Genetic architecture and shared mechanisms of common 'neglected' diseases

Maik Pietzner[1,2,3], Spiros Denaxas[4,5,6,7], Summaira Yasmeen[1], Maria Wörheide[8], Hannah E. Schmidt[1], Burulça Uluvar[1], Aakash Nepal[1], Mine Koprulu[3], Julia Carrasco-Zanini[2,3], Matthias Arnold[8,9], Annika Vogt[10], Gabi Kastenmüller[8], Harry Hemingway[4,5,7,*], Claudia Langenberg[1,2,3,*]

[1]Computational Medicine, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany
[2]Precision Healthcare Institute, Queen Mary University of London, London, UK
[3]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
[4]Institute of Health Informatics, University College London, London, UK
[5]Health Data Research UK, London, UK
[6]British Heart Foundation Data Science Centre, London, UK
[7]National Institute of Health Research University College London Hospitals Biomedical Research Centre
[8]Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
[9]Department of Psychiatry and Behavioral Sciences, Duke University, Durham, NC, USA
[10]Department of Dermatology and Allergy, Charité–Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany
*these authors jointly supervised the work

Corresponding authors
Dr Maik Pietzner (maik.pietzner@bih-charite.de)
Prof Harry Hemingway (h.hemingway@ucl.ac.uk)
Prof Claudia Langenberg (claudia.langenberg@qmul.ac.uk)

ABSTRACT

Many diseases with significant health burden and healthcare costs have been neglected in biomedical research, partly due to a lack of data. Here, we systematically harmonized >12 million primary care and hospitalisation health records from ~440,000 UK Biobank participants into 1445 collated disease terms for genetic analyses. This included ~200 diseases with >10,000 cases that are predominantly managed in primary care, like skin and non-serious infectious diseases. We quantify the heritability for these common diseases, identify novel loci with extreme effect sizes, and highlight novel roles of poorly characterised genes, e.g. *PNLIPRP3* as a sebaceous gland cell-specific gene for rosacea. We characterise the substantial regional pleiotropy with more than 300 independent genomic regions associated with multiple, often seemingly unrelated diseases and use these insights to generate a pan-genome disease network of shared disease loci to prioritise potential pathways contributing to multiple diseases and onset of multimorbidity. We demonstrate the value of primary care data to improve and guide genetic approaches for drug selection, repurposing, and adverse event prediction, e.g., *IGFR1* as a putative multi-disease target for, among others, gout and atrial fibrillation, through network augmentation and integration of molecular quantitative trait loci. We make all results publicly available via an interactive webserver (https://www.omicscience.org/apps/phecodes). Our results provide new insights for diseases across diverse clinical specialties and provide a resource for drug discovery and mechanistic understanding.

54 **INTRODUCTION**

55 An inordinate amount of biomedical research and funding has been dedicated to conditions
56 with high mortality and need for hospitalisations, largely neglecting hundreds of diseases that
57 are less serious and predominantly managed in primary care, but that still pose a major
58 burden for patients and the healthcare system[1,2]. Such common, but often understudied
59 diseases further substantially contribute to one of the greatest health challenges in high- but
60 also low- and middle-income countries, the increasing prevalence of multimorbidity[3].

61 Genome-wide association studies (GWAS) that examine associations between millions of
62 human DNA sequence variants and (serious) diseases have transformed biomedical research
63 over the past 15 years[4,5] and characterised the polygenic architecture of many complex
64 diseases. They revealed novel disease pathways and drug candidates with support from
65 human genetic studies are twice as likely to achieve final approval[6,7]. However, the common
66 strategy of ever-increasing sample sizes for common traits and diseases now starts reaching
67 saturation[8]. A greater translational potential will therefore lie in efforts that target patient
68 populations that have been so far neglected, addressing common conditions mostly treated
69 in primary care, as we recently demonstrated[9]. Such single disease efforts, however, need to
70 be set into the context of an ever aging population that shifts demographics globally with
71 many patients now presenting with not one but many diseases[3], that requires the
72 development of novel treatment strategies. Biobank-scale studies that link deep
73 characterization of participants with virtually complete health care trajectories through
74 electronic health records (EHRs) only now provide the ability to systematically address this
75 unmet medical need by generating a standardized set of disease definitions that captures the
76 breadth of medical diagnosis and patient burden in a single cohort.

77 Here, we pooled EHRs for >440,000 participants of the UK Biobank (UKB) study from multiple
78 data resources into 1445 medical disorder concepts ('phecodes'[10,11]) or 'diseases', with
79 records from primary care more than doubling sample sizes for most conditions enabling first-
80 ever GWAS for a number of diseases. We provide unique genetic insights that revealed
81 unreported high-impact loci, novel risk genes for diseases predominantly managed in primary
82 care, and extensive regional pleiotropy that links multiple, even seemingly unrelated diseases,
83 providing potential pathways to multimorbidity. We finally demonstrate how our results can
84 identify opportunities for drug repurposing and potential adverse effects of drugs on diseases
85 currently missed.

86 **RESULTS**

87 Here, we describe the genetic architecture of 1445 diseases, based on medical disorder
88 concepts[10,11] defined and collated from >12 million medical records from primary (general
89 practice) and secondary care (hospital admissions and cancer registry), death certificates, and
90 self-report among 441,671 white European UKB participants. For half of these diseases
91 (50.1%, n=724), primary care was the only source of diagnosis data in more than 50% of
92 participants (**Fig. 1**). Dermatological and (mild) infectious diseases showed the largest
93 increase in sample size, with common conditions such as rosacea (n=25,307; 93.3% primary
94 care) or acute sinusitis (n=58,140; 99.4% primary care) almost exclusively being identified
95 through and managed in primary care (**Fig. 1 and Supplementary Tab. 1**).

96 For 363 diseases, common single nucleotide polymorphisms (SNPs) explained a significant
97 ($p<3.5x10^{-5}$) amount (median: 7.2%; IQR:4.1-9.0%) of the heritability, reaching values as high
98 as 23.0% for retinoschisis or 9.0% for rosacea, illustrating a substantial heritable component
99 of common, but not fatal conditions managed in primary care (**Supplementary Tab. 1**). We
100 further observed little evidence of genomic inflation (median LDSC-intercept 1.01; IQR: 1.00-
101 1.02) across 695 diseases with at least one genome-wide significant locus ($p<5x10^{-8}$).

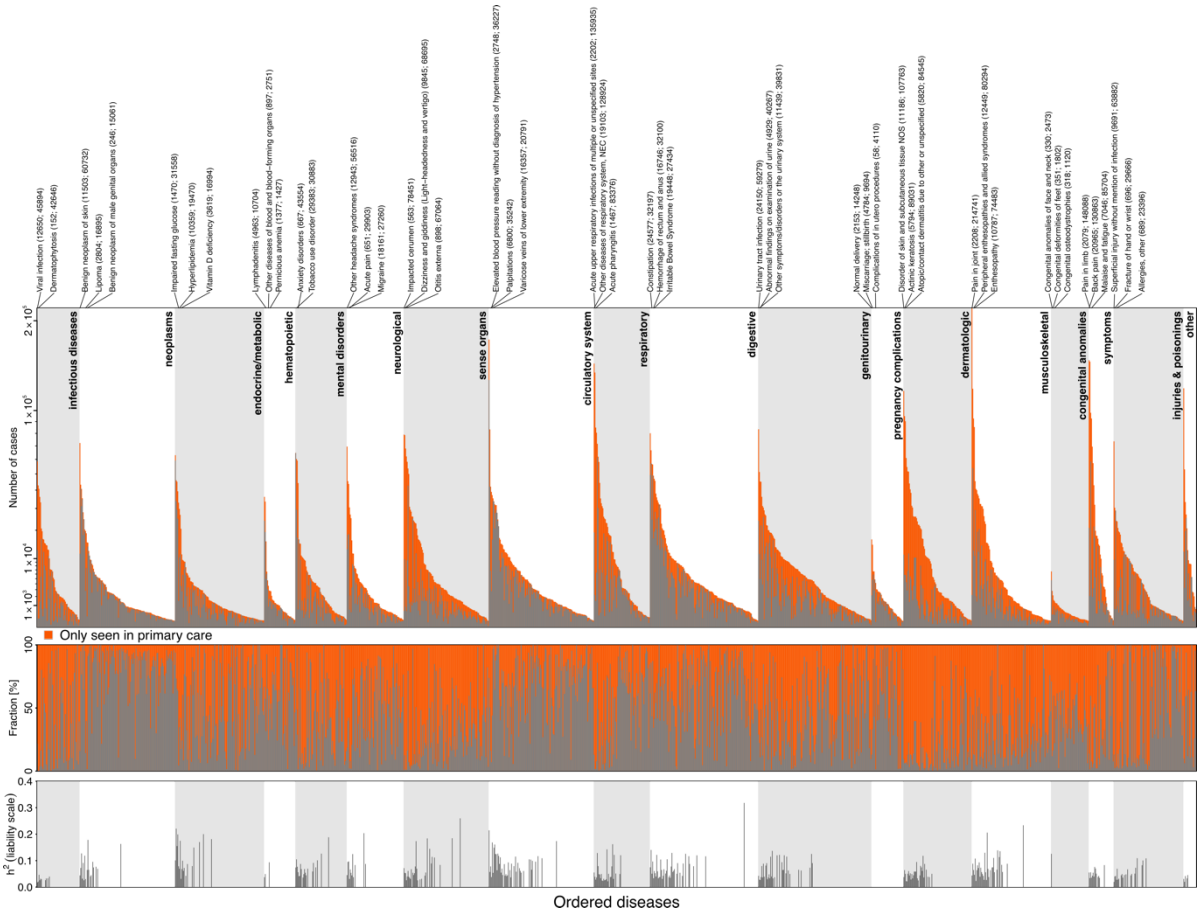


Figure 1 Distribution of case numbers. The *upper panel* displays the number of participants with one of 1445 diseases considered. Diseases were grouped according to the chapters of the International Classification of Diseases 10 (ICD-10) system. Within each chapter, diseases were ordered by case numbers. The contribution of records only seen in primary care is indicated in orange. The y-axis has been square root transformed to increase visibility. Selected diseases mostly (>50%) seen and managed in primary care have been annotated with numbers in brackets indicating non-primary care and primary care cases. The *middle panel* displays the proportion of cases for which primary care was the only data resource. The *lower panel* displays the estimated single nucleotide polymorphism-based heritability on the liability scale for diseases. Estimates are only shown if they passed multiple testing correction ($p<3.6x10^{-5}$). The underlying data can be found in Supplementary Tab. 1.

Novel loci for common conditions

We identified 6188 genome-wide (GW) significant ($p<5x10^{-8}$) genetic variant – disease associations, including 695 unique diseases (351 mostly recorded in primary care) and 4422 unique loci, through a combination of GWAS and Bayesian fine-mapping (**Fig. 2 and Supplementary Tab. 2**, see **Methods**). About 16% (n=1028; including 251 at corrected GW-significance) of all association pairs have not been reported so far ($r^2<0.1$) in the GWAS catalog, of which two-thirds (n=673) are currently also missed by large biobank engines as summarized in the Open Targets platform[12,13] (download 27/04/2022). For more than half of the reported loci (n=2326; n=3542 genetic variant – disease pairs), we identified unreported

trait associations, whereas the remainder (n=1199; n=1555 genetic variant – disease pairs) mapped to previously reported phenotype associations in the GWAS catalog[14] (download 27/04/2022). We replicated 80.1% (2053 out of 2560; directionally concordant and $p<0.05$) of all identified variant – disease pairs with matching entries in the FinnGen release 8[15], including 89 novel loci (**Supplementary Tab. 3**).

A very small fraction (0.5%, n=31 out of 6188) of all variant – disease associations showed evidence of significant ($p<8.7\times10^{-6}$) sex differences, with only three showing evidence for sexual dimorphic effects (**Supplementary Fig. 3 and Tab. 4**). This included a stronger effect of the low-frequent missense variant rs75495843 (MAF=3.0%; p.Ser460Leu within *PLCD1*) on the risk for sebaceous cysts in women (OR: 3.17; 95%-CI: 3.03-3.30; $p<1\times10^{-300}$) compared to men (OR: 2.22; 95%-CI: 2.12-2.34; $p<9.6\times10^{-226}$). The same missense variant has already been demonstrated to cause hereditary trichilemmal cyst formation in a two-hit model, in which the germline missense variant predisposes to or exaggerates the effect of somatic mutations in *PLCD1*[16].

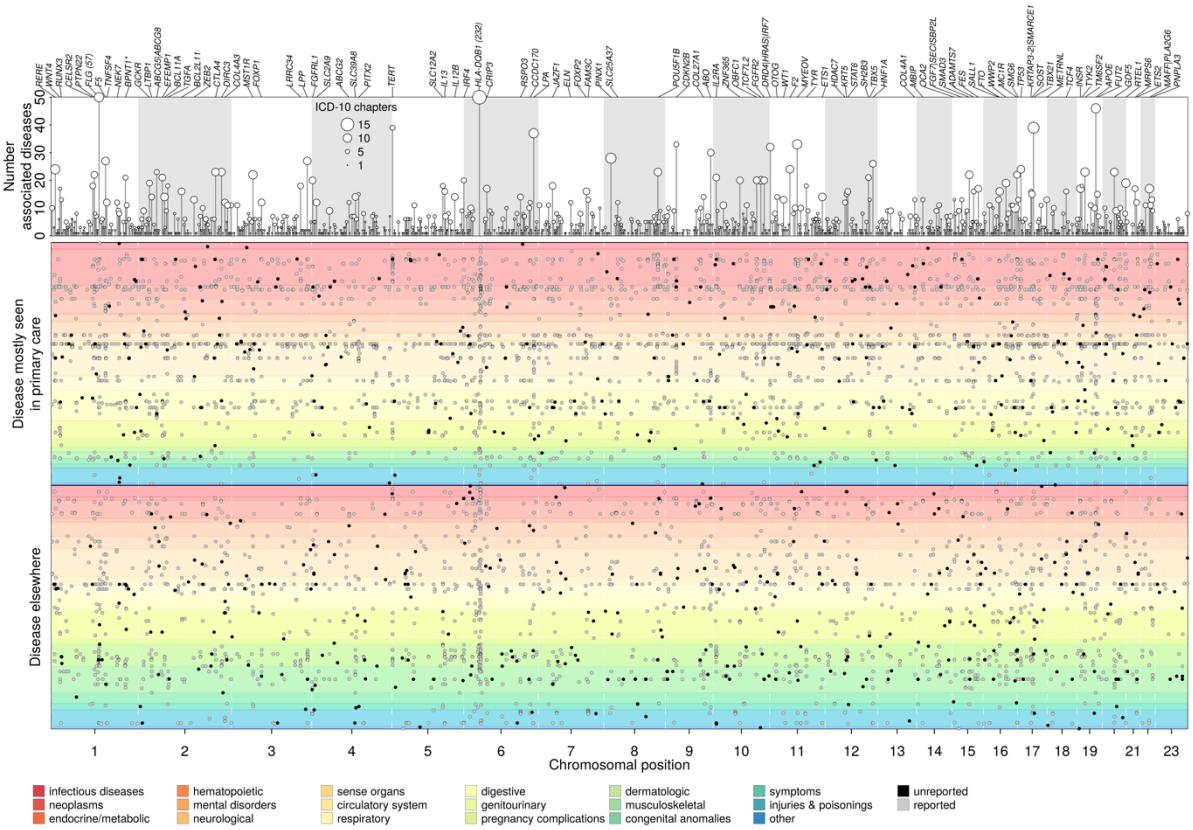


Figure 2 Summary of genome-wide association results across 1445 diseases. The *upper panel* displays the number of associated diseases for each of 1104 independent genomic regions, with dot sizes

140 indicating the number of distinct ICD-10 chapters associated diseases belonged to. The most
141 frequently prioritised effector genes are annotated for regions with ≥12 disease associations. The
142 *middle panel* shows the genomic location of significant ($p<5x10^{-8}$) genetic variants associated with at
143 least one out of 724 diseases (y-axis) that are mostly recorded in primary care. The *lower panel* shows
144 the genomic location of genetic variants associated with at least one disease that was recorded in
145 multiple resources or resources other than primary care. Disease chapters are highlighted by
146 background colours, whereas dot colours indicate whether the genetic variant has not been reported
147 yet ($r^2>0.1$; black) or previously in the GWAS catalog or large-scale biobank engines (grey).
148
149 We fine-mapped more than half (n=377) of the unreported loci to ≤5 genetic variants in 95%-
150 credible sets, including 186 with a single causal variant (**Supplementary Tab. 2**). This included
151 low frequency variants (1%≤ minor allele frequency (MAF) <5%) with strong effects (|odds
152 ratio (OR)| > 2.0) on 19 diseases including eight mainly seen in primary care. For example,
153 rs145341070 (MAF=1.0%) increasing the risk for the inflammatory skin disease panniculitis by
154 almost 15-fold (OR: 14.96; 95%-CI: 5.94-37.66; $p< 9.4x10^{-09}$, candidate gene: *MYEOV*) or
155 rs138740163 (MAF=1.0%) increasing the risk for 'Mittelschmerz' (ovulation pain)[17] by more
156 than 5-fold among women (OR: 6.00; 95%-CI: 3.18-11.30, $p<3.02x10^{-8}$, *RBM19*) with *RBM19*
157 as a potential candidate gene already been linked to preimplantation development in animal
158 models[18,19]. We further identified 96 rare (MAF<1%) variant – disease associations, of which
159 13 out of 15 imputed exonic variants were concordant with exome sequencing data available
160 in UKB. This included a more than 50% decreased risk for polymyalgia rheumatica associated
161 with the rare variant rs184097503 (MAF=0.4%; OR: 0.43, 95%-CI: 0.32-0.57, $p<4.1x10^{-9}$).
162 rs184097503 is in high LD ($r^2=0.92$) with the transthyretin stabilising missense variant
163 rs28933981 (p.T139M) the effect of which is mimicked with Tafamidis in patients suffering
164 from transthyretin amyloid cardiomyopathy[20]. These results illustrate strong locus-specific
165 contributions to diseases that predominantly present in primary care.

166 ***Assigning novel effector genes for common conditions***

167 To derive biologically, and possibly clinically, tangible insights, we systematically integrated
168 functional annotations, gene[21] and protein[22] expression quantitative trait loci (QTL), an open-
169 source machine-learning pipeline[13] as well as a biomedical knowledge graph[23] to map effector
170 genes with high confidence (see **Methods**). We successfully assigned candidate causal genes
171 for more than half of all genetic variant – disease associations (n=3271) with at least moderate
172 evidence (candidate gene score ≥ 2), including 531 high-confidence assignments (candidate
173 gene score ≥ 4; **Supplementary Tab. 2**). These gene sets provide a rich resource to improve

our understanding of common diseases typically seen and treated in primary care. For example, genes associated with the inflammatory ear disease otitis externa were enriched (FC: 95.5; $p<7.2x10^{-5}$) for members of the local control region (LCR)-associated chromatin remodelling complex, responsible for the transcriptional regulation of the β-globin locus[24] (**Supplementary Tab. 5**).

Among the loci with high-confidence gene assignments, ten were so far not reported. For example, the common frameshift variant rs145843135 (MAF=5.4%, *PNLIPRP3*) was associated with a reduced risk for rosacea (OR: 0.88; 95%-CI 0.85 - 0.92, $p<4.9x10^{-10}$) and further reduced expression of *PNLIPRP3* in skin tissue from the GTEx project[21], with statistical colocalization confirming a shared genetic signal with high confidence (posterior probability(PP) = 99.3%; **Fig. 3**). *PNLIPRP3* encodes for pancreatic lipase-related protein 3 (PNLIPRP3) and is, despite its name, highly expressed in skin[21,25]. More specifically, we observed that *PNLIPRP3* was almost exclusively expressed in sebaceous cells (**Fig. 3**) that produce sebum, rich in triglycerides, contributing to the integrity of the skin barrier. The exact function of PNLIPRP3 has not been shown so far, but it is predicted to have lipase activity, contributing to the breakdown of triglycerides[26]. Reduced expression of *PNLIPRP3* in carriers of the risk allele, and hence less activity of PNLIPRP3, may therefore alter sebum composition, but not necessarily production, in line with a previous study in patients with papulopustular rosacea[27].

Another line of evidence for a role of lipid and intermediate metabolism in rosacea originated from a common (MAF=45%) genetic signal at rs10882555 (OR: 0.90; 95%-CI: 0.9 - 0.92; $p<8.8x10^{-27}$) likely tagging a missense variant (rs591157; r^2=0.94) in *C10orf129* encoding for Acyl-CoA Synthetase Medium Chain Family Member 6 that has previously been shown to alter plasma levels of 3-hydroxy-2-ethylpropionate, a break down product of leucine[28]. It remains to be determined how altered sebum composition contributes to the chronic facial inflammatory dermatosis characteristic for rosacea, but the protective, cell-type specific effect of *PNLIPRP3* downregulation warrants further investigation as a putative non-steroidal treatment option with putatively less adverse systemic effects.

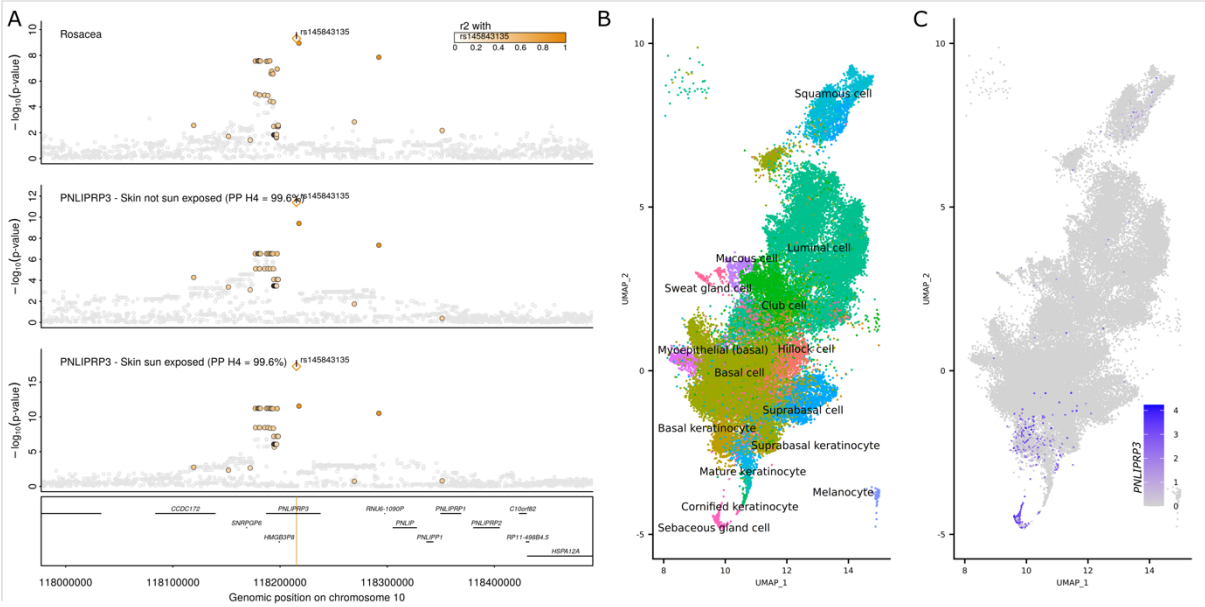


Figure 3 *PNPLIPR3* is a sebaceous gland cell-specific risk gene for rosacea. A Regional association plot for the risk of rosacea (top) and expression of *PNLIPRP3* in skin tissue (not sun exposed – middle; sun exposed – lower) from the GTEx v8 project[21]. Numbers in brackets indicate posterior probabilities (PP) of a shared genetic signal between gene expression and rosacea risk. Each genetic variant is coloured according to its linkage disequilibrium with the lead risk variant rs145843135. Association statistics for rosacea are based on logistic regression models from the present study, whereas association statistics for *PNLIPRP3* were obtained from linear regression models published by the GTEx consortium. **B** UMAP plot of epithelial cells based on single-nuclei RNA sequencing data from the GTEx project[29]. Different cell-types are coloured and annotated accordingly. **C** Gene expression of *PNLIPRP3* in the same cell-types. A full version of PNLIPRP3 expression across various cell types can be found in Supplementary Fig. 4.

Shared loci reveal molecular pathways of multimorbidity

Genomic regions that harbour multiple variants associated with multiple diseases can highlight pathways underlying the co-occurrence of even seemingly unrelated diseases[3]. About a third (339 out of 1103) of all non-overlapping genomic regions outside the MHC region included one or more clusters of genetic variants (r^2>0.8) associated with five or more diseases (**Fig. 2**). This included extremely complex regions like chr1:150Mb, covering 58 genetic variants associated with 57 diseases across 11 chapters of the ICD-10 mapping to highly pleiotropic genes like *FLG* or *IL6R* (**Fig. 4A**). Some regional association patterns were only poorly explained by the size or complexity of the region, that is, number of contained genes and independent variants (**Fig. 4A**). For example, a region on chromosome 9 (chr9:22Mb) centred around the cellular senescence genes *CDKN2A/CDKN2B* included 19 genetic variants associated with 33 diseases across six ICD-10 chapters, covering cardiovascular, eye, skin, and metabolic diseases as well as cancers of different organs (**Fig. 4B**). While the locus has been described for the severe end of cardiovascular and metabolic

229 diseases, we observed in part opposing effect direction for primary care conditions like rectal
230 polyps but also a protective effect for cancers of the brain and skin (**Fig. 4B**).
231 To go beyond such regional pleiotropy, we partitioned the entire set of 756 loci associated
232 with at least two out of 371 diseases into 77 locus – disease clusters[30] (see **Methods** and
233 **Supplementary Tab. 6**), that is, diseases tightly linked by a shared set of genetic loci across
234 the genome. The network can also be explored interactively
235 (https://www.omicscience.org/apps/phecodes). Most clusters (n=57) contained diseases
236 belonging to at most two ICD-10 chapters with genes enriched for highly plausible pathways,
237 such as 'cardiac muscle contraction' for atrial fibrillation and arrhythmias, or 'bile acid
238 signalling' for non-malignant diseases of the exocrine pancreas and gall bladder
239 (**Supplementary Tab. 6**).
240 In contrast, clusters covering multiple organ systems can identify the interplay of pathways
241 determining multimorbidity. For example, patients that develop atopic dermatitis early in life
242 are known to be at high risk for other allergic diseases, most commonly asthma[31], but the role
243 of other inflammatory skin diseases is less clear. We identified a cluster including various
244 inflammatory skin diseases and asthma (**Fig. 4C**) that was enriched for genes involved in
245 leucocyte and T-cell activation, as well as other immune processes involving cytokine
246 signalling and T-cell differentiation (**Supplementary Tab. 6**). While loci shared between atopic
247 dermatitis and asthma were directionally concordant, only one out of four loci shared
248 between asthma and the inflammatory skin disease rosacea did so (**Fig. 4C**). Specifically, effect
249 directions for rs6498021 (OR_{asthma}: 0.95; 95%-CI: 0.93 - 0.96; $OR_{rosacea}$= 1.08; 95%-CI: 1.05 -
250 1.11; *IL21R*), rs115008099 (OR_{asthma}: 1.10; 95%-CI: 1.08 – 1.12; $OR_{rosacea}$: 0.89; 95%-CI: 0.87 –
251 0.91; *IL13;* **Fig. 3D**), and rs61816761 (OR_{asthma}:1.23; 95%-CI: 1.18 – 1.28; $OR_{rosacea}$: 0.82; 95%-
252 CI: 0.77 – 0.87; *FLG*) were in opposing directions. This suggests that unspecific inflammatory
253 stimuli (via *TLR1*) may increase the risk for both, asthma and rosacea, while genes modulating
254 T cell activation and differentiation via interleukins as well as maintenance of the skin barrier
255 (via *FLG*) distinguishes the aetiology of atopic dermatitis and rosacea, and eventually asthma.
256 In general, our results help to dissect the role of different, e.g., inflammatory, pathways to
257 inform the development of tailored treatment strategies for commonly co-occurring
258 conditions. This is of relevance when targeting shared disease mechanisms that likely operate
259 in opposing directions. For example, nine interleukin-13 antibodies are currently in clinical

trials, including for asthma, with atopic dermatitis, pruritis, dry skin, and erythema being common side effects of drugs like Tralokinumab[32].

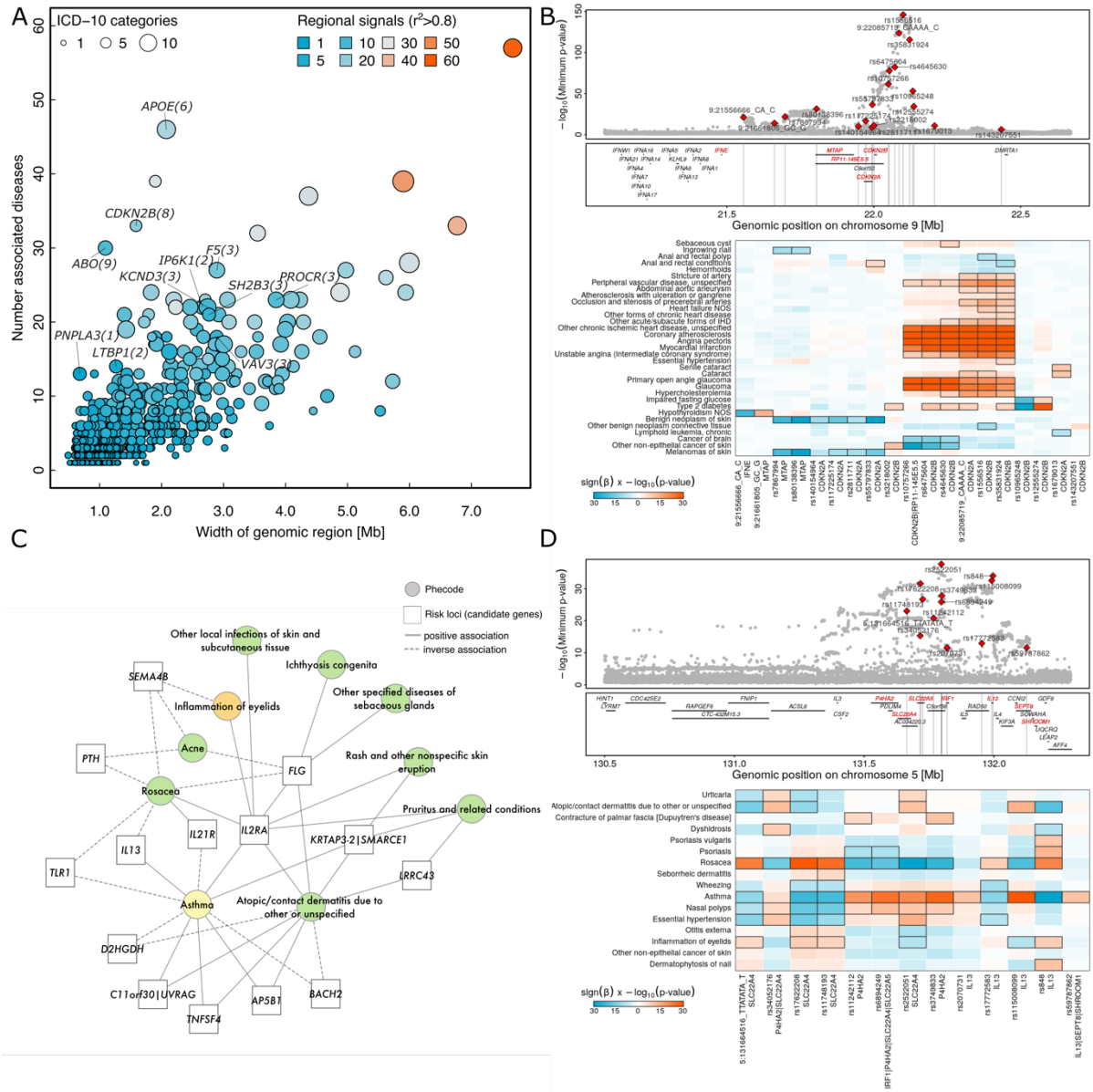


Figure 4 Regional and global pleiotropy of genetic loci. A For each of 1103 identified genomic regions, excluding the MHC region, the width is shown against the number of associated diseases with at least one variant in the respective region. Colouring was done based on the number of variant clusters (r^2>0.8) and the size of the dots reflects the number of ICD-10 chapters the diseases belonged to. Regions for which the number of associated diseases was poorly explained by region width, gene density, or number of weakly independent signals are annotated with the gene most frequently prioritized. **B** Regional association plot centred around *CDKN2A/CDKN2B* showing the minimum p-value across all diseases named in the heatmap. P-values were derived from logistic regression models. All lead, weakly independent variants in the region (r^2<0.8) are annotated and the heatmap underneath displays transformed p-values and effect directions. Candidate causal genes are highlighted in red and given next to lead variants. Variant – disease associations passing genome-wide significance are indicated by black frames. **C** Subnetwork from a larger locus – disease network based

276 on global, e.g., across the genome, pleiotropy centred around asthma and skin diseases. This network
277 has been selected as one of 77 clusters in a bi-clustering approach. The entire network can be explored
278 interactively (https://omicscience.org/apps/phecodes/). **D** Same as B but centred around *IL13*.
279
280 ***COVID-19 and diseases mainly recorded in primary care***
281 We next asked whether our results can help contextualize emerging diseases, namely COVID-
282 19, for which a high morbidity burden is an important risk factor for critical illness[33]. We
283 identified a total of 487 significant genetic correlations (r_G; false discovery rate (FDR) < 5%)
284 with at least one of 236 diseases (**Supplementary Tab. 7 and Fig. 5**) based on three different
285 COVID-19 outcomes provided by the COVID-19 Host Genetics Initiative[34]. More than half of
286 the genetically correlated diseases (n=129; 54.7%) are predominantly managed in primary
287 care, spanning a wide range of clinical specialties. For example, urticaria (r_G=0.37; $p<1.0x10^{-}$
288 ^{5}), viral enteritis (r_G=0.31; $p<6.5x10^{-3}$) and pyelonephritis (r_G=0.37; $p<7.5x10^{-3}$), erectile
289 dysfunction (r_G=0.28; $p<2.2x10^{-4}$), or gout (r_G=0.18; $p<5.7x10^{-5}$) were significantly, although
290 moderately, correlated with critical illness following COVID-19 (**Supplementary Fig. 5 and**
291 **Tab. 7**). However, in Mendelian randomization (MR) analysis we did not obtain robust
292 evidence (FDR<5%) that any of the 162 significantly correlated diseases for which we had
293 sufficient genetic instruments was likely causative for any of the COVID-19 outcomes
294 (**Supplementary Tab. 7**). In contrast, we found evidence that genetic liability to all three
295 COVID-19 outcomes was significantly associated with viral pneumonia (e.g., critical illness:
296 beta=0.32; $p<7.3x10^{-8}$) in a reverse MR setting (**Supplementary Tab. 7**). Even if not causal,
297 these results motivate to consider diseases predominantly managed in primary care in the
298 risk assessment for COVID-19, and our results can be immediately repurposed once large-
299 scale GWAS for post-acute sequelae of COVID-19[35] become available.
300
301 ***Drug repurposing and adverse effect prediction***
302 To demonstrate the translational potential of our results, we next cross-referenced all
303 candidate causal genes with drug target annotations using the Open Targets platform[36], and
304 observed a 1.8-fold enrichment (409 out of 3088; $p<3.2x10^{-21}$) of genes matching to drug
305 targets, in line with previous observations[6,7]. This included 136 gene-drug(s)-disease triplets
306 matching current clinical trials or approved medications, a number that almost tripled (n=387
307 gene-drug(s)-disease triplets) once we considered experimentally identified gene and protein
308 interaction partners[37]. This included well-known examples, such as RANKL inhibitors for

309 osteoporosis (via *TNFSF11*), potassium channel blockers for type 2 diabetes (via *KCNJ11*), or
310 more recent developments like alpha-synuclein inhibitors for Parkinson's disease (e.g.,
311 NCT02459886; via *SCNA*) (**Supplementary Tab. 8**).
312 To identify potentially immediate examples to explore drug repurposing, we selected a total
313 of 133 approved drugs (n=39 distinct loci) for which we 1) had matching GWAS evidence to
314 establish effect directions of target modulation, and 2) observed additional diseases not yet
315 indicated for the drug (**Supplementary Tab. 9** and https://omicscience.org/apps/phecodes/).
316 While most newly suggested disease indications were similar to the established indication,
317 we identified noteworthy exceptions. For example, the disease spectrum for Janus kinase
318 (JAK) 2-inhibitors or TNF-alpha antibodies indicated for rheumatoid arthritis might be
319 augmented to autoimmune thyroid disease (**Fig. 5**). In detail, variants likely tagging *IRF5*
320 (rs142738614) and *STAT4* (rs11889341) were consistently associated with a higher risk for
321 rheumatoid arthritis as well as hypothyroidism, and *STAT4* further with thyrotoxicosis. Both
322 gene products act as transcription factors regulating the targets for JAK2-inhibitors and TNF-
323 alpha antibodies (**Fig. 5**). The association between both variants and hypothyroidism
324 persisted upon exclusion of patients with a history of thyrotoxicosis, while thyrotoxicosis was
325 solely associated with the *STAT4* variant. This might suggest that while JAK2-inhibitors and
326 TNF-alpha antibodies are candidates for repurposing for, likely autoimmune, hypothyroidism,
327 only JAK2-inhibitors might be explored for autoimmune thyrotoxicosis such as Graves'
328 disease. The co-occurrence of rheumatoid arthritis and autoimmune thyroid disease has long
329 been recognized[38,39], with few, but suggestive reports that TNF-alpha inhibitors might
330 improve thyroid autoimmunity in patients with rheumatic arthritis[40] and inhibition of TNF-
331 alpha signalling in a mouse model ameliorated thyroiditis[41]. While we provide here genetically
332 anchored evidence for targeted treatment strategies, the best timepoint for the treatment
333 needs to be established in preclinical models and human trials as both medications might be
334 most effective early in the disease process before much of the thyroid gland is damaged.
335

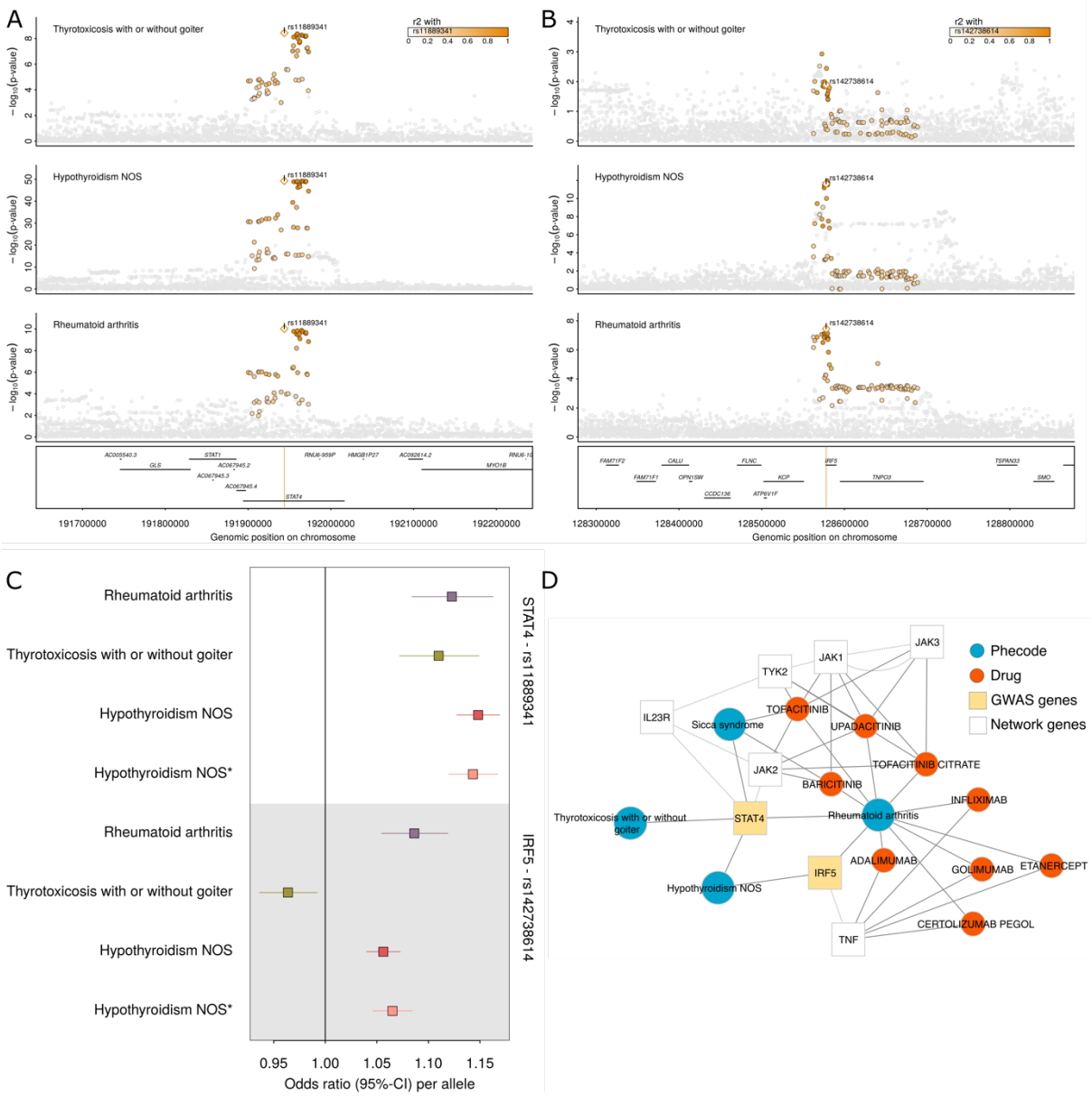


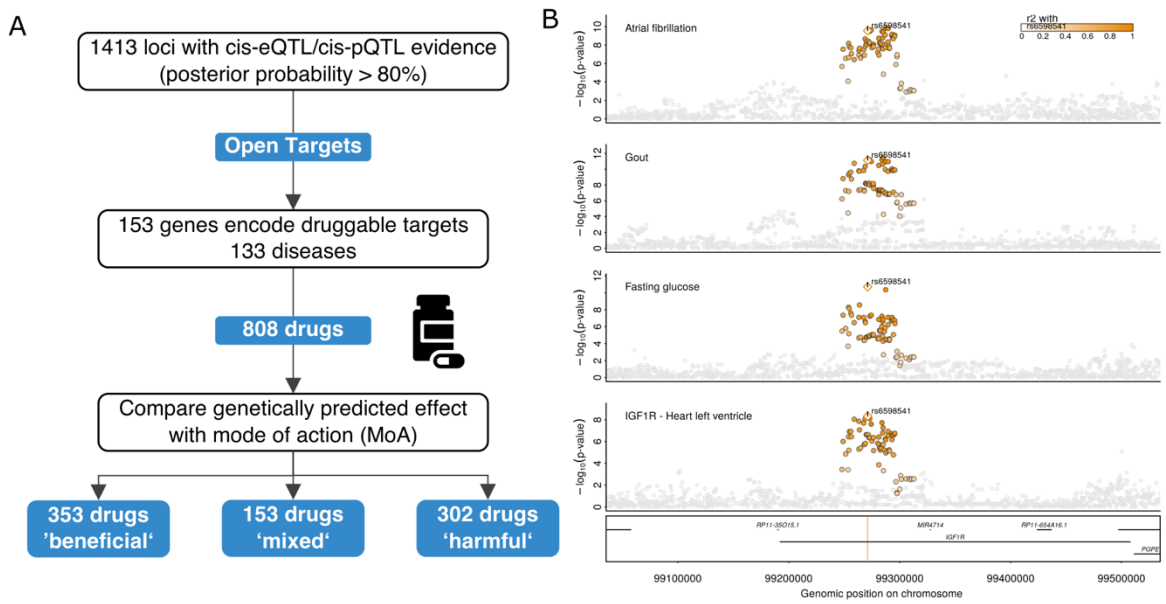
Figure 5 Evidence for repurposing of drugs indicated for rheumatoid arthritis for thyroid disease. A Stacked regional association plot centred around *STAT4* showing regional association statistics from logistic regression models for rheumatoid arthritis, hypothyroidism not elsewhere specified (NOS), and thyrotoxicosis with or without goitre. Each dot represents a single nucleotide polymorphism and those in high to moderate linkage disequilibrium (r^2) with the lead signal at the locus are colour in orange. **B** Same as **A** but centred around *IRF5*. **C** Odds ratios and 95%-confidence intervals for the two lead variants at each locus. *statistics for hypothyroidism NOS after excluding participants recorded for thyrotoxicosis with or without goitre. **D** Disease-gene-drug network displaying how JAK-2 inhibitors and TNF-alpha antibodies are linked to risk loci identified for rheumatoid arthritis and thyroid disease (light orange rectangles). Genes reported to interact with genes at risk loci are indicated by dotted lines and drugs were linked to target genes based on the Open Targets platform.

Molecular informed treatment options and adverse drug effects

To identify further opportunities for drug repurposing, to inform drug selection in early clinical trial phases, but also to predict potential adverse drug effects more systematically, we cross-

352 referenced expression of candidate causal genes or proteins (*cis*-acting) with strong evidence
353 of a shared genetic signal at risk loci (PP>80%) with recorded drug targets from Open
354 Targets[36]. We identified a total of 153 genes (n=142 by cis-eQTL; n=20 by cis-pQTL) across 133
355 diseases for which the product is the target of at least one out of 808 unique drugs
356 (**Supplementary Tab. 10**). These drugs segregated into three different categories when we
357 compared the genetically predicted effect of the target with the mechanism of action (MoA)
358 (**Supplementary Tab. 10 and 11**): 1) 353 drugs with a putative 'beneficial' effect, i.e., the
359 matching drug supposedly reverses the genetically predicted adverse effect of the target, 2)
360 302 drugs with putative 'harmful' effects, i.e., the drug might exaggerate the genetically
361 predicted mechanism, and 3) 153 drugs with both, putative 'beneficial' and 'adverse' effects.
362 The latter two highlighted hypothyroidism as a shared adverse effect for 53 drugs (10 targets),
363 including the well-described thyrotoxic effect of immune checkpoint inhibitors like
364 Ipilimumab[42].
365 Among the drugs with exclusively putative 'beneficial' effects in at least this study, were
366 multiple opportunities for drug repurposing for diseases mainly treated in primary care. For
367 example, the insulin-like growth factor I receptor (IGF1R) agonist Mecasermin might be
368 explored for the treatment of gout (*IGF1R* expression in heart left ventricle: PP=96.8%; OR
369 per 1 s.d. increase in gene expression: 0.75, 95%-CI: 0.69-0.81; $p<3.24x10^{-12}$; **Fig. 6**). A finding
370 also supported by a previous GWAS on serum urate[43], a biomarker for gout, candidate gene
371 studies[44], and more recently the Global Biobank Meta-Analysis Initiative[45,46]. We further
372 identified putative beneficial effects on the risk of atrial fibrillation (PP=96.1%; OR: 0.82, 95%-
373 CI: 0.78 – 0.88; $p<2.7x10^{-10}$), glomerular filtration (PP=96.6%; beta=1.61; $p<7.3x10^{-8}$), body
374 composition (e.g., hip circumference: PP=94.8; beta=-0.05; $p<9.9x10^{-10}$), and fasting glucose
375 levels (PP=99.0%; beta=-0.05; $p<2.0x10^{-11}$) in a phenome-wide colocalization screen. In
376 contrast, we observed little evidence for potential adverse effects, with haematocrit
377 (PP=98.2%; beta=0.05; $p<7.9x10^{-10}$) and red blood cell distribution width (PP=85.2%,
378 beta=0.05, $p<1.1x10^{-8}$) being the sole examples (**Supplementary Tab. 12**).
379 Low plasma glucose levels are a common side effect of Mecasermin[47] and a potential role of
380 *IGF1R* in type 2 diabetes more generally has been recently proposed[48]. There is further
381 evidence that the small molecule BGP-15, that stabilizes IGF1R, protects against experimental
382 models of atrial fibrillation in mice[49]. While these results demonstrate the translational

383 potential of integrating genetics with EHRs, they should be treated as preliminary and
384 evaluation in appropriate preclinical models is needed before human trials are considered.
385



386

387 **Figure 6 Summary of molecular drug target prioritization.** A Flowchart for the prioritization for drug
388 targets mapping to genetic risk loci in the present study. eQTL = expression quantitative trait loci; pQTL
389 = protein quantitative trait loci B Stacked regional association plot centred around *IGF1R*. Summary
390 statistics from logistic regression models for gout are from the present study and those for Atrial
391 fibrillation and fasting glucose have been obtained from the OpenGWAS database[50]. Gene expression
392 summary statistics were obtained from GTEx v8[21]. Colouring indicates linkage disequilibrium with the
393 lead cis-eQTL at the locus.
394

395 **DISCUSSION**

396 Despite their significant impact on patients' lives and healthcare expenses[2], numerous
397 common diseases predominantly managed in primary care have been largely neglected and
398 are underrepresented in biomedical research. We identified hundreds of genetic loci
399 associated with such diseases by harmonizing millions of EHRs towards a common set of
400 medical concepts and 1) demonstrated a substantial genetic control for most, in part driven
401 by strong single loci, 2) identified new diseases mechanisms, 3) showed the ubiquitous
402 presence of pleiotropy by which single loci or groups of loci are linked to multiple diseases or
403 multimorbidity, and 4) showcased how our results can guide treatment strategies by
404 highlighting opportunities for drug repurposing or to predict potentially adverse effects. We
405 created an interactive webserver (https://omicscience.org/apps/phecodes/) for fast
406 dissemination and exploration of our results to benefit the scientific community.

407 Most previous studies[51–53] focussed either on self-report or EHRs from secondary care to
408 perform GWAS with only few starting to explore the potential of primary care records[15]. This
409 focus on serious diseases is in stark contrast to the impaired quality of life patients with non-
410 serious diseases experience. For example, about 70% of patients with rosacea suffer from low
411 self-esteem and one-third reported stigmatization due to the persistent facial erythema and
412 flushing, that led them to reduce social contacts and they even missed work[54]. We confirm
413 known genetic causes of rosacea, like low-frequent variations in *STAT1*[55,56] (**Supplemental**
414 **Tab. 2**), provide new effector genes, like *PNLIPRP3*, and augment current concepts of how
415 toll-like receptors (TLR) contribute to the disease aetiology, with genetic support for TLR1
416 (**Supplemental Tab. 2**), but not TLR2[57]. These results clearly demonstrated the immense
417 potential of EHRs to immediately provide conceptual advances to understand the aetiology
418 of many common conditions managed in primary care. Such strong and distinct genetic
419 findings will also help to increase the recognition of diseases and medical conditions that
420 significantly impair patients' quality of life even without fatal consequences. However, the
421 medical concept ontology we used, phecodes, occasionally comprised a large set of often
422 diverse medical codes and more work on refined definitions is needed to transition from the
423 'phenome' to the 'diseasome'.
424 In an ever-aging population, most patient will develop not one but multiple diseases, so-called
425 multimorbidity[3], but it is currently unclear what drives multimorbidity, e.g., whether
426 commonly co-occurring conditions are genetically linked, and to what extend less severe
427 diseases contribute to the burden. We observed substantial pleiotropy at more than a third
428 of the genomic regions identified, with many genetic variants having effects on multiple
429 diseases, sometimes in opposing directions, that eventually revealed clusters of, sometimes
430 seemingly unrelated, diseases that seem to share common antecedents and might therefore
431 provide a mechanistic angle to understand and tackle the burden of multimorbidity.
432 Human genetic studies have now a pivotal role in the drug discovery process and we
433 replicated the enrichment of druggable targets among candidate causal genes[6,7] and further
434 provided support for multiple examples to explore drug repurposing. Most 'rediscoveries' did
435 thereby not relate to the candidate causal gene at the risk locus but rather indirectly via
436 protein – protein interaction networks, in line with recent observations[58,59]. State-of-the-art
437 genetic prediction tools, like MR, are not necessarily suited to formally support such
438 predictions, like JAK2-inhibitors for autoimmune thyroid disease, and we therefore anchored

439 such findings on shared disease aetiology instead. Our results can therefore provide a
440 resource for similar efforts incorporating GWAS summary statistics from large consortia with
441 higher power for less common outcomes. The same applies to the genetic prediction of
442 adverse events.
443 There are a few limitations that need to be considered when interpreting our results. First,
444 while we integrated multiple EHR resources that doubled the sample size for many outcomes
445 compared to previous studies, billing records do not represent clinical diagnosis and only
446 contextualization of genetic findings with disease biology, like gene expression in relevant
447 cell-types, can aid translation of our results. Second, our findings require replication in
448 independent and, importantly, ethnically diverse studies, which will hopefully be possible in
449 the near future. Third, we applied a GW-threshold to declare statistical significance that did
450 not take the number of diseases tested into account. While this may have increased the false-
451 positive rate, the replication rate of GW-significant loci is high in general[60] and our strategy
452 aligns with recent successful examples[15]. Most importantly it allowed us to provide first
453 biologically plausible and clinically relevant loci for diseases never looked at with still
454 comparatively small sample sizes given the population-based character of UK Biobank, and
455 further to investigate their role in a larger disease context based on genetic pleiotropy.
456 Fourth, studying rare (MAF<0.1%) genetic variation with putatively stronger effects will likely
457 reveal additional findings and potentially further support herein identified candidate causal
458 genes.
459 Our results demonstrate that integration of genetic data with billing information from EHRs,
460 including primary care, can yield biologically tangible and clinically relevant insights to
461 mitigate the health burden from common but often understudied diseases. We hope that our
462 results will serve as a resource to the wider research community to further accelerate the
463 development of much needed safe and effective treatments.
464

465 **ACKNOWLEDGEMENT**

466 The authors acknowledge the Scientific Computing of the IT Division at the Charité -
467 Universitätsmedizin Berlin for providing computational resources that have contributed to
468 the research results reported in this paper
469 (https://www.charite.de/en/research/research_support_services/research_infrastructure/s
470 cience_it/#c30646061). We want to acknowledge the participants and investigators of the

FinnGen study. M.K. is supported by Gates Cambridge Trust. J.C.-Z. is supported by a 4-year Wellcome Trust PhD Studentship and the Cambridge Trust.

Data availability

All individual level data is publicly available to bona fide researchers from the UK Biobank (https://www.ukbiobank.ac.uk/). This research has been conducted under the application 44448. GWAS summary statistics are available from the interactive webserver upon publication (https://omicscience.org/apps/phecodes/).

Code availability

Associated code and scripts for the analysis will be made available upon publication.

Declaration of Interest

None of the authors have a conflict of interest.

Author contributions

Conceptualization: MP, HH, CL

Data curation/Software: MP, SD, MW, YS, AN, MA

Formal Analysis: MP, YS, HS, BU, MK

Methodology: MP, SD

Visualization: MP, BU

Funding acquisition: CL, HH

Project administration: CL, HH

Supervision: MP, CL, HH

Writing – original draft: MP, CL, HH

Writing – review & editing: SD, SY, MK, HS, JCZ, AV, GK

498 **METHODS**
499
500 *Study population*
501 UK Biobank (UKB) is a prospective cohort study from the UK, which contains more than
502 500,000 volunteers between 40 and 69 years of age at inclusion. The study design, sample
503 characteristics and genome-wide genotype data have been described in Sudow *et al.*[61] and
504 Bycroft *et al.*[62]. The UKB was approved by the National Research Ethics Service Committee
505 Northwest Multi-Centre Haydock and all study procedures were performed in accordance
506 with the World Medical Association Declaration of Helsinki ethical principles for medical
507 research. We included 441,671 individuals in the genome-wide association analyses for whom
508 inclusion criteria (given consent to further usage of the data, availability of genetic data and
509 passed quality control of genetic data) applied.
510
511 *Disease ascertainment*
512 We collated electronic health records from primary and secondary care, cancer registries, and
513 death certificates based on tables provided by UK Biobank (gp_clinical.txt,
514 covid19_emis_gp_clinical.txt, covid19_tpp_gp_clinical.txt, hesin_diag.txt, death.txt)
515 downloaded in June 2021. We parsed all records to exclude codes with a recorded date before
516 or within the year of birth of the participant to minimize coding errors from EHRs. For each
517 data set separately we generated mapping tables that link ICD-10, ICD-9, Read version 2,
518 Clinical Terms Version 3 (CTV3) terms, or SNOMED-CT codes to a set of 1560 summarized
519 clinical entities called phecodes[11,63] (**Supplementary Tab. 1**). For example, more than 90 ICD-
520 10 codes can indicate participants with type 1 diabetes that are here collectively summarized
521 under the phecode 'type 1 diabetes'[10]. We subsequently fused all data sources based on a
522 common set of phecodes and retained for each participant and each phecode only the earliest
523 entry across all EHR resources. We adjusted all analyses for assessment centre to account for
524 regional differences in coding systems and case ascertainment. To increase the accessibility
525 of our results, we used the term 'disease' instead of 'phecode' throughout the paper.
526
527 *Genotyping, quality control and participant selection*
528 Details on genotyping for UKBB have been reported in detail by Bycroft et al.[62]. Briefly, we
529 used data from the 'v3' release of UKBB containing the full set of Haplotype Reference

530 Consortium (HRC) and 1000 Genomes imputed variants. We applied recommended sample
531 exclusions by UKBB including low quality control values, sex mismatch, and heterozygosity
532 outliers. We defined a subset of 'white European' ancestry by clustering participants based
533 on the first four genetic principal component derived from the genotyped data using a k-
534 means clustering approach with k=5. We classified all participants who belonged to the
535 largest cluster and self-identified as of being 'white,' 'British', 'Any other white background',
536 or 'Irish' as 'white European'. After application of quality control criteria and dropping
537 participants who have withdrawn their consent, a total of 444,441 UKBB participants were
538 available for analysis with genotype and phenotype data.
539 We used only called or imputed genotypes and short insertions/deletions (here commonly
540 referred to as SNPs for simplicity) with a minor allele frequency (MAF) > 0.001%, imputation
541 score >0.4 for common (MAF≥0.5%) and >0.9 for rare (MAF<0.5%), within Hardy-Weinberg
542 equilibrium ($p_{HWE}>10^{-15}$), and minor allele count (MAC) > 10. This left us with 15,519,342
543 autosomal and X-chromosomal variants for statistical analysis. GRCh37 was used as reference
544 genome assembly.
545
546 *Confirming rare, imputed variants with whole exome sequencing data*
547 To assess the quality of the assigned genotypes through imputation, we have examined the
548 correlation of overlapping rare variants with at least one significant disease association
549 between the imputed and whole exome sequencing (WES) data available over 200,000
550 individuals in UK Biobank (available from October 2020). For the comparison, we first lifted
551 over the positions of rare, imputed variants from b37 to b38 using LiftOver and converted the
552 SNP dosages of the imputed variants from bgen files to hard-call genotypes using plink v.2.
553 Subsequently, we have assessed the correlation of overlapping rare variants between the
554 imputed and WES data for 200,039 consented individuals. We have defined variants with
555 $r^2>0.8$ as concordant.
556
557 *Genome-wide association studies*
558 We performed genome-wide association studies (GWAS) for a total of 1,445 phecodes with
559 at least 80 cases (n>100 prior genetic exclusions) using REGENIE v2.2.4 via a two-step
560 procedure to account for population structure as described in detail elsewhere[64]. We used a
561 set of high-quality genotyped variants (MAF>1%, MAC>100, missingness <10%, $p_{HWE}>10^{-15}$) in

562 the first step for individual trait predictions using the leave one chromosome out (LOCO)
563 scheme. These predictions were used in the second step as offset to run logistic regression
564 models with saddle point approximation to account for case/control imbalance and rare
565 variant associations. Each model was adjusted for age, sex, genotyping batch, assessment
566 centre, and the first ten genetic principal components. We used a standard genome-wide
567 significance threshold to declare significant findings ($p<5x10^{-8}$). For diseases reported in only
568 one sex (n=113 in women, n=26 in men), we excluded the respective set from GWAS to avoid
569 inflation by inappropriate controls. In general, we included all participant with a disease in
570 their records as case and treated all other participants as controls to make best use of the
571 computational efficacy of REGENIE. Testing for reported SNPs showed highly consistent
572 results whether related diseases were included as controls rather than omitted. We used LD-
573 score regression to test for genomic inflation and calculated the SNP-based heritability (LDSC
574 v1.0.1)[65]. In the absence of general prevalence estimates across diseases, we used the
575 prevalence of diseases in UKBB to convert SNP-based heritability estimates to the liability
576 scale.
577
578 *Signal selection and fine-mapping*
579 We used regional clumping (±500kb) to select independent genomic regions associated with
580 each disease treating the extended MHC region as a whole (chr6:25.5-34.0Mb) and collapsed
581 neighbouring regions using BEDtools v1.5.
582 Within each region, apart from the MHC region, we performed statistical fine-mapping using
583 "Sum of single effects" model (SuSiE)[66] as implemented in the R package *susieR* (v.0.11.92).
584 Briefly, SuSiE employs a Bayesian framework for variable selection in a multiple regression
585 problem with the aim to identify sets of independent variants each of which likely contains
586 the true causally underlying genetic variant. We implemented the workflow using default
587 prior and parameter settings, apart from the minimum absolute correlation, which we set to
588 0.1. Since SuSiE is implemented in a linear regression framework, we used the GWAS
589 summary statistics with a matching correlation matrix of dosage genotypes instead of
590 individual level data to implement fine-mapping (*susie_rss()*) as recommended by the
591 authors[66]. We computed the correlation matrix with LDstore v1.0. In situations where SuSiE
592 did not deliver a credible set, we used the Wakefield approximation[67] to compute 95%-
593 credible sets.

594 We additionally performed sensitivity analyses by computing joint models for each locus –
595 disease combination, that is, we obtained all lead SNPs across the credible sets provided by
596 SuSiE and fitted one logistic regression model jointly modelling the effect of all distinct
597 credible sets in the locus. We subsequently retained only lead variants from each credible set
598 that reached genome-wide significance in either marginal or conditional statistics. We
599 implemented the logistic regression models in R v.4.1.2 using the *glm()* function and used only
600 unrelated white-European participants and the same adjustment set as described above.
601 We finally used LD-clumping (r^2>0.8) to identify credible sets shared across diseases.
602
603 *Definition of unreported loci and previous associations*
604 To define genetic loci that have not been reported so far, we first systematically checked
605 whether any of the identified variants or proxies (r^2>0.1) has already been reported in the
606 OpenTargets database[12] or the GWAS catalog[14] (download: 27/04/2022). We mapped
607 phecodes to EFO terms using ICD-10 and SNOMED-CT codes to define 'matching' phenotypic
608 entries.

609 *Replication in FinnGen*
610 In addition to a rigorous comparison to the GWAS catalog and other databases, we sought to
611 replicate our findings in the latest release from the FinnGen consortium[15] (release 8;
612 download 20/01/2023). We mapped phecodes to disease codes provided by FinnGen using
613 common ICD-10 codes and kept only GWAS summary statistics for diseases with at least half
614 the sample size in FinnGen (n=342,499) and for each mapping entry only the one with the
615 highest sample size. We note that due to different codes considered to group diseases, these
616 mappings should be regarded as approximate which may have lowered the success for
617 replication. After download, we lifted FinnGen statistics over to genome build 37 and
618 obtained summary statistics for a total of 2560 genetic variant – disease pairs with significant
619 marginal effects in UKB and approximately similar allele frequencies in both cohorts (≤20%
620 relative difference). We refer to concordant associations for all pairs with the same
621 association direction and at least nominal statistical significance (p<0.05) in FinnGen.
622
623 *Sex-specific SNP effect*

624 We tested for a potential modulating effect of sex on SNP – phecode association using an
625 additional interaction term in logistic regression models with the same adjustment set as
626 described above. We implemented this analysis in R v.4.1.2 using a set of 361,781 unrelated
627 white European individuals.
628
629 *Candidate Causal Gene assignment*
630 We used six different resources to provide candidate causal genes at each of the loci
631 identified.
632 1) Statistical colocalization with gene expression (cis-eQTL) in 49 tissues from the GTEx
633 project (v8)[21]. Briefly, we considered all protein coding genes or processed transcripts
634 encoded in a 1Mb window around risk loci and performed statistical colocalization[68]
635 to test for a shared genetic signal between a disease and gene expression. We adopted
636 a recently recommended prior setting[69] with $p_{12}=5x10^{-6}$.
637 2) Statistical colocalization with plasma protein levels (cis-pQTL) from our previous
638 study[22] implemented in the same way as for eQTL.
639 3) Annotation of lead genetic variants and potential functional proxies ($r^2>0.6$) using the
640 Variant Effect Predictor software (version 106). We considered all genetic variants as
641 'functional' that had a score of twelve or lower based on the VEP recommendations
642 (https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html).
643 4) Nearest gene based on the gene body or transcription start site.
644 5) The Open Targets variant-to-gene pipeline[13].
645 6) A biomedical knowledge graph[23] that comprised multiple, often manually curated,
646 resources of gene – disease and other links. We assigned candidate genes based on
647 whether any of the ten nearest genes was no more than two nodes away from the
648 matching disease code ('phecode').
649 For each candidate gene at a locus, we computed a weighted score across all six criteria. We
650 divided colocalization evidence based on moderate (50%>PP<80%; 0.5 points) or high
651 (PP>80%; 1 point) confidence and functional variants were assigned highest priority (2
652 points). The final score is the sum across all criteria. We reported gene(s) with the highest
653 score as candidate causal gene(s) at each locus.

654 *Visualising cell-type specific gene expression*

655 We queried cell-type specific expression patterns for *PNLIPRP3* in the CELLxGENE browser[70]
656 and subsequently obtained the pre-processed Seurat from the solely single-nuclei RNA
657 sequencing study that showed expression levels of *PNPLIPRP3*. The study was based on eight
658 healthy organ donors from the GTEx project and contained a total of 209,126 cells[29]. We
659 imported the data into R using the Seurat package (v.4.0.0) and restricted the plotting to
660 epithelial and other cells of the skin to produce a Uniform Manifold Approximation and
661 Projection (UMAP) plot. We excluded alveolar, ciliated, and neuroendocrine cells from the
662 plot to increase visibility. The full UMAP plot is available as Supplementary Fig. 4.

663 *Bi-clustering of locus – disease network*
664 To identify highly connected co-clusters of loci and disease, we performed BiLouvain
665 clustering algorithm on a bipartite network. The bipartite network consisted of independent
666 sets of loci and diseases, thereby loci were connected to diseases through genome-wide
667 significant associations and diseases are only connected through shared loci. We used the
668 following settings to implement the BiLouvain algorithm[30]: cut-off for iterations = 0.01, cut-
669 off for phases = 0.0, order = 1, fuse = 0, CF = 0.0 and alpha = 1.0. The result of the clustering
670 had a Murata+ modularity score of 0.759. We used Cytoscape v3.9.0 to visualize the
671 network[71].
672
673 *COVID-19 genetic correlation and Mendelian randomization*
674 We downloaded GWAS summary statistics for three different endpoints related to COVID-19
675 (A2 – critical illness; B2 – hospitalisation; C2 – SarsCov-2 infection) provided by the COVID-19
676 Host Genetics Initiative (release 7)[34]. We used summary statistics excluding UK Biobank to
677 avoid sample overlap. We computed genetic correlations as implemented by LD-score
678 regression (LDSC v1.0.1)[65] with precomputed LD-scores, excluding the extended MHC region.
679 To test for potentially causal associations of phecodes onto COVID-19, we used genetic
680 instruments identified in the present study for a total of 176 phecodes with at least five
681 genetic variants in a two-sample MR setting. We used MR-PRESSO[72] as a first line tool as
682 previously suggested[73] to account for possible pleiotropy and subsequently report effect
683 estimates from inverse-variance weighted analysis as the primary results. We flagged MR
684 results that showed signs of heterogeneity across instruments using Cochran Q statistic.
685 Conversely, we obtained regional lead variants ($p<5x10^{-8}$; 1Mb window) across all three

686 COVID-19 outcomes to test for a putative causal effect of genetic liability towards COVID-19
687 and 850 phecodes with at least 1000 cases. We excluded any variants mapping to the MHC
688 regions for all analysis and implemented MR using the R packages *MendelianRandomization*
689 (v0.6.0)[74] and *TwoSampleMR* (v0.5.6)[75].
690
691 *Drug target annotation*
692 To identify putative druggable targets and test for evidence for drugs already in clinical trials,
693 we queried the Open Target database[36] for all genes likely linked to GWAS loci based on our
694 candidate causal gene assignment. We augmented this gene list by functionally validated
695 interaction partners from the SIGNOR network[37] and retained all matching drugs for further
696 investigation. We further mapped phecodes to different disease ontologies to obtain drugs
697 already in clinical trials and to define indication matchings.
698 To provide evidence whether a drug might be beneficial or harmful based on genetic
699 predictions, we compared genetically predicted effect estimates based on Wald ratios
700 between colocalising (PP>80%) gene/proteins and phecodes with mechanisms of action again
701 obtained from the Open Targets database.
702
703 *Phenome-wide colocalization screen at IGF1R*
704 To explore potential effects of genetic variation affecting gene expression of *IGF1R* beyond
705 gout, we obtained corresponding summary statistics (±250kb around the gene body) for the
706 expression in left ventricle of the heart (positive colocalization results for gout) from GTEx
707 v8[21]. We then queried the lead variant (rs6598541) and proxies in strong LD (r^2>0.8; based on
708 UK Biobank) using the OpenGWAS database[50] to identify all associations with $p<10^{-5}$. For all
709 49 traits meeting this criterion, we performed statistical colocalization as described above.
710 Since some anthropometric traits harboured multiple independent signals in the region,
711 violating the single causal variant assumption, we further used a more recent development
712 of colocalization[76] that first uses SuSiE to fine-map[77] each trait and afterwards performs
713 colocalization among all possible credible sets. We used the same LD-backbone as described
714 in the fine-mapping section and further restricted the maximum number of credible sets to
715 five and pruned correlated credible sets if needed (r^2>0.2). We eventually used the Wald ratio
716 to provide an estimate about the effect of increasing *IGF1R* expression by 1 s.d. on all
717 outcomes examined.

718 **REFERENCES**
719

720 1. Abbafati, C. *et al.* Global burden of 369 diseases and injuries in 204 countries and
721 territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study
722 2019. *Lancet* **396**, 1204–1222 (2020).
723 2. Dieleman, J. L. *et al.* US Health Care Spending by Payer and Health Condition, 1996-
724 2016. *JAMA* **323**, 863–884 (2020).
725 3. Whitty, C. J. M. *et al.* Rising to the challenge of multimorbidity. *The BMJ* vol. 368
726 (2020).
727 4. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Prim.* **1**, 59
728 (2021).
729 5. Abdellaoui, A., Yengo, L., Verweij, K. J. H. H. & Visscher, P. M. 15 years of GWAS
730 discovery: Realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).
731 6. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug
732 indications. *Nat. Genet.* **47**, 856–860 (2015).
733 7. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as
734 likely to be approved? Revised estimates of the impact of genetic support for drug
735 mechanisms on the probability of drug approval. *PLOS Genet.* **15**, e1008489 (2019).
736 8. Yengo, L. *et al.* A saturated map of common genetic variants associated with human
737 height. *Nature* **610**, 704–712 (2022).
738 9. Hartmann, S. *et al.* Genome-wide association study identifies ADRA2A and IRX1 as
739 novel risk genes for Raynaud's phenomenon. *medRxiv* 2022.10.19.22281276 (2022)
740 doi:10.1101/2022.10.19.22281276.
741 10. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From
742 PheWAS to PheRS. *Annu. Rev. Biomed. Data Sci.* **4**, 1–19 (2021).
743 11. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow
744 Development and Initial Evaluation. *JMIR Med. informatics* **7**, e14325 (2019).
745 12. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-
746 associated genes using large-scale genetics and functional genomics. *Nucleic Acids*
747 *Res.* **49**, D1311–D1320 (2021).
748 13. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and
749 genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533
750 (2021).
751 14. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition
752 resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
753 15. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated
754 population. *Nature* **613**, 508–518 (2023).
755 16. Hörer, S. *et al.* A Monoallelic Two-Hit Mechanism in PLCD1 Explains the Genetic
756 Pathogenesis of Hereditary Trichilemmal Cyst Formation. *J. Invest. Dermatol.* **139**,
757 2154-2163.e5 (2019).
758 17. Hann, L. E., Hall, D. A., Black, E. B. & Ferrucci, J. T. Mittelschmerz. Sonographic
759 demonstration. *JAMA* **241**, 2731–2 (1979).
760 18. Schall, P. Z., Ruebel, M. L., Midic, U., VandeVoort, C. A. & Latham, K. E. Temporal
761 patterns of gene regulation and upstream regulators contributing to major
762 developmental transitions during Rhesus macaque preimplantation development.
763 *Mol. Hum. Reprod.* **25**, 111–123 (2019).
764 19. Zhang, J., Tomasini, A. J. & Mayer, A. N. RBM19 is essential for preimplantation

765 development in the mouse. *BMC Dev. Biol.* **8**, 115 (2008).
766 20. Maurer, M. S. *et al.* Tafamidis Treatment for Patients with Transthyretin Amyloid
767 Cardiomyopathy. *N. Engl. J. Med.* **379**, 1007–1016 (2018).
768 21. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across
769 human tissues. *Science* **369**, 1318–1330 (2020).
770 22. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases.
771 *Science* **374**, eabj1541 (2021).
772 23. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision
773 medicine. *Sci. data* **10**, 67 (2023).
774 24. Mahajan, M. C., Narlikar, G. J., Boyapaty, G., Kingston, R. E. & Weissman, S. M.
775 Heterogeneous nuclear ribonucleoprotein C1/C2, MeCP1, and SWI/SNF form a
776 chromatin remodeling complex at the beta-globin locus control region. *Proc. Natl.*
777 *Acad. Sci. U. S. A.* **102**, 15012–7 (2005).
778 25. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**,
779 1260419 (2015).
780 26. No Title. https://www.uniprot.org/uniprotkb/Q17RR3/entry
781 https://www.uniprot.org/uniprotkb/Q17RR3/entry (2023).
782 27. Dajnoki, Z. *et al.* Sebaceous Gland-Rich Skin Is Characterized by TSLP Expression and
783 Distinct Immune Surveillance Which Is Disturbed in Rosacea. *J. Invest. Dermatol.* **137**,
784 1114–1125 (2017).
785 28. Feofanova, E. V *et al.* A Genome-wide Association Study Discovers 46 Loci of the
786 Human Metabolome in the Hispanic Community Health Study/Study of Latinos. *Am. J.*
787 *Hum. Genet.* **107**, 849–863 (2020).
788 29. Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps toward
789 understanding disease gene function. *Science* **376**, eabl4290 (2022).
790 30. Pesantez-Cabrera, P. & Kalyanaraman, A. Detecting communities in biological
791 bipartite networks. *ACM-BCB 2016 - 7th ACM Conf. Bioinformatics, Comput. Biol.*
792 *Heal. Informatics* 98–107 (2016) doi:10.1145/2975167.2975177.
793 31. Ravnborg, N. *et al.* Prevalence of asthma in patients with atopic dermatitis: A
794 systematic review and meta-analysis. *J. Am. Acad. Dermatol.* **84**, 471–478 (2021).
795 32. Gallagher, A. *et al.* Anti-interleukin-13 and anti-interleukin-4 agents versus placebo,
796 anti-interleukin-5 or anti-immunoglobulin-E agents, for people with asthma. *Cochrane*
797 *database Syst. Rev.* **10**, CD012929 (2021).
798 33. Williamson, E. J. *et al.* Factors associated with COVID-19-related death using
799 OpenSAFELY. *Nature* **584**, 430–436 (2020).
800 34. The COVID-19 Host Genetics Initiative. A second update on mapping the human
801 genetic architecture of COVID-19. *medRxiv* **608**, 2022.12.24.22283874 (2022).
802 35. Al-Aly, Z., Xie, Y. & Bowe, B. High-dimensional characterization of post-acute sequalae
803 of COVID-19. *Nature* **594**, (2021).
804 36. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimagined, redesigned,
805 rebuilt. *Nucleic Acids Res.* **51**, D1353–D1359 (2023).
806 37. Licata, L. *et al.* SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update.
807 *Nucleic Acids Res.* **48**, D504–D510 (2020).
808 38. Liu, Y.-J., Miao, H.-B., Lin, S. & Chen, Z. Association between rheumatoid arthritis and
809 thyroid dysfunction: A meta-analysis and systematic review. *Front. Endocrinol.*
810 *(Lausanne).* **13**, 1015516 (2022).
811 39. Wu, D. *et al.* Graves' Disease and Rheumatoid Arthritis: A Bidirectional Mendelian

Randomization Study. *Front. Endocrinol. (Lausanne).* **12**, 1–9 (2021).
40. Bliddal, S., Borresen, S. W. & Feldt-Rasmussen, U. Thyroid Autoimmunity and Function after Treatment with Biological Antirheumatic Agents in Rheumatoid Arthritis. *Front. Endocrinol. (Lausanne).* **8**, 179 (2017).
41. Di Dalmazi, G., Chalan, P. & Caturegli, P. MYMD-1, a Novel Immunometabolic Regulator, Ameliorates Autoimmune Thyroiditis via Suppression of Th1 Responses and TNF-α Release. *J. Immunol.* **202**, 1350–1362 (2019).
42. Wright, J. J., Powers, A. C. & Johnson, D. B. Endocrine toxicities of immune checkpoint inhibitors. *Nat. Rev. Endocrinol.* **17**, 389–399 (2021).
43. Köttgen, A. *et al.* Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.* **45**, 145–54 (2013).
44. Phipps-Green, A. J. *et al.* Twenty-eight loci that influence serum urate levels: analysis of association with gout. *Ann. Rheum. Dis.* **75**, 124–30 (2016).
45. Namba, S., Konuma, T., Wu, K. H., Zhou, W. & Okada, Y. A practical guideline of genomics-driven drug discovery in the era of global biobank meta-analysis. *Cell Genomics* **2**, 100190 (2022).
46. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**, (2022).
47. Kemp, S. F., Fowlkes, J. L. & Thrailkill, K. M. Efficacy and safety of mecasermin rinfabate. *Expert Opin. Biol. Ther.* **6**, 533–538 (2006).
48. Gardner, E. J. *et al.* Damaging missense variants in IGF1R implicate a role for IGF-1 resistance in the etiology of type 2 diabetes. *Cell genomics* **2**, None (2022).
49. Sapra, G. *et al.* The small-molecule BGP-15 protects against heart failure and atrial fibrillation in mice. *Nat. Commun.* **5**, 5705 (2014).
50. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv* 2020.08.10.244293 (2020) doi:10.1101/2020.08.10.244293.
51. Cortes, A. *et al.* Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. *Nat. Genet.* **49**, 1311–1318 (2017).
52. Cortes, A., Albers, P. K., Dendrou, C. A., Fugger, L. & McVean, G. Identifying cross-disease components of genetic risk across hospital data in the UK Biobank. *Nat. Genet.* **52**, 126–134 (2020).
53. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
54. Oussedik, E., Bourcier, M. & Tan, J. Psychosocial Burden and Other Impacts of Rosacea on Patients' Quality of Life. *Dermatol. Clin.* **36**, 103–113 (2018).
55. Baghad, B. *et al.* Pediatric Demodicosis Associated with Gain-of-Function Variant in STAT1 Presenting as Rosacea-Type Rash. *J. Clin. Immunol.* **41**, 698–700 (2021).
56. Deng, Z. *et al.* Keratinocyte-Immune Cell Crosstalk in a STAT1-Mediated Pathway: Novel Insights Into Rosacea Pathogenesis. *Front. Immunol.* **12**, 1–13 (2021).
57. van Zuuren, E. J. *et al.* Rosacea: New Concepts in Classification and Treatment. *Am. J. Clin. Dermatol.* **22**, 457–465 (2021).
58. Barrio-Hernandez, I. *et al.* Network expansion of genetic associations defines a pleiotropy map of human cell biology. *Nat. Genet.* (2023) doi:10.1038/s41588-023-01327-9.
59. MacNamara, A. *et al.* Network and pathway expansion of genetic disease associations identifies successful drug targets. *Sci. Rep.* **10**, (2020).
60. Marigorta, U. M., Rodríguez, J. A., Gibson, G. & Navarro, A. Replicability and

Prediction: Lessons and Challenges from GWAS. *Trends Genet.* **34**, 504–517 (2018).
61. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
62. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
63. Denaxas, S. Mapping the Read2/CTV3 controlled clinical terminologies to Phecodes in UK Biobank primary care electronic health records: implementation and evaluation. *Proc. Am. Med. Informatics Assoc. Annu. Symp. 2021* (2021).
64. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
65. Bulik-sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Publ. Gr.* **47**, (2015).
66. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
67. Wakefield, J. Bayes Factors for Genome-Wide Association Studies : Comparison with P -values. **86**, 79–86 (2009).
68. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, (2014).
69. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* **16**, 1–20 (2020).
70. Discover., C. I. (n. d.). C. Cell. No Title. *Retrieved (20/01/2023), from https://cellxgene.cziscience.com/*.
71. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003).
72. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
73. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472–477 (2021).
74. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).
75. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).
76. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* **17**, e1009440 (2021).
77. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the " Sum of Single Effects " model. 1–33 (2021).