frontiers | Frontiers in Immunology

TYPE Mini Review
PUBLISHED 05 August 2022
DOI 10.3389/fimmu.2022.972107

A survey of genome-wide association studies, polygenic scores and UK Biobank highlights resources for autoimmune disease genetics

Rochi Saurabh[1], Césaire J. K. Fouodo[2], Inke R. König[2], Hauke Busch[1] and Inken Wohlers[1]*

[1]Medical Systems Biology, Lübeck Institute for Experimental Dermatology (LIED) and Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany, [2]Institute of Medical Biometry and Statistics (IMBS), University of Lübeck, Lübeck, Germany

Check for updates

OPEN ACCESS

EDITED BY
Kyle T. Amber,
Rush University, United States

REVIEWED BY
Elias Eliopoulos,
Agricultural University of Athens,
Greece

*CORRESPONDENCE
Inken Wohlers
Inken.Wohlers@uni-luebeck.de

SPECIALTY SECTION
This article was submitted to Autoimmune and Autoinflammatory Disorders, a section of the journal Frontiers in Immunology

RECEIVED 17 June 2022
ACCEPTED 12 July 2022
PUBLISHED 05 August 2022

CITATION
Saurabh R, Fouodo CJK, König IR, Busch H and Wohlers I (2022) A survey of genome-wide association studies, polygenic scores and UK Biobank highlights resources for autoimmune disease genetics.
Front. Immunol. 13:972107.
doi: 10.3389/fimmu.2022.972107

Autoimmune diseases share a general mechanism of auto-antigens harming tissues. Still. they are phenotypically diverse, with genetic as well as environmental factors contributing to their etiology at varying degrees. Associated genomic loci and variants have been identified in numerous genome-wide association studies (GWAS), whose results are increasingly used for polygenic scores (PGS) that are used to predict disease risk. At the same time, a technological shift from genotyping arrays to next generation sequencing (NGS) is ongoing. NGS allows the identification of virtually all - including rare - genetic variants, which in combination with methodological developments promises to improve the prediction of disease risk and elucidate molecular mechanisms underlying disease. Here we review current, publicly available autoimmune disease GWAS and PGS data based on information from the GWAS and PGS catalog, respectively. We summarize autoimmune diseases investigated, respective studies conducted and their results. Further, we review genetic data and autoimmune disease patients in the UK Biobank (UKB), the largest resource for genetic and phenotypic data available for academic research. We find that only comparably prevalent autoimmune diseases are covered by the UKB and at the same time assessed by both GWAS and PGS catalogs. These are systemic (systemic lupus erythematosus) as well as organ-specific, affecting the gastrointestinal tract (inflammatory bowel disease as well as specifically Crohn’s disease and ulcerative colitis), joints (juvenile ideopathic arthritis, psoriatic arthritis, rheumatoid arthritis, ankylosing spondylitis), glands (Sjögren syndrome), the nervous system (multiple sclerosis), and the skin (vitiligo).

KEYWORDS

autoimmune disease, genetics, genome-wide association study, GWAS, polygenic scores, genetic risk, UK Biobank, experimental factor ontology

Introduction

Autoimmune diseases are a range of diseases in which the immune response to self-antigens results in damage or dysfunction of tissues. It can be systemic or can affect specific organs or body systems. Autoimmune diseases are characterized by a multifactorial etiology, in which genetic factors interplay with environmental factors. Estimates of heritability, that is, variability in occurrence of autoimmune disease explained by genetic factors, vary considerably and have been reported to be between 42 and 91% for pediatric age-of-onset and lower for adult onset cases (1). Variation in the human major histocompatibility complex (MHC) regions harboring the human leukocyte antigen (HLA) genes is most strongly linked to autoimmune disease (2). Beside HLA, other genetic loci are shared between autoimmune diseases, with first investigations finding 47/107 (44%) immune-mediated disease risk variants associated with multiple, but not all such diseases (3), and later work identifying 244 shared disease loci (4). Accordingly, efforts are ongoing to unravel shared disease mechanisms based on shared genetic profiles (5, 6). Such genetics-driven systems approaches to autoimmune disease can largely benefit from public resources of genome and phenotype data as well as derived information. Here we perform a survey of autoimmune disease-related content in three such resources: (i) The NHGRI-EBI GWAS catalog (7) reporting on genome-wide association studies (GWAS), (ii) the newly established PGS catalog (8) having information on polygenic scores (PGS) and (iii) the UK Biobank (UKB) holding genetic and phenotypic data of ~500,000 people from the UK (9) (for details on data processing, see https://github.com/iwohlers/2022_autoimmune_review). While GWAS aim to identify associations of a large number of genetic variants with phenotypes or traits (10), the main goal of PGS studies is to estimate the risk of developing a disease or of the presence of a specific trait depending on the genetic profiles (11).

Autoimmune diseases and their relationships within biomedical ontologies

An ontology is a controlled vocabulary, formalizing domain knowledge into terms and their relationships. The Experimental Factor Ontology (EFO) is a biomedical ontology curated by the European Bioinformatics Institute (EBI) and used by the GWAS and PGS catalog for the purpose of disease classification (12). The part of EFO relating to autoimmune disease is shown in Suppl. Table S1 (EFO version 3.42.0; https://bioportal.bioontology.org/ontologies/EFO). It contains “is a” relationships between “parent” and “child” terms, e.g., rheumatoid arthritis (child term) is an autoimmune disease (parent term). The ontology branch of child terms for autoimmune disease contains 120 terms organized in up to five levels (Suppl. Table S1). For the disease-related part of EFO, some terms have been taken from other ontologies (denoted by IDs not starting with “EFO”). Within the autoimmune disease sub-branch, 13 terms are taken from the Mondo disease ontology that is also curated by EBI [https://bioportal.bioontology.org/ontologies/MONDO; (13)]. Further, three terms are taken from Orphanet, an online database of rare diseases and orphan drugs (Copyright, INSERM 1997. available at http://www.orpha.net) and one term from the disease ontology [www.disease-ontology.org; (14)], cf. Suppl. Table S2.

Genomics data, genetic variation and notable reference resources

Differences between genomic sequences are called genetic variations. They are classified into single nucleotide variants (SNVs), for which the base at a single position differs, indels, which are insertions or deletions of size up to 50 bases, and structural variants (SVs), which are genomic alterations of a size larger than 50 bases. SNVs commonly found in a population are also called single nucleotide polymorphisms (SNPs). Genotyping arrays assess predefined, common variants, i.e., SNPs. Genetic variations are specified with respect to a specific reference genome. For humans, this is GRCh38, the genome of the Genome Reference Consortium, with its latest version 38.

Human SNVs have been well characterized. The first milestone was the whole genome sequencing performed as part of the 1000 Genomes Project, which resulted in “a global reference of human genetic variation” based on the genomic data of 2,504 individuals from 26 populations (15). The 1000G-based genetic variation with respect to the reference genome was overall 84.7 million SNVs, 3.6 million indels and ~60,000 structural variants; each individual carried 4.1 million to 5.0 million sites that differed from the reference genome. This first comprehensive catalog of genetic variation was later extended by whole genome sequencing of the Human Genome Diversity Project (n=929) (16). Of the many, often national, genome sequencing initiatives, gnomAD [n=71,702 whole genome sequenced; (17)], Topmed [n=53,831 whole genome sequenced; (18)] and the UK Biobank [UKB, n=150,119 whole genomes sequenced; (19)] stand out in terms of sample size. Within the UKB cohort, 585.0 million SNVs, representing 7.0% of all possible human SNVs, 58.7 million indels and ~900,000 SVs have been identified (19). Many of the SNVs, 46%, are carried by only one sequenced individual (called “singletons”) and only 3.4% (~20 million) have a frequency of more than 0.1% (19).

Autoimmune disease genome-wide association studies (GWAS)

The main goal of GWAS is to identify associations of genetic variants with a phenotype or trait without prior knowledge about their genomic location. Although GWAS could in principle use different kinds of genetic variants, to date almost always SNPs are utilized (11). GWAS then consist of testing for associations of SNPs with the phenotype or trait of interest. Since the first GWAS about twenty years ago (20), more than 5,700 analyses have been conducted, yielding more than 3,300 traits established to be statistically associated with genetic variants (10).

Testing for associations between a phenotype or a trait and genetic variants is based on a statistical model, and the type of the model used depends on whether the phenotype or trait is continuous (e.g. Body Mass Index (BMI)) or dichotomous (e.g. presence or absence of an autoimmune disease). In the case of a continuous phenotype or trait, a linear regression model is most commonly used, whereas logistic regression is mostly applied for dichotomous ones. Typically, the models are estimated for each single variant separately. The typical GWAS output comprises, for each variant, a report giving the ID of the variant according to dbSNP (21), the effect allele, the statistical effect and the corresponding p-value. Since GWAS test a large number of genetic variants at the same time, the statistical significance threshold has to be corrected to avoid false positive results. The widely used approach for this aim is the so-called Bonferroni correction (10, 22), consisting of dividing the overall statistical significance threshold by the total number of independent tests, in this case, the tested independent variants. As a consequence, a threshold of $5*10^{-8}$ is commonly used in practice, since the human genome contains approximately one million common, independent variants (10).

The GWAS catalog is a publicly available, manually curated resource, which contains published GWAS and association results and is developed by the NHGRI and EMBL-EBI (7). Catalog data is provided for the latest reference genome version (GRCh38.p13) and variant database version (dbSNPBuild 154). GWAS catalog source files of studies and associations have been used in our survey (files gwas-catalog-studies_ontology-annotated.tsv and gwas-catalog-associations_ontology-annotated.tsv from http://ftp.ebi.ac.uk/pub/databases/gwas/releases/2022/05/23/), and entries with “MAPPED_TRAIT_URI” an autoimmune EFO IDs (Suppl. Table S2) were extracted. Overall, the GWAS catalog studies contain 442 autoimmune disease GWAS (“STUDY ACCESSION”) published between 2006 and 2022 in 58 different journals with 221 unique PubMed IDs (Suppl. Table S3); these studies have been conducted on 377 different datasets (according to column “INITIAL SAMPLE SIZE” in Supplementary Table S3). A subset of studies (n=179 (47%)) reported no genome-wide significant variants.

The GWAS catalog contains 5,023 associations that cover 41 autoimmune diseases (according to EFO ID) based on 253 datasets (according to column “INITIAL SAMPLE SIZE”) relating to 200 unique PubMed IDs (Suppl. Table S4). These associations correspond to 3,212 unique SNPs (according to column “SNPS”) and 1,760 unique genes or gene combinations reported in the literature (column “REPORTED GENE(S)”; Supplementary Table S4).

Polygenic scores (PGS) developed for autoimmune diseases

GWAS are typically used for traits with an underlying polygenic architecture, that is, many genetic variants just show small effect sizes on the phenotype or trait of interest. As a result, prediction performances of single associated variants are generally poor. Therefore, polygenic scores, also termed “risk scores” if applied to a disease, are used to overcome these limitations. The main idea of predictive models based on polygenic scores is to combine effects of single genetic variants to expect a stronger association with the response phenotype or trait. The standard approach used for quantifying genetic liability in the prediction of disease risks are weighted polygenic scores (11). Based on this, PGSs are generally obtained as weighted sum scores of risk alleles using effect sizes from GWAS. More recently, new statistical machine learning approaches have emerged as a powerful approach for the computation of PGS (23).

The PGS catalog is a database and website established with the aim of making published PGS easily available and allowing their systematic evaluation (8). We obtained PGS catalog source files from https://www.pgscatalog.org/downloads and extracted from the respective files *via* EFO IDs all information related to autoimmune diseases: polygenic scores (Suppl. Table S5), score development samples (Suppl. Table S6), performance evaluation metrics (Suppl. Table S7) and evaluation samples (Suppl. Table S8). The database contains 18 autoimmune diseases for which 47 polygenic scores are published in 14 papers between 2018 and 2022 (cf. Suppl. Table S5). These have been developed with 15 different computational methods, mostly with the tools snpnet [n=18 scores; (24)], using genome-wide significant GWAS variants (n=7 scores), LDPred [n=6; version 1 and 2; (25, 26)] or by applying pruning and threshold (n=4). Corresponding to the method or tool applied for PGS constructions, the number of variants considered in the scores ranges from 3 to 6,907,112. Many cohorts of mainly European and/or East Asian (largely Chinese) ancestry have been used for score development, mainly as source GWAS underlying the respective score, but also for parameter training (Suppl. Table S6). Further, a large number of PGS have been developed using the UK Biobank. Autoimmune PGS have been evaluated in 124 data sets (Suppl. Table S7) yielding 225 performance assessments in 16 publications (Suppl. Table S8). The most common performance measure is the Area Under the Receiver-Operating Characteristic Curve (AUROC), which shows the fraction of individuals incorrectly classified as having the disease (false positive rate) versus the fraction of individuals correctly classified as having the disease (true positive rate) at different PGS score thresholds. An AUROC value of 0.5 corresponds to a random and a value of 1.0 to a perfect classification. Typically, AUROC classification performance of autoimmune PGS are in the range from 0.56 to 0.99 (provided for n=124; Suppl. Table S8). Overall, according to the PGS catalog, 16 different publications either constructed and/or evaluated an autoimmune disease PGS (columns "PGS Publication (PGP) ID" of Suppl. Table S5 and Suppl. Table S8).

Genetic data and autoimmune diseases covered by the UK Biobank

The UK Biobank (UKB) is the largest resource for human genomic and phenotypic data available for global academic health research, containing data from approximately 500,000 individuals from the United Kingdom (9). Its first release in 2018 contained UK Biobank Axiom Array-based genotypes (m=825,927) imputed to m~96 million variants of these individuals (9). In 2021, whole exome sequencing data of 454,787 of its individuals was released (with m~2 million exonic SNVs) (27). In 2022 whole genome sequencing-based variants for n=150,119 individuals resulted in overall m~585 million SNVs, ~59 million indels, 2.5 million microsatellites and 900k structural variants (19), representing a nearly complete variant profile for these individuals.

UK Biobank provides medical diagnosis according to the International Statistical Classification of Diseases and Related Health Problems (ICD) of the World Health Organization (WHO), whose current version is ICD-10. Besides ICD-10 codes gathered from medical records, UK Biobank provides diagnoses that are self-reported by participants, referred to by dedicated UKB-internal IDs (starting with "20002_"). To extract autoimmune disease information from UK Biobank, we used the mapping file internal to the ontology mapping tool Zooma of the EMBL-EBI ontology lookup service (28) which was generated as part of a large-scale, comprehensive mapping of UK Biobank ICD-10 codes and self-reported diseases to EFO terms (https://github.com/EBISPOT/EFO-UKB-mappings). We find that 20 of 120 autoimmune EFO IDs (cf. Suppl. Table S2) correspond to patients and patient genotypes within UK Biobank (Suppl. Table S9), of which 9 have both self-reported and ICD-10 diagnosis, 6 are only self-reported and 5 only have ICD-10-based information. The number of respective patients ranges from 13 (reactive arthritis with ICD-10 code M03) to 12,556 (rheumatoid arthritis with ICD-10 code M06) (Suppl. Table S9).

Overlap between autoimmune diseases assessed by GWAS, PGS and UKB

We investigated for which autoimmune diseases the GWAS catalog, the PGS catalog and UK Biobank contain information by comparing the respective autoimmune EFO IDs covered by each resource. Overall, there are 120 autoimmune disease EFO IDs (Suppl. Table S2) representing different levels of diagnosis (Suppl. Table S1). Of those, the GWAS catalog covers 41 EFO IDs (Suppl. Table S4) and the PGS catalog 18 EFO IDs (Suppl. Table S5). As the UK Biobank does not use EFO IDs, we used a published mapping of EFO IDs to UKB data fields instead, as described in the last section and provided in Suppl. Table S9. This assigned 20 EFO IDs to traits in UKB (Suppl. Table S9). The overlap of autoimmune diseases covered by the three resources is shown in Figure 1. Out of 120 EFO IDs, only 48 autoimmune diseases are present in any of the three databases, most in the GWAS catalog. The GWAS catalog is sharing 15 autoimmune disease EFO IDs with the PGS catalog and 15 can be mapped to UK Biobank. Further, 12 disease EFO IDs are shared between PGS catalog and UK Biobank. There are 11 autoimmune disease EFO IDs common to all three databases. They relate to: ankylosing spondylitis, appendicitis, Crohn's disease, inflammatory bowel disease, juvenile idiopathic arthritis, psoriatic arthritis, rheumatoid arthritis, Sjögren syndrome, systemic lupus erythematosus, ulcerative colitis and vitiligo. Several of the autoimmune diseases related to EFO IDs that are not shared by all three resources are cases that highlight limitations with respect to the definition of terms and relationships within the EFO and in the mapping of EFO terms to external codes and identifiers, which may not be one-on-one and needs disease-specific knowledge (for details see caption of Figure 1).

We have investigated more closely the 10 autoimmune diseases with most GWAS studies according to GWAS catalog (Table 1). They are systemic lupus erythematosus (29), rheumatoid arthritis (30), multiple sclerosis (31), inflammatory bowel disease (32) with its two subtypes Crohn's disease and ulcerative colitis, vitiligo (33), Sjögren syndrome (34), Grave's disease (35), and Behcet's syndrome

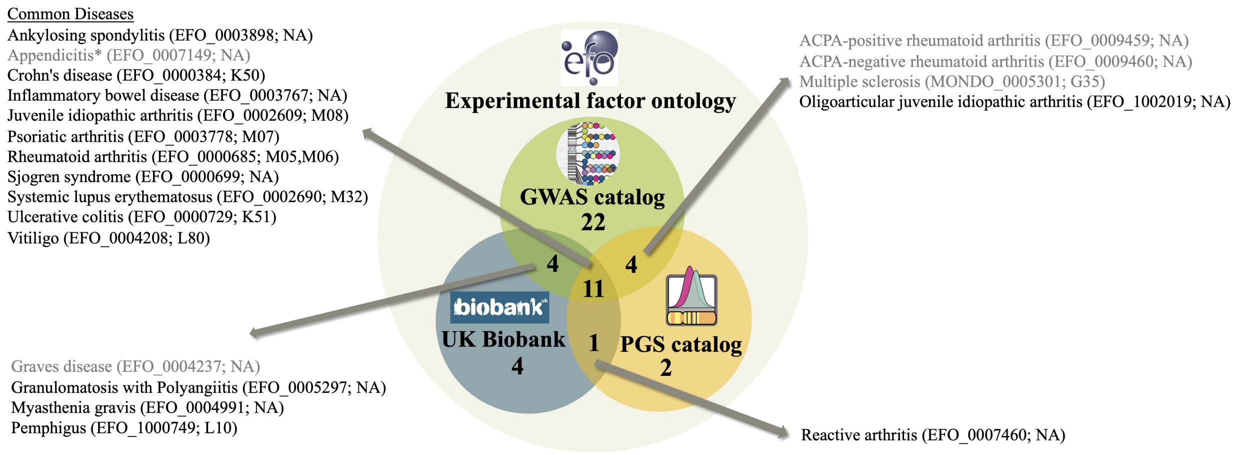
Common Diseases
Ankylosing spondylitis (EFO_0003898; NA)
Appendicitis* (EFO_0007149; NA)
Crohn's disease (EFO_0000384; K50)
Inflammatory bowel disease (EFO_0003767; NA)
Juvenile idiopathic arthritis (EFO_0002609; M08)
Psoriatic arthritis (EFO_0003778; M07)
Rheumatoid arthritis (EFO_0000685; M05,M06)
Sjogren syndrome (EFO_0000699; NA)
Systemic lupus erythematosus (EFO_0002690; M32)
Ulcerative colitis (EFO_0000729; K51)
Vitiligo (EFO_0004208; L80)
Experimental factor ontology
GWAS catalog
22
4
4
11
biobank
UK Biobank
4
1
PGS catalog
2
ACPA-positive rheumatoid arthritis (EFO_0009459; NA)
ACPA-negative rheumatoid arthritis (EFO_0009460; NA)
Multiple sclerosis (MONDO_0005301; G35)
Oligoarticular juvenile idiopathic arthritis (EFO_1002019; NA)
Graves disease (EFO_0004237; NA)
Granulomatosis with Polyangiitis (EFO_0005297; NA)
Myasthenia gravis (EFO_0004991; NA)
Pemphigus (EFO_1000749; L10)
Reactive arthritis (EFO_0007460; NA)

FIGURE 1
A Venn diagram representing the number of autoimmune diseases in the experimental factor ontology (EFO) overlapping with GWAS catalog and PGS catalog covered EFO IDs as well as UK Biobank data field matched according to Suppl. Table S9. EFO IDs of diseases in more than two resources are listed together with their name according to EFO and their UKB ICD-10 code according to Suppl. Table S9. Diseases shaded gray are affected by issues with disease definition and classification compromising the mapping. These issues are: (i) Appendicitis is classified as an autoimmune disease in EFO because it is a child term of inflammatory bowel disease, however, it is not considered an autoimmune disease. (ii) Grave's disease EFO child terms are in PGS, not the EFO ID of Grave's disease itself though. (iii) ACPA-positive and ACPA-negative rheumatoid arthritis is not mappable to UKB. UKB, however, contains seropositive and other rheumatoid arthritis, a distinction not covered by EFO. (iv) UKB has information on multiple sclerosis, yet since the recent EFO version was updated to using the MONDO ID for multiple sclerosis, the mapping to UKB data fields failed. "NA" denotes that mapping to UKB is not available for the respective EFO ID.

(36). The GWAS catalog association data of the top 10 autoimmune diseases (underlying Table 1) are provided in Supplementary Tables S10-S19. The most GWAS-studied autoimmune disease is systemic lupus erythematosus, for which 37 different GWAS have been performed, the largest one using 13,377 cases and 194,993 controls. These studies have reported 788 unique SNPs and 439 unique genes or gene combinations. In the PGS catalog, six studies are noted on systemic lupus erythematosus, which report six different risk scores. These PGS have been evaluated in 32 settings. The largest number of cases has been analyzed for inflammatory bowel disease (n=25,042), the lowest for Sjögren syndrome (n=1,599). Overall, the number of independent genomic loci associated with disease increases with the number of studies and cases (Table 1).

Discussion

We systematically reviewed the autoimmune disease-related content of the GWAS catalog of variant associations, the PGS catalog of polygenic scores and the UK Biobank of genomic and phenotypic data. These curated data sources and the ease of obtaining and querying them have already and will continue to unravel genetic and molecular underpinnings of autoimmune disease (37). An example are the currently ongoing 61 UKB-approved projects that are related to autoimmune disease (keyword search "autoimmune disease", June 13th, 2022). Our survey shows that the catalog of autoimmune GWAS studies and associations is already very comprehensive and generated in more than a decade. PGS for autoimmune diseases are rather few, very novel and largely developed within the last three years. Accordingly, in the polygenic disease genetics field, research efforts go into two interrelated directions: (i) unraveling specific functional effects of variants and (ii) combining effect estimates for a better personalized risk prediction. Computational approaches toward the first aim are the association of risk alleles with molecular traits (38) and the identification of functional variants *via* so-called fine-mapping (39). Although there is progress in the field, it is still a long way from variant associations to molecular disease mechanisms as well as treatments (37). Toward the second aim, polygenic scores are still being improved, for example by considering rare variants (40) or inclusion of functional information (41). Optimizing prediction performance is non-trivial, since machine learning models need to be calibrated to generalize to unseen data, i.e. overfitting of training data prevented (42). The AUROC of current autoimmune disease PGS varies widely (range 0.56-0.99; typically 0.6-0.8). Further

TABLE 1 The ten autoimmune diseases (defined by EFO term) which have the highest number of GWAS studies registered at the GWAS catalog. Displayed is the summary of information obtained from GWAS catalog, PGS catalog and UK Biobank. With respect to GWAS catalog, this is the number of unique studies (according to column "STUDY ACCESSION" of Suppl. Table S3), the highest number of cases with corresponding number of controls, the number of unique variants reported (according to column "SNP_ID_CURRENT" of Suppl. Table S4), the number of independent, associated genomic loci reported in the literature, the number of unique genes or gene combinations reported in the respective publications (according to column "REPORTED GENE(S)" of Suppl. Table S4). With respect to PGS catalog, reported are the number of unique studies (according to column "PGS Publication (PGP) ID" of Suppl. Table S5 and Suppl. Table S8), unique scores developed (according to column "Polygenic Score (PGS) ID" of Suppl. Table S5), the range of variants utilized in the scores for the respective disease and the number of performance evaluations in independent samples (according to column "PGS Performance Metric (PPM) ID" of Suppl. Table S8). Finally, for the UK Biobank, the UKB data field, ICD-10 code (if available in UKB) and patient number according to Suppl. Table S9 is provided.

Trait	EFO IDs	GWAS catalog						PGS catalog				UK Biobank		
		# Studies	# Cases	# Controls	# SNPs	# Loci	# Gene	# Studies	# PGS	# Variants	# PGS Eval.	Data Field	ICD10 Code	# Indiv.
Systemic lupus erythematosus	EFO_0002690	37	13,377	194,993	788	132[1]	439	6	6	41- 293,684	32	131894	M32	1,053
Rheumatoid arthritis	EFO_0000685	37	22,628	288,664	421	>150[2]	249	3	6	3- 95,083	33	131850	*M06	12,556
Multiple sclerosis	MONDO_0005301	27	14,802	26,703	603	233[3]	479	3	5	36- 129,077	25	131042	G35	2,518
Crohn's disease	EFO_0000384	27	12,924	21,442	411	>200[4]	265	1	2	220-257	9	131626	K50	3,355
Ulcerative colitis	EFO_0000729	25	12,366	33,609	295	>200[4]	184	2	4	179-566,637	26	131628	K51	6,451
Inflammatory bowel disease	EFO_0003767	12	25,042	34,915	387	>200[4]	238	3	2	195-690,7112	7	_**	_**	_**
Vitiligo	EFO_0004208	10	2,853	37,405	91	49[5]	80	3	3	42-77	10	131802	L80	1,201
Sjogren syndrome	EFO_0000699	10	1,599	658,316	48	25[6]	42	1	1	7	5	20002_1382	_	572***
Grave's disease	EFO_0004237	8	4,487	629,598	74	12[7]	27	_	_	_	_	20002_1522	_	183***
Behcet's syndrome	EFO_0003780	8	3,197	5,785	40	21[8]	35	_	_	_	_	41202	_	18****

[1] (29); [2] (30); [3] (31); [4] (32); [5] (33); [6] (34); [7] (35); [8] (36); *Excludes seropositive rheumatoid arthritis (M05) with 1,401 patients ** K50+K51 *** Self-reported **** Based on medical history of hospital patients.

evaluation of polygenic scores in more cohorts and systematic comparisons, facilitated by the PGS catalog, will help gaining further insights into PGS predictive performance for individual autoimmune diseases. Perspectively, PGS can be amended with other biomedical, clinical and behavioral data. Such rich, combined data sources together with recent developments in artificial intelligence promise to improve prediction of disease and personalized treatment options (43). Finally, polygenic scores can be used to investigate interactions of genetic and environmental factors, which is particularly relevant for autoimmune diseases, in which environmental factors play a key role (44).

Author contributions

RS and IW summarized data and generated tables and figures. RS, CF, and IW wrote the first draft of the manuscript. All authors contributed to writing and editing the manuscript. All authors approved the submitted version.

Funding

HB and IW acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 22167-390884018. This work was funded by the Research Training Group 2633 Autoimmune Pre-Disease project A9.

Acknowledgments

The authors acknowledge support through the high-performance computer cluster (OMICS-Cluster) of the University of Lübeck.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.972107/full#supplementary-material

References

1. Li YR, Zhao SD, Li J, Bradfield JP, Mohebnasab M, Steel L, et al. Genetic sharing and heritability of paediatric age of onset autoimmune diseases. *Nat Commun* (2015) 6(1):8442. doi: 10.1038/ncomms9442

2. Seldin MF. The genetics of human autoimmune disease: A perspective on progress in the field and future directions. *J Autoimmun* (2015) 64:1–12. doi: 10.1016/j.jaut.2015.08.015

3. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. Pervasive sharing of genetic effects in autoimmune disease. *PloS Genet* (2011) 7(8): e1002254. doi: 10.1371/journal.pgen.1002254

4. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, Han B, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet* (2016) 48(5):510–8. doi: 10.1038/ng.3528

5. Gokuladhas S, Schierding W, Golovina E, Fadason T, O'Sullivan J. Unravelling the shared genetic mechanisms underlying 18 autoimmune diseases using a systems approach. *Front Immunol* (2021) 12:693142. doi: 10.3389/fimmu.2021.693142

6. Lincoln MR, Connally N, Axisa PP, Gasperi C, Mitrovic M, van Heel D, et al. Joint analysis reveals shared autoimmune disease associations and identifies common mechanisms. *medRxiv* (2021). doi: 10.1101/2021.05.13.21257044

7. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* (2019) 47 (D1):D1005–12. doi: 10.1093/nar/gky1120

8. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* (2021) 53(4):420–5. doi: 10.1038/s41588-021-00783-5

9. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature* (2018) 562 (7726):203–9. doi: 10.1038/s41586-018-0579-z

10. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. *Nat Rev Methods Primers* (2021) 1(1):1–21. doi: 10.1038/s43586-021-00056-9

11. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* (2020) 15(9):2759–72. doi: 10.1038/s41596-020-0353-1

12. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample variables with an experimental factor ontology. *Bioinformatics* (2010) 26(8):1112–8. doi: 10.1093/bioinformatics/btq099

13. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res* (2017) 45(D1):D712–22. doi: 10.1093/nar/gkw1128

14. Schriml LM, Munro JB, Schor M, Olley D, McCracken C, Felix V, et al. The human disease ontology 2022 update. *Nucleic Acids Res* (2022) 50(D1):D1255–61. doi: 10.1093/nar/gkab1063

15. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature* (2015) 526(7571):68–74. doi: 10.1038/nature15393

16. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* (2020) 367(6484):1–10. doi: 10.1126/science.aay5012

17. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* (2020) 581(7809):434–43. doi: 10.1038/s41586-020-2308-7

18. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* (2021) 590(7845):290–9. doi: 10.1038/s41586-021-03205-y

19. Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, et al. The sequences of 150,119 genomes in the UK biobank. *bioRxiv* (2022). doi: 10.1101/2021.11.16.468246

20. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet* (2002) 32(4):650–4. doi: 10.1038/ng1047

21. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* (2001) 29 (1):308–11. doi: 10.1093/nar/29.1.308

22. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilit `a. pubblicazioni del r istituto superiore di scienze economiche e commerciali di firenze*, Vol. 8. (1936). Florence, Italy: Libreria Internazionale Seeber. pp. 3–62.

23. Gola D, Erdmann J, Müller-Myhsok B, Schunkert H, König IR. Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genet Epidemiol* (2020) 44(2):125–38. doi: 10.1002/gepi.22279

24. Qian J, Tanigawa Y, Du W, Aguirre M, Chang C, Tibshirani R, et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK biobank. *PloS Genet* (2020) 16(10):e1009141. doi: 10.1371/journal.pgen.1009141

25. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet* (2015) 97(4):576–92. doi: 10.1016/j.ajhg.2015.09.001

26. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics* (2020)16;36(22–23):5424–31. doi: 10.1093/bioinformatics/btaa1029

27. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome sequencing and analysis of 454,787 UK biobank participants. *Nature* (2021) 599(7886):628–34. doi: 10.1038/s41586-021-04103-z

28. Jupp S, Burdett T, Leroy C, Parkinson HE. A new ontology lookup service at EMBL-EBI. In J Malone, R Stevens, K Forsberg, A Splendiani, editors. *Proceedings of the 8th International Conference on Semantic Web Applications and Tools for Life Sciences (SWAT4LS)*. 1546 Cambridge EUR: Workshop Proceedings (2015). p. 118–119.

29. Wang YF, Zhang Y, Lin Z, Zhang H, Wang TY, Cao Y, et al. Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups. *Nat Commun* (2021) 12(1):772. doi: 10.1038/s41467-021-21049-y

30. Padyukov L. Genetics of rheumatoid arthritis. *Semin Immunopathol* (2022) 44(1):47–62. doi: 10.1007/s00281-022-00912-0

31. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* (2019) 365(6460):eaav7188. doi: 10.1126/science.aav7188

32. Garcia-Etxebarria K, Merino O, Gaite-Reguero A, Rodrigues PM, Herrarte A, Etxart A, et al. Local genetic variation of inflammatory bowel disease in Basque population and its effect in risk prediction. *Sci Rep* (2022) 12(1):3386. doi: 10.1038/s41598-022-07401-2

33. Jin Y, Roberts GHL, Ferrara TM, Ben S, van Geel N, Wolkerstorfer A, et al. Early-onset autoimmune vitiligo associated with an enhancer variant haplotype that upregulates class II HLA expression. *Nat Commun* (2019) 10(1):391. doi: 10.1038/s41467-019-08337-4

34. Imgenberg-Kreuz J, Rasmussen A, Sivils K, Nordmark G. Genetics and epigenetics in primary sjögren's syndrome. *Rheumatology (Oxford)* (2021) 60 (5):2085–98. doi: 10.1093/rheumatology/key330

35. Zhao SX, Xue LQ, Liu W, Gu ZH, Pan CM, Yang SY, et al. Robust evidence for five new graves' disease risk loci from a staged genome-wide association analysis. *Hum Mol Genet* (2013) 22(16):3347–62. doi: 10.1093/hmg/ddt183

36. Ortiz-Fernández L, Sawalha AH. Genetics of behçet's disease: Functional genetic analysis and estimating disease heritability. *Front Med (Lausanne)* (2021) 8:625710. doi: 10.3389/fmed.2021.625710

37. Makin S. Cracking the genetic code of autoimmune disease. *Nature* (2021) 595(7867):S57–9. doi: 10.1038/d41586-021-01839-6

38. Gutierrez-Arcelus M, Rich SS, Raychaudhuri S. Autoimmune diseases — connecting risk alleles with molecular traits of the immune system. *Nat Rev Genet* (2016) 17(3):160–74. doi: 10.1038/nrg.2015.33

39. Caliskan M, Brown CD, Maranville JC. A catalog of GWAS fine-mapping efforts in autoimmune disease. *Am J Hum Genet* (2021) 108(4):549–63. doi: 10.1016/j.ajhg.2021.03.009

40. Lali R, Chong M, Omidi A, Mohammadi-Shemirani P, Le A, Cui E, et al. Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories. *Nat Commun* (2021) 12(1):5852. doi: 10.1038/s41467-021-26114-0

41. Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot WJ, Khera AV, et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat Genet* (2022) 54(4):450–8. doi: 10.1038/s41588-022-01036-9

42. Gonzalez G. *An intro to machine learning for biomedical scientists* (2021). Available at: https://towardsdatascience.com/storytelling-machine-learning-intro-d46e339cb6de.

43. Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit Med* (2020) 3(1):1–11. doi: 10.1038/s41746-020-0229-3

44. Ellis JA, Kemp AS, Ponsonby AL. Gene-environment interaction in autoimmune disease. *Expert Rev Mol Med* (2014) 6:e4. doi: 10.1017/erm.2014.5