**Full Title**: Ancestrally and Temporally Diverse Analysis of Penetrance of Clinical Variants in 72,434 Individuals

**Running Title**: Penetrance of Clinical Variants

**Authors**: Iain S. Forrest[1,2,3,4], Kumardeep Chaudhary[1,3,4,5], Ha My T. Vy[1,4], Shantanu Bafna[1,4], Daniel M. Jordan[1,4], Ghislain Rocheleau[1,4], Ruth J.F. Loos[1,4,6], Judy H. Cho[1,3,4,7], Ron Do[1,3,4].

**Author Affiliations**:

1. The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
2. Medical Scientist Training Program, Icahn School of Medicine at Mount Sinai, New York, NY, USA
3. The Bio*Me* Phenomics Center, Icahn School of Medicine at Mount Sinai, New York, NY, USA
4. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
5. The Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA
6. The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

7. Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

**Corresponding Author**:

Ron Do, PhD

Annenberg Building, Floor 18 Room 80A

1468 Madison Ave

New York, NY-10029

Phone Number: 212-241-6206 | Fax Number: 212-849-2643

ron.do@mssm.edu

**Word Count**: 3,840

**No. of Tables and Figures**: 6

**No. of Supplementary Figures**: 5

**No. of Supplementary Tables**: 10

**ABSTRACT**

A major goal of genomic medicine is to quantify the disease risk of genetic variants. Here, we report the penetrance of 37,772 clinically relevant variants (including those reported in ClinVar[1] and of loss-of-function consequence) for 197 diseases in an analysis of exome sequence data for 72,434 individuals over five ancestries and six decades of ages from two large-scale population-based biobanks (Bio*Me* Biobank and UK Biobank). With a high-quality set of 5,359 clinically impactful variants, we evaluate disease prevalence in carriers and non-carriers to interrogate major determinants and implications of penetrance. First, we associate biomarker levels with penetrance of variants in known disease-predisposition genes and illustrate their clear biological link to disease. We then systematically uncover large numbers of ClinVar pathogenic variants that confer low risk of disease, even among those reviewed by experts, while delineating stark differences in variant penetrance by molecular consequence. Furthermore, we ascertain numerous variants present in non-European ancestries and reveal how increasing carrier age modifies penetrance estimates. Lastly, we examine substantial heterogeneity of penetrance among variants in known disease-predisposition genes for conditions such as familial hypercholesterolemia and breast cancer. These data indicate that existing categorical systems for variant classification do not adequately capture disease risk and warrant consideration of a more quantitative system based on population-based penetrance to evaluate clinical impact.

70 The advent of high-throughput sequencing has led to an exponential increase in
71 genetic tests for health-related purposes[2,3]. The American College of Medical Genetics
72 & Genomics (ACMG) recommends clinical action if a pathogenic genetic variant is
73 found in one of 59 genes (hereafter ACMG59)[4]. This "genetics-first" approach is feasible
74 if variant pathogenicity is ascertained with high confidence and veracity. A large
75 database of human variation and phenotypes, ClinVar[1], classifies variant pathogenicity
76 in a scheme that informs clinical interpretation of genetic test results (e.g., "pathogenic",
77 "uncertain significance", "benign")[5]. However, most ClinVar variants are of uncertain
78 clinical significance, and misclassification of 12% of pathogenic and 90% of conflicting
79 variants has inflated their pathogenicity[6]. Variants implicated in diseases from breast
80 cancer to cardiomyopathy have been scrutinized for overestimated claims of
81 pathogenicity[7] and many variants classified as pathogenic were recently downgraded to
82 lower ClinVar classes[8]. Any unreliability of pathogenicity, already a categorical rather
83 than quantitative metric of disease risk, further diminishes its clinical value. There is
84 therefore a critical need to provide accurate information for a variant's effect on disease
85 to guide patient care.
86 Penetrance, the probability of a disease phenotype given a particular genotype[9],
87 is unknown for the vast majority of variants in ClinVar and genetic databases. Well-
88 known exceptions include ~60% of carriers of certain *BRCA1* and *BRCA2* variants
89 develop breast cancer by 70 years of age[10] and 73% of individuals heterozygous for
90 *LDLR* variants present with hypercholesterolemia[11]. Findings of highly penetrant
91 variants immediately guide treatment, such as statins, and screening, such as

92 mammograms[12,13]. Penetrance thus gives meaningful and actionable information for
93 variants by quantifying a carrier's disease risk.

94 Until recently, penetrance has been primarily derived from family-based or
95 clinical cohort studies[14–17]. These typically focus on a small number of genes and
96 inherently maximize penetrance estimates by recruiting patients with disease or family
97 history of disease, and are therefore susceptible to ascertainment bias[18]. Small sample
98 size and genetic or environmental modifiers further limit their ability to reliably appraise
99 penetrance[19]. In contrast, the recent introduction of the UK Biobank (UKB) and other
100 population-based biobanks has created a trove of genetic and phenotypic data for large
101 numbers of unrelated individuals[20,21]. Exome sequences coupled to electronic health
102 records (EHRs) provide an unprecedented opportunity for using a population-based
103 method to measure penetrance on a large scale with less ascertainment bias than
104 traditional studies[22]. In addition, multi-ethnic biobanks such as Mount Sinai's Bio*Me*
105 Biobank (Bio*Me*)[23] enable the analysis of penetrance across diverse ancestries. Recent
106 studies have begun to use a population-based approach to explore penetrance of genes
107 in breast cancer[24] and familial hypercholesterolemia (FaH)[25], yet the population-based
108 penetrance of most variants, genes, and diseases remains uncharacterized.

109 Here, we perform a comprehensive analysis of variant penetrance observed in
110 two large-scale EHR-linked population-based biobanks (Bio*Me* and UKB). We
111 measured the penetrance of 37,772 variants, including a high-quality set of 5,359
112 clinically impactful variants, for 197 diseases of dominant inheritance using exome
113 sequences from 72,434 individuals. Penetrance was analyzed in all individuals and
114 separately in five ancestries across six decades of ages. We investigated determinants

of penetrance, including ClinVar pathogenicity, variant molecular consequence, and carrier ancestry and age. The clinical manifestation of variants was scrutinized by examining metabolite levels, biological measurements, and physician notes for carriers, and clinically significant variants and genes were highlighted. This highly scalable, population-based approach of generating penetrance estimates for thousands of clinically relevant variants expands our knowledge of the genetic influences on disease and has the potential to advance genomic medicine.

## RESULTS

Analysis of penetrance of clinical variants

A graphical abstract is provided (**Supplementary Figure 1**). Whole-exome sequencing, quality control, and filtering were performed on 80,773 samples from two population-based EHR-linked biobanks (**Methods**) to generate a final dataset with 72,434 individuals (**Table 1**). Demographics, self-reported ancestry, biological measurements, and International Classification of Disease 10 (ICD-10) diagnosis codes were available for individuals in the final dataset. We defined case-control status for over 400 non-recessive diseases in ClinVar sourced from the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)[26] using ICD-10 codes[27], of which 197 had at least one case in the final dataset (**Supplementary Table 1**). We analyzed 37,772 variants that were either reported in ClinVar for at least one of the aforementioned diseases or were of predicted loss-of-function (LoF) consequence (**Supplementary Table 2**). A stringent set of 5,359 clinically impactful (impactful) variants—defined by pathogenic/likely pathogenic classification in ClinVar or LoF annotation (splice acceptor/donor, stop gained/lost, frameshift, start lost) with Variant Effect Predictor[28] in a gene mediating disease via LoF mechanism, and in a gene with non-recessive inheritance—was used for downstream analyses unless otherwise stated (**Supplementary Figure 2**). We identified carriers with at least one allele (hereafter referred to as carriers) for each variant and computed the proportion of carriers affected with disease to determine penetrance. Variant penetrance was estimated in all carriers and separately in carriers of five diverse ancestries and six age ranges. The proportion

160 of non-carriers affected with disease was also quantified for each variant to ascertain
161 disease prevalence among the population in the absence of a scorable variant.
162
163 Validation of phenotyping and penetrance approach
164 We performed a series of analyses to validate our population-based approach of
165 phenotyping and computing penetrance. First, we selected nine diseases from a variety
166 of systems—age-related macular degeneration (AMD), arrhythmogenic right ventricular
167 cardiomyopathy (ARVC), FaH, familial breast cancer (FBC), type 2 diabetes (T2D),
168 etc.—and identified cases using both ICD-10 diagnosis codes and previously published
169 clinical algorithms[29–37]. We then evaluated and compared the penetrance of 208 ClinVar
170 pathogenic variants corresponding to the diseases in Bio*Me* using both phenotype
171 approaches (**Supplementary Table 3**). Similar penetrance values were observed for all
172 diseases. Three diseases (Brugada syndrome, hepatocellular carcinoma, and
173 pulmonary arterial hypertension) had equivalent mean variant penetrance with both
174 approaches; mean variant penetrance of the other six diseases varied by 8% or less. A
175 tabulated list of these results is provided (**Supplementary Table 4**).
176 Second, we further assessed the accuracy of our phenotyping approach by
177 manually reviewing physician notes in the problem list (PL) for cases and controls of six
178 representative diseases: Alzheimer's disease (AD), AMD, FaH, FBC, stroke, and T2D.
179 We randomly sampled 50 ICD-10-based cases and controls for each disease and
180 examined their PL for symptoms or findings indicative of a diagnosis while blinded to
181 ICD-10 disease status (**Supplementary Table 5**). We observed high concordance in
182 case-control classification between ICD-10- and PL-based phenotyping, ranging from

183 100% (for AD) to 93% (AMD, FaH). Case concordance ranged from 88% (AMD, FaH)–
184 100% (AD, FBC, T2D), while control concordance ranged from 94% (T2D)–100% (AD,
185 stroke).
186
187 Biomarker levels linked to variant penetrance
188 We investigated metabolite and clinical measurements of carriers of impactful
189 variants with different penetrance in genes with clear biological links to three diseases:
190 FaH[12], maturity-onset diabetes of the young (MODY)[38], and obesity[39,40] (**Figure 1**).
191 Carriers of penetrant variants for FaH in *LDLR* had greater low-density lipoprotein
192 cholesterol (LDL-C) and total cholesterol levels than carriers of incompletely and
193 nonpenetrant variants (**Figure 1a**). Using linear regression, we observed robust
194 association between variant penetrance and LDL-C (effect size [β]=0.53 mg/dL per 1%
195 increase in penetrance, standard error [SE]=0.12; *P*=0.004) and total cholesterol levels
196 (β=0.98 mg/dL, SE=0.18; *P*=0.002) adjusting for clinical covariates and statin use.
197 Analogously, carriers of penetrant variants for MODY in hepatocyte nuclear factor-1α
198 (*HNF1A*) had elevated glucose and hemoglobin A1c (HbA1c) levels compared to
199 carriers of incompletely penetrant variants (**Figure 1b**). Variant penetrance was
200 significantly associated with glucose (β=0.76 mg/dL, SE=0.19; *P*=0.008) and HbA1c
201 levels (β=0.083% [2.4 mg/dL], SE=0.020 [0.57 mg/dL]; *P*=0.01) adjusted for clinical
202 covariates and diabetes medication use. Lastly, carriers of penetrant variants for obesity
203 in uncoupling protein 3 (*UCP3*) had higher body mass index (BMI) than carriers of
204 incompletely and nonpenetrant variants (**Figure 1c**). Variant penetrance was strongly
205 associated with BMI (β=0.24 kg/m$^2$, SE=0.044; *P*=6x10$^{-4}$) and weight (β=1.3 lbs [0.59

206 kg], SE=0.26 [0.12 kg]; $P$=8x10$^{-4}$) adjusting for clinical covariates. Together, these
207 analyses using published phenotype algorithms, physician notes, and biological
208 measurements support the validity of our population-based penetrance method and
209 illustrate the clinical presentation of impactful variants.
210
211 Distribution of observed variant penetrance and disease risk
212 We examined the distribution of disease risk conferred by 5,359 impactful
213 variants and 157 corresponding diseases. Risk difference (RD) is the difference
214 between the prevalence of disease in carriers and non-carriers, and represents the
215 excess risk of disease attributed to the variant of interest. A summary of mean variant
216 RD for all 157 diseases is provided (**Supplementary Table 6**). We observed 565 (11%)
217 variants with RD exceeding 0.05 for 55 (35%) diseases (**Figure 2**). In contrast, we
218 detected a very large number of weakly penetrant and nonpenetrant variants that confer
219 little to no disease risk (4,794 [89%] variants with RD≤0.05), which can be attributed in
220 part to ClinVar pathogenic variants from traditional studies that maximize penetrance
221 estimates and have biased ascertainment[18]. Recent studies have demonstrated
222 overestimation of the disease rates of these variants[6–8], indicating a need for
223 population-based approaches to more unbiasedly gauge penetrance.
224 Next, we investigated the prevalence of disease in carriers (i.e., penetrance) and
225 non-carriers of impactful variants according to ClinVar pathogenicity (pathogenic,
226 uncertain, conflicting, benign) (**Figure 3a**), ClinVar review status (expert reviewed,
227 multiple submitters, single submitter, no assertion criteria) (**Figure 3b**), and molecular
228 consequence (**Figure 3c**). We hypothesized that variants reported as pathogenic in

229 ClinVar, reviewed by experts, or with LoF consequence have greater *a priori* evidence
230 of disease risk and are expected to have higher penetrance than variants reported as
231 benign, without criteria assertion, or of synonymous consequence. In agreement with
232 these expectations, we observed that mean variant penetrance was highest among
233 pathogenic (6.9% vs. 0.86% next highest mean for uncertain class; $P=5x10^{-270}$), expert
234 reviewed (18% vs. 6.4% next highest mean for multiple submitters; $P=9x10^{-17}$), and
235 frameshift variants (10% vs. 4.1% for missense; $P=2x10^{-13}$). The mean penetrance of
236 ClinVar variants with conflicting (0.75%), uncertain (0.86%), and benign clinical
237 significance (0.085%) was similar, warranting caution when interpreting variants that do
238 not fully meet the ACMG's strict pathogenic criteria[41]. We also note that the mean
239 penetrance of ClinVar pathogenic variants was only 6.9% and only 18% for those with
240 the highest review status (expert reviewed), indicating that these broad classifications
241 do not adequately capture all variants with large effect on disease as measured by
242 penetrance.

243 We conducted several sensitivity analyses to account for differing sample sizes
244 of penetrance estimates. Penetrance distributions according to ClinVar pathogenicity,
245 review status, and molecular consequence remained similar when stratified by
246 increasing thresholds of sample sizes (**Supplementary Figure 3**). Singletons (very rare
247 variants appearing only once) comprised a large portion of the impactful variants
248 (3,507/5,359 [65%]) and statistically are completely penetrant or nonpenetrant. To test
249 the validity of singleton penetrance, we evaluated the proportion of penetrant singletons
250 by ClinVar pathogenicity, review status, and molecular consequence (**Supplementary**
251 **Figure 4**). If singleton penetrance were accurate, the proportion of penetrant singletons

252 among pathogenic, expert reviewed, or LoF singletons would be expected to exceed
253 that of benign, non-expert reviewed, or non-LoF singletons, respectively. We observed
254 that the proportion of penetrant singletons was in fact greatest among pathogenic (13%
255 vs. 2.6% in next highest class of conflicting; $P=2x10^{-19}$), expert reviewed (41% vs. 12%
256 in next highest group of multiple submitters; $P=4x10^{-11}$), and frameshift singletons (16%
257 vs. 7.8% for missense; $P=1x10^{-3}$). We therefore retained penetrance estimates of
258 smaller sample sizes in our analyses.
259
260 Ancestral and temporal dimensionality of penetrance data
261 Most penetrance studies to date have focused on one ancestry, typically
262 European[42,43], with few notable exceptions[11,23,44]. Yet it is well-known that allele
263 frequency (AF) and disease risk of variants can vary substantially in different
264 populations[45,46]. Thus, we computed penetrance among multiple self-reported
265 ancestries, including European, African, Hispanic, and Asian ancestries. To
266 demonstrate this diversity, we identified numerous ancestry-specific impactful variants.
267 Out of 2,209 impactful variants in Bio*Me*, we observed 75 (3.4%), 43 (1.9%), 17
268 (0.77%), and 71 (3.2%) variants exclusively present in African, Hispanic, Asian, and
269 European ancestry, respectively. Out of 3,408 impactful variants in UKB, we observed 8
270 (0.23%), 11 (0.32%), and 235 (6.9%) variants specific to African, Asian, and European
271 ancestry, respectively. Furthermore, we identified ancestry-specific variants that were
272 highly penetrant, such as an Asian ancestry-specific pathogenic frameshift variant in
273 *HBB* (NC_000011.10:c.5226994_5226995insC) associated with a greatly increased risk
274 of thalassemia (RD=0.99; $P=9x10^{-6}$) and a European ancestry-specific pathogenic

275 frameshift variant in *PALB2* (NC_000016.9:c.23647358_23647359del) associated with
276 a significantly elevated risk of breast cancer (RD=0.92; *P*=0.007).
277 We also delineated variant penetrance based on different age thresholds of
278 carriers ranging from at least 20 years (≥20) to at least 70 years (≥70). Age of disease
279 onset is pertinent for estimating penetrance: congenital or early onset diseases will have
280 manifested in older carriers of penetrant variants, whereas later onset diseases may not
281 have presented yet in younger carriers. To address temporality, we characterized the
282 observed change in variant penetrance with increasing carrier age for 157 diseases
283 corresponding to the impactful variants, stratified by age of disease onset (**Figure 4**;
284 **Supplementary Figure 5**). We first annotated and grouped the diseases by age of
285 onset: Earlier (congenital, childhood, or adolescent), Later (adulthood), or Any
286 (**Supplementary Table 7**). For each disease group, we then calculated the change in
287 variant penetrance (ΔPenetrance) between the lowest carrier age threshold and
288 increasing age thresholds (e.g., variant penetrance for ≥20 versus ≥30, ≥20 versus ≥40,
289 etc.). As expected, the mean ΔPenetrance increased in the Later disease group with
290 higher carrier age thresholds in Bio*Me* (**Figure 4a**). The largest mean ΔPenetrance was
291 +1.8% in the Later disease group when comparing ≥20 and ≥70, while mean
292 ΔPenetrance remained ~0% in the Earlier and Any disease groups. In UKB, no
293 significant differences in mean ΔPenetrance were observed in the Later disease group
294 (**Figure 4b**) likely due to its older age range of carriers (40-69 years) whereby Later
295 diseases would have manifested. Thus, penetrance estimates of diseases with early or
296 any age of onset remained stable over time, while penetrance estimates of diseases
297 with later age of onset predictably increased with higher carrier ages.

298

299 Clinical utility of variant penetrance data

300 A genotype-first approach in genomic medicine must identify penetrant variation
301 in clinically significant genes to inform screening, management, and treatment. The
302 ACMG recommends reporting secondary findings of pathogenic variants in the
303 ACMG59, but acknowledges that insufficient data on penetrance requires ongoing study
304 and revision[49]. *BRCA1* and *BRCA2* are both in the ACMG59, and a recent study
305 estimated the population-based penetrance of pathogenic variants in these genes for
306 FBC using data from the UKB[50]. However, penetrance was estimated at the gene level,
307 and non-ClinVar variants were omitted as were variants in other genes that have been
308 shown to confer significant risk of breast cancer, such as *PALB2*, *CHEK2*, *ATM*, *PTEN*,
309 and others[24]. Thus, we investigated the penetrance of impactful variants in 10 known or
310 suspected breast cancer-predisposition genes (*BRCA1*, *BRCA2*, *PALB2*, *CHEK2*, *ATM*,
311 *PTEN*, *CDH1*, *BARD1*, *BRIP1*, and *RAD51D*) to illustrate the clinical value of a
312 population-based approach of determining penetrance at the variant level (**Figure 5a**).
313 The highest disease risk on average was conferred by impactful variants in *BRCA1*
314 (mean variant penetrance=38%, mean RD=0.32; $P=2x10^{-6}$), *BRCA2* (mean variant
315 penetrance=38%, mean RD=0.32; $P=1x10^{-10}$), and *PALB2* (mean variant
316 penetrance=26%, mean RD=0.20; $P=0.009$).

317 These data clarify the disease risk of previously reported variants in ClinVar and
318 novel variants not yet reported in ClinVar or the literature. We highlighted examples of
319 variants that are significantly associated with risk of seven diseases, including FBC, to
320 demonstrate this dual utility (**Supplementary Table 8**). Many impactful variants in

321 *BRCA2* were strongly associated with FBC, such as a known pathogenic frameshift
322 variant (NC_000013.11:c.32340301del; penetrance=47%, RD=0.39; $P$=9x10$^{-6}$) and a
323 previously unreported frameshift variant (NC_000013.11:c.32340630_32340631del;
324 penetrance=100%, RD=0.96; $P$=4x10$^{-13}$). Evidence of high penetrance strengthens a
325 candidate variant's case for clinical significance, while assertions of pathogenicity are
326 refined by quantitative disease risk estimates.
327 Notably, we observed substantial heterogeneity in the penetrance of variants
328 even within the same disease-predisposition gene. Impactful variants in *LDLR*, for
329 instance, exhibited a wide range of penetrance for FaH (**Figure 5b**). Similarly, there was
330 a large distribution of penetrance associated with impactful variants in *BRCA1* (variant
331 penetrance standard deviation [SD]=45%, RD SD=0.44), *BRCA2* (variant penetrance
332 SD=46%, RD SD=0.46), and *PALB2* (variant penetrance SD=42%, RD SD=0.42) for
333 FBC (**Figure 6a**). Instead of coarsely categorizing variants as simply pathogenic versus
334 non-pathogenic and collapsing all variants in a gene, these data reveal granularity and
335 nuance in the penetrance of individual variants.
336

337 **DISCUSSION**

338 A major goal of genomic medicine is to tailor clinical care of patients to their
339 unique genetic composition, especially penetrant variants. Scalability and accuracy are
340 crucial for determining penetrance and employing its information in a clinical setting.
341 Conventional penetrance estimates from family-based or clinical cohort studies typically
342 focus on one variant or gene at a time[14–17] (limited scalability), and have small sample
343 size, ascertainment bias, inconsistent carrier screening, and genetic/environmental
344 confounders[18,19] (limited accuracy). A proliferation of population biobanks, such as UKB
345 and Bio*Me*, has made available thousands of genetic and phenotypic records[20,21,23].
346 This raises the possibility of using a population-based method to measure penetrance
347 whereby large numbers of unrelated carriers are assessed[22]. Here, we performed a
348 comprehensive assessment of variant penetrance using 72,434 exomes from two large-
349 scale population-based biobanks, with the dataset freely accessible and provided in
350 **Supplementary Tables 9** and **10**.

351 Reliable penetrance estimates depend on reliable phenotyping of carriers. We
352 performed robust validation analyses to support our phenotyping and penetrance
353 strategy. First, we computed penetrance for nine diseases using both diagnosis codes
354 and previously published clinical algorithms, finding high concordance in all diseases
355 tested. Second, we manually curated physician notes for six diseases to verify ICD-10-
356 based diagnoses. Third, we extracted laboratory and clinical test results (e.g., levels of
357 lipids, glucose, HbA1c, and BMI) for carriers and observed that penetrance was strongly
358 associated with relevant quantitative traits, adjusted for medications and clinical
359 covariates. Hence, by validating our phenotyping and penetrance method, and

360 evaluating penetrance for thousands of variants in a high-throughput manner, we
361 ensure both accuracy and scalability. As penetrance studies grow, similar validation
362 analyses should be implemented to certify fidelity of penetrance estimates.

363 While several efforts have begun to probe the upward bias of penetrance
364 estimates in traditional studies, there has not yet been a systematic investigation of the
365 pervasiveness of overestimated penetrance. In the present study, we examined the
366 distribution of disease risk associated with a high-quality set of 5,359 impactful variants
367 and 157 diseases. Persistently lower disease risk was observed for many diseases, with
368 a few exceptions of highly penetrant variants such as in *BRCA1* and *BRCA2*. This can
369 be interpreted in light of a few important considerations: 1) small sample size and
370 stochasticity for rarer variants may contribute to variable penetrance estimates;
371 however, a number of analyses supported the accuracy of our penetrance
372 measurements with respect to sample size, including for singletons (**Supplementary**
373 **Figure 3**, **Supplementary Figure 4**); 2) ascertainment of cases in conventional studies
374 has been shown to inflate the disease risk of variants[18,19] and recent population-based
375 studies with lower penetrance estimates for FaH and developmental disorders bolster
376 this explanation[11,22]; 3) reported pathogenicity does not equate with penetrance, as we
377 found by interrogating whether existing qualitative classification systems in ClinVar
378 capture the quantitative disease risk of variants. ClinVar pathogenic variants were more
379 penetrant than variants of other ClinVar classes yet still weakly penetrant overall, in line
380 with previous studies[6–8] and consistent with ClinVar's definition of pathogenic that
381 includes "low penetrance" variants (https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/).
382 We then examined a major driver of variation in penetrance among ClinVar pathogenic

383 variants: when stratified by ClinVar review status (evidence for clinical significance of a
384 variant), pathogenic variants that were not expert reviewed had lower penetrance.
385 In light of these findings, it is incumbent to discuss whether variants reported as
386 pathogenic but empirically shown to have low penetrance should be classified
387 differently, or whether categorical systems of disease risk (pathogenic versus non-
388 pathogenic) should be complemented with a quantitative system based on penetrance[7].
389 Recent commentary by Khera and Hegele proposed a new paradigm for classifying FaH
390 based on two parameters: carrier status for a pathogenic *LDLR* variant and severity of
391 hypercholesterolemia[51]. This advanced schema for disease classification, and others
392 like it, would greatly benefit from knowledge of penetrance for pathogenic *LDLR* variants
393 to better stratify disease risk and personalize medical care.
394 We also emphasize the importance of including diverse ancestries and age
395 ranges when characterizing variant penetrance. Populations differ by AF and disease
396 factors[45,46], yet most genetic studies have focused on Europeans[52,53]. Here, we capture
397 ancestry-specific penetrance in detail, identifying over a hundred variants in non-
398 European ancestries. While past studies have reported age-dependent penetrance in
399 age-related diseases such as FBC[17,54], amyotrophic lateral sclerosis[55,56], and obesity[57],
400 these typically estimate gene-based penetrance whereby all variation in a gene is
401 aggregated. In contrast, we evaluate age-dependent penetrance at the variant level. As
402 expected, early onset diseases showed stable penetrance over time while later onset
403 diseases had increasing penetrance with older carriers. For example, the observed
404 penetrance of an expert reviewed ClinVar pathogenic frameshift variant in *BRCA2*

405 (NC_000013.11:c.32340301del) increased from 38% in European ancestry carriers ≥40
406 years of age to 56% in those ≥60 years of age.

407 There were several study limitations. First, ICD-10 diagnosis codes from the EHR
408 were used to define case-control status[26,27] (**Supplementary Table 1**). While commonly
409 used in EHR-linked biobank studies[61,62], there may be some misclassification[63–66].
410 Though several validation analyses reinforced the soundness of our phenotyping, these
411 were completed for a subset of the diseases. Second, we cannot exclude the possibility
412 of potential bias in our datasets. Bio*Me* is predominantly composed of individuals
413 recruited from the Mount Sinai Health System and may have a higher burden of
414 diseases and therefore penetrance estimates. In contrast, the preponderance of healthy
415 volunteers in UKB may lead to conservative estimates of penetrance. Third, though our
416 sample size exceeded 72,000 exomes and enabled us to ascertain rare variants, many
417 penetrance estimates for rare variants (e.g., singletons) are based on low numbers of
418 carriers and may produce variable estimates. We included raw counts of carriers and
419 non-carriers for each variant in the dataset so that estimates may be interpreted
420 accordingly. Fourth, we mapped variants to diseases based on disease genes reported
421 in ClinVar from pathogenic variant submissions. Inaccuracies in ClinVar submissions,
422 and therefore our mapping from variant to disease, are possible, though we manually
423 checked mappings extensively against the literature for accuracy.

424 In conclusion, we present a large-scale, systematic investigation of variant
425 penetrance that leveraged thousands of exomes linked to the EHR. While we made
426 available the penetrance measurements of all 37,772 variants for full transparency, we
427 utilized a stringent set of 5,359 impactful variants that were ClinVar pathogenic or LoF in

428 a non-recessive gene for many analyses. We demarcated differences in disease risk
429 among variants of distinct ClinVar pathogenicity classes, review status, and molecular
430 consequences. We also accessed a rich resource of clinical phenotypes to thoroughly
431 explore penetrance: detailed physician notes, medications, biological measurements,
432 and laboratory results. Critically, we investigated multiple dimensions of penetrance
433 data spanning five ancestries and six decades of age thresholds. This study provides a
434 blueprint for future studies to efficiently and accurately determine variant penetrance,
435 the results of which will greatly improve our understanding of the genetic underpinnings
436 of human disease.
437
438
439
440
441
442
443
444
445
446
447
448
449
450

451 **METHODS**

452 Study populations and sample filtering

453 A flowchart of the study design is provided (**Supplementary Figure 1**). We
454 evaluated penetrance from individuals in two large-scale electronic health record (EHR)-
455 linked population-based biobanks: The Bio*Me* Biobank (Bio*Me*) and UK Biobank (UKB).
456 The study protocols were approved by the Institutional Review Board (IRB) of the Icahn
457 School of Medicine at Mount Sinai. Use of data from UKB was completed and approved
458 using the UK Biobank Resource under Application Number 16218. Informed consent
459 was obtained for all study participants in both Bio*Me* and UKB through the approved
460 IRB protocols. An overview of the demographics and clinical traits for both study
461 populations is provided (**Table 1**). Bio*Me* is an EHR-linked biobank for ~50,000 patients
462 of African, Hispanic, European, and Other (Asian, Native American, and miscellaneous)
463 self-reported ancestry who are recruited from the Mount Sinai Health System in
464 Manhattan, NYC from 2007 onwards. All Bio*Me* participants consented to providing
465 biological and DNA samples linked to de-identified EHRs. A subset of the individuals
466 (n=31,250) was exome sequenced before undergoing quality control. A total of 229
467 samples with discordance between genetic sex and sex listed in the manifest, low
468 coverage, contamination, low call rate, or duplications were excluded, leaving 30,813
469 samples. In addition, samples lacking complete demographic data (n=345), younger
470 than 20 years of age (n=610), or without International Classification of Diseases-Clinical
471 Modification 10 (ICD-10) diagnosis data (n=819) were removed to generate the final
472 study set of 29,039 samples.

473 The UKB is a population-based longitudinal cohort of ~500,000 individuals chiefly
474 of British self-reported ancestry between 40-69 years of age who were enrolled at
475 various sites across the United Kingdom between 2006-2010[21,67]. All individuals
476 consented to providing medical history, demographic data, and DNA samples. A subset
477 of 49,960 individuals had their exome sequenced and passed standard quality control
478 (QC) filters, described extensively elsewhere[68]. We further excluded samples that
479 lacked complete demographic information (n=2), or did not have ICD-10 diagnosis
480 codes available (n=6,563), leaving a final study set of 43,395 samples for analysis.
481
482 Whole-exome sequencing and quality control
483 In Bio*Me*, variant call files (VCFs) produced by Illumina v4 HiSeq 2500 contained
484 9,202,884 variants that were called in the samples. Goldilocks Filter (GF) was
485 implemented on the VCFs. For single nucleotide polymorphisms (SNPs), cells with
486 depth-normalized quality scores <3 or depth of coverage <7 were set to missing. For
487 insertions and deletions (indels), cells with depth-normalized quality scores <5 or depth
488 of coverage <10 were set to missing. Variant sites were then filtered, whereby sites of
489 heterozygous variation failed the Allele Balance (AB) cutoff and were removed. SNP
490 sites required ≥1 sample to carry an alternate AB ≥15% and indel sites required ≥1
491 sample to carry an alternate AB ≥20%. Together, these site filters removed 441,406
492 sites, leaving 8,761,478 variants after GF. Next, sites with missing genotypes for >2% of
493 individuals in the dataset (267,955 sites) were removed. AB was calculated for biallelic
494 SNPs and 320,877 sites with AB <0.3 or >0.8 were removed, leaving 8,172,646 sites.
495 Lastly, the dataset was filtered to regions within the target regions of the exome capture

496 platform (4,256,827 sites) and separated into 2 file sets for biallelic and multiallelic sites
497 (3,948,623 and 308,204, respectively) due to differences in QC procedures.
498 In UKB, exome data from the first tranche of exome sequence data generated
499 with the Functional Equivalence pipeline[68] was used. Sequence data and QC for UKB
500 are described elsewhere[21,67].
501
502 Variant curation and annotation
503 Using exome data from the 72,434 study samples, 37,772 variants associated
504 with ClinVar[1] diseases were ascertained using PLINK version 2.0[69]. To enrich for
505 penetrant variants, we selected a strict subset of 5,359 clinically impactful (impactful)
506 variants for most downstream analyses. These were defined by curating variant
507 summary information in ClinVar VCF files released in December 2019, functional
508 annotations from Variant Effect Predictor (VEP)[28] version 99.2, and genic mode of
509 inheritance from Online Mendelian Inheritance in Man (OMIM)[70]. An overview of variant
510 selection is provided in **Supplementary Figure 2**. First, we included variants of
511 pathogenic and/or likely pathogenic classification in ClinVar, and previously unreported
512 variants with a damaging molecular consequence (splice acceptor/donor, stop
513 gained/lost, frameshift, or start lost; collectively defined as LoF) annotated by VEP. LoF
514 variants in a gene were mapped to disease based on prior pathogenic variant
515 submissions in ClinVar linking genes to diseases (e.g., *BRCA1* LoF variants were
516 mapped to breast cancer based on prior pathogenic variant submissions in ClinVar
517 linking *BRCA1* to breast cancer). This filtered out benign variants or variants with
518 uncertain or conflicting clinical significance in ClinVar and variants with synonymous

519 molecular consequence. As missense variants have varying and uncertain degrees of
520 pathogenicity, non-ClinVar missense variants were also excluded. We also noted the
521 review status for each ClinVar variant, which summarizes the level of evidence
522 supporting a variant's claim of clinical significance ranging from the lowest level of no
523 assertion criteria (review status=0), to single submitter or multiple submitters with
524 conflicting interpretation (review status=1), to multiple submitters with no conflict of
525 interpretation (review status=2), to reviewed by an expert panel such as ClinGen[71]
526 (review status=3), to the highest level of practice clinical guidelines (review status=4).
527 Second, we excluded variants in genes with exclusively recessive mode of inheritance
528 reported in OMIM. The gene for each variant was retrieved from NCBI reference
529 sequences (RefSeq[72]) and corroborated with genomic coordinates in OMIM. The mode
530 of inheritance was mined for all phenotypes reported for each gene in OMIM and genes
531 were then summarized as Dominant (only dominantly inherited phenotypes), Both (both
532 dominantly and recessively inherited phenotypes), or Recessive (only recessively
533 inherited phenotypes).
534
535 Phenotyping of carriers
536 In both UKB and Bio*Me*, we obtained case status using ICD-10 diagnosis codes.
537 ICD-10 codes are commonly used in genetic population studies to define cases of a
538 disease[61,62] and they map directly to ClinVar diseases in the Systematized
539 Nomenclature of Medicine Clinical Terms (SNOMED CT)[27]. All of the 29,039 Bio*Me*
540 samples and 43,395 UKB samples in the final dataset had ICD-10 diagnosis codes
541 available. Cases were identified by the presence of a corresponding ICD-10 code while

542 controls were identified by the absence of all corresponding ICD-10 codes. We thereby
543 defined case-control status for over 400 SNOMED CT diseases of non-recessive
544 inheritance in ClinVar (disease inheritance was retrieved from NCBI's MedGen
545 database at https://www.ncbi.nlm.nih.gov/medgen/); 197 of the diseases were present
546 with at least one case in the final dataset. A complete list of the cases, controls, and
547 ICD-10 diagnosis codes for each disease is provided in **Supplementary Table 1**.
548 Validation of the ICD-10-based phenotyping method was performed in a set of
549 three analyses in Bio*Me*. First, penetrance estimates using the ICD-10 phenotypes were
550 compared against results using clinical algorithms from the literature for nine
551 diseases[29–37]: age-related macular degeneration (AMD), arrhythmogenic right
552 ventricular cardiomyopathy (ARVC), Brugada syndrome (BrS), familial
553 hypercholesterolemia (FaH), familial breast cancer (FBC), hepatocellular carcinoma
554 (HCC), idiopathic pulmonary arterial hypertension (IPAH), prostate cancer (PCa), and
555 type 2 diabetes (T2D) (**Supplementary Table 3**, **Supplementary Table 4**). Second, we
556 verified ICD-10 phenotypes by manually reviewing physician notes in the problems list
557 (PL) for six diseases: Alzheimer's disease (AD), AMD, FaH, FBC, ischemic stroke, and
558 T2D (**Supplementary Table 5**). For each disease, we extracted physician notes for a
559 random sample of 50 ICD-10-based cases and 50 ICD-10-based controls. The
560 presence of symptoms or findings indicative of a diagnosis corroborated ICD-10-based
561 cases, while the absence of symptoms or findings corroborated ICD-10-based controls.
562 Third and lastly, we examined the association of biomarker levels with penetrance
563 estimates for variants in genes with clear biological roles in three diseases: *LDLR*[12] in
564 FaH (measured by LDL-C and total cholesterol), *HNF1A*[38] in MODY (measured by

565 glucose and hemoglobin A1c), and *UCP3*[39,40] in obesity (measured by BMI and weight).
566 We extracted median biological measurements for carriers and medication usage
567 relevant to FaH (statins) and T2D (insulin, insulin analogs, pramlintide, glucagon-like
568 peptide 1 agonists, metformin, sulfonylurea, dipeptidyl peptidase 4 inhibitors, glitazone,
569 sodium-glucose transport protein 2 inhibitors, or alpha-glucosidase inhibitors) to control
570 for the effect of medications on biological measurements in association analyses.
571
572 Statistical analysis
573 All statistical tests and plots were made using R statistical software version
574 3.5.3[75]. Differences in categorical variables were assessed with a Fischer's exact test,
575 while differences in continuous variables were tested with a two-sided t-test.
576 Significance level was set at $P<0.05$ for comparisons between two groups. Risk
577 difference (RD) between the prevalence of disease in carriers and non-carriers was
578 computed, and significance was evaluated with Fisher's exact test. A strict Bonferroni
579 correction was applied to *P*-values in analyses involving multiple comparisons, including
580 comparisons of mean penetrance for five ClinVar pathogenicity classes, mean
581 penetrance for four ClinVar review status levels, and mean penetrance for eight
582 molecular consequences. Phenotype validation using PL physician notes was
583 performed by computing the ratio of PL-based cases to ICD-10-based cases and PL-
584 based controls to ICD-10-based controls to determine case and control concordance,
585 respectively. Association analyses of variant penetrance with biomarker levels were
586 performed using a multivariable linear regression, adjusted for clinical covariates of age,

587 sex, BMI (except in the analysis of obesity), 10 genetic principal components (PCs), and
588 medication usage where relevant.
589
590 Data availability
591 The UKB data may be browsed at http://biobank.ndph.ox.ac.uk/showcase/ and
592 access to data can be requested at https://www.ukbiobank.ac.uk/register-apply/. More
593 information about Bio*Me* can be found at
594 https://icahn.mssm.edu/research/ipm/programs/biome-biobank/researcher-faqs. The
595 complete penetrance dataset used for all analyses is provided (**Supplementary Tables**
596 **9** and **10**) with no restrictions on the data released.
597
598
599
600
601
602
603
604
605
606
607
608

609 **References**


610 1. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence
611 variation and human phenotype. doi:10.1093/nar/gkt1113.
612 2. Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine–Progress,
613 Pitfalls, and Promise. *Cell* vol. 177 45–57 (2019).
614 3. Yang, Y. *et al.* Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian
615 Disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
616 4. Watson, M. S. ACMG policy statement: Updated recommendations regarding
617 analysis and reporting of secondary findings in clinical genome-scale sequencing.
618 *Genetics in Medicine* vol. 17 68–69 (2015).
619 5. Landrum, M. J. & Kattman, B. L. ClinVar at five years: Delivering on the promise.
620 *Hum. Mutat.* **39**, 1623–1630 (2018).
621 6. Shah, N. *et al.* Identification of Misclassified ClinVar Variants via Disease
622 Population Prevalence. *Am. J. Hum. Genet.* **102**, 609–619 (2018).
623 7. Manrai, A. K., Ioannidis, J. P. A. & Kohane, I. S. Clinical genomics: From
624 pathogenicity claims to quantitative risk estimates. *JAMA - Journal of the*
625 *American Medical Association* vol. 315 1233–1234 (2016).
626 8. Xiang, J. *et al.* Reinterpretation of common pathogenic variants in ClinVar
627 revealed a high proportion of downgrades. *Sci. Rep.* **10**, 1–5 (2020).
628 9. Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C. & Gelbart, W. M. *An*
629 *Introduction to Genetic Analysis*. (W. H. Freeman, 2000).
630 10. Mavaddat, N. *et al.* Cancer Risks for BRCA1 and BRCA2 Mutation Carriers:
631 Results From Prospective Analysis of EMBRACE. **105**, 812–822 (2013).
632 11. Khera, A. V. *et al.* Diagnostic Yield and Clinical Utility of Sequencing Familial
633 Hypercholesterolemia Genes in Patients With Severe Hypercholesterolemia. *J.*
634 *Am. Coll. Cardiol.* **67**, 2578–2589 (2016).
635 12. Goldberg, A. C. *et al.* Familial Hypercholesterolemia: Screening, diagnosis and
636 management of pediatric and adult patients. *J. Clin. Lipidol.* **5**, S1–S8 (2011).
637 13. Saslow, D. *et al.* American Cancer Society Guidelines for Breast Screening with
638 MRI as an Adjunct to Mammography. *CA. Cancer J. Clin.* **57**, 75–89 (2007).
639 14. Hobbs, H. H. *et al.* Evidence for a dominant gene that suppresses
640 hypercholesterolemia in a family with defective low density lipoprotein receptors.
641 *J. Clin. Invest.* **84**, 656–664 (1989).
642 15. Wohlgemuth, M. *et al.* A family-based study into penetrance in
643 facioscapulohumeral muscular dystrophy type 1. *Neurology* **91**, e444–e454
644 (2018).
645 16. Gong, G., Hannon, N. & Whittemore, A. S. Estimating gene penetrance from
646 family data. *Genet. Epidemiol.* **34**, 373–381 (2010).
647 17. Kuchenbaecker, K. B. *et al.* Risks of breast, ovarian, and contralateral breast
648 cancer for BRCA1 and BRCA2 mutation carriers. *JAMA - J. Am. Med. Assoc.* **317**,
649 2402–2416 (2017).
650 18. Park, S. *et al.* Adjusting heterogeneous ascertainment bias for genetic association
651 analysis with extended families. *BMC Med. Genet.* **16**, 62 (2015).
652 19. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-
653 Sawatzki, H. Where genotype is not predictive of phenotype: Towards an

654 understanding of the molecular basis of reduced penetrance in human inherited
655 disease. *Hum. Genet.* **132**, 1077–1130 (2013).
656 20. De Souza, Y. G. & Greenspan, J. S. Biobanking past, present and future:
657 Responsibilities and benefits. *AIDS* vol. 27 303–312 (2013).
658 21. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the
659 Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS*
660 *Med.* **12**, e1001779 (2015).
661 22. Wright, C. F. *et al.* Assessing the Pathogenicity, Penetrance, and Expressivity of
662 Putative Disease-Causing Variants in a Population Setting. *Am. J. Hum. Genet.*
663 **104**, 275–286 (2019).
664 23. Abul-Husn, N. S. *et al.* Exome sequencing reveals a high prevalence of BRCA1
665 and BRCA2 founder variants in a diverse population-based biobank. *Genome*
666 *Med.* **12**, 2 (2019).
667 24. Hu, C. *et al.* A Population-Based Study of Genes Previously Implicated in Breast
668 Cancer. *N. Engl. J. Med.* **384**, 440–451 (2021).
669 25. Fahed, A. C. *et al.* Polygenic background modifies penetrance of monogenic
670 variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 1–9 (2020).
671 26. SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/index.html.
672 27. SNOMED CT to ICD-10-CM Map. *National Library of Medicine*
673 https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.
674 html.
675 28. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122
676 (2016).
677 29. Ning Shang, George Hripcsak, Chunhua Weng, Wendy K. Chung, K. C. Breast
678 Cancer | PheKB. *PheKB* https://phekb.org/phenotype/breast-cancer (2018).
679 30. Safarova, M. S., Liu, H. & Kullo, I. J. Rapid identification of familial
680 hypercholesterolemia from electronic health records: The SEARCH study. *J. Clin.*
681 *Lipidol.* **10**, 1230–1239 (2016).
682 31. Papani, R. *et al.* Validation of claims-based algorithms for pulmonary arterial
683 hypertension: *https://doi.org/10.1177/2045894018759246* (2018)
684 doi:10.1177/2045894018759246.
685 32. Ye, J. Z., Delmar, M., Lundby, A. & Olesen, M. S. Reevaluation of genetic variants
686 previously associated with arrhythmogenic right ventricular cardiomyopathy
687 integrating population-based cohorts and proteomics data. *Clin. Genet.* **96**, 506–
688 514 (2019).
689 33. Goldberg, D. S., Lewis, J. D., Halpern, S. D., Weiner, M. G. & Re, V. Lo.
690 Validation of a coding algorithm to identify patients with hepatocellular carcinoma
691 in an administrative database. *Pharmacoepidemiol. Drug Saf.* **22**, 103–107
692 (2013).
693 34. Wolf, A. M. D. *et al.* American Cancer Society Guideline for the Early Detection of
694 Prostate Cancer: Update 2010. *CA. Cancer J. Clin.* **60**, 70–98 (2010).
695 35. Ritchie, M. D. *et al.* Genome- and phenome-wide analyses of cardiac conduction
696 identifies markers of arrhythmia risk. *Circulation* **127**, 1377–1385 (2013).
697 36. Simonett, J. M. *et al.* A validated phenotyping algorithm for genetic association
698 studies in age-related macular degeneration. *Sci. Rep.* **5**, 1–8 (2015).
699 37. Kho, A. N. *et al.* Use of diverse electronic medical record systems to identify

genetic risk for type 2 diabetes within a genome-wide association study. *J. Am. Med. Informatics Assoc.* **19**, 212–218 (2012).
38. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-1α gene in maturity-onset diabetes of the young (MODY3). *Nature* **384**, 455–458 (1996).
39. Vidal-Puig, A., Solanes, G., Grujic, D., Flier, J. S. & Lowell, B. B. UCP3: An uncoupling protein homologue expressed preferentially and abundantly in skeletal muscle and brown adipose tissue. *Biochem. Biophys. Res. Commun.* **235**, 79–82 (1997).
40. Musa, C. V. *et al.* Four novel UCP3 gene variants associated with childhood obesity: Effect on fatty acid oxidation and on prevention of triglyceride storage. *Int. J. Obes.* **36**, 207–217 (2012).
41. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
42. Nugent, A. *et al.* Reporting of race in genome and exome sequencing studies of cancer: a scoping review of the literature. *Genet. Med.* **21**, 2676–2680 (2019).
43. Fouchier, S. W., Defesche, J. C., Umans-Eckenhausen, M. A. & Kastelein, J. J. The molecular basis of familial hypercholesterolemia in the Netherlands. *Hum. Genet.* **109**, 602–615 (2001).
44. Rader, D. J. *et al.* Association of the V122I Hereditary Transthyretin Amyloidosis Genetic Variant with Heart Failure among Individuals of African or Hispanic/Latino Ancestry. *JAMA - J. Am. Med. Assoc.* **322**, 2191–2202 (2019).
45. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37**, 161–165 (2005).
46. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).
47. Bonnycastle, L. L. *et al.* Autosomal dominant diabetes arising from a wolfram syndrome 1 mutation. *Diabetes* **62**, 3943–3950 (2013).
48. Yang, Y. & Chan, L. Monogenic diabetes: What it teaches us on the common forms of type 1 and type 2 diabetes. *Endocr. Rev.* **37**, 190–222 (2016).
49. Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574 (2013).
50. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
51. Khera, A. V. & Hegele, R. A. What Is Familial Hypercholesterolemia, and Why Does It Matter? *Circulation* **141**, 1760–1763 (2020).
52. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
53. Genetics for all. *Nat. Genet.* **51**, 579 (2019).
54. Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. *Am. J. Hum. Genet.* **72**, 1117–1130 (2003).
55. De Belleroche, J., Orrell, R. & King, A. Familial amyotrophic lateral sclerosis/motor neurone disease (FALS): A review of current developments. *J. Med. Genet.* **32**, 841–847 (1995).

746 56. Murphy, N. A. *et al.* Age-related penetrance of the C9orf72 repeat expansion. *Sci.*
747 *Rep.* **7**, (2017).
748 57. Stutzmann, F. *et al.* Prevalence of melanocortin-4 receptor deficiency in
749 europeans and their age-dependent penetrance in multigenerational pedigrees.
750 *Diabetes* **57**, 2511–2518 (2008).
751 58. Kobayashi, Y. *et al.* Pathogenic variant burden in the ExAC database: An
752 empirical approach to evaluating population data for clinical variant interpretation.
753 *Genome Med.* **9**, 13 (2017).
754 59. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**,
755 179–189 (2020).
756 60. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency
757 genetic variants in common disease. *Genome Biol.* **18**, 1–17 (2017).
758 61. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and
759 putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 1–14
760 (2018).
761 62. Choquet, H. *et al.* A multiethnic genome-wide association study of primary open-
762 angle glaucoma identifies novel risk loci. *Nat. Commun.* **9**, 1–14 (2018).
763 63. Bradley, W. G. Inaccuracy of the ICD-9-CM in identifying the diagnosis of
764 ischemic cerebrovascular disease. *Neurology* **51**, (1998).
765 64. Öhman, L., Johansson, M., Jansson, J. H., Lind, M. & Johansson, L. Positive
766 predictive value and misclassification of diagnosis of pulmonary embolism and
767 deep vein thrombosis in swedish patient registries. *Clin. Epidemiol.* **10**, 1215–
768 1221 (2018).
769 65. Doktorchik, C. *et al.* Validation of a case definition for depression in administrative
770 data against primary chart data as a reference standard 11 Medical and Health
771 Sciences 1117 Public Health and Health Services. *BMC Psychiatry* **19**, 1–8
772 (2019).
773 66. Cozzolino, F. *et al.* A diagnostic accuracy study validating cardiovascular ICD-9-
774 CM codes in healthcare administrative databases. The Umbria Data-Value
775 Project. *PLoS One* **14**, e0218919 (2019).
776 67. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic
777 data. *Nature* **562**, 203–209 (2018).
778 68. Hout, C. V. Van *et al.* Whole exome sequencing and characterization of coding
779 variation in 49,960 individuals in the UK Biobank. *bioRxiv* 572347 (2019)
780 doi:10.1101/572347.
781 69. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger
782 and richer datasets. *Gigascience* **4**, 7 (2015).
783 70. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A.
784 OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of
785 human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2014).
786 71. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *N. Engl. J. Med.*
787 **372**, 2235–2242 (2015).
788 72. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq):
789 A curated non-redundant sequence database of genomes, transcripts and
790 proteins. *Nucleic Acids Res.* **35**, D61 (2007).
791 73. Ferrer, J. *et al.* Mitochondrial glycerol-3-phosphate dehydrogenase: Cloning of an

alternatively spliced human islet-cell cDNA, tissue distribution, physical mapping, and identification of a polymorphic genetic marker. *Diabetes* **45**, 262–266 (1996).

74. Meyre, D. *et al.* Variants of ENPP1 are associated with childhood and adult obesity and increase the risk of glucose intolerance and type 2 diabetes. *Nat. Genet.* **37**, 863–867 (2005).

75. R Core Team. *R: a language and environment for statistical computing*. (R Foundation for Statistical Computing, 2018).

819 **Author Contributions**

820 IF and RD designed the study. IF, RL, JC, and RD obtained the data. IF, KC, HV, SB,
821 DJ, and GR analyzed and/or interpreted the data. IF and RD drafted and revised the
822 manuscript. All authors read and provided critical comments on the manuscript.

823

824 **Acknowledgements**

825 The Bio*Me* healthcare delivery cohort at Mount Sinai was founded and maintained with
826 a generous gift from the Andrea and Charles Bronfman Philanthropies. We thank the
827 individuals who were involved in the quality control and/or file handling for the exome
828 sequencing and genome-wide genotyping data, including Aayushee Jain, Lisheng Zhou,
829 Michael Preuss, Quingbin Song, Stephane Wenric, and Steve Ellis. We also thank the
830 thesis advisory committee of IF, including Bruce D. Gelb, Sander Houten, Paz Polak,
831 and Stuart Scott, for their critical feedback and expertise. Research reported in this
832 paper was supported by the Office of Research Infrastructure of the National Institutes
833 of Health under award numbers S10OD018522 and S10OD026880. The content is
834 solely the responsibility of the authors and does not necessarily represent the official
835 views of the National Institutes of Health.

836

837 **Funding**

838 IF is supported by T32GM007280 the Medical Scientist Training Program Training
839 Grant from the National Institute of General Medical Sciences of the National Institutes
840 of Health. RD is supported by R35GM124836 from the National Institute of General
841 Medical Sciences of the National Institutes of Health, and R01HL139865 from the

National Heart, Lung, and Blood Institute of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

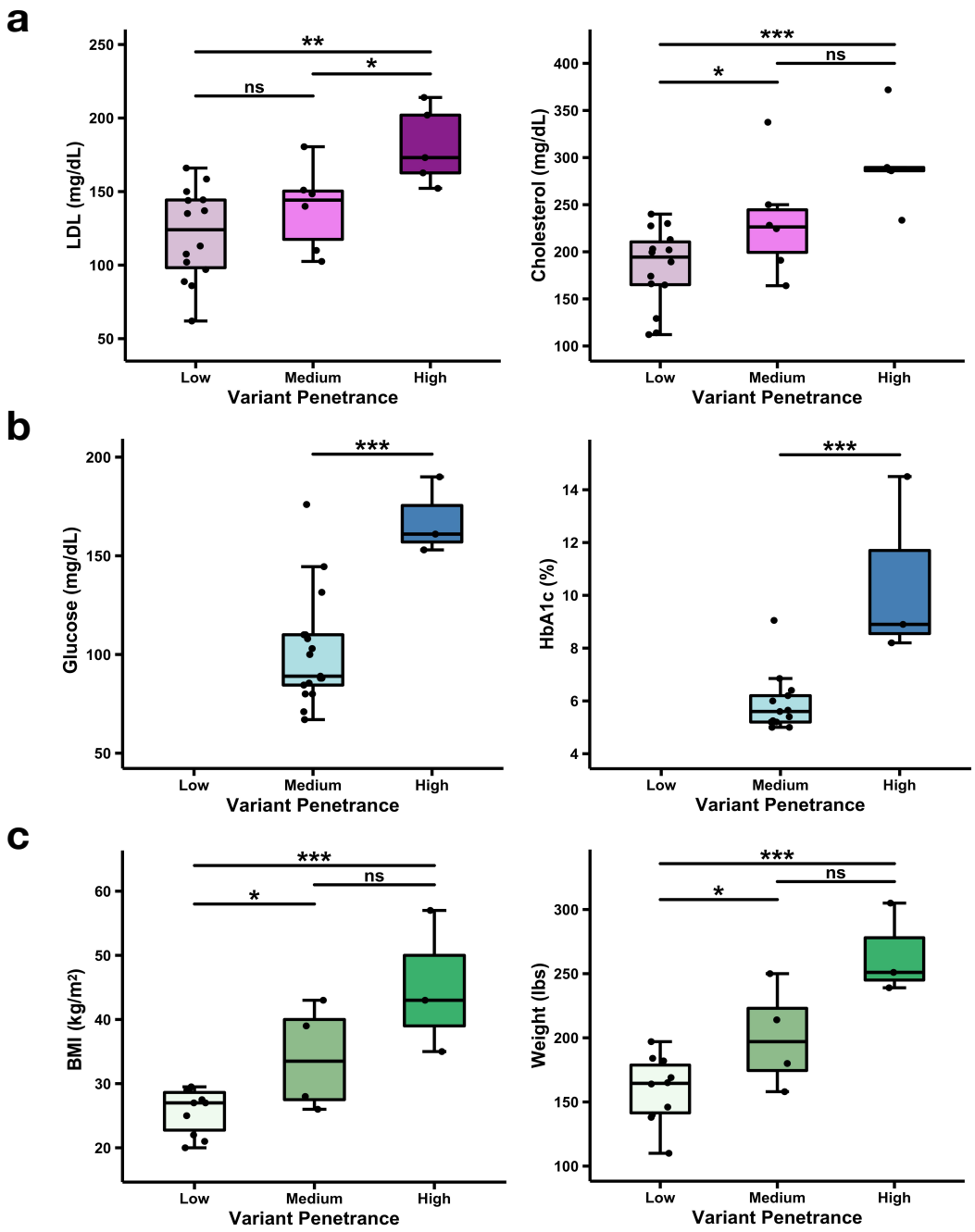
**Competing interests**

RD received grants from AstraZeneca, grants and nonfinancial support from Goldfinch Bio, is a scientific co-founder, consultant and equity holder for Pensieve Health, and is a consultant for Variant Bio.

865 **FIGURES**

866 **Figure 1**. Biomarker levels in carriers of clinically impactful variants with varying

867 penetrance for three diseases in the Bio*Me* Biobank.




868

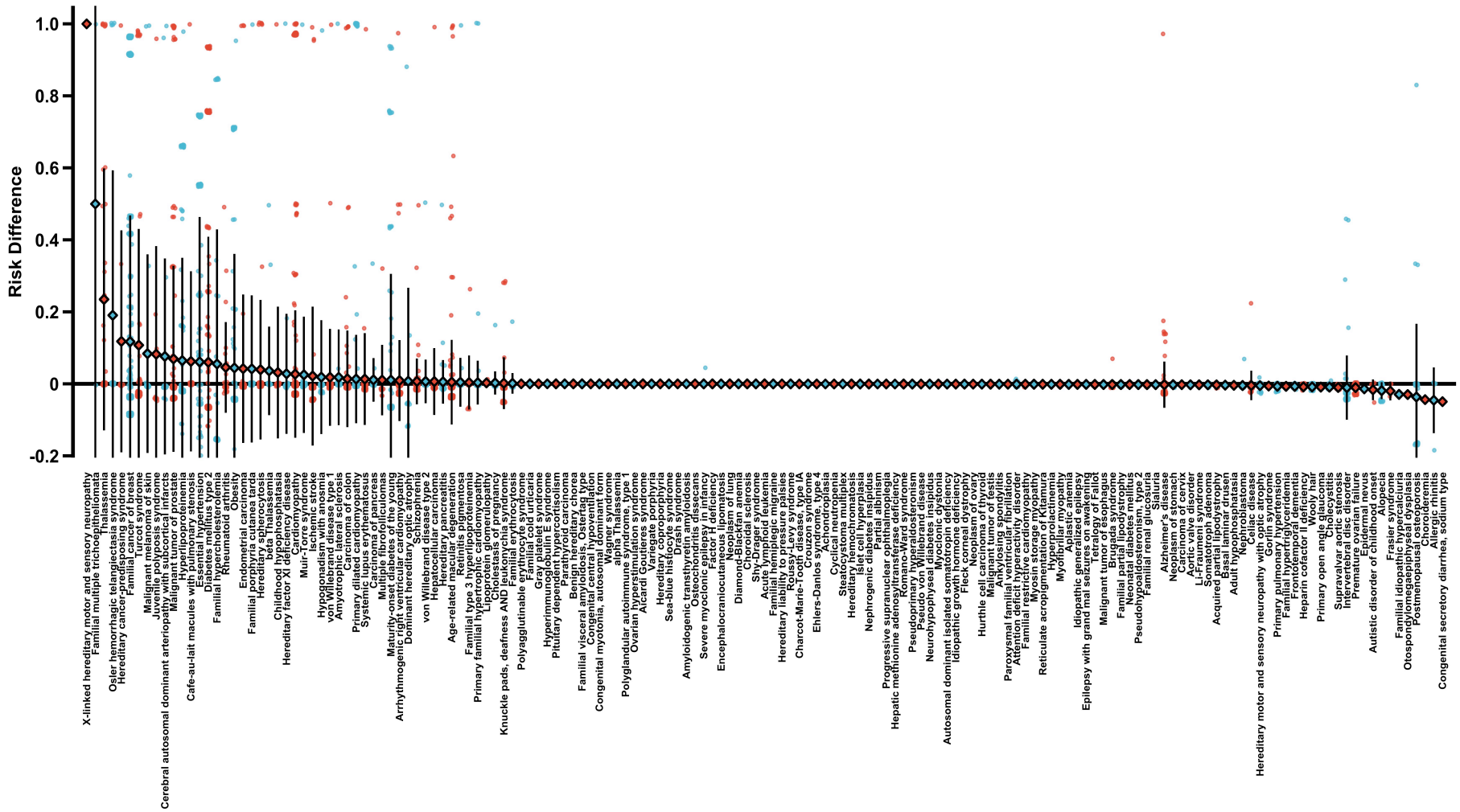
869

870 Metabolite and clinical measurements for carriers of clinically impactful variants with

871 varying penetrance for three diseases in the Bio*Me* Biobank. Clinically impactful

872 variants are reported as pathogenic/likely pathogenic in ClinVar or are loss-of-function
873 in a gene that mediates disease via loss-of-function mechanism, and in a gene with
874 non-recessive inheritance. Mean of measurements is compared between penetrance
875 strata with two-tailed t-test: *, $P<0.05$; **, $P<0.01$; ***, $P<0.001$. **a**, Variants in low-
876 density lipoprotein receptor (*LDLR*) are stratified by penetrance for familial
877 hypercholesterolemia (Low [0%], Medium [greater than 0 and less than 100%], and
878 High [100%]) and shown as box plots with measurements of low-density lipoprotein
879 cholesterol (LDL) and total cholesterol (cholesterol) levels in carriers. **b**, Variants in
880 hepatocyte nuclear factor 1-α (*HNF1A*) are stratified by penetrance for maturity-onset
881 diabetes of the young (Low [0%], Medium [greater than 0 and less than or equal to
882 50%], and High [greater than 0 and less than 100%]) and depicted as box plots with
883 levels of glucose and hemoglobin A1c (HbA1c) in carriers. No impactful variants in
884 *HNF1A* were of Low penetrance. **c**, Variants in uncoupling protein 3 (*UCP3*) are
885 stratified by penetrance for obesity (Low [0%], Medium [greater than 0 and less than
886 100%], and High [100%]) and displayed as box plots with body mass index (BMI) and
887 weight of carriers.
888
889
890
891
892
893

**Figure 2**. Distribution of observed disease risk for 5,359 clinically impactful variants in 72,434 individuals.

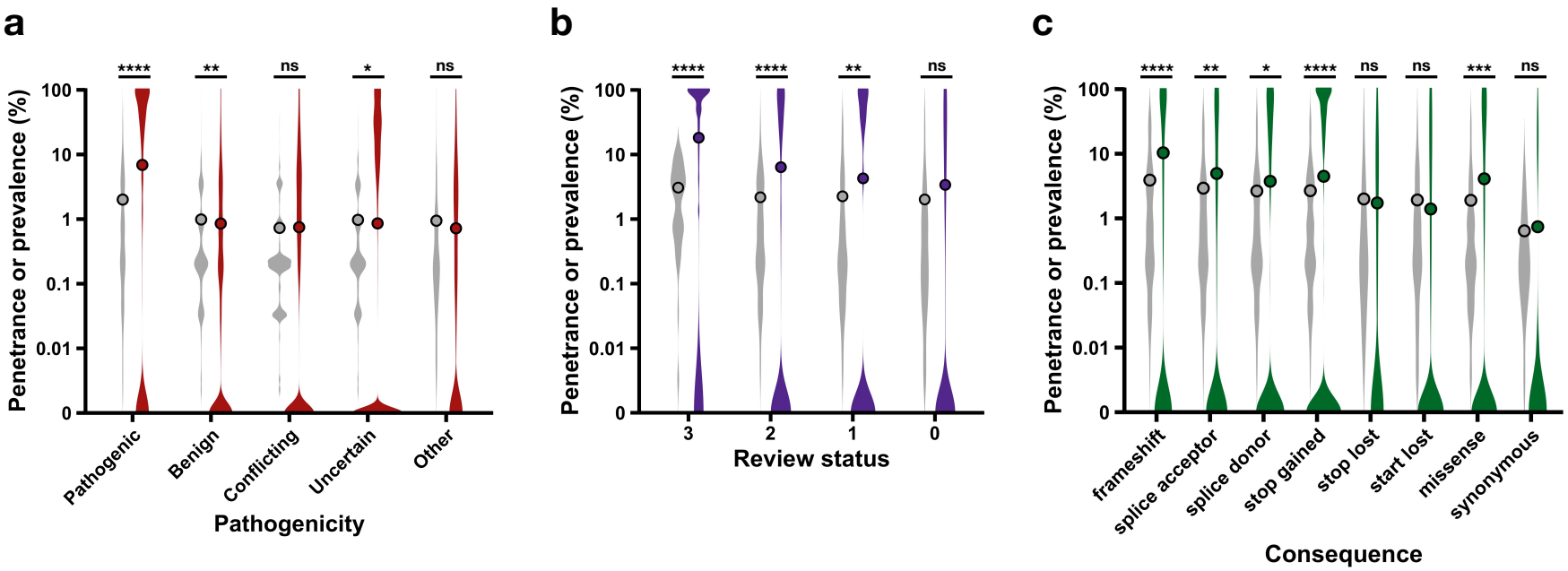



Distribution of observed disease risk for 5,359 clinically impactful variants in 72,434 individuals. Risk difference is the difference in disease prevalence between carriers and non-carriers of a variant of interest. Clinically impactful variants are

899 reported as pathogenic/likely pathogenic in ClinVar or are loss-of-function in a gene that mediates disease via loss-of-
900 function mechanism, and in a gene with non-recessive inheritance. Diseases are sorted by descending mean variant risk
901 difference (diamonds), with all variant risk difference estimates plotted (points) along with the standard deviation (error
902 bars) per disease. A complete tabulated list of results is provided in **Supplementary Table 6**.
903
904
905
906
907
908
909
910
911
912
913
914

915 **Figure 3**. Penetrance of 34,301 variants stratified by ClinVar pathogenicity, ClinVar review status, and molecular

916 consequence.




917
918

919 Penetrance of 34,301 variants in non-recessive genes stratified by ClinVar pathogenicity, ClinVar review status, and

920 molecular consequence. Penetrance distributions are shown as violin plots in red, purple, and green color on a base-10

921 logarithmic scale with the mean penetrance overlaid as points, alongside disease prevalence in non-carriers shown as

922 violin plots in grey color with the mean disease prevalence superimposed as points. Pathogenic/likely pathogenic variants

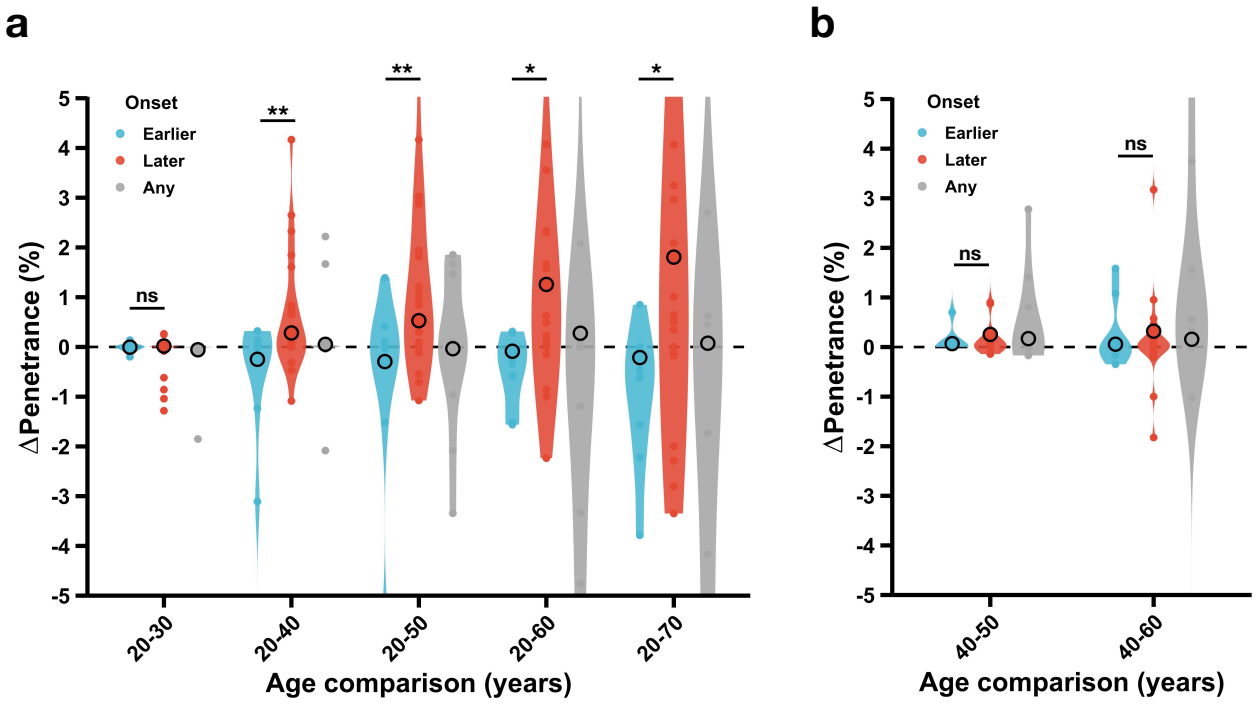
923 are grouped as pathogenic and benign/likely benign variants are grouped as benign. *, two-tailed t-test $P<0.05$; **,

924 $P<0.01$; ***, $P<0.001$; ****, $P<0.0001$. **a**, Penetrance is stratified by classification of variant pathogenicity in ClinVar.

925 Pathogenic variants on average have the highest penetrance (6.9% vs. 0.86% next highest mean penetrance in uncertain
926 class; $P$=5x10$^{-270}$). Only pathogenic variants confer a significantly increased risk of disease on average to carriers
927 compared to baseline disease risk in non-carriers (risk difference [RD]=0.046; $P$=1x10$^{-22}$). Other pathogenicity, variants
928 with a ClinVar pathogenicity classification other than pathogenic, benign, conflicting, or uncertain. **b**, Penetrance is
929 stratified by variant review status in ClinVar as reviewed by experts (review status=3), multiple submitters (review
930 status=2), single submitter (review status=1), or no assertion criteria (review status=0). Variants reviewed by experts
931 (review status=3) have the highest penetrance on average (18% vs. 6.3% next highest mean penetrance in review
932 status=2 group; $P$=9x10$^{-17}$). Pathogenic variants of review status=3, =2, and =1 confer a significantly increased risk of
933 disease on average, with RD=0.15 ($P$=1x10$^{-13}$), RD=0.042 ($P$=3x10$^{-10}$), and RD=0.020 ($P$=4x10$^{-3}$), respectively. **c**,
934 Penetrance is stratified by molecular consequence annotated with Variant Effect Predictor (VEP). Frameshift variants on
935 average have the highest penetrance (10% vs. 4.1% for missense; $P$=2x10$^{-13}$). There was a significantly elevated risk of
936 disease on average with frameshift (RD=0.064; $P$=3x10$^{-19}$), splice acceptor (RD=0.020; $P$=5x10$^{-3}$), splice donor
937 (RD=0.011; $P$=4x10$^{-2}$), stop gained (RD=0.018; $P$=3x10$^{-7}$), and missense variants (RD=0.022; $P$=4x10$^{-4}$).

938 **Figure 4**. Association between age of disease onset and age-dependent change in

939 penetrance for 157 diseases.




940
941

942 Association between age of disease onset and age-dependent change in penetrance for

943 157 diseases. Diseases correspond to the 5,359 clinically impactful variants and are

944 grouped according to their age of onset: Earlier, Later, or Any. Change in penetrance is

945 displayed as a violin plot for each age of onset group when comparing two carrier age

946 thresholds with the mean change in penetrance superimposed as a point. ΔPenetrance

947 (%), change in variant penetrance represented as a percent (+ values indicate greater

948 penetrance estimate with the older age threshold and - values indicate greater

949 penetrance estimate with the younger age threshold); age comparison, two carrier age

950 thresholds for which penetrance is compared (e.g., 20-70 compares penetrance with

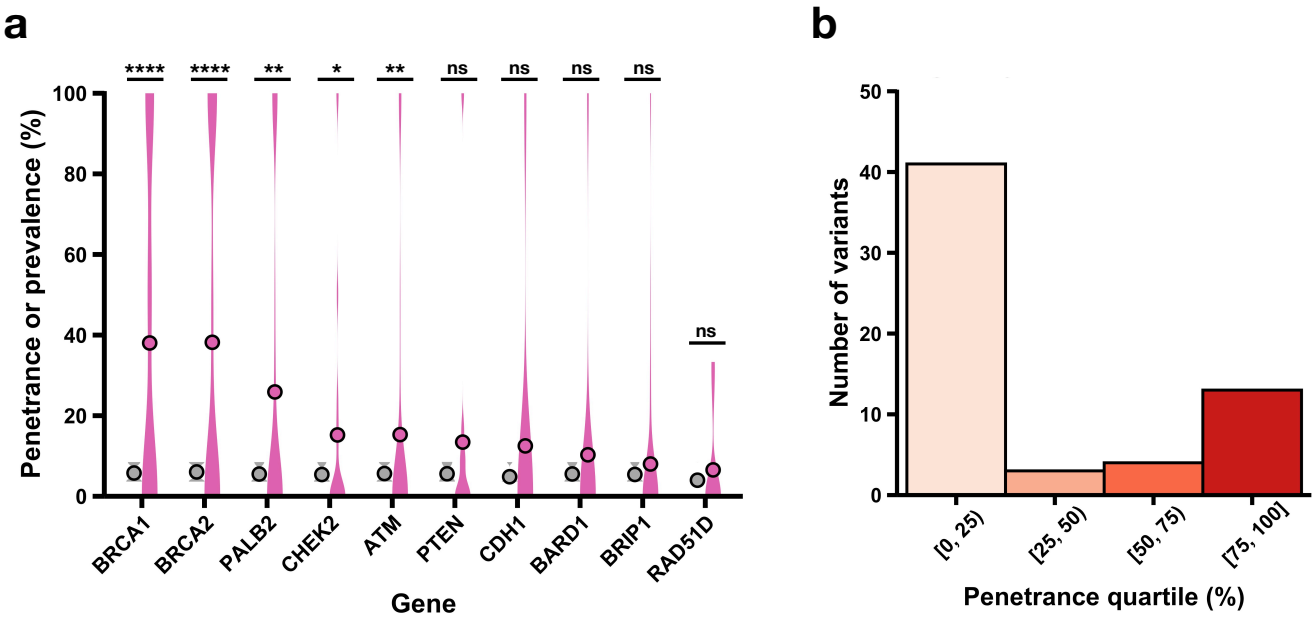
951 carriers ≥20 years old and penetrance with carriers ≥70 years old); onset, disease

952 groups according to age of onset; *, two-tailed t-test $P<0.05$; **, $P<0.01$; ***, $P<0.001$;

953 ****, *P*<0.0001. A complete plot including diseases with ΔPenetrance >5% or
954 ΔPenetrance <-5% is provided (**Supplementary Figure 5**). **a**, The Bio*Me* Biobank has
955 carriers of age 20-90 years old and ΔPenetrance is assessed over five carrier age
956 thresholds: Later diseases have greater ΔPenetrance on average than Earlier diseases
957 for 20-40 (*P*=0.006), 20-50 (*P*=0.002), 20-60 (*P*=0.02), and 20-70 (*P*=0.02) age
958 comparisons. **b**, UK Biobank has carriers of age 40-69 years old and ΔPenetrance is
959 evaluated over two carrier age thresholds: there is no difference in ΔPenetrance
960 between Later and Earlier diseases for either 40-50 (*P*=0.2) or 40-60 (*P*=0.2) age
961 comparisons.
962
963
964
965
966
967
968
969
970
971
972
973
974
975

976 **Figure 5**. Disease risk associated with clinically impactful variants in disease-

977 predisposition genes for breast cancer and hypercholesterolemia.




978
979

980 Disease risk associated with clinically impactful variants in known breast cancer- and

981 hypercholesterolemia-predisposition genes. Clinically impactful (impactful) variants are

982 reported as pathogenic/likely pathogenic in ClinVar or are loss-of-function in a gene that

983 mediates disease via loss-of-function mechanism, and in a gene with non-recessive

984 inheritance. **a**, Penetrance of impactful variants in 10 known or suspected breast

985 cancer-predisposition genes is displayed as pink violin plots with the mean penetrance

986 overlaid as a point, alongside disease prevalence in non-carriers shown as grey violin

987 plots with the mean prevalence superimposed as a point. Violin plots are sorted by

988 genes with descending mean penetrance, with the highest mean variant penetrance of

989 38% in *BRCA1* (n=48 variants, mean risk difference [RD]=0.32; $P$=2x10$^{-6}$) and 38% in

990 *BRCA2* (n=92 variants, mean RD=0.32; $P$=1x10$^{-10}$). *, two-tailed t-test $P$<0.05; **,

991 $P$<0.01; ***, $P$<0.001; ****, $P$<0.0001. **b**, Heterogeneity in the penetrance of 59 impactful

992 low-density lipoprotein receptor (*LDLR*) variants is shown. The number of variants is
993 plotted as a histogram across four quartiles of penetrance. [0, 25), penetrance less than
994 25%; [25, 50), penetrance greater than or equal to 25% and less than 50%; [50, 75),
995 penetrance greater than or equal to 50% and less than 75%; [75, 100], penetrance
996 greater than or equal to 75%.
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014

1015 **TABLES**

1016

1017 **Table 1.** Overview of baseline demographic and clinical traits for 72,434 individuals

1018 from two population-based biobanks.

| Trait | Bio*Me* (n=29,039) | UKB (n=43,395) |
|---|---|---|
| Male, n (%) | 11,684 (40%) | 19,330 (45%) |
| Age, mean (SD) | 59 (16) | 57 (8) |
| European ancestry, n (%) | 9,376 (32%) | 40,447 (93%) |
| African ancestry, n (%) | 7,190 (25%) | 916 (2%) |
| Hispanic ancestry, n (%) | 8,528 (29%) | -- |
| Asian ancestry, n (%) | 1,349 (5%) | 1,088 (3%) |
| Mixed ancestry, n (%) | 2,028 (7%) | 344 (0.8%) |
| BMI in kg/m2, mean (SD) | 28 (7) | 28 (5) |
| SBP in mmHg, mean (SD) | 127 (20) | 140 (19) |
| DBP in mmHg, mean (SD) | 73 (12) | 82 (11) |
| LDL-C in mg/dL, mean (SD) | 98 (36) | 63 (16) |
| Total cholesterol in mg/dL, mean (SD) | 177 (45) | 102 (21) |
| HbA1c in %, mean (SD) | 6 (2) | 6 (3) |
| Glucose in mg/dL, mean (SD) | 106 (47) | 93 (20) |
| HTN, n (%) | 12,962 (45%) | 11,094 (26%) |
| T2D, n (%) | 7,044 (24%) | 2,820 (6%) |
| FBC, n (%) | 1,488 (5%) | 1,581 (4%) |
| AMD, n (%) | 1,041 (4%) | 352 (0.8%) |

1019

1020 Overview of baseline demographic and clinical traits for 72,434 individuals from two

1021 population-based biobanks: Bio*Me* Biobank (Bio*Me*) and UK Biobank (UKB). n, number;

1022 SD, standard deviation; SBP, systolic blood pressure; DBP, diastolic blood pressure;

1023 LDL-C, low-density lipoprotein cholesterol; HbA1c, hemoglobin A1c; HTN, hypertension;

1024 T2D, type 2 diabetes; FBC, familial breast cancer; AMD, age-related macular

1025 degeneration; --, no Hispanic ancestry individuals are in UKB.