:ACC	0.01418	30

Supplementary Table 3

Run number	Corr	Split	Npresel	Nsel
1	0.8	10	1016	254
2	0.8	10	1016	302
3	0.8	10	1016	350
4	0.8	10	1016	398
5	0.8	10	1016	446
6	0.8	10	1016	494
7	0.8	10	968	242
8	0.8	10	968	290
9	0.8	10	968	338
10	0.8	10	968	386
11	0.8	10	968	434
12	0.8	10	968	482
13	0.8	10	920	230
14	0.8	10	920	278
15	0.8	10	920	326
16	0.8	10	920	374
17	0.8	10	920	422
18	0.8	10	872	218
19	0.8	10	872	266
20	0.8	10	872	314
21	0.8	10	872	362
22	0.8	10	872	410
23	0.8	10	824	206
24	0.8	10	824	254
25	0.8	10	824	302
26	0.8	10	824	350
27	0.8	10	824	398
28	0.8	10	776	194
29	0.8	10	776	242
30	0.8	10	776	290
31	0.8	10	776	338
32	0.8	10	776	386
33	0.8	10	728	182
34	0.8	10	728	230
35	0.8	10	728	278
36	0.8	10	728	326
37	0.8	10	680	170
38	0.8	10	680	218
39	0.8	10	680	266

40	0.8	10	680	314
41	0.8	10	632	158
42	0.8	10	632	206
43	0.8	10	632	254
44	0.8	10	632	302
45	0.8	10	584	146
46	0.8	10	584	194
47	0.8	10	584	242
48	0.8	10	584	290
49	0.8	10	536	134
50	0.8	10	536	182
51	0.8	10	536	230
52	0.8	10	488	122
53	0.8	10	488	170
54	0.8	10	488	218
55	0.8	10	440	110
56	0.8	10	440	158
57	0.8	10	440	206
58	0.8	10	392	98
59	0.8	10	392	146
60	0.8	10	392	194
61	0.8	10	344	86
62	0.8	10	344	134
63	0.8	10	296	74
64	0.8	10	296	122
65	0.8	10	248	62
66	0.8	10	248	110
67	0.8	10	200	50
68	0.8	10	200	98
69	0.7	10	1016	254
70	0.7	10	1016	302
71	0.7	10	1016	350
72	0.7	10	1016	398
73	0.7	10	1016	446
74	0.7	10	1016	494
75	0.7	10	968	242
76	0.7	10	968	290
77	0.7	10	968	338
78	0.7	10	968	386
79	0.7	10	968	434
80	0.7	10	968	482
81	0.7	10	920	230
82	0.7	10	920	278
83	0.7	10	920	326
84	0.7	10	920	374

85	0.7	10	920	422
86	0.7	10	872	218
87	0.7	10	872	266
88	0.7	10	872	314
89	0.7	10	872	362
90	0.7	10	872	410
91	0.7	10	824	206
92	0.7	10	824	254
93	0.7	10	824	302
94	0.7	10	824	350
95	0.7	10	824	398
96	0.7	10	776	194
97	0.7	10	776	242
98	0.7	10	776	290
99	0.7	10	776	338
100	0.7	10	776	386
101	0.7	10	728	182
102	0.7	10	728	230
103	0.7	10	728	278
104	0.7	10	728	326
105	0.7	10	680	170
106	0.7	10	680	218
107	0.7	10	680	266
108	0.7	10	680	314
109	0.7	10	632	158
110	0.7	10	632	206
111	0.7	10	632	254
112	0.7	10	632	302
113	0.7	10	584	146
114	0.7	10	584	194
115	0.7	10	584	242
116	0.7	10	584	290
117	0.7	10	536	134
118	0.7	10	536	182
119	0.7	10	536	230
120	0.7	10	488	122
121	0.7	10	488	170
122	0.7	10	488	218
123	0.7	10	440	110
124	0.7	10	440	158
125	0.7	10	440	206
126	0.7	10	392	98
127	0.7	10	392	146
128	0.7	10	392	194
129	0.7	10	344	86

130	0.7	10	344	134
131	0.7	10	296	74
132	0.7	10	296	122
133	0.7	10	248	62
134	0.7	10	248	110
135	0.7	10	200	50
136	0.7	10	200	98
137	0.9	10	1016	254
138	0.9	10	1016	302
139	0.9	10	1016	350
140	0.9	10	1016	398
141	0.9	10	1016	446
142	0.9	10	1016	494
143	0.9	10	968	242
144	0.9	10	968	290
145	0.9	10	968	338
146	0.9	10	968	386
147	0.9	10	968	434
148	0.9	10	968	482
149	0.9	10	920	230
150	0.9	10	920	278
151	0.9	10	920	326
152	0.9	10	920	374
153	0.9	10	920	422
154	0.9	10	872	218
155	0.9	10	872	266
156	0.9	10	872	314
157	0.9	10	872	362
158	0.9	10	872	410
159	0.9	10	824	206
160	0.9	10	824	254
161	0.9	10	824	302
162	0.9	10	824	350
163	0.9	10	824	398
164	0.9	10	776	194
165	0.9	10	776	242
166	0.9	10	776	290
167	0.9	10	776	338
168	0.9	10	776	386
169	0.9	10	728	182
170	0.9	10	728	230
171	0.9	10	728	278
172	0.9	10	728	326
173	0.9	10	680	170
174	0.9	10	680	218

175	0.9	10	680	266
176	0.9	10	680	314
177	0.9	10	632	158
178	0.9	10	632	206
179	0.9	10	632	254
180	0.9	10	632	302
181	0.9	10	584	146
182	0.9	10	584	194
183	0.9	10	584	242
184	0.9	10	584	290
185	0.9	10	536	134
186	0.9	10	536	182
187	0.9	10	536	230
188	0.9	10	488	122
189	0.9	10	488	170
190	0.9	10	488	218
191	0.9	10	440	110
192	0.9	10	440	158
193	0.9	10	440	206
194	0.9	10	392	98
195	0.9	10	392	146
196	0.9	10	392	194
197	0.9	10	344	86
198	0.9	10	344	134
199	0.9	10	296	74
200	0.9	10	296	122
201	0.9	10	248	62
202	0.9	10	248	110
203	0.9	10	200	50
204	0.9	10	200	98

Supplementary Table 4

Configuration	1
set	test
errmax	0,05
Accuracy	71,670276
Matthews	0,289932
Youden	0,21597
Sensitivity	0,288288
Specificity	0,927858
PPV	0,663212
NPV	0,725694
Positive_LR	3,996119
Negative_LR	0,767048
Hyperparameter_Set	0.8_872_266_10

Supplementary Table 5

		Covering of the condition	Error of the condition
Condition 1	14:92932650_2 = "C"	0.90	0.11
Condition 2	17:43906828_1 = "G"	0.00	0.11
Condition 3	17:43912635_1 = "A"	0.00	0.11
Condition 4	17:44038536_1 = "CA"	0.45	0.22
Condition 5	17:44040823_1 = "C"	0.00	0.11
Condition 6	17:44065263_1 = "T"	11.49	5.44
Condition 7	17:44183317_1 = "C"	0.45	0.11
Condition 8	17:44185431_1 = "T"	0.00	0.11
Condition 9	17:44222335_1 = "G"	0.00	0.11
Condition 10	17:44283022_1 = "A"	0.00	0.11
Condition 11	3:119111870_1 = "T"	4.95	3.66
Condition 12	7:128705730_1 = "T"	1.58	0.44
Condition 13	Sex = F	0.90	5.88

Supplementary Table 6

		Covering of the condition	Error of the condition
Condition 1	17:38020058_2 = "AC"	0.00	0.11
Condition 2	17:38049589_2 = "T"	0.45	0.55
Condition 3	17:38070071_2 = "C"	0.23	0.11
Condition 4	17:43933579_1 = "C"	7.88	6.33
Condition 5	2:135188248_1 = "A"	1.13	0.67
Condition 6	2:25332696_2 = "C"	1.80	1.33
Condition 7	3:159726324_1 = "A"	0.45	1.11

References

1. Schüssler-Fiorenza Rose, S. M. *et al.* A longitudinal big data approach for precision health. *Nat. Med.* **25**, 792–804 (2019).
2. Gerussi, A. & Carbone, M. Primary Biliary Cholangitis. *Autoimmune Liver Disease* 123–141 (2020). doi:<https://doi.org/10.1002/9781119532637.ch7>
3. Gerussi, A., Carbone, M., Asselta, R. & Invernizzi, P. Genetics of Autoimmune Liver Diseases BT - Liver Immunology : Principles and Practice. in (eds. Gershwin, M. E., M. Vierling, J., Tanaka, A. & P. Manns, M.) 69–85 (Springer International Publishing, 2020). doi:10.1007/978-3-030-51709-0_5
4. Gerussi, A., Carbone, M., Corpechot, C. & Schramm, C. The genetic architecture of primary biliary cholangitis. *Eur. J. Med. Genet.* **64**, 104292 (2021).
5. Gulamhussein, A. F. & Hirschfield, G. M. Primary biliary cholangitis: pathogenesis and therapeutic opportunities. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 93–110 (2020).
6. Carbone, M. *et al.* Primary biliary cholangitis: a multifaceted pathogenesis with potential therapeutic targets. *J. Hepatol.* **73**, 965–966 (2020).
7. Cordell, H. J. *et al.* An international genome-wide meta-analysis of primary biliary cholangitis: novel risk loci and candidate drugs. *J. Hepatol.* (2021). doi:<https://doi.org/10.1016/j.jhep.2021.04.055>

8. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **3**, 11–13 (2020).
9. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
10. Ho, D. S. W., Schierding, W., Wake, M., Saffery, R. & O'Sullivan, J. Machine Learning SNP Based Prediction for Precision Medicine. *Front. Genet.* **10**, 1–10 (2019).
11. Sun, J. *et al.* Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction. *Nat. Commun.* **12**, 1–9 (2021).
12. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
13. de los Campos, G., Vazquez, A. I., Hsu, S. & Lello, L. Complex-Trait Prediction in the Era of Big Data. *Trends Genet.* **34**, 746–754 (2018).
14. Lindor, K. D. *et al.* Primary biliary cirrhosis. *Hepatology* **50**, 291–308 (2009).
15. de Hond, A. A. H. *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit. Med.* **5**, 2 (2022).
16. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
17. Gunning, D. *et al.* XAI—Explainable artificial intelligence.

Sci. Robot. **4**, eaay7120 (2019).

18. Price, W. N. Big data and black-box medical algorithms. *Sci. Transl. Med.* **10**, eaao5333 (2018).
19. Azodi, C. B., Tang, J. & Shiu, S.-H. Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends Genet.* (2020). doi:<https://doi.org/10.1016/j.tig.2020.03.005>
20. Cangelosi, D. *et al.* Logic Learning Machine creates explicit and stable rules stratifying neuroblastoma patients. *BMC Bioinformatics* **14**, S12 (2013).
21. Cangelosi, D. *et al.* Use of Attribute Driven Incremental Discretization and Logic Learning Machine to build a prognostic classifier for neuroblastoma patients. *BMC Bioinformatics* **15**, S4 (2014).
22. Cangelosi, D. *et al.* Hypoxia predicts poor prognosis in neuroblastoma patients and associates with biological mechanisms involved in telomerase activation and tumor microenvironment reprogramming. *Cancers (Basel)*. **12**, 1–45 (2020).
23. Mordenti, M. *et al.* Validation of a new multiple osteochondromas classification through Switching Neural Networks. *American Journal Of Medical Genetics. Part A* **161A**, 556–560 (2013).
24. Parodi, S. *et al.* Differential diagnosis of pleural mesothelioma using Logic Learning Machine. *BMC Bioinformatics* **16**, S3 (2015).
25. Parodi, S., Dosi, C., Zambon, A., Ferrari, E. & Muselli, M.

Identifying Environmental and Social Factors Predisposing to Pathological Gambling Combining Standard Logistic Regression and Logic Learning Machine. *J. Gambl. Stud.* **33**, 1121–1137 (2017).

26. Verda, D., Parodi, S., Ferrari, E. & Muselli, M. Analyzing gene expression data for pediatric and adult cancer diagnosis using logic learning machine and standard supervised methods. *BMC Bioinformatics* **20**, 1–13 (2019).
27. Skotko, B. G. *et al.* A predictive model for obstructive sleep apnea and Down syndrome. *Am. J. Med. Genet. Part A* **173**, 889–896 (2017).
28. Caruana, R. & Freitag, D. Greedy attribute selection. in *Machine Learning Proceedings 1994* 28–36 (Elsevier, 1994).
29. Omiecinski, E. R. Alternative Interest Measures for Mining Associations in Databases. *IEEE Trans. Knowl. Data Eng.* **15**, 57–69 (2003).
30. Verda, D., Parodi, S., Ferrari, E. & Muselli, M. Analyzing gene expression data for pediatric and adult cancer diagnosis using logic learning machine and standard supervised methods. *BMC Bioinformatics* **20**, 390 (2019).
31. Wei, W. H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15**, 722–733 (2014).
32. Paulus, J. K. & Kent, D. M. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *npj Digit.*

Med. **3**, 99 (2020).

33. Zou, J. *et al.* A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18 (2019).
34. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
35. Hartl, J. *et al.* Risk of primary biliary cholangitis relatives: a prospective cohort study. in *International Liver Congress (ILC) 2021* (2021).
36. McGee, S. Simplifying likelihood ratios. *J. Gen. Intern. Med.* **17**, 646–649 (2002).
37. Khera, A. V *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
38. Kuo, R. J. & Zulvia, F. E. The gradient evolution algorithm: A new metaheuristic. *Inf. Sci. (Ny)*. **316**, 246–265 (2015).

CHAPTER 6

The archaic mutational load predicts the fate of introgressed fragments in humans

Alessio Gerussi^{1,2,3,4}, Rosanna Asselta^{5,6}, Viviane Slon^{1,2}, Pietro Invernizzi^{3,4}, Fabrizio Mafessoni⁷

¹Department of Anatomy and Anthropology and Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

²The Dan David Center for Human Evolution and Biohistory Research, Tel Aviv University, Tel Aviv, Israel

³Division of Gastroenterology, Center for Autoimmune Liver Diseases, Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy

⁴European Reference Network on Hepatological Diseases (ERN RARE-LIVER), San Gerardo Hospital, Monza, Italy

⁵Humanitas Clinical and Research Center, Rozzano, Milan, Italy

⁶Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy

⁷Department of Plant and Environmental Sciences, Weizmann Institute of Science, Herzl Street 234, Rehovot 7610001, Israel

Manuscript in preparation

PhD Candidate contribution: leader in the conceptualization of the study, data curation, analytic process, interpretation of the results and writing the manuscript.

Abstract

Background and Aims: About 1-4% of the genome of humans living out of Africa entered in the human gene pool through interbreeding with Neanderthals. It has been suggested that Neanderthals accumulated deleterious variants that were swiftly selected against when entered the human gene pool, explaining the depletion of fragments originated from Neanderthals in genes of present-day individuals. Here we leverage the diversity of archaic genomes and deleteriousness measures of mutations to characterize the archaic mutational load along the genome.

Methods: We combined publicly available genomic datasets of present-day humans and archaic hominins and measures of phylogenetic conservations to develop population-genetics aware statistics. These were used to test hypotheses using generalized linear mixed models and resampling-based methods.

Results: We show that regions with more putatively deleterious mutations in archaic populations than in humans were more efficiently removed after introgression than regions with a lower mutational load. We found a similar pattern for variants influencing gene expression and immune-related variants, despite these are overrepresented in fragments of Neanderthal origin.

Conclusion: Fragments carrying an excess of Neanderthal-derived mutations were largely purged by natural selection.

Introduction

Upon entering the human gene pool after introgression, the fate of Neanderthal fragments has been shaped by demographic events and natural selection, acting to increase or decrease their frequency in human population for adaptive^{1–4} or deleterious fragments^{5–7}, respectively. It has been suggested that Neanderthals had a higher mutational load, i.e. a higher proportion of deleterious variants, than modern humans, because of their long-term small effective population size^{5,8,9}. Thus, the deleterious variants present in Neanderthal fragments in the human population would have caused them to decrease in allele frequency and often to disappear in large stretches of the genome, often referred to as “deserts of Neanderthal ancestry”^{6,7,10}. Previous research has showed that these regions are rich in genes, supporting the view that natural selection underlies the disappearance of Neanderthal fragments from the human gene pool⁶. Despite this, causal evidence linking specific variants or parts of the genome – within Neanderthal deserts - with the loss of Neanderthal ancestry is missing. In this study, we explore the possibility of leveraging the ancient genomes of archaic hominins and genomic estimates of deleteriousness at nucleotide-resolution to estimate the mutational load in Neanderthals along the genome, in order to develop a framework to identify the variants which were mostly responsible for the reduced fitness in humans carrying Neanderthal DNA.

Methods

To build our measures of archaic mutational load we subdivided the genome in windows of 100kb, in turn grouped in blocks of 5Mb. We used four high-coverage archaic genomes^{11–14} to obtain a site frequency spectrum (**Supplementary Figure 1**). Note that this site frequency spectrum only loosely reflects the allele frequencies in Neanderthals and Denisovans, as the specimens were sampled across more than 4000 km and 100 thousand years¹². In addition, 5–20% of these genomes is covered by stretches of runs of homozygosity >2cM, likely due to recent inbreeding.

We thus performed analyses using either the full observed site frequency spectrum, a randomly sample allele from each genome or exclusively the genome of Vindija.39, which is the closer Neanderthal genome to the introgressing Neanderthal population into modern humans¹⁴. Unless specified differently, analyses are reported for the full site frequency spectrum, as we found indications that it captures patterns likely due to selection (**Supplementary Figure 2**).

For each polymorphism, the ancestral and derived states were assigned on the basis of the state in five different primate reference genomes (chimpanzee, bonobo, gorilla, orangutan, and macaque). 1000 Genomes¹⁵ and SGDP¹⁶ dataset were retrieved from <https://www.internationalgenome.org/> and <https://reichdata.hms.harvard.edu/pub/datasets/sgdp/>, respectively.

The Variant Call Format (VCF) files for the archaic genomes were downloaded from <http://ftp.eva.mpg.de/Neanderthal/>. All VCFs were combined and manipulated using bedtools, bcftools and vctools. The deleteriousness scores were retrieved from <https://kircherlab.bihealth.org/download/CADD/v1.6/GRCh37/>, from the whole genome Single Nucleotide Variant (SNV) files¹⁷. Combined Annotation Dependent Depletion (CADD) scores¹⁸ were discarded as they are built using information specific of modern humans, thus they might be biased when analyzing Neanderthals. The scores on which we focused were Gerp scores, PhyloP, PhastCons and b-statistic^{19–23}. We selected these scores as they are supposed to be unbiased towards humans. While Gerp scores are calculated using deep phylogenetic comparisons, for PhyloP and PhasCons we used scores calculated using multiple genome alignments of primates, excluding humans.

We used introgression maps obtained with different methods, from Vernot *et al*⁷, Sankararam *et al*⁶, Skov *et al*²⁴, and Browning *et al*²⁵. Unless specified differently in specific analyses, we defined as introgressed fragments those identified as archaic in Browning *et al*. We also tested the robustness of our results to the different maps chosen, by considering an intersection map of all studies. When the Browning introgression map, for which the archaic allele was directly retrieved from the identified introgressed fragments, was not used, we defined alleles of certain archaic origin as those that fall in introgressed regions, have frequency in Africans smaller than 1.5 %, a higher

frequency in Europeans or Asians than Africans according to the 1000 Genomes allele frequencies, and are either directly observed in the four high-coverage archaic genomes or show high linkage disequilibrium (LD) ($R^2 > 0.9$) in the 1000 Genomes dataset with variants observed in the high coverage archaic genomes.

We computed our statistics in 100kb windows, a length chosen to reduce the number of introgressed fragments overlapping multiple windows. For all statistical analyses (both bootstraps and linear models) we included a random categorical factor *block* to group the 100kb windows in 5Mb-blocks, aiming at further correcting for linkage and genomic proximity. Analyses were performed in R²⁶ using the R packages `data.table` and `lme4` for the Generalized Linear Mixed Model analyses. Specifically, we built models with the syntax `(0+covariates|block)`. To test significance we used the function `anova` to compare full models including the archaic load versus null models not including it.

Results

We first built several measures of the archaic and modern human mutational load, leveraging different conservation scores and measures of purifying selection, i.e. Gerp scores, PhyloP, PhastCons and b-statistic^{19–23}. We computed a different measure of mutational load for each score, weighing the score by the frequency of each derived mutation in archaics and Africans. We then compared the mutational load in archaics and Africans in regions where archaic fragments are identified, to test if we observed a higher mutational load in non-introgressed regions, which likely experienced purifying selection. In **Figure 1A** we show results for Gerp Scores, while other scores are shown in **Supplementary Figures 3 and 4**. For all our measures of mutational load we observed a higher load in non-introgressed fragments. For most of our measures we observed significantly higher values in archaics compared to Africans, compatibly with the fact that Neanderthal populations had a small effective size for at least 200.000 years^{11,14}. The only exception is when the load was calculated using b-statistics (**Supplementary Figure 3**). This could be due to the fact that the b-statistic measure background selection and for this reason varies only on the wide scale, preventing comparisons of lineage-specific mutations since their effects are average out in wide regions²². Thus, in general, we observed a larger difference between the mutational load in introgressed versus non-introgressed regions in archaics, compatibly with selection acting to remove fragments carrying an

excess of archaic deleterious variants. This pattern was particularly apparent for loads calculated using PhastCons scores, where the archaic load in non-introgressed regions was much higher than introgressed while the opposite was observed for the modern human load (**Supplementary Figure 3**).

Note that the archaic load in non-introgressed regions was higher than in introgressed ones even also when stratified by the African mutational load, pointing to the direct effect of archaic variants in determining the deleteriousness of introgressed fragments (**Figure 1B**). To test this, we regressed the mutational load in archaic and that in Africans to predict the frequency of introgression in the 100kb windows. For all scores, we detected a significant negative effect of the archaic load ($p\text{-value} < 0.05$), while the African load was not significant. Results were analogous when a logistic model was used to predict whether a window overlapped an introgressed fragment at least partially, or when the average potential deleteriousness (measured as the average score for all potential mutations in a given regions) was used instead of the African load. Note that in this analysis, we aimed at testing the causality of archaic variants. A potential caveat is the fact that we cannot directly measure selection, but only some scores, which might lead to spurious effects due to the collinearity in our measures of archaic and African load. An alternative would be to use the direct count of mutations in introgressed fragments. However, methods to identify fragments to date are biased towards regions with more mutations, either because of their conditioning on archaic genomes⁶, or because

the density of sites differing between archaics and modern humans is directly used to identify fragments²⁷. Note that this effect cannot be responsible for our observation, as we observed higher load in non-introgressed fragments, which goes in the opposite direction.

We then asked whether the difference in mutational load between archaic and Africans, used as a proxy of the mutational load of non-introgressed alleles in the corresponding genomic regions, predicts the fate of alleles (**Figure 2**). We found a negative relationship between difference in mutational load and introgression frequency, namely regions with the largest difference in mutational load were the regions with the lowest chance of being introgressed. This suggests that regions in which the effects of deleterious fragments brought by archaic introgression were partially masked by the deleterious variation in modern humans were purged less efficiently. This effect was remarkably strong for Gerp and PhastCons scores, which were also able to capture the increased load in archaics compared to Africans (p-value < 10⁻⁵ for both scores). The pattern was instead only visible as a weaker trend for PhyloP scores, with only marginal support (p-value 0.09). The load measured by the presence of variants in regions with strong background selection (b-statistics) also showed a highly significant (p-value<10⁻⁵) trend towards less introgression for regions with an excess of deleterious load in archaic versus Africans. However, it is important to notice that the intermediate 40-60% bin of differences show a noticeably smaller proportion of introgressed

fragments, conferring an unexpected U-shaped to the relationship between b-statistic and introgression. While we do not have an explanation for this phenomenon, we already noticed that b-statistic did not provide a measure of load able to capture the higher load in archaics (**Supplementary Figure 3**).

We finally turned our attention to regions of biological interest. Immune genes seem to behave differently than other genes, showing higher allele frequencies and being less depleted in Neanderthal ancestry^{28,29}. We thus leverage the DICE dataset (dice-database.org)³⁰, which contains expression quantitative trait loci (eQTL) for different cell types involved in immunity.

First, we showed that immune-mediated and autoimmune traits (**Supplementary Figure 5**) and eQTLs regulating the expression of immune cells (**Figure 3**) are enriched in Neanderthal fragments, in line with previous evidence. Interestingly, using the DICE dataset we observed some nuanced differences between immune cell subtypes, such as little or absent enrichment for monocytes, B naïve lymphocytes and NK cells, supporting previous findings obtained on different expression datasets on these cells (**Figure 3**). In addition, the granularity of DICE dataset revealed that the already known enrichment in T helper (Th) lymphocytes is present in all major Th subtypes, namely Th1, Th2, Th1/Th17, Th17 and T follicular helper (TFH) cells (**Figure 3**).

Second, we compared immune eQTLs, using DICE, and non-immune eQTLs, as found in the GTEx dataset^{31,32}. Overall, we found that immune eQTLs show a lower mutational load, with

more variation being either neutral or fast evolving than non-immune eQTLs, both in introgressed and non-introgressed fragments (**Figure 4**). This result is compatible with the enrichment of immune eQTLs in introgressed fragments, further supporting the fact that lower mutational load, particularly in archaics, corresponds to higher chance of introgression. However, it is important to notice that lower load, especially when considering scores obtained from phylogenetic conservation, might indicate faster evolution rather than neutrality, preventing a distinction between relaxed purifying selection and balancing selection.

Discussion

We showed that the measure of mutational load in Neanderthals is predictive of the fate of Neanderthal alleles along the genome. Namely, regions with more deleterious variation in Neanderthals are depleted in Neanderthal fragments. Remarkably, we found that regions with the highest difference between the mutational load in Neanderthal and in African populations, which are nearly devoid of Neanderthal ancestry, are the most depleted in Neanderthal ancestry. This shows that Neanderthal fragments which brought variants that were more deleterious compared to those in the human gene pool were more rapidly removed.

We anticipate that this information could be leveraged to devise methods to pinpoint regions which affected to a larger degree the fitness of the early Neanderthal-modern humans' offspring. Note that it has been shown, both via theoretical model and genomic evidence, that much of the loss of Neanderthal variation occurred within the first generations that archaic fragments entered the human gene pool^{5,33,34}. This is expected to be due to the rapid loss of long fragments of archaic ancestry, which carry a large number of deleterious alleles³⁴. Hence, this process is highly stochastic and subjected to chance, possibly masking the effect of selection at specific regions. However, evidence of multiple encounters between modern humans and archaic has been increasing^{33,35,36}, suggesting that repeated inflows of archaic alleles might reduce this stochasticity. In addition, specific regions of the genome influencing particularly important

phenotypes and differing between the two populations, might have been particularly responsible for the fate of early human-archaic offspring, and our study aims at moving forward in identifying these regions. For instance, it has been shown that the X-chromosome was likely subjected to this force, being highly depleted in archaic ancestry and being subjected to recurrent selection in primates^{6,37–39}.

Conclusions

We bring further evidence that archaic variation was subjected to purifying selection, and indications that regions with lower load in Africans more efficiently purged the incoming archaic DNA fragments.

Figure Legends

Figure 1. The mutational load in introgressed and non-introgressed fragments measured through Gerp Scores.

A) The mutational load in non introgressed is significantly higher in archaics than in Africans (p -value $<10^{-3}$ based on 1000 5Mb block-bootstraps).

B) The mutational load in archaics stratified by the mutational load in Africans, divided in percentiles. For all bins, the non-introgressed load is significantly higher (p -value <0.05) than the introgressed load.

Figure 2. Proportion of windows containing introgressed fragments (y-axis) along increasing differences in mutational load between archaics and Africans (x-axis).

For b-statistic, the difference in mutational load increases leftwards. The difference in mutational load for the different scores has been grouped in 20% quantiles.

Figure 3. Observed number of expression QTL in different immune cell types overlapped by different maps of Neanderthal ancestry (red dot) versus the distribution of 100 simulated sets eQTLs.

The simulated eQTLs were generated by rotating the coordinates of original eQTLs by controlling for the allele frequencies of the original eQTLs, selecting the new location of the SNPs to the closest one with a frequency difference smaller than 2%.

Different columns indicate different maps of Neanderthal ancestry, from left to right: the intersection of all maps, then Skov et al., 2018²⁴, Sankararaman et al., 2014⁶, the union of all maps and Vernot et al., 2014⁷.

Figure 4. The mutational load of immune genes.

- A)** The mutational load in introgressed and non-introgressed fragments measured through Gerp Scores, in immune and non-immune eQTLs.
- (B)** The mutational load in introgressed fragments carrying immune and non-immune eQTLs, stratified by the mutational load in Africans, divided in percentiles.

Supplementary Figure Legends

Supplementary Figure 1. Average frequency in the 1000 Genomes stratified by the site frequency spectrum in Neanderthals (Nea_Count_Factor, total of 6 chromosomes) for non-introgressed (**A**) and introgressed alleles (**B**).

Supplementary Figure 2. Average frequency in the 1000 Genomes stratified by the site frequency spectrum in Neanderthals (Nea_Count_Factor, total of 6 chromosomes) for non-introgressed (**A**) and introgressed alleles (**B**), for coding (red) and intergenic regions (blue).

Supplementary Figure 3. The mutational load in introgressed and non-introgressed fragments measured through PhyloP (top), PhastCons (middle) and b-statistics (bottom).

Note that b-statistics measure the reduction in effective population size due to background selection, hence lower values indicate higher load. The violin plots show the distribution of 1000 5Mb block-bootstraps.

Supplementary Figure 4. The archaic mutational load in introgressed and non-introgressed fragments measured through PhyloP (top), PhastCons (middle) and b-statistics (bottom), stratified in bins of 20% percentiles of the mutational load in Africans, calculated with the same score.

Note that b-statistics measure the reduction in effective population size due to background selection, hence lower values indicate higher load. The violin plots show the distribution of 1000 5Mb block-bootstraps. All comparisons between introgressed and non-introgressed show archaic stronger mutational load in non-introgressed regions (p -values < 0.05), with the exception of the percentiles (20-40 % and 60-80 % in the PhyloP scores).

Supplementary Figure 5. Observed candidates overlapping Neanderthal fragments for different immune and non-immune GWAS studies.

Immune GWAS candidates are defined as the union of three studies^{40–42}, and are represented in the plot as FARMARELL. Other non-autoimmune GWAS were either downloaded from <https://www.ebi.ac.uk/gwas/> or taken from Farh et al⁴⁰. The observed overlap is represented as a red dot, versus the distribution of 100 simulated sets GWAS, using the same resampling scheme described in **Figure 3**. Columns represent different maps of Neanderthal ancestry, as described in **Figure 3**.

Figure 1

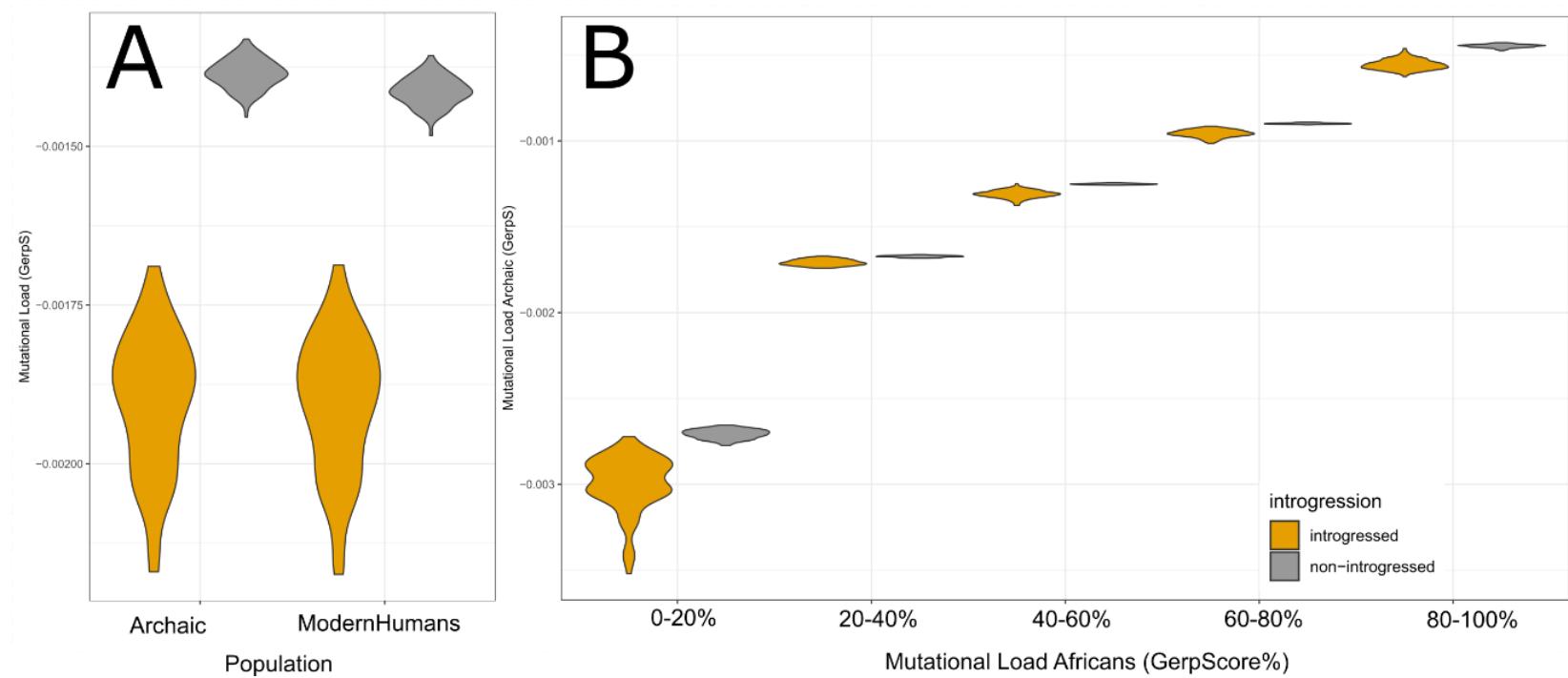


Figure 2

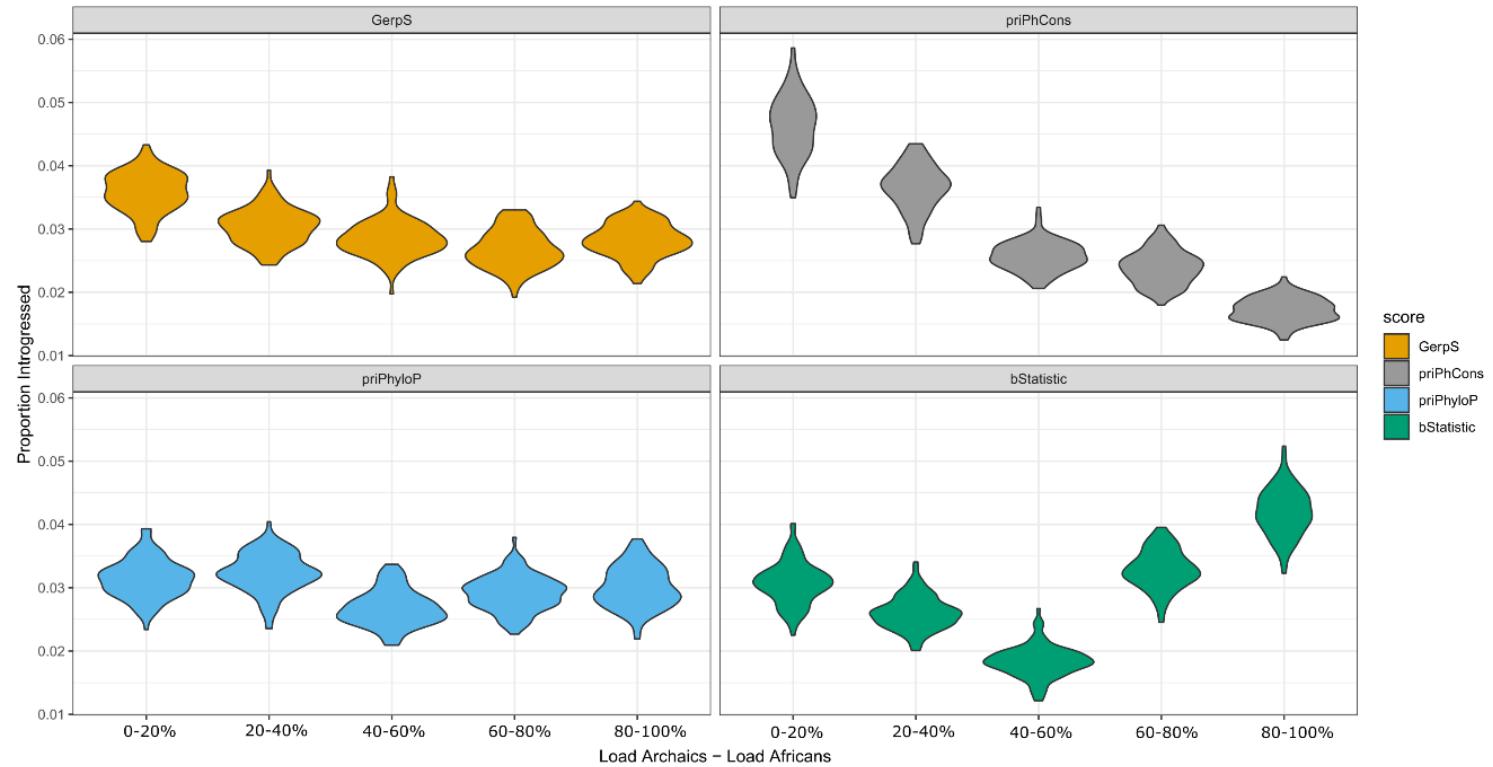


Figure 3

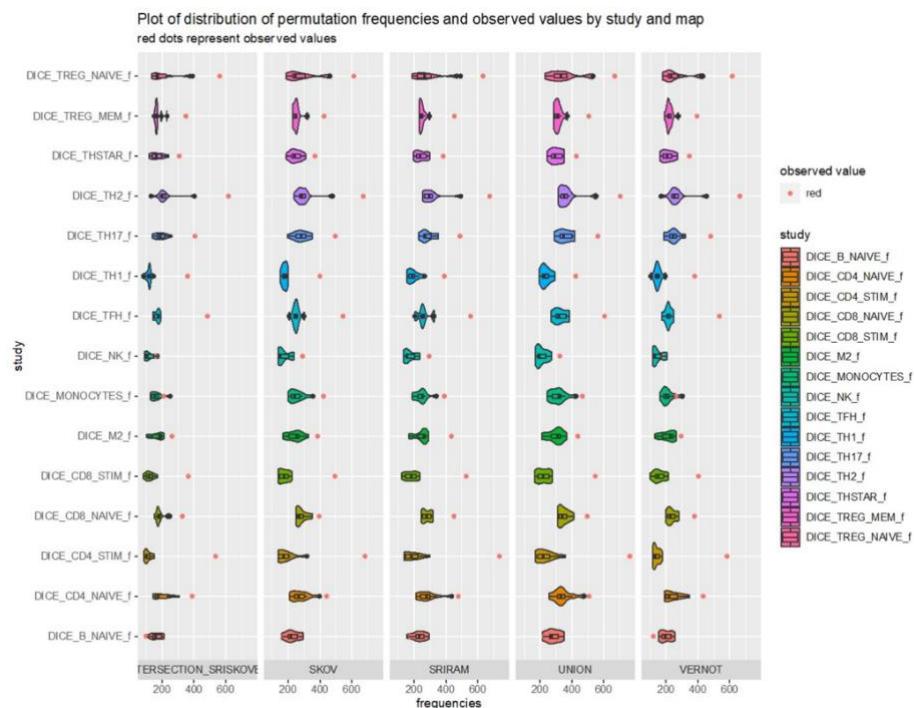
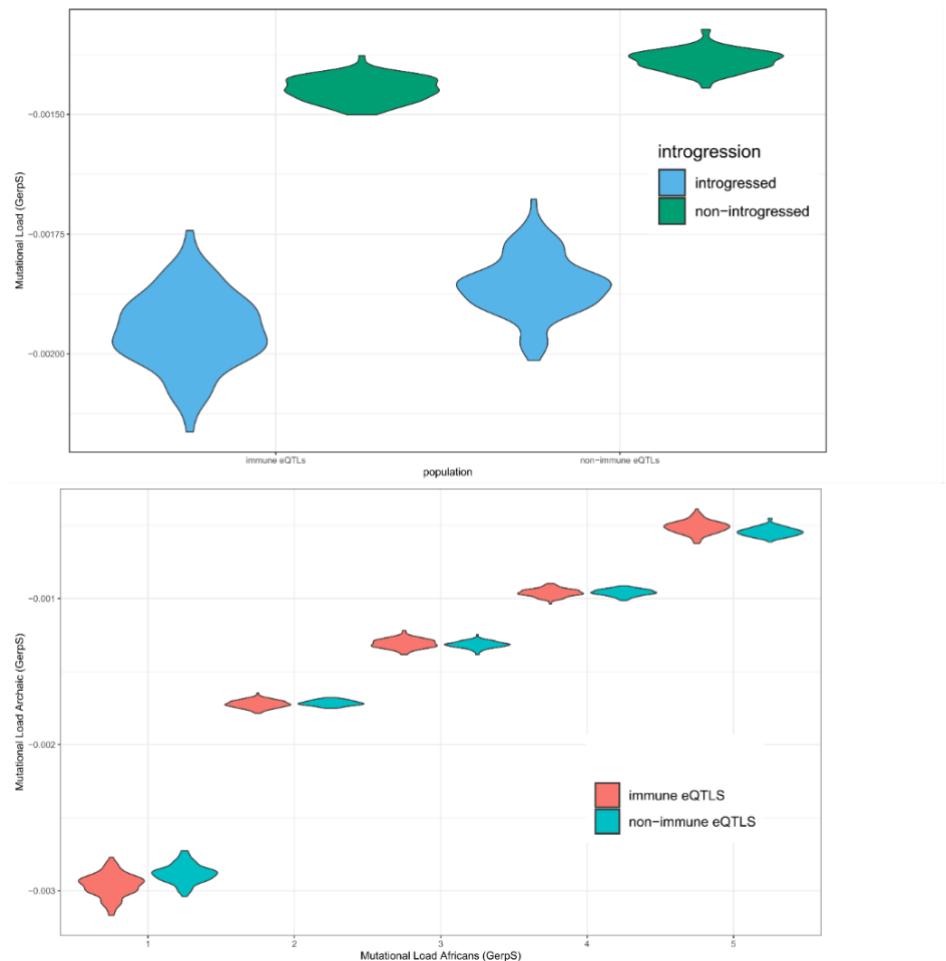
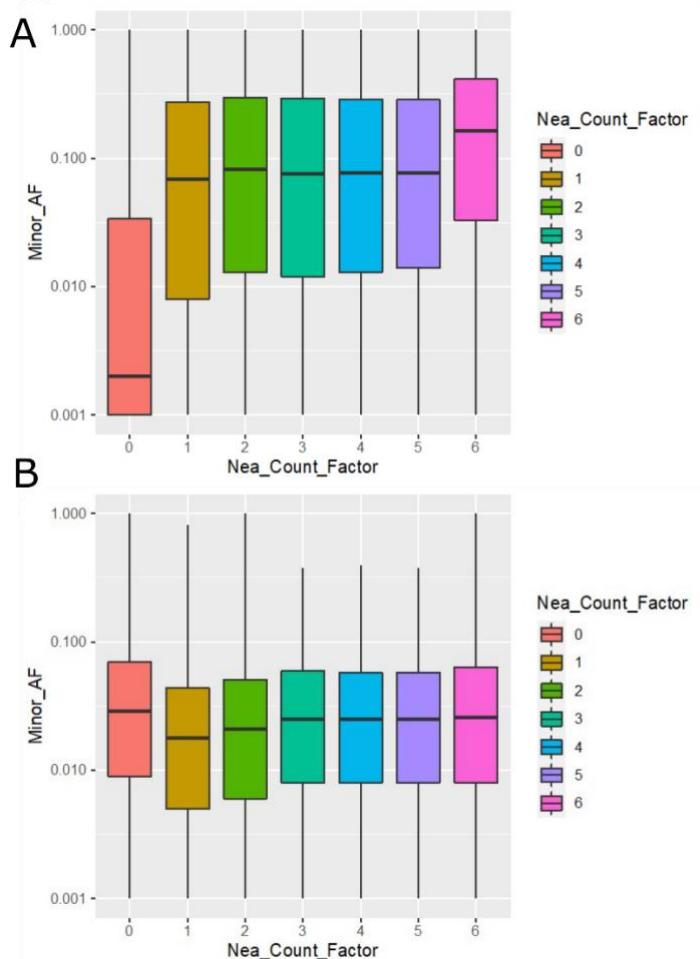


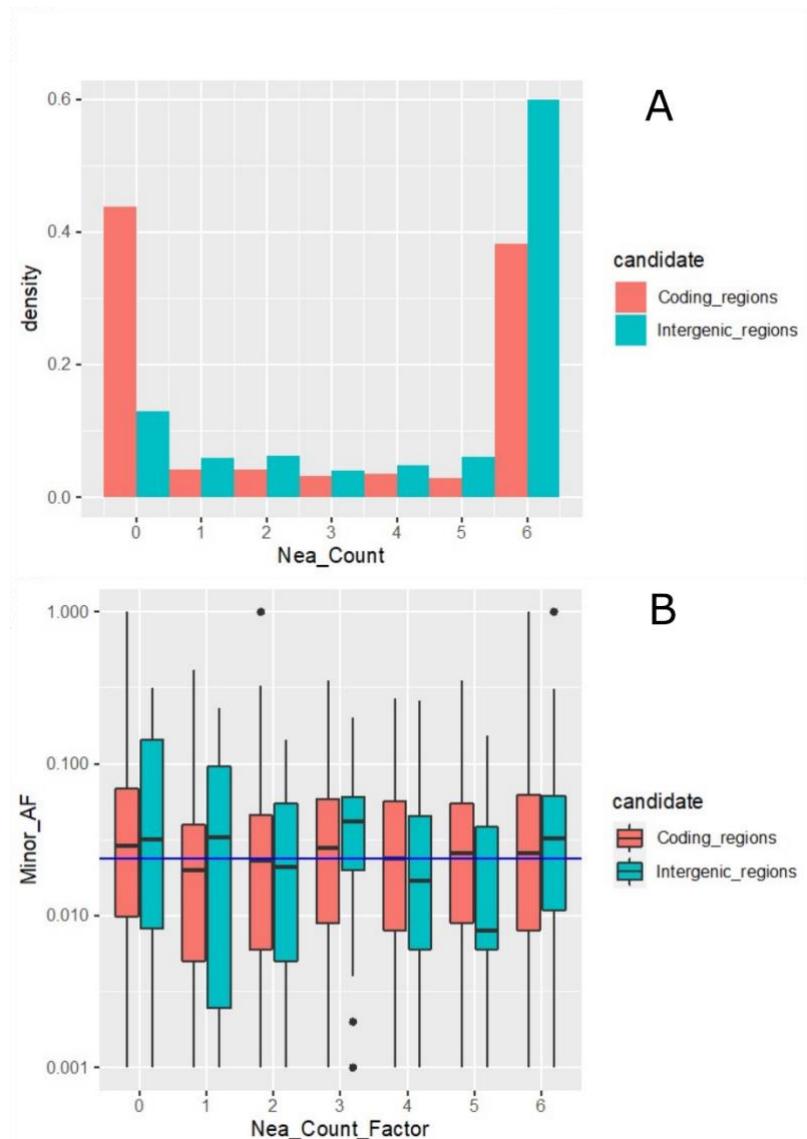
Figure 4



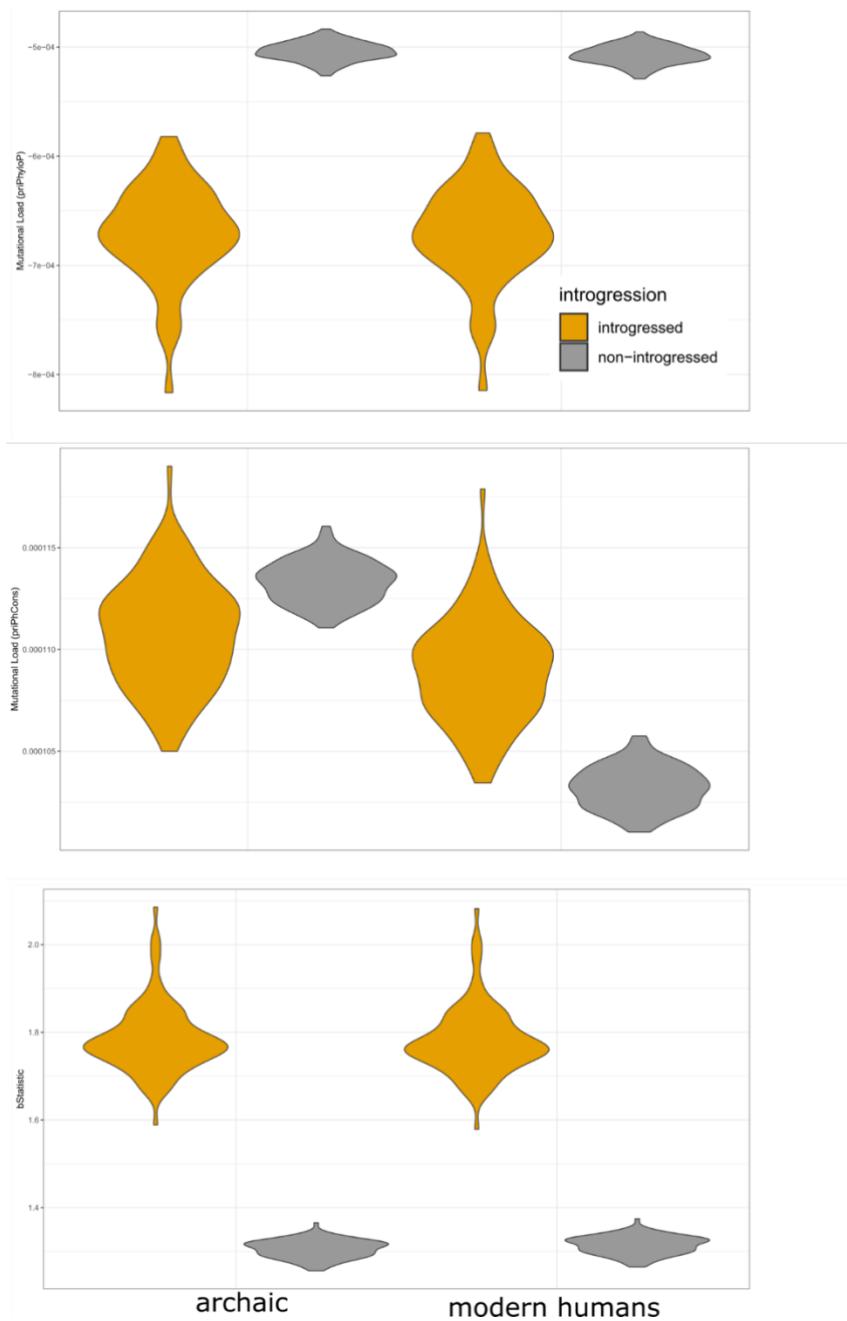
Supplementary Figure 1



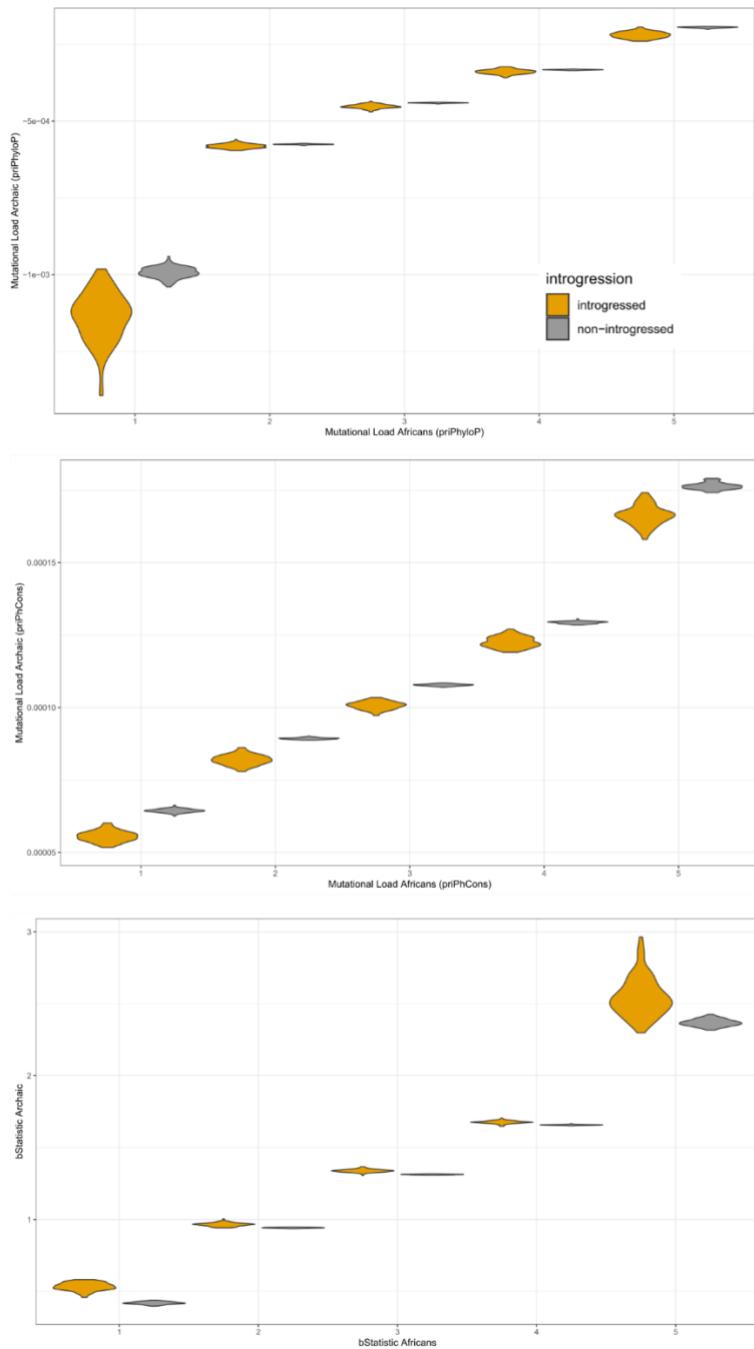
Supplementary Figure 2



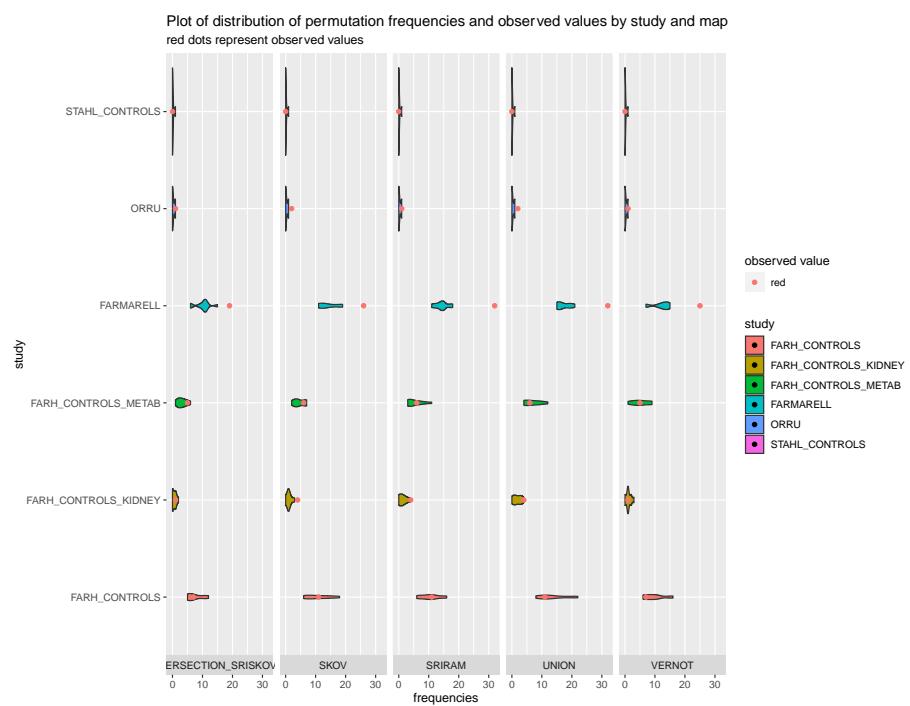
Supplementary Figure 3



Supplementary Figure 4



Supplementary Figure 5



Acknowledgements

Dr. Alessio Gerussi was the recipient of an EMBO Scientific Exchange Grant (number 8854) to carry out research in the lab of Dr. Viviane Slon (Dan David Center for Human Evolution and Biohistory Research, Tel Aviv University, Klausner St 12, Tel Aviv, Israel). The topic of the research project was "Defining the role of archaic genetic variants in the susceptibility to autoimmune diseases". The fellowship started on 2 November 2020, and EMBO funding had been granted for a total period of 90 days.

References

1. Dannemann, M., Andrés, A. M. & Kelso, J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *Am. J. Hum. Genet.* **98**, 22–33 (2016).
2. Gower, G., Picazo, P. I., Fumagalli, M. & Racimo, F. Detecting adaptive introgression in human evolution using convolutional neural networks. *Elife* **10**, (2021).
3. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* **16**, 359–371 (2015).
4. Racimo, F., Marnetto, D. & Huerta-Sánchez, E. Signatures of archaic adaptive introgression in present-day human populations. *Mol. Biol. Evol.* **34**, 296–317 (2017).
5. Harris, K. & Nielsen, R. The genetic cost of neanderthal introgression. *Genetics* **203**, 881–891 (2016).
6. Sankararaman, S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354 (2014).
7. Vernot, B. et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* (80-.). **352**, 235 LP – 239 (2016).
8. Castellano, S. et al. Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6666–6671 (2014).
9. Mafessoni, F. & Prüfer, K. Better support for a small

effective population size of Neandertals and a long shared history of Neandertals and Denisovans. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E10256–E10257 (2017).

10. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Curr. Biol.* **26**, 1241–1247 (2016).
11. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
12. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci.* **117**, 15132–15136 (2020).
13. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
14. Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science (80-.).* **358**, 655–658 (2017).
15. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
16. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
17. Kircher, M. *et al.* A general framework for estimating the

- relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
- 18. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
 - 19. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
 - 20. Davydov, E. V *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
 - 21. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
 - 22. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
 - 23. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
 - 24. Skov, L. *et al.* Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet.* **14**, e1007641 (2018).
 - 25. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**, 53–

- 61.e9 (2018).
26. Team, R. C. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* (2016).
 27. Skov, L. *et al.* The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. *Nature* (2020). doi:10.1038/s41586-020-2225-9
 28. Quach, H. *et al.* Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell* **167**, 643-656.e17 (2016).
 29. Rotival, M. & Quintana-Murci, L. Functional consequences of archaic introgression and their impact on fitness. *Genome Biol.* **21**, 3 (2020).
 30. Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715.e16 (2018).
 31. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
 32. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348**, 648–660 (2015).
 33. Hajdinjak, M. *et al.* Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry. *Nature* **592**, 253–257 (2021).
 34. Petr, M., Pääbo, S., Kelso, J. & Vernot, B. Limits of long-term selection against Neandertal introgression. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 1639–1644 (2019).
 35. Fu, Q. *et al.* An early modern human from Romania with a

- recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
- 36. Villanea, F. A. & Schraiber, J. G. Multiple episodes of interbreeding between Neanderthal and modern humans. *Nat. Ecol. Evol.* **3**, 39–44 (2019).
 - 37. Juric, I., Aeschbacher, S. & Coop, G. The Strength of Selection against Neanderthal Introgression. *PLoS Genet.* **12**, e1006340 (2016).
 - 38. Dutheil, J. Y., Munch, K., Nam, K., Mailund, T. & Schierup, M. H. Strong Selective Sweeps on the X Chromosome in the Human-Chimpanzee Ancestor Explain Its Low Divergence. *PLoS Genet.* **11**, e1005451 (2015).
 - 39. Nam, K. *et al.* Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6413–6418 (2015).
 - 40. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337 (2014).
 - 41. Márquez, A. *et al.* Meta-analysis of Immunochip data of four autoimmune diseases reveals novel single-disease and cross-phenotype associations. *Genome Med.* **10**, 97 (2018).
 - 42. Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* **48**, 510–518 (2016).

CHAPTER 7: Summary, Conclusions and Future Perspectives

Summary

High-throughput sequencing and genome-wide association studies have generated a large amount of data that has helped to better understand the genetic architecture of complex traits, including PBC¹. Although association studies have elucidated several novel loci and linked pathways, the functional characterization is still in its infancy and much of the information remains to be explored².

High-throughput sequencing technology has also been applied to characterize the genome of ancient hominins like Neandertals. There is a growing interest in the synthesis of genetic medicine with the growing knowledge of evolutionary history of genetic variants, also following the progressive increase in the available ancient genomes sequenced at high coverage³. Most, if not all, the genetic variants that contribute to the susceptibility of complex traits belong to conserved molecular pathways essential for cellular life. Furthermore, diseases do not show the same prevalence in all populations, and genetic history does explain a part of this variability³. The results derived from the 1000 Genomes project are an example of the magnitude of the genomic variation present across populations of present-day humans⁴.

The main goal of this PhD project was to leverage this great body of genetic knowledge to better characterize the genetic architecture of immune diseases, with a main focus on PBC. Our work has employed established statistical methods such as

meta-analysis as well as novel data mining tools, such as packages dedicated to the study of chrX and ML softwares. In addition, we have also applied established computational methods and ML tools to shed light on the evolutionary history of variants associated with immune-mediated traits.

More specifically, the meta-analysis of previous GWAS in PBC has identified additional risk loci and, by means of functional annotation of credible causal variants and multi-omic analysis, has produced a list of candidate genes together with several drugs that are potentially suitable for re-purposing to PBC.

The extensive analysis of chrX has identified several suggestive new loci associated with PBC located on this chromosome. The major finding of the XWAS approach has been the identification of a genome-wide significantly associated locus characterized by the presence of different genes and of a superenhancer possibly involved in their co-regulation, as well as in the regulation of FOXP3 (which is located in the same TAD).

Based on the new set of variants identified in the international meta-analysis, we have generated a novel PRS that has been incorporated in a new integrative risk model. This model, that has included 22 non-HLA variants, one HLA variant and sex, is accurate (AUC 0.83 and 0.81 in the two cohorts under study) and well calibrated. The PRS has pinpointed a subgroup of subjects at strikingly higher risk (OR ~ 14) of developing the disease that should be the target of tailored follow-up strategies. Prospective studies are needed to evaluate the potential role of PRS in first-degree relatives of patients with PBC.

We have also presented the first example of a successful, proof-of-concept analysis of GWAS data with ML to study genetic liability of PBC. ML is computationally feasible and generates accurate information that can be leveraged for disease prediction (AUC 0.73). Our ML-based model predicts genetic susceptibility to PBC through a methodologically innovative and explainable method. The innovation relies on the explainability of the rules for disease prediction, predicting genetic liability at the individual level; in addition, rules have considered groups of variants instead of single variants alone, paving the way for integrating gene-gene interactions in genetic predictive models. Yet, the accuracy of the ML-based model was lower than the integrative risk model including PRS, sex and HLA; we provide an example of robustness of PRSs in a specific disease model of interest and also show the trade-off required between explainability, computational burden and accuracy.

Finally, to study the evolutionary determinants of variants related to immune-mediated complex traits, we have leveraged information about genetic variants of archaic hominins and genomic estimates of deleteriousness at nucleotide-resolution to estimate the mutational load in Neanderthals along the genome. We have shown that the allele frequency in high coverage Neanderthal genomes is informative of the fate of alleles in the human population. While the common identified pattern is that Neandertal fragments bringing more deleterious variants have been purged away, several fragments involved in immunity are observed nowadays at high frequencies in human populations.

After confirming the enrichment of Neanderthal ancestry in immune-related traits and genes, we also observed that regions of the human genome carrying putatively deleterious variants and involved in immunity do not show traces of purifying selection but rather the opposite, reinforcing the notion that immune-related genes are under balancing selection.

Conclusion and application to translational medicine

Translational medicine has at its core interdisciplinarity, and aims to combine disciplines, resources, expertise, and techniques to improve prevention, diagnosis, and therapies of human diseases. In line with the mission of the PhD Program in Translational and Molecular Medicine (DIMET), this project has touched several fields, from paleogenetics to biomedical informatics, from statistical genetics to genomic medicine.

Translational medicine has an intimate link with Personalised medicine, which refers to a medical paradigm where individuals' phenotypes and genotypes are characterized to improve targeted prevention strategies and/or tailored treatment allocation⁵. Although most of the variation between individuals has no effect on health, an individual's health stems from genetic variation with influences from the environment.

This PhD work has generated several lines of evidence that involve different aspects of PBC, which worked as a proxy for a complex autoimmune human trait. These new data can have an impact to better clarify the disease pathogenesis, to improve risk stratification and for drug discovery/repurpose. Moreover, the study of evolutionary determinants of immunity and autoimmunity has brought concepts inherently part of evolutionary biology and medicine to the field of autoimmunity. To know evolutionary history of variants provides an explanatory framework which goes beyond description of disease association, contributing to

understand why a variant is crucial for a specific biological function and why complex traits have different incidence across populations³.

The results of the meta-analysis identified new loci that add additional insights into the pathogenesis of PBC⁶. Most of the new signals are related to chemokine/cytokine signaling, inflammation and immune cell trafficking and differentiation. Further, the study of the chrX has allowed us to focus the attention to a novel locus of high biological interest. The best association signal points to a unique LD region characterized by the presence of 7 genes (TIMM17B, PQBP1, PIM2, SLC35A2, OTUD5, KCND1, and GRIPAP1) and a SE, GH0XJ048933 (within OTUD5), which presents features with a potential impact on PBC pathogenesis.

It is well established that the study of the chrX enlarges the explained heritability of complex traits⁷. In addition, there is a specific line of evidence that supports the role of chrX in PBC specifically^{8–10}. Therefore, we aim to characterize the functional role of the superenhancer GH0XJ048933 to understand how the chrX can influence the development of PBC. Next steps will be the functional dissection of the molecular mechanism linking the identified chrX super-enhancer GH0XJ048933 with the pathogenesis of PBC and the comparison of gene expression levels of GH0XJ048933-associated enhancer RNA in CD4+ and CD8+ lymphocytes and monocytes/macrophages from PBC patients and healthy controls.

In terms of risk stratification strategies, there is a growing interest in the application of PRSs and ML models into clinical medicine^{11,12}.

PRSs capture independent risk, which is complementary to classical risk factors and clinical risk scores¹³. We have shown that a novel version of a PRS for PBC is accurate and well-calibrated and identifies subjects with high risk of developing PBC. Thus, PRS can improve the accuracy of early and targeted prevention. There is already evidence in other autoimmune conditions, such as celiac disease, that the application of a PRS in clinical practice can improve disease detection as compared to HLA testing¹⁴. We foresee the application of PRS in PBC as a novel risk stratifier that can identify people at risk to develop the disease that need tailored follow-up.

As far as ML is concerned, our proof-of-concept study has clearly demonstrated that there is potential for this novel way to approach genetic data, because we have shown the feasibility of a ML analysis of genetic variants to generate intelligible rules for disease prediction. Moreover, our preliminary data support the use of ML to study epistatic interactions from a novel perspective. In terms of accuracy of prediction, our ML pipeline was less accurate than the PRS applied to the same cohorts. Yet, the two analyses are only partially comparable, since for the former one cohort was used as the training set and the other one as the validation set, while for the latter the effect sizes were derived from a third study (the meta-analysis) and applied independently to the two Italian cohorts. With this caveat in mind, and taking

into account the limitation that this is an example of a specific trait of interest, so that generalization is possible only to some extent, the debate whether ML really brings something more than PRS is open. It is conceivable that ML algorithms may create more powerful predictive models when they are applied to large individual-level dataset¹⁵; unfortunately, that was not our case, and it is likely that the limited sample size of our training and validation cohorts have hampered the power of Rulex LLM. Conversely, it is true that PRSs model genetic liability in a more simplistic way (since they adopt fixed linear models), but have the advantage to use summary statistics to get effect size estimates¹³. In addition, from a computational perspective, PRSs have now established methods with extensive validation, while guidelines for ML models are still in their infancy^{12,13}. A possible advantage of Rulex LLM is the opportunity to forecast genetic liability of a specific individual at risk in the clinic. PRSs generate a numeric score for each subject, which is difficult to put in context by itself, because it is not clear which threshold a clinician should use to put the number in the right clinical context. To overcome this issue, it is common practice to use the empirical method, i.e. individuals are divided into strata such as quartiles, and we also adopted this standard method when presenting our PRS^{13,16}. On a different note, Rulex LLM can forecast whether the subject is a case or a control telling the user the level of confidence of this prediction and the rule that has been used to make this prediction. Differently from the PRS score number, the confidence of a prediction represents a value that has more

immediate intelligibility and practical translation. A recent proof-of-concept study introduced a novel ML method called Mondrian Cross-Conformal Prediction (MCCP) capable of estimating confidence levels for an individual's predicted risk. MCCP provides estimated probability values for assigning an individual as a case or control, which has direct utility in clinics rather than group-wise estimates, which arbitrarily define the top 10%, 5%, or 1% of samples as the high-risk group¹⁶. Thus, ML models might be superior to PRSs in terms of clinical translation of predictions. Moreover, ML algorithms employ multivariate, non-parametric methods that can recognize patterns from non-normally distributed and strongly correlated data, which is a hurdle typically complex for linear models of statistical genetics¹⁵. Besides accuracy of predictions, ML models have foreseeable applications to study highly interactive complex data structures, such as gene-gene interaction networks¹⁵. There is growing preliminary evidence that joint-modeling of gene interactions can increase the power to identify pathways relevant for a trait¹⁷. However, it is still to be determined whether our specific workflow is scalable when more variants are exponentially included as input features.

We have clearly shown in our study that the incorporation of information about Neanderthal ancestry improves the characterization of the deleteriousness of genomic variants present in introgressed fragments. Individuals have different genetic backgrounds based on their ancestry and these different histories alter the relationships between genotypes,

environmental factors and risk of disease, to inform personalized predictions about disease risk the field needs considering evolution by quantifying genetic ancestry³. One of the most important gaps that genetics needs to fill is the ongoing under-representation of minorities¹⁸. Incorporating evolutionary information could help to this end.

Despite our work being focused only on genomic data, the future of personalized medicine relies on multi-omics integration. To do so, especially for rare diseases, international endeavors should be implemented, to cut costs and address many, if not all, immune-mediated diseases altogether, breaking cultural barriers between clinicians and computational scientists, and educating the public about the benefit of the access to big personal health datasets.

Functional characterization of GWAS often can point toward druggable pathways. Our large, trans-ethnic GWMA of PBC has identified candidate genes to be prioritised together with drugs potentially suitable for re-purposing. In the past, some authors have advocated the failure of the GWAS approach in PBC as tool to confidently identify therapeutic targets, following the disappointing results of Ustekinumab, a monoclonal antibody targeting IL-12/IL-23, whose gene had been tagged by one of the first GWAS¹⁹. The negative result of Ustekinumab in PBC provided an example of the methodological limitations of GWAS. The genomic community of scientists has progressively learnt over time how to face these limitations by different strategies to dissect tagged loci². It is likely that genes spatially located close

to the GWAS signal played a much more complex role than predicted. The complexity behind locus dissection in GWAS has been brilliantly shown in other fields of medicine such as obesity, where variants in the FTO region harbor the strongest genetic association but FTO locus does not directly control adipocyte homeostasis. Rather, the causal variant rs1421085 disrupts a repressor binding site, which leads to depression of other two genes essential in early adipocyte differentiation located very far from the FTO locus²⁰. This study reinforced the notion that dissection of regulatory control of GWAS variants requires integration of several layers of information that in the first years of GWAS was probably not fully captured by investigators²¹. A remarkable example of this strategy is the recent work by the Manolis Kellis' lab, which released a compendium of 10,000 epigenomic maps across 800 samples to annotate 30,000 genetic loci that were associated with 540 traits, revealing pervasive pleiotropy in human traits²². This study showed the significance of high-resolution epigenomic annotations to study genetics of complex traits. We should also bear in mind the difference between disease susceptibility and disease progression, two different and possibly unrelated processes, so that a good therapeutic target should not necessarily be found immediately downstream the susceptibility genetic locus. To this end, drug re-purposing in PBC is not of immediate translation. The evaluation of immunomodulators in PBC would likely require a modification in clinical trial design, adopting different endpoints

than those used for anti-cholestatic medications and different selection criteria.

References

1. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
2. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 1–21 (2020).
3. Benton, M. L. *et al.* The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* (2021). doi:10.1038/s41576-020-00305-9
4. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
5. Zeggini, E., Gloyn, A. L., Barton, A. C. & Wain, L. V. Translational genomics and precision medicine: Moving from the lab to the clinic. *Science (80-)*. **365**, 1409 LP – 1413 (2019).
6. Cordell, H. J. *et al.* An international genome-wide meta-analysis of primary biliary cholangitis: novel risk loci and candidate drugs. *J. Hepatol.* (2021). doi:<https://doi.org/10.1016/j.jhep.2021.04.055>
7. Wise, A. L., Gyi, L. & Manolio, T. A. EXclusion: Toward integrating the X chromosome in genome-wide association analyses. *Am. J. Hum. Genet.* **92**, 643–647 (2013).
8. Invernizzi, P. *et al.* Frequency of monosomy X in women with primary biliary cirrhosis. *Lancet* **363**, 533–535 (2004).
9. Gerussi, A., Carbone, M., Corpechot, C. & Schramm, C.

The genetic architecture of primary biliary cholangitis. *Eur. J. Med. Genet.* **64**, 104292 (2021).

10. Gerussi, A., Cristoferi, L., Carbone, M., Asselta, R. & Invernizzi, P. The immunobiology of female predominance in primary biliary cholangitis. *J. Autoimmun.* **95**, 124–132 (2018).
11. Wand, H., Lambert, S. A., Tamburro, C. & Iacocca, M. A. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**, (2021).
12. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. **0123456789**,
13. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **3**, 11–13 (2020).
14. Sharp, S. A. *et al.* A single nucleotide polymorphism genetic risk score to aid diagnosis of coeliac disease: a pilot study in clinical care. *Aliment. Pharmacol. Ther.* **52**, 1165–1173 (2020).
15. de los Campos, G., Vazquez, A. I., Hsu, S. & Lello, L. Complex-Trait Prediction in the Era of Big Data. *Trends Genet.* **34**, 746–754 (2018).
16. Sun, J. *et al.* Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction. *Nat. Commun.* **12**, 1–9 (2021).
17. Silver, M. *et al.* Pathways-driven sparse regression identifies pathways and genes associated with high-

density lipoprotein cholesterol in two Asian cohorts. *PLoS Genet.* **9**, e1003939 (2013).

18. Green, E. D. *et al.* Strategic vision for improving human health at The Forefront of Genomics. *Nature* **586**, 683–692 (2020).
19. Hirschfield, G. M. *et al.* Ustekinumab for patients with primary biliary cholangitis who have an inadequate response to ursodeoxycholic acid: A proof-of-concept study. *Hepatology* **64**, 189–199 (2016).
20. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans . *N. Engl. J. Med.* **373**, 895–907 (2015).
21. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* (2020). doi:<https://doi.org/10.1016/j.tig.2020.08.009>
22. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, (2021).

Publications

- 1: Gerussi A, Verda D, Bernasconi DP, Carbone M, Komori A, Abe M, Inao M, Namisaki T, Mochida S, Yoshiji H, Hirschfield G, Lindor K, Pares A, Corpechot C, Cazzagon N, Floreani A, Marzioni M, Alvaro D, Vespasiani-Gentilucci U, Cristoferi L, Valsecchi MG, Muselli M; Japan PBC Study Group; Global PBC Study Group; Italian PBC study group, Hansen BE, Tanaka A, Invernizzi P. Machine learning in primary biliary cholangitis: a novel approach for risk stratification. *Liver Int.* 2021 Dec 24. doi: 10.1111/liv.15141. Epub ahead of print. PMID: 34951722.
- 2: Cordell HJ, Fryett JJ, Ueno K, Darlay R, Aiba Y, Hitomi Y, Kawashima M, Nishida N, Khor SS, Gervais O, Kawai Y, Nagasaki M, Tokunaga K, Tang R, Shi Y, Li Z, Juran BD, Atkinson EJ, Gerussi A, Carbone M, Asselta R, Cheung A, de Andrade M, Baras A, Horowitz J, Ferreira MAR, Sun D, Jones DE, Flack S, Spicer A, Mulcahy VL, Byun J, Han Y, Sandford RN, Lazaridis KN, Amos CI, Hirschfield GM, Seldin MF, Invernizzi P, Siminovitch KA, Ma X, Nakamura M, Mells GF; PBC Consortia; Canadian PBC Consortium; Chinese PBC Consortium; Italian PBC Study Group; Japan-PBC-GWAS Consortium; US PBC Consortium; UK-PBC Consortium. Corrigendum to 'An international genome-wide meta-analysis of primary biliary cholangitis: Novel risk loci and candidate drugs' [J Hepatol 2021;75(3):572-581]. *J Hepatol.* 2021 Dec 9:S0168-8278(21)02219-4. doi: 10.1016/j.jhep.2021.11.015. Epub ahead of print. Erratum for: *J Hepatol.* 2021 Sep;75(3):572-581. PMID: 34895949.
- 3: Efe C, Lammert C, Taşçılar K, Dhanasekaran R, Ebik B, Higuera-de la Tijera F, Çalışkan AR, Peralta M, Gerussi A, Massoumi H, Catana AM, Purnak T, Rigamonti C, Aldana AJG, Khakoo N, Nazal L, Frager S, Demir N, Irak K, Melekoğlu-Ellik Z,

Kacmaz H, Balaban Y, Atay K, Eren F, Alvares-da-Silva MR, Cristoferi L, Urzua Á, Eşkazan T, Magro B, Snijders R, Barutçu S, Lytvyyak E, Zazueta GM, Demirezer-Bolat A, Aydin M, Heurgue-Berlot A, De Martin E, Ekin N, Yıldırım S, Yavuz A, Bıyük M, Narro GC, Kıyıcı M, Akyıldız M, Kahramanoğlu-Aksoy E, Vincent M, Carr RM, Günşar F, Reyes EC, Harputluoğlu M, Aloman C, Gatselis NK, Üstündağ Y, Brahm J, Vargas NCE, Güzelbulut F, Garcia SR, Aguirre J, Anders M, Ratusnu N, Hatemi I, Mendizabal M, Floreani A, Fagioli S, Silva M, Idilman R, Satapathy SK, Silveira M, Drenth JPH, Dalekos GN, N Assis D, Björnsson E, Boyer JL, Yoshida EM, Invernizzi P, Levy C, Montano-Loza AJ, Schiano TD, Ridruejo E, Wahlin S. Effects of immunosuppressive drugs on COVID-19 severity in patients with autoimmune hepatitis. *Liver Int.* 2021 Nov 30. doi: 10.1111/liv.15121. Epub ahead of print. PMID: 34846800.

4: Cappadona C, Paraboschi EM, Ziliotto N, Bottaro S, Rimoldi V, Gerussi A, Azimonti A, Brenna D, Brunati A, Cameroni C, Campanaro G, Carloni F, Cavardini G, Ciravegna M, Composto A, Converso G, Corbella P, D'Eugenio D, Dal Rì G, Di Giorgio SM, Grondelli MC, Guerrera L, Laffoucriere G, Lando B, Lopedote L, Maizza B, Marconi E, Mariola C, Matronola GM, Menga LM, Montorsi G, Papatolo A, Patti R, Profeta L, Rebasti V, Smidili A, Tarchi SM, Tartaglia FC, Tettamanzi G, Tinelli E, Stuani R, Bolchini C, Pattini L, Invernizzi P, Degenhardt F, Franke A, Duga S, Asselta R. MEDTEC Students against Coronavirus: Investigating the Role of Hemostatic Genes in the Predisposition to COVID-19 Severity. *J Pers Med.* 2021 Nov 9;11(11):1166. doi: 10.3390/jpm11111166. PMID: 34834519; PMCID: PMC8622845.

5: Cristoferi L, Gerussi A, Invernizzi P. Anti-gp210 and other anti-nuclear pore complex autoantibodies in primary biliary cholangitis: What we know and what we should know. *Liver Int.*

2021 Mar;41(3):432-435. doi: 10.1111/liv.14791. PMID: 34542229.

6: Gerussi A, Carbone M, Corpechot C, Schramm C, Asselta R, Invernizzi P. The genetic architecture of primary biliary cholangitis. *Eur J Med Genet*. 2021 Sep;64(9):104292. doi: 10.1016/j.ejmg.2021.104292. Epub 2021 Jul 23. PMID: 34303876.

7: Balsano C, Alisi A, Brunetto MR, Invernizzi P, Burra P, Piscaglia F; Special Interest Group (SIG) Artificial Intelligence and Liver Diseases; Italian Association for the Study of the Liver (AISF). The application of artificial intelligence in hepatology: A systematic review. *Dig Liver Dis*. 2021 Jul 12:S1590-8658(21)00317-0. doi: 10.1016/j.dld.2021.06.011. Epub ahead of print. PMID: 34266794.

8: COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature*. 2021 Dec;600(7889):472-477. doi: 10.1038/s41586-021-03767-x. Epub 2021 Jul 8. PMID: 34237774; PMCID: PMC8674144.

9: Cordell HJ, Fryett JJ, Ueno K, Darlay R, Aiba Y, Hitomi Y, Kawashima M, Nishida N, Khor SS, Gervais O, Kawai Y, Nagasaki M, Tokunaga K, Tang R, Shi Y, Li Z, Juran BD, Atkinson EJ, Gerussi A, Carbone M, Asselta R, Cheung A, de Andrade M, Baras A, Horowitz J, Ferreira MAR, Sun D, Jones DE, Flack S, Spicer A, Mulcahy VL, Byan J, Han Y, Sandford RN, Lazaridis KN, Amos CI, Hirschfield GM, Seldin MF, Invernizzi P, Siminovitch KA, Ma X, Nakamura M, Mells GF; PBC Consortia; Canadian PBC Consortium; Chinese PBC Consortium; Italian PBC Study Group; Japan-PBC-GWAS Consortium; US PBC Consortium; UK-PBC Consortium. An international genome-wide meta-analysis of primary biliary cholangitis: Novel

risk loci and candidate drugs. *J Hepatol.* 2021 Sep;75(3):572-581. doi: 10.1016/j.jhep.2021.04.055. Epub 2021 May 23. Erratum in: *J Hepatol.* 2021 Dec 9;; PMID: 34033851.

10: Saettini F, Fazio G, Moratto D, Galbiati M, Zucchini N, Ippolito D, Dinelli ME, Imberti L, Mauri M, Melzi ML, Bonanomi S, Gerussi A, Pinelli M, Barisani C, Bugarin C, Chiarini M, Giacomelli M, Piazza R, Cazzaniga G, Invernizzi P, Giliani SC, Badolato R, Biondi A. Case Report: Hypomorphic Function and Somatic Reversion in DOCK8 Deficiency in One Patient With Two Novel Variants and Sclerosing Cholangitis. *Front Immunol.* 2021 Apr 16;12:673487. doi: 10.3389/fimmu.2021.673487. PMID: 33936120; PMCID: PMC8085392.

11: Gerussi A, Natalini A, Antonangeli F, Mancuso C, Agostinetto E, Barisani D, Di Rosa F, Andrade R, Invernizzi P. Immune-Mediated Drug-Induced Liver Injury: Immunogenetics and Experimental Models. *Int J Mol Sci.* 2021 Apr 27;22(9):4557. doi: 10.3390/ijms22094557. PMID: 33925355; PMCID: PMC8123708.

12: Cristoferi L, Calvaruso V, Overi D, Viganò M, Rigamonti C, Degasperi E, Cardinale V, Labanca S, Zucchini N, Fichera A, Di Marco V, Leutner M, Venere R, Picciotto A, Lucà M, Mulinacci G, Palermo A, Gerussi A, D'Amato D, Elisabeth O'Donnell S, Cerini F, De Benedittis C, Malinverno F, Ronca V, Mancuso C, Cazzagon N, Ciaccio A, Barisani D, Marzioni M, Floreani A, Alvaro D, Gaudio E, Invernizzi P, Carpino G, Nardi A, Carbone M; Italian PBC Registry. Accuracy of Transient Elastography in Assessing Fibrosis at Diagnosis in Naïve Patients With Primary Biliary Cholangitis: A Dual Cut-Off Approach. *Hepatology.* 2021 Sep;74(3):1496-1508. doi: 10.1002/hep.31810. Epub 2021 May 28. PMID: 33724515; PMCID: PMC8518641.

13: Efe C, Dhanasekaran R, Lammert C, Ebik B, Higuera-de la Tijera F, Aloman C, Rıza Çalışkan A, Peralta M, Gerussi A, Massoumi H, Catana AM, Torgutalp M, Purnak T, Rigamonti C, Gomez Aldana AJ, Khakoo N, Kacmaz H, Nazal L, Frager S, Demir N, Irak K, Ellik ZM, Balaban Y, Atay K, Eren F, Cristoferi L, Batibay E, Urzua Á, Snijders R, Kiyıcı M, Akyıldız M, Ekin N, Carr RM, Harputluoğlu M, Hatemi I, Mendizabal M, Silva M, Idilman R, Silveira M, Drenth JPH, Assis DN, Björnsson E, Boyer JL, Invernizzi P, Levy C, Schiano TD, Ridruejo E, Wahlin S. Outcome of COVID-19 in Patients With Autoimmune Hepatitis: An International Multicenter Study. *Hepatology*. 2021 Jun;73(6):2099-2109. doi: 10.1002/hep.31797. Erratum in: *Hepatology*. 2021 Dec 8;; PMID: 33713486; PMCID: PMC8250536.

14: D'Amato D, De Vincentis A, Malinverno F, Viganò M, Alvaro D, Pompili M, Picciotto A, Palitti VP, Russello M, Storato S, Pigozzi MG, Calvaruso V, De Gasperi E, Lleo A, Castellaneta A, Pellicelli A, Cazzagon N, Floreani A, Muratori L, Fagioli S, Niro GA, Feletti V, Cozzolongo R, Terreni N, Marzioni M, Pellicano R, Pozzoni P, Baiocchi L, Chessa L, Rosina F, Bertino G, Vinci M, Morgando A, Vanni E, Scifo G, Sacco R, D'Antò M, Bellia V, Boldizzoni R, Casella S, Omazzi B, Poggi G, Cristoferi L, Gerussi A, Ronca V, Venere R, Ponziani F, Cannavò M, Mussetto A, Fontana R, Losito F, Frazzetto E, Distefano M, Colapietro F, Labanca S, Marconi G, Grassi G, Galati G, O'Donnell SE, Mancuso C, Mulinacci G, Palermo A, Claar E, Izzi A, Picardi A, Invernizzi P, Carbone M, Vespasiani-Gentilucci U; Italian PBC Registry and the Club Epatologi Ospedalieri (CLEO)/Associazione Italiana Gastroenterologi ed Endoscopisti Digestivi Ospedalieri (AIGO) PBC Study Group. Real-world experience with obeticholic acid in patients with primary biliary cholangitis. *JHEP Rep.* 2021 Jan 27;3(2):100248. doi: 10.1016/j.jhepr.2021.100248. PMID: 33681748; PMCID: PMC7930359.

- 15: Asselta R, Paraboschi EM, Gerussi A, Cordell HJ, Mells GF, Sandford RN, Jones DE, Nakamura M, Ueno K, Hitomi Y, Kawashima M, Nishida N, Tokunaga K, Nagasaki M, Tanaka A, Tang R, Li Z, Shi Y, Liu X, Xiong M, Hirschfield G, Siminovitch KA; Canadian-US PBC Consortium; Italian PBC Genetics Study Group; UK-PBC Consortium; Japan PBC-GWAS Consortium, Carbone M, Cardamone G, Duga S, Gershwin ME, Seldin MF, Invernizzi P. X Chromosome Contribution to the Genetic Architecture of Primary Biliary Cholangitis. *Gastroenterology*. 2021 Jun;160(7):2483-2495.e26. doi: 10.1053/j.gastro.2021.02.061. Epub 2021 Mar 4. PMID: 33675743; PMCID: PMC8169555.
- 16: Montali L, Gragnano A, Miglioretti M, Frigerio A, Vecchio L, Gerussi A, Cristoferi L, Ronca V, D'Amato D, O'Donnell SE, Mancuso C, Lucà M, Yagi M, Reig A, Jopson L, Pilar S, Jones D, Pares A, Mells G, Tanaka A, Carbone M, Invernizzi P. Quality of life in patients with primary biliary cholangitis: A cross-geographical comparison. *J Transl Autoimmun*. 2021 Jan 6;4:100081. doi: 10.1016/j.jtauto.2021.100081. PMID: 33554101; PMCID: PMC7843515.
- 17: Mulinacci G, Cristoferi L, Palermo A, Lucà M, Gerussi A, Invernizzi P, Carbone M. Risk stratification in primary sclerosing cholangitis. *Minerva Gastroenterol Dietol*. 2020 Dec 10. doi: 10.23736/S1121-421X.20.02821-4. Epub ahead of print. PMID: 33300753.
- 18: Gerussi A, Halliday N, Carbone M, Invernizzi P, Thorburn D. Open challenges in the management of autoimmune hepatitis. *Minerva Gastroenterol Dietol*. 2020 Dec 3. doi: 10.23736/S1121-421X.20.02805-6. Epub ahead of print. PMID: 33267568.

- 19: Gerussi A, Rigamonti C, Elia C, Cazzagon N, Floreani A, Pozzi R, Pozzoni P, Claar E, Pasulo L, Fagioli S, Cristoferi L, Carbone M, Invernizzi P. Coronavirus Disease 2019 (COVID-19) in autoimmune hepatitis: a lesson from immunosuppressed patients. *Hepatol Commun.* 2020 Jun 9;4(9):1257–62. doi: 10.1002/hep4.1557. Epub ahead of print. PMID: 32838102; PMCID: PMC7300554.
- 20: Gerussi A, Restelli U, Croce D, Bonfanti M, Invernizzi P, Carbone M. Cost of illness of Primary Biliary Cholangitis - a population-based study. *Dig Liver Dis.* 2021 Sep;53(9):1167–1170. doi: 10.1016/j.dld.2020.07.029. Epub 2020 Aug 21. PMID: 32830065.
- 21: Palermo A, Gerussi A, Mulinacci G, Invernizzi P, Carbone M. Identifying Racial Disparities in Primary Biliary Cholangitis Patients: A Step Toward Achieving Equitable Outcomes Among All. *Dig Dis Sci.* 2021 May;66(5):1386–1387. doi: 10.1007/s10620-020-06528-4. PMID: 32789729.
- 22: Gerussi A, Bernasconi DP, O'Donnell SE, Lammers WJ, Van Buuren H, Hirschfield G, Janssen H, Corpechot C, Reig A, Pares A, Battezzati PM, Zuin MG, Cazzagon N, Floreani A, Nevens F, Gatselis N, Dalekos G, Mayo MJ, Thorburn D, Bruns T, Mason AL, Verhelst X, Kowdley K, van der Meer A, Niro GA, Beretta-Piccoli BT, Marzoni M, Belli LS, Marra F, Valsecchi MG, Lindor KD, Invernizzi P, Hansen BE, Carbone M; Italian PBC Study Group and the GLOBAL PBC Study Group. Measurement of Gamma Glutamyl Transferase to Determine Risk of Liver Transplantation or Death in Patients With Primary Biliary Cholangitis. *Clin Gastroenterol Hepatol.* 2021 Aug;19(8):1688–1697.e14. doi: 10.1016/j.cgh.2020.08.006. Epub 2020 Aug 7. PMID: 32777554.

- 23: Carbone M, Milani C, Gerussi A, Ronca V, Cristoferi L, Invernizzi P. Primary biliary cholangitis: a multifaceted pathogenesis with potential therapeutic targets. *J Hepatol.* 2020 Oct;73(4):965-966. doi: 10.1016/j.jhep.2020.05.041. Epub 2020 Jul 21. PMID: 32709365.
- 24: Gerussi A, Halliday N, Saffioti F, Bernasconi DP, Roccarina D, Marshall A, Thorburn D. Normalization of serum immunoglobulin G levels is associated with improved transplant-free survival in patients with autoimmune hepatitis. *Dig Liver Dis.* 2020 Jul;52(7):761-767. doi: 10.1016/j.dld.2020.04.012. Epub 2020 May 27. PMID: 32473882.
- 25: Carbone M, Kodra Y, Rocchetti A, Manno V, Minelli G, Gerussi A, Ronca V, Malinverno F, Cristoferi L, Floreani A, Invernizzi P, Conti S, Tarusco D. Primary Sclerosing Cholangitis: Burden of Disease and Mortality Using Data from the National Rare Diseases Registry in Italy. *Int J Environ Res Public Health.* 2020 Apr 29;17(9):3095. doi: 10.3390/ijerph17093095. PMID: 32365682; PMCID:PMC7246900.
- 26: Gerussi A, Lucà M, Cristoferi L, Ronca V, Mancuso C, Milani C, D'Amato D, O'Donnell SE, Carbone M, Invernizzi P. New Therapeutic Targets in Autoimmune Cholangiopathies. *Front Med (Lausanne).* 2020 Apr 7;7:117. doi: 10.3389/fmed.2020.00117. PMID: 32318580; PMCID: PMC7154090.
- 27: Bossen L, Rebora P, Bernuzzi F, Jepsen P, Gerussi A, Andreone P, Galli A, Terzioli B, Alvaro D, Labbadia G, Aloise C, Baiocchi L, Giannini E, Abenavoli L, Toniutto P, Marra F, Marzoni M, Niro G, Floreani A, Møller HJ, Valsecchi MG, Carbone M, Grønbaek H, Invernizzi P. Soluble CD163 and mannose receptor as markers of liver disease severity and prognosis in patients with primary biliary

cholangitis. *Liver Int.* 2020 Jun;40(6):1408-1414. doi: 10.1111/liv.14466. Epub 2020 Apr 24. PMID: 32279422.

28: Corpechot C, Chazouillères O, Belnou P, Montano-Loza AJ, Mason A, Ebadi M, Eurich D, Chopra S, Jacob D, Schramm C, Sterneck M, Bruns T, Reuken P, Rauchfuss F, Roccarina D, Thorburn D, Gerussi A, Trivedi P, Hirschfield G, McDowell P, Nevens F, Boillot O, Bosch A, Giostra E, Conti F, Poupon R, Parés A, Reig A, Donato MF, Malinverno F, Floreani A, Russo FP, Cazzagon N, Verhelst X, Goet J, Harms M, van Buuren H, Hansen B, Carrat F, Dumortier J; Global PBC Study Group. Long-term impact of preventive UDCA therapy after transplantation for primary biliary cholangitis. *J Hepatol.* 2020 Sep;73(3):559-565. doi: 10.1016/j.jhep.2020.03.043. Epub 2020 Apr 7. PMID: 32275981.

29: Hedin CRH, Sado G, Ndegwa N, Lytvyak E, Mason A, Montano-Loza A, Gerussi A, Saffioti F, Thorburn D, Nilsson E, Larsson G, Moum BA, van Munster KN, Ponsioen CY, Levy C, Nogueira NF, Bowlus CL, Gotlieb N, Shibolet O, Lynch KD, Chapman RW, Rupp C, Vesterhus M, Jørgensen KK, Rorsman F, Schramm C, Sabino J, Vermeire S, Zago A, Cazzagon N, Marschall HU, Ytting H, Ben Belkacem K, Chazouilleres O, Almer S; International PSC Study Group, Bergquist A. Effects of Tumor Necrosis Factor Antagonists in Patients With Primary Sclerosing Cholangitis. *Clin Gastroenterol Hepatol.* 2020 Sep;18(10):2295-2304.e2. doi: 10.1016/j.cgh.2020.02.014. Epub 2020 Feb 15. PMID: 32068151.

30: Gerussi A, D'Amato D, Cristoferi L, O'Donnell SE, Carbone M, Invernizzi P. Multiple therapeutic targets in rare cholestatic liver diseases: Time to redefine treatment strategies. *Ann Hepatol.* 2020 Jan-Feb;19(1):5-16. doi: 10.1016/j.aohep.2019.09.009. Epub 2019 Oct 31. PMID: 31771820.

- 31: Manno V, Gerussi A, Carbone M, Minelli G, Taruscio D, Conti S, Invernizzi P. A National Hospital-Based Study of Hospitalized Patients With Primary Biliary Cholangitis. *Hepatol Commun.* 2019 Jul 15;3(9):1250-1257. doi: 10.1002/hep4.1407. PMID: 31497745; PMCID: PMC6719751.
- 32: Raggi C, Fiaccadori K, Pastore M, Correnti M, Piombanti B, Forti E, Navari N, Abbadessa G, Hall T, Destro A, Di Tommaso L, Roncalli M, Meng F, Glaser S, Rovida E, Peraldo-Neia C, Olaizola P, Banales JM, Gerussi A, Elvevi A, Droz Dit Busset M, Bhoori S, Mazzaferro V, Alpini G, Marra F, Invernizzi P. Antitumor Activity of a Novel Fibroblast Growth Factor Receptor Inhibitor for Intrahepatic Cholangiocarcinoma. *Am J Pathol.* 2019 Oct;189(10):2090-2101. doi: 10.1016/j.ajpath.2019.06.007. Epub 2019 Jul 24. PMID: 31351075.
- 33: Ronca V, Gerussi A, Cristoferi L, Carbone M, Invernizzi P. Precision medicine in primary biliary cholangitis. *J Dig Dis.* 2019 Jul;20(7):338-345. doi: 10.1111/1751-2980.12787. Epub 2019 Jul 10. PMID: 31099953.
- 34: Bombaci M, Pesce E, Torri A, Carpi D, Crosti M, Lanzafame M, Cordigliero C, Sinisi A, Moro M, Bernuzzi F, Gerussi A, Geginat J, Muratori L, Terracciano LM, Invernizzi P, Abrignani S, Grifantini R. Novel biomarkers for primary biliary cholangitis to improve diagnosis and understand underlying regulatory mechanisms. *Liver Int.* 2019 Nov;39(11):2124-2135. doi: 10.1111/liv.14128. Epub 2019 May 15. PMID: 31033124.
- 35: Gerussi A, Carbone M, Invernizzi P. Editorial: liver transplantation for primary biliary cholangitis-the need for timely and more effective treatments. *Aliment Pharmacol Ther.* 2019 Feb;49(4):472-473. doi: 10.1111/apt.15095. PMID: 30689256.

36: Gerussi A, Invernizzi P. Better end points needed in primary sclerosing cholangitis trials. *Nat Rev Gastroenterol Hepatol*. 2019 Mar;16(3):143-144. doi: 10.1038/s41575-019-0110-5. PMID: 30655632.

37: Gerussi A, Cristoferi L, Carbone M, Asselta R, Invernizzi P. The immunobiology of female predominance in primary biliary cholangitis. *J Autoimmun*. 2018 Dec;95:124-132. doi: 10.1016/j.jaut.2018.10.015. Epub 2018 Oct 25. PMID: 30509386.

Acknowledgements

“Tu eri il più teorico di tutti. Dovevi agganciare tutto alle tue idee, anche allora. Valutare la situazione, trarre conclusioni. Esercitavi una rigida sorveglianza su te stesso. Le mattane restavano dentro. Un ragazzo ragionevole”, *Pastorale Americana*, Philip Roth.

“Un paese vuol dire non essere soli, sapere che nella gente, nelle piante, nella terra c’è qualcosa di tuo, che anche quando non ci sei resta ad aspettarti. Ma non è facile starci tranquillo. [...] Queste cose si capiscono con il tempo e l’esperienza. Possibile che a quarant’anni, e con tutto il mondo che ho visto, non sappia ancora cos’è il mio paese?”, *La luna e i falò*, Cesare Pavese.

“Allora ero così giovane e poco lungimirante che non capivo le ripercussioni che quelle scelte - o piuttosto quel rifiuto di fare delle scelte, quell’estenuante indecisione - avrebbero potuto avere sulla mia vita in futuro. Partire, come scoprii presto, non bastava a curare il senso di spaesamento, ma al contrario lo rafforzava. Mi mancava Israele proprio come un tempo mi mancava il resto del mondo, e scimmiettavo lo struggimento dei miei nonni con un esilio auto-inflitto. Credevo di poter continuare a provare posti nuovi, proprio come si continua a cambiare abito nel camerino di un negozio d’abbigliamento, e se qualcosa non mi andava bene, potevo sempre tornare indietro al punto di partenza, come se niente fosse. Pensavo che tutto mi avrebbe aspettato immobile e immutato nel tempo.”, *L’arte di partire*, Ayelet Tsabari.

“Vedi Sara, vorrei dirle, la corsa assomiglia più a un’arte marziale che a uno sport. Chi la ama compie una scelta estetica, accede a una disciplina interiore che c’entra pochissimo con l’attività sportiva. [...] Lo sforzo del maratoneta è, al contrario, precipuamente apollineo: restare lucidi dal primo all’ultimo metro, gestire la follia del corpo, tenerla sempre a un soffio dal suo impazzimento definitivo.”, Sulla corsa, Mauro Covacich.

“L’uomo vive ogni cosa subito per la prima volta, senza preparazioni. Come un attore che entra in scena senza aver mai provato [...] La vertigine potremmo anche chiamarla ebbrezza della debolezza. Ci si rende conto della propria debolezza e invece di resisterle, ci si vuole abbandonare ad essa.”, L’insostenibile leggerezza dell’essere, Milan Kundera.

“Non lasciatevi ingannare che sia poca cosa la vita! Bevetela a grandi sorsi! Non vi sarà bastata quando la dovrete perdere”, da “Contro la seduzione”, Bertold Brecht.

Ringrazio tutte le persone che mi hanno accompagnato in questo percorso fantastico, dai miei genitori ai tutor/supervisor, dai colleghi agli amici. Sono stati anni intensi, di grande crescita professionale e personale. Nominarvi personalmente facendo un elenco non è cosa che ben si adatta a questi miei tempi. Voi che ci siete stati già sapete; vi porto tutti nel cuore.