Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting

Caroline F. Wright,[1,*] Ben West,[1] Marcus Tuke,[1] Samuel E. Jones,[1] Kashyap Patel,[1] Thomas W. Laver,[1] Robin N. Beaumont,[1] Jessica Tyrrell,[1] Andrew R. Wood,[1] Timothy M. Frayling,[1] Andrew T. Hattersley,[1] and Michael N. Weedon[1,*]

More than 100,000 genetic variants are classified as disease causing in public databases. However, the true penetrance of many of these rare alleles is uncertain and might be over-estimated by clinical ascertainment. Here, we use data from 379,768 UK Biobank (UKB) participants of European ancestry to assess the pathogenicity and penetrance of putatively clinically important rare variants. Although rare variants are harder to genotype accurately than common variants, we were able to classify as high quality 1,244 of 4,585 (27%) putatively clinically relevant rare (MAF < 1%) variants genotyped on the UKB microarray. We defined as "clinically relevant" variants that were classified as either pathogenic or likely pathogenic in ClinVar or are in genes known to cause two specific monogenic diseases: maturity-onset diabetes of the young (MODY) and severe developmental disorders (DDs). We assessed the penetrance and pathogenicity of these high-quality variants by testing their association with 401 clinically relevant traits. 27 of the variants were associated with a UKB trait, and we were able to refine the penetrance estimate for some of the variants. For example, the HNF4A c.340C>T (p.Arg114Trp) (GenBank: NM_175914.4) variant associated with diabetes is <10% penetrant by the time an individual is 40 years old. We also observed associations with relevant traits for heterozygous carriers of some rare recessive conditions, e.g., heterozygous carriers of the ERCC4 c.2395C>T (p.Arg799Trp) variant that causes Xeroderma pigmentosum were more susceptible to sunburn. Finally, we refute the previous disease association of *RNF135* in developmental disorders. In conclusion, this study shows that very large population-based studies will help refine our understanding of the pathogenicity of rare genetic variants.

Introduction

One of the ongoing challenges in genetic medicine is that of variant interpretation. Many variants and genes have been erroneously associated with disease as a result of study design problems (including ascertainment bias and inadequate cohort size),[1–3] as well as biological phenomena such as genetic heterogeneity, reduced penetrance, variable expressivity, composite phenotypes, pleiotropy, and epistasis.[4–13] These issues have resulted in ambiguity over how to interpret clinically ascertained variants found in individuals with no known family history or symptoms of the disease.[14] Although there has traditionally been a division between rare disease genetics (studied in small disease cohorts and individual high-risk families) and common disease genetics (studied in large disease cohorts and population biobanks), in reality a continuum of causality is likely for many human disorders.[15] Fortunately, rare and common disease studies suffer from opposing ascertainment biases. Clinical and family-based cohorts ascertained as a result of a specific clinical presentation will tend to overestimate the penetrance of any identified disease-causing variants.[16] In contrast, population cohorts tend to be enriched for healthy individuals (the so-called "healthy volunteer" selection bias) who have both the time and ability to volunteer for a study,[17,18] and they will therefore tend to underestimate penetrance. Population cohorts that have high-resolution genetic and clinical data are therefore invaluable for establishing minimum penetrance estimates, exploring variable expressivity, and challenging pathogenicity assertions made in the clinical arena.

Several studies have already started to bridge this gap by using population data to evaluate rare disease-causing variants,[19,20] refine penetrance estimates,[21] and refute reportedly pathogenic variants.[22,23] These previous studies were mostly limited to a very specific set of variants (e.g. protein-truncating variants) or one particular disease, or they were too small to statistically test phenotypic penetrance. With its wealth of linked phenotypic and clinical information on ~500,000 genotyped individuals, UK Biobank (UKB)[24] offers a powerful dataset in which to systematically evaluate the pathogenicity, penetrance, and expressivity of clinically important variants in the population. However, differences in the technologies used for assaying genetic variation can hinder these analyses. A particular concern is the use of genotyping arrays (such as those currently used by UKB),[25] which have been designed primarily to assay common variation. In contrast, rare single-nucleotide variants (SNVs) and small insertions/deletions (indels) have typically been detected through sequencing assays.[26] A method is therefore needed to select well-genotyped, rare variants in UKB; this method can then be used in addressing biological and clinical questions.

[1] Institute of Biomedical and Clinical Science, University of Exeter Medical School, Research, Innovation, Learning and Development building, Royal Devon & Exeter Hospital, Barrack Road, Exeter EX2 5DW, UK
*Correspondence: caroline.wright@exeter.ac.uk (C.F.W.), m.n.weedon@exeter.ac.uk (M.N.W.)
https://doi.org/10.1016/j.ajhg.2018.12.015.

Here we describe a systematic method for evaluating the analytical validity of rare-variant genotyping data from the UKB arrays, investigate the relationship between data quality and minor-allele frequency (MAF), and evaluate the association of a subset of clinically interesting, well-genotyped coding variants with relevant phenotypes in UKB. We focus on ClinVar variants that have been classified as "pathogenic" or "likely pathogenic" by at least one submitter,[27] as well as variants in genes known to cause two specific monogenic diseases, maturity-onset diabetes of the young (MODY [MIM: 606391]) and developmental disorders (DDs), in which we have some expertise.

Subjects and Methods

UKB Cohort

Between 2006 and 2010, UKB recruited more than 500,000 individuals aged 37–73 years from across the UK. Participants provided a range of information (e.g., demographics, health status, lifestyle) via questionnaires and interviews. Additionally, anthropometric measurements; blood-pressure readings; and blood, urine, and saliva samples were taken for future analysis. Genotypes for SNVs and indels were generated from the Affymetrix Axiom UKB array (~450,000 individuals) and the UKBiLEVE array (~50,000 individuals) in 106 batches of ~4,700 samples. This dataset underwent extensive central quality control (see Web Resources).[25] We limited our analysis to 379,768 QC-passed white Europeans.

Variant Prioritization

We annotated variants by using Annovar[28] and calculated MAFs by using PLINK.[29] To prioritize variants of potential clinical importance, we selected those with at least one classification of pathogenicity (pathogenic or likely pathogenic) in the ClinVar database;[27] variants with conflicting classifications were not excluded. In addition, irrespective of their presence in ClinVar, we selected predicted protein-truncating variants (PTVs; stopgain SNVs and frameshift indels) and known pathogenic functional variants (nonsynonymous SNVs and inframe indels) in genes known to be associated with MODY[30,31] and dominant DD[32,33] for detailed evaluation. These diseases and genes were selected on the basis of our own prior experience, the availability of well-curated gene lists that include the mode of inheritance and mechanism of action, and the different prior probabilities associated with finding diabetes (a common disease) and severe DD (a rare disease) in UKB. We excluded common variants (MAF $> 1\%$) because these have already been thoroughly investigated through genome-wide association studies,[34,35] and we further refined the list of variants to include only those where the Hardy-Weinberg equilibrium (HWE) had a $p > 0.05$ and the proportion of missing genotypes across all samples was < 0.01 ($n = 4,585$).

Assessing Analytical Validity

To assess the analytical validity of these variants, we used Evoker Lite (see Web Resources) to generate cluster plots of intensities, and we combined data from all the batches into one plot for each variant. Cluster plots were manually assessed and ranked in quality from 1–5, where 1 = poor quality, no discernible separate clusters; 2 = poor quality, no discernible separate clusters but noisy data; 3 = unclear or uncertain; 4 = good quality, clearly separable clusters but noisy data; and 5 = good quality, clear separation between clusters (Figure S1). In an initial 750-variant subset that was independently evaluated by two scientists (Figure S2), correlation between the two independent scorers was high ($R^2 = 0.8$), and there was a 95% agreement in low quality (score = 1 or 2) versus high quality (score = 4 or 5) variants. All remaining variants of interest were evaluated by one scientist, and those with high quality scores were checked by the second scientist. Only variants with an average score of ≥ 4 were retained for further analysis. For all 1,244 high-quality variants, we assessed whether the rare genotype calls were unusually distributed across the 107 genotype batches. None of the rare genotype calls at these variants were entirely due to calls from a single batch. Across the 1,244 variants, the highest proportion of rare genotype calls in a single batch was four from a total of 13 for Affx-89007317. A plot of total allele count for each variant versus maximum allele count across each individual batch demonstrated a linear association with no clear outlying variants.

Assessing Clinical Relevance

Using PLINK, we ran a phenome-wide association in 379,768 QC-passed white Europeans for all of our 1,244 high-quality rare variants against a curated list of 401 clinically relevant traits in UKB (Table S1)[29] and those variants with a Bonferroni-corrected $p < 1\times10^{-7}$ (0.05/[401*1244]) were prioritized for detailed evaluation. For continuous traits, we used linear regression wherein we adjusted for age, sex (unless the trait was sex-specific), center, genotyping chip, and ten ancestry principal components. For binary traits, we used Fisher's exact test as the primary association method. We excluded three variants that had been reclassified in ClinVar as benign since our variant annotation (November 2017). To assess the potential clinical implications of high-quality rare variants, we compared the UKB traits with the clinical presentation of the disease for each gene and the evidence supporting the assertion of pathogenicity of the variant by using ClinVar,[27] DECIPHER,[36] and OMIM.[37] For high-quality rare variants in MODY genes and PTVs in DD genes, we had no p-value cut-off for investigating diabetes and developmental traits (cognitive function, educational attainment, body mass index, height, hearing, and albumin creatinine ratios). Conditional analysis of the most-associated regional variant (within a 1 Mb window) from each trait led us to remove one trait-variant association that was explained by linkage disequilibrium with a common causal variant.

Results

Variants below 0.001% Frequency Are Not Reliably Genotyped

Across all the variants evaluated for analytical validity via combined cluster plots (n unique = 4,585; see Subjects and Methods), we categorized 27% as high quality (average score $\geq$4), 64% as low quality, most likely false positives (average score $\leq$2.5), and 9% as unclear (Table 1). There was a strong correlation between the analytical-validity quality score and both the MAF (Table 1 and Figure 1) and the presence of the variant in either gnomAD[38] or the 1000 genomes project.[39] For low- versus high-quality variants, a nonparametric regression analysis estimated the area under the ROC curve to be 0.95 (95% confidence interval (CI) = 0.943–0.956); the false-positive rate (FPR) at

Table 1. Evaluated Variants

MAF Bin (%)	FP	TP	Unclear	Total
0–0.0005	511	0	11	522
0.0005–0.001	607	8	59	674
0.001–0.005	1,598	218	210	2,026
0.005–0.01	138	204	73	415
0.01–0.05	66	456	48	570
0.05–0.1	2	129	5	136
0.1–0.5	6	189	7	202
0.5–1	0	40	0	40
Total	2,928	1,244	413	4,585

Number of variants manually evaluated for analytical validity in different MAF bins; quality scores are grouped into false positives (FP, score = 1 or 2), unclear scores (score = 3), and true positives (TP, score = 4 or 5).

MAF > 0.005% was ~20%, but the FPR was ~60% at MAF > 0.001%.

Estimates of Minimum Effect Size for Known Pathogenic Variants

The 1,244 high-quality, putatively pathogenic rare variants, along with their ClinVar-associated disease and the allele frequencies in UKB and gnomAD, are shown in Table S2. Of these variants, only 27 were associated ($p < 1 \times 10^{-7}$) with one of the 401 traits we tested against in UKB (Table 2 and Table S3). Of these, 13 have previously been linked with a dominant disease, although most are considered to be only risk factors or low-penetrance variants rather than true highly pathogenic monogenic variants. Specifically, we observed well-established associations between variants in *HOXB13*[40] (MIM: 604607) and *PALB2*[41] (MIM: 610355) and prostate cancer (MIM: 176807) and breast cancer (MIM: 114480), respectively. The effect sizes are broadly in line with estimates from population-based studies and consistently lower than those of family-based studies (e.g.[40,41]). The HOXB13 c.251G>A (p.Gly84Glu) variant was originally suggested to increase risk 20-fold,[40] but a population-based study[42] of this variant reported an odds ratio of 3.5, 95% CI ([2.4, 5.2]), similar to the odds ratio observed in the UKB of 4.09, 95% CI ([3.24, 5.17]). For *PALB2*, a similar loss-of-function variant found at relatively high frequency in Finland had an odds ratio of 11.3 (95% CI ([1.8, 57.81])) when family-based cases were used, but it had an odds ratio of only 3.94, 95% CI ([1.5, 12.1]) when unselected cases were used;[43] this ratio is comparable to the estimate of 4.55, 95% CI ([3.05, 6.79]) in the UKB. The other 11 variants were causally linked to disease, and we observed that these variants were associated with a related trait in our population-based cohort (Table 2). In *FLG* (MIM: 135940), two PTVs that cause ichthyosis vulgaris[44] (MIM: 146700) were associated with increased odds of eczema (MIM: 603165) [odds ratios were 1.66 (95% CI [1.40, 1.98] and 1.96 (95% CI [1.69, 2.27])], consistent with effect sizes for loss-of-function mutations from previous studies.[45] A *TSHR* (MIM: 603372) PTV that causes nonautoimmune hyperthyroidism[46] (MIM: 609152) was associated with an increased odds of hypothyroidism (odds ratio 3.34, 95% CI [2.47, 4.51]). A nonsynonymous *LRRK2* (MIM: 609007) variant that causes Parkinson disease (MIM: 607060) was associated with an odds ratio of 4.76 (95% CI [3.25, 6.96]) that a person would have a parent with Parkinson disease; this estimate is consistent with studies of affected family members.[47–49] A nonsynonymous *PER3* (MIM: 603427) variant

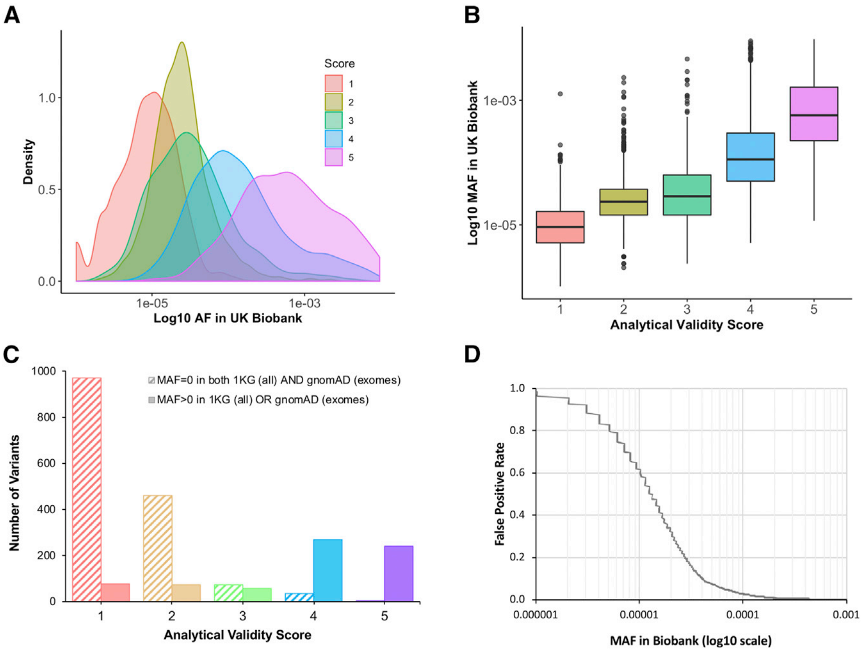

Figure 1. Correlation between Minor Allele Frequency and Analytical Validity Quality Score

(A and B) Density plot (A) and boxplot (B) of manual quality scores (from 1–5, see Figure S1) of genotype data in UKB versus minor allele frequency (MAF) for 4,585 putatively clinically important variants, where MAF < 1%, Hardy–Weinberg equilibrium (HWE) > 0.05, and missingness < 0.01.

(C) Histogram of the number of variants at each quality score versus presence or absence of the variant in gnomAD (exome data) or the 1000 Genomes Project.

Red = score 1; gold = score 2; green = score 3; blue = score 4; purple = score 5.

(D) Estimation of the false-positive rate (FPR) versus MAF for variants assayed with the UKB genotyping arrays, calculated by the grouping of quality scores into low (score = 1 or 2) and high (score = 4 or 5) and use of the rocreg command in Stata for fitting a ROC curve.

Table 2. Pathogenic Variants

Gene	UKB ID	Position (GRCh37)	HGVS	MAF White British (%)	Significantly Associated Trait(s) in UKB (Units)	Odds Ratio or Beta [95% CI]	p value	Linked Disease (Mode of Inheritance)
ACSF3	dbSNP: rs141090143	chr16: 89220556 C>T	GenBank: NM_174917: c.C1672T:p.R558W	0.632	ease of sunburn (number of episodes)	0.31 [0.20, 0.42]	4×10^{-10}	combined malonic and methylmalonic aciduria (AR)
AR	dbSNP: rs137852591	chrX: 66941751 C>G	GenBank: NM_000044: c.C2395G:p.Q799E	0.129	skeletal mass (SD)	−0.16 [−0.21, −0.11]	1×10^{-10}	partial androgen insensitivity syndrome (XLR)
					height (cm)	−0.85 [−1.27, −0.43]	1×10^{-8}	
	dbSNP: rs1800053	chrX: 66931295 C>A	GenBank: NM_000044: c.C1937A:p.A646D	0.269	balding pattern (males only)	−0.13 [−0.17, −0.08]	1×10^{-8}	partial androgen insensitivity syndrome (XLR)
ERCC4	dbSNP: rs121913049	chr16: 14041848 C>T	GenBank: NM_005236: c.C2395T:p.R799W	0.060	ease of sunburn (number of episodes)	0.98 [0.64, 1.33]	2×10^{-8}	xeroderma pigmentosum (AR)
FLG	dbSNP: rs150597413	chr1: 152277622 G>T	GenBank: NM_002016: c.C9740A:p.S3247X	0.369	eczema	1.66 [1.40, 1.98]	9×10^{-8}	ichthyosis vulgaris (AD)
	dbSNP: rs138726443	chr1: 152280023 G>A	GenBank: NM_002016: c.C7339T:p.R2447X	0.446	eczema	1.96 [1.69, 2.27]	5×10^{-16}	ichthyosis vulgaris (AD)
GCK	dbSNP: rs104894006	chr7: 44189591 G>A	GenBank: NM_000162: c.C556T:p.R186X	0.001	maturity-onset diabetes of the young	68 [14, 325]	2×10^{-8}	diabetes mellitus (AD)
HBB	Affx-52141620	chr11: 5248004 G>A	GenBank: NM_000518: c.C118T:p.Q40X	0.005	mean corpuscular volume (SD)	−2.92 [−3.26, −2.57]	6×10^{-63}	beta-thalassemia (AR)*
					red blood cell distribution width (SD)	1.87 [1.53, 2.21]	$5 \text{ x } 10^{-27}$	
HOXB13	dbSNP: rs138213197	chr17: 46805705 C>T	GenBank: NM_006361: c.G251A:p.G84E	0.160	prostate cancer	4.09 [3.24, 5.17]	1×10^{-23}	prostate cancer susceptibility (AD)
					father with prostate cancer	1.75 [1.47, 2.09]	4×10^{-9}	
HNF4A	dbSNP: rs137853336	chr20: 43042354 C>T	GenBank: NM_175914: c.340C>T:R114W	0.015	diabetes	2.9 [1.7, 5]	3×10^{-4}	maturity-onset diabetes of the young (AD)
HSPA9	Affx-80274027	chr5: 137902404 CT>-	GenBank: NM_004134:c.882_883del:p.T294fs	0.017	mean corpuscular volume (SD)	−0.49 [−0.67, −0.32]	2×10^{-8}	even-plus syndrome (AR)
					red blood cell distribution width (SD)	1.17 [0.99, 1.34]	9×10^{-40}	
KLF1	Affx-80299186	chr19: 12995833 ->C	GenBank: NM_006563: c.954dupG:p.R319fs	0.017	mean corpuscular volume (SD)	−1.27 [−1.45, −1.1]	9×10^{-48}	blood group Lutheran inhibitor (AD)
					red blood cell distribution width (SD)	1.48 [1.3, 1.65]	2×10^{-63}	
LRRK2	dbSNP: rs34637584	chr12: 40734202 G>A	GenBank: NM_198578: c.G6055A:p.G2019S	0.032	parent with Parkinson disease	4.76 [3.25, 6.96]	1×10^{-11}	Parkinson disease (AD)
MYH7	Affx-86888962	chr14: 23887458 C>T	GenBank: NM_000257: c.C4130T:p.T1377M	0.117	pulse rate (beats per minute)	−4.75 [−5.47, −4.01]	4×10^{-41}	primary familial hypertrophic cardiomyopathy (AD)

(*Continued on next page*)

Table 2. *Continued*

Gene	UKB ID	Position (GRCh37)	HGVS	MAF White British (%)	Significantly Associated Trait(s) in UKB (Units)	Odds Ratio or Beta [95% CI]	p value	Linked Disease (Mode of Inheritance)
NPC1	dbSNP: rs80358259	chr18: 21116700 A>G	GenBank: NM_000271: c.T3182C:p.I1061T	0.075	mean corpuscular volume (SD)	−0.24 [−0.32, −0.15]	2×10^{-8}	Niemann-Pick disease (AR)
OCA2	dbSNP: rs28934272	chr15: 28230247 C>T	GenBank: NM_000275: c.G1327A:p.V443I	0.834	ease of sunburn (number of episodes)	0.49 [0.40, 0.58]	1×10^{-47}	oculocutaneous albinism (AR)
	dbSNP: rs121918170	chr15: 28228529 T>C	GenBank: NM_000275: c.A1465G:p.N489D	0.094	ease of sunburn (number of episodes)	0.91 [0.64, 1.18]	1×10^{-14}	oculocutaneous albinism (AR)
PALB2	dbSNP: rs180177132	chr16: 23632683 C>T	GenBank: NM_024675: c.G3113A:p.W1038X	0.033	breast cancer	4.55 [3.05, 6.79]	2×10^{-10}	familial breast cancer (AD)
					mother with breast cancer	2.62 [1.92, 3.59]	5×10^{-8}	
PER3	dbSNP: rs139315125	chr1: 7869960 A>G	GenBank: NM_001289862: c.A1250G:p.H417R	0.438	morning person	1.37 [1.27, 1.47]	2×10^{-16}	advanced sleep phase syndrome (AD)
	dbSNP: rs150812083	chr1: 7869953 C>G	GenBank: NM_001289862: c.C1243G:p.P415A	0.458	morning person	1.35 [1.25, 1.46]	7×10^{-15}	
SEC23B	dbSNP: rs121918221	chr20: 18496339 G>A	GenBank: NM_006363: c.G325A:p.E109K	0.027	red blood cell distribution width (SD)	0.39 [0.25, 0.52]	3×10^{-8}	congenital dyserythropoietic anemia (AR)
SLC6A19	dbSNP: rs121434346	chr5: 1212453 G>A	GenBank: NM_001003841: c.G517A:p.D173N	0.442	red blood cell distribution width (SD)	−0.15 [−0.18, −0.11]	2×10^{-16}	neutral 1 amino acid transport defect (AR)
TACR3	dbSNP: rs144292455	chr4: 104577415 C>T	GenBank: NM_001059: c.G824A:p.W275X	0.054	reproductive age at menarche (yr)	0.66 [0.45, 0.87]	2×10^{-10}	hypogonadotropic hypogonadism (AR)
TMPRSS6	dbSNP: rs137853120	chr22: 37469593 C>T	GenBank: NM_153609: c.G1561A:p.D521N	0.019	mean corpuscular volume (SD)	−0.67 [−0.83, −0.51]	3×10^{-16}	microcytic anemia (AR)
					red blood cell distribution width (SD)	0.69 [0.53, 0.85]	5×10^{-17}	
TSHR	dbSNP: rs121908866	chr14: 81610039 G>A	GenBank: NM_000369: c.G1637A:p.W546X	0.041	hypothyroid	3.34 [2.47, 4.51]	7×10^{-12}	hypothyroidism (AD, AR)
					autoimmune disease	2.31 [1.76, 3.04]	4×10^{-8}	
TNFRSF13B	dbSNP: rs34557412	chr17: 16852187 A>G	GenBank: NM_012452: c.T310C:p.C104R	0.703	mean corpuscular volume (SD)	−0.09 [−0.12, −0.07]	4×10^{-11}	common variable immunodeficiency (AD, AR)

Reduced penetrance, variable expressivity, and carrier phenotypes for rare (MAF < 1%) ClinVar pathogenic variants with genome-wide significant associations in UKB. Abbreviations are as follows: UKB = UK Biobank, HGVS = Human Genome Variation Society, MAF = minor allele frequency, SD = standard deviations, cm = centimeters, yr = years, CI = confidence interval, AD = autosomal dominant, AR = autosomal recessive, XLR = X-liniked recessive.

previously classified as pathogenic for advanced sleep phase syndrome (MIM: 616882) had an odds ratio of only 1.38 for being a morning person and advanced sleep timing by only 8 minutes, 95% CI ([4, 13])[50,51] compared to a reported 4.2 hour shift in midpoint sleep. Height, skeletal weight, and male pattern baldness were negatively associated with two nonsynonymous variants in *AR* (MIM: 313700) that cause partial androgen insensitivity syndrome.[52] Finally, a nonsynonymous *MYH7* variant (MIM: 160760), which has been classified by a ClinGen expert panel as pathogenic for hypertrophic cardiomyopathy[53] (MIM: 192600), was associated with a reduced pulse rate of 5 (95% CI ([4, 6])) beats per minute.

Reduced Penetrance for the Common HNF4A p.Arg114Trp MODY Mutation

We specifically investigated known pathogenic variants and PTVs in MODY genes, where we found two rare

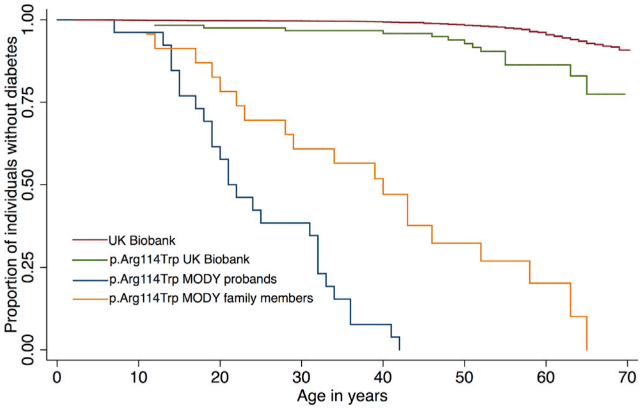

Figure 2. Comparison of Penetrance Estimate for *HNF4A* p.Arg114Trp in UK Biobank versus Previously Published Estimates from MODY Cohort Studies

A Kaplan-Meier plot of the proportion of individuals who are diabetes free at various ages for 379,768 individuals from UK Biobank (red line), 122 UK Biobank individuals who are heterozygous for HNF4A p.Arg114Trp (green line), 26 MODY referral probands (blue line), and 24 family members of the probands (yellow line) from Laver *et al.*[54]

variants that were high quality, definitely pathogenic, and strongly associated with diabetes (Table 2): a very rare stop-gain variant in *GCK* (MIM: 138079) (OR = 68, 95% CI [14, 328], $p = 2 \times 10^{-8}$), and a nonsynonymous variant (p.Arg114Trp) in *HNF4A* (MIM: 600281) (OR = 2.9, 95% CI [1.7, 5.0], $p = 3 \times 10^{-4}$). Both were associated with diabetes in UKB, in line with previous findings.[54–56] However, the penetrance of the *HNF4A* variant was previously estimated on the basis of a large MODY diabetes cohort to be up to 75% at age 40 years,[54] although we estimate the minimum penetrance to be <10% from UKB (Figure 2). This has important implications for the attributable risk associated with the variant in different cohorts and for the interpretation of genetic test results: if the p.Arg114Trp variant was found in an affected individual after clinical testing, it might still be the primary cause of that person's diabetes, although incidental discovery of the variant in an unaffected individual would not be predictive.

Related Mild Heterozygous Phenotypes in Autosomal-Recessive Disorders

Of our 27 high-quality, rare putatively pathogenic variants associated with a trait in UKB, 16 have previously been linked with a recessive disease (Table 2). We observed associations with milder or related traits in the heterozygous carriers of these monogenic recessive diseases in our population cohort. A nonsynonymous *ERCC4* (MIM: 133520) variant that causes recessive xeroderma pigmentosum[57] and two nonsynonymous *OCA2* (MIM: 611409) variants that cause oculocutaneous albinism[58,59] (MIM: 203200) were associated with ease of sunburn. A stopgain *TACR3* (MIM: 162332) variant, which causes recessive hypogonadotropic hypogonadism[60,61] (MIM: 614840), was associated with an 8 month increase in age at which a girl experiences menarche. In addition, variants in six genes known to be associated with different recessive blood-related disorders were also associated with decreased mean corpuscular volume and/or increased red blood cell distribution width (such genes included *HBB* [MIM: 141900], variants in which cause β-thalassemia (MIM: 612985), but the carrier state is already known to cause the much milder β-thalassemia minor[62]).

Benign Protein-Truncating Variants in Monogenic Genes

We focused our clinical analysis of variants in DD genes on just PTVs, of which six (including two variants in one gene) were of high quality and were in genes that are reported to cause disease via a haploinsufficiency mechanism (Table 3). None of these variants were associated with developmentally relevant traits in UKB ($p > 0.1$), suggesting they are all benign. For three variants, the location of the variant in the gene is notably different from that of known pathogenic variants. *GNAS* (MIM: 139320) is the only one of the five genes with a high probability of being loss-of-function intolerant (pLI)[38] on the basis of the frequency of loss-of-function variants in the Exome Aggregation Consortium (ExAC) browser.[38] The stop-gain variant in *GNAS* is present in the highly variable first exon of the gene and is likely to result in nonsense-mediated RNA decay; in contrast, pathogenic *GNAS* variants that cause Albright hereditary osteodystrophy (MIM: 103580) are located in later, highly constrained exons.[63] Similarly, the stop-gain variant in *TGIF1* (MIM: 602630) is located in the first exon, where multiple PTVs in gnomAD[38] are also located, but *TGIF1* pathogenic variants causing holoprosencephaly are located in the final exons, where they affect DNA binding affinity.[64] Finally, a frameshift deletion in *HIST1H1E* (MIM: 142220) is located near the start of the single exon of this gene; however, pathogenic *HIST1H1E* frameshift deletions that cause child overgrowth and intellectual disability are located near the end of the exon, where they result in a truncated histone protein with lower net charge that is less effective at binding DNA.[65] Hence, we believe that these three rare PTVs are benign because of their locations, despite the fact that they occur in genes that cause dominant DD via haploinsufficiency.

Refuting Previous Disease Associations

For the other three DD variants, our findings are not consistent with the genes' causing a dominant DD via haploinsufficiency. First, there was no association between a frameshift variant in the middle of *COL4A3* (MIM: 120070)—where pathogenic variants are thought to cause a rare dominant form of Alport syndrome (MIM: 104200) (as well as benign familial hematuria [MIM: 141200])[66,67]—and albumin creatinine ratios, hearing, or any of the development traits in UKB. Similarly, there was no association between either stop-gain or frameshift variants in *RNF135* (MIM: 611358) —where haploinsufficiency is thought to cause macrocephaly, macrosomia,

Table 3. Benign Variants

Gene	UKB ID	Position (GRCh37)	HGVS	MAF (%)	Biobank Trait	Beta [95% CI]	p Value	Linked Monogenic Disease
COL4A3	Affx-80270894	chr2: 228148945 G>GAGTAAAGGGCC	GenBank: NM_000091: c.2766_2776del: p.G922fs	0.01813	education years	−0.007 [−0.17, 0.156]	0.93	DD (Alport syndrome, autosomal dominant)
					fluid intelligence	−0.085 [−0.3, 0.133]	0.45	
					BMI	0.096 [−0.07, 0.259]	0.25	
					height	−0.097 [−0.26, 0.066]	0.24	
					albumin creatine ratio	0.941 [0.44, 2.015]	0.88	
					hearing left	−0.13 [−0.41, 0.144]	0.35	
					hearing right	0.076 [−0.2, 0.35]	0.59	
GNAS	dbSNP: rs200910410	chr20: 57428858 T>C	GenBank: NM_080425: c.C538T:p.Q180X	0.03063	education years	−0.053 [−0.18, 0.072]	0.41	DD (Albright hereditary osteodystrophy)
					fluid intelligence	0.062 [−0.12, 0.239]	0.49	
					BMI	0.048 [−0.08, 0.174]	0.46	
					height	−0.105 [−0.23, 0.021]	0.10	
HIST1H1E	Affx-89024826	chr6: 26156672 T>TC	GenBank: NM_005321: c.55delC:p.P19fs	0.02438	education years	0.098 [−0.05, 0.24]	0.18	DD (Childhood overgrowth)
					fluid intelligence	0.1 [−0.1, 0.302]	0.33	
					BMI	−0.014 [−0.16, 0.129]	0.85	
					height	0.059 [−0.08, 0.202]	0.42	
RNF135	dbSNP: rs121918161	chr17: 29324307 T>C	GenBank: NM_032322: c.C727T:p.Q243X	0.00215	education years	0.341 [−0.11, 0.79]	0.14	DD (macrocephaly, macrosomia, facial dysmorphism syndrome)
					fluid intelligence	0.16 [−0.41, 0.726]	0.58	
					BMI	0.177 [−0.27, 0.626]	0.44	
					height	0.082 [−0.37, 0.532]	0.72	
	Affx-80285705	chr17: 29325809 G>GC		0.05265	education years	0.02 [−0.08, 0.115]	0.68	
					fluid intelligence	−0.097 [−0.23, 0.032]	0.14	
					BMI	0.017 [−0.08, 0.113]	0.72	
					height	0.03 [−0.07, 0.125]	0.54	

(*Continued on next page*)

Table 3. *Continued*

Gene	UKB ID	Position (GRCh37)	HGVS	MAF (%)	Biobank Trait	Beta [95% CI]	p Value	Linked Monogenic Disease
TGIF1	dbSNP: rs202123354	chr18: 3452067 A>G	GenBank: NM_170695: c.G90A:p.W30X	0.01241	education years	0.023 [−0.23, 0.278]	0.86	DD (holoprosencephaly)
					fluid intelligence	0.169 [−0.2, 0.539]	0.37	
					BMI	0.134 [−0.12, 0.389]	0.30	
					height	0.053 [−0.2, 0.308]	0.68	

Classification of likely pathogenic variants in maturity-onset diabetes of the young (MODY) and developmental disorders (DD) from UKB. Abbreviations are as follows: UKB = UK Biobank, RSID = Reference SNP cluster ID, HGVS = Human Genome Variation Society, MAF = minor allele frequency, CI = confidence interval, BMI = body mass index, DD = developmental disorder.

and facial dysmorphism syndrome[68] (MIM:614192)—with any development traits in UKB. In both cases, given the high-quality genotyping of these variants in UKB and a lack of association with any clinically relevant traits, together with a pLI of zero for both genes, the age of the original publications, and the lack of enrichment of *de novo* mutations within the DDD study,[33] we suggest that haploinsufficiency in these genes is not a cause of a severe DD.

Discussion

Previous studies have been unable to analyze rare variants in sufficiently large population-based studies to establish pathogenicity and lower bounds for penetrance. Large population cohorts such as UKB provide an opportunity to investigate the relationship between genes and disease. However, the absence of genome-wide sequencing data has thus far minimized the impact of UKB in the rare disease community. We have established a method, using combined intensity plots for individual variants across all genotyping batches, for evaluating the analytical validity of rare variants genotyped by microarray. Although we initially tried to examine variant cluster plots for each batch separately, as recommended by UKB, this proved impossible because of the rarity of most clinically important variants. MAF was an extremely good predictor of the likelihood that a variant would be genotyped well by the UKB arrays (Figure 1). At MAF $>$ 0.005% ($\sim$50 heterozygous individuals out of 500,000 in UKB) the FPR was $\sim$7%, and most variants were well genotyped, but the FPR was $\sim$60% at MAF $>$ 0.001% ($\sim$10 heterozygous individuals), and we classified all variants at MAF $<$ 0.0005% ($\sim$5 heterozygous individuals) as being low quality. This has important implications for epidemiological research carried out uncritically with these data. Although many rare variants in UKB are well-genotyped with the arrays, the rarer the variant, the more likely it is to be poor quality and therefore yield false associations.

A limitation of our work is that we did not attempt to confirm the variants by using an independent assay. However, most researchers using data from UKB will be similarly unable to attempt independent variant confirmations, and thus a method for evaluating the genotyping quality of rare variants directly from the data has widespread utility. The validity of our method is supported by our ability to replicate numerous previous findings of well-known, clinically important variants classified as pathogenic in ClinVar (Table 2 and Table S3, plus additional well-established associations for variants where MAF $>$ 1%). In addition, our analyses of likely pathogenic variants in two disease subtypes (MODY and DD) were independent of any potential biases or misclassification errors associated with ClinVar, and the findings were consistent with our prior expectations. We expected there to be a small number of individuals in UKB with monogenic subtypes of diabetes, and we found two pathogenic variants that were associated with appropriate traits in UKB (Table 2); we were thus able to lower the previous penetrance estimate for a pathogenic variant in *HNF4A* (Figure 2). In contrast, we did not expect there to be any instances of severe DD, in light of the rarity of the condition, the relatively senior age of the UKB population, and the inherent challenges of obtaining consent from individuals with severe DD to be added to population biobanks.[69] We therefore believe that the PTVs identified in dominant DD genes in UKB are benign (Table 3) and are confident in refuting previous associations between haploinsufficiency in *RNF135* and *COL4A3* and dominant DD (note that this refutation has no bearing on the asserted relationship between the latter and either recessive DD or alternative mechanisms of disease).

In this study, we have shown that population genetic data can be used for estimating lower bounds for the effect size and penetrance of pathogenic, disease-causing variants and refined our understanding of the links between rare variants (MAF <1%) and monogenic diseases. Performing a similar analysis on very rare variants (MAF <0.001%) will require large-scale sequencing data rather than genotyping arrays. Although population-based studies will be biased in the opposite direction from clinical studies, i.e. towards healthy individuals, they are nonetheless crucial for informing minimum and age-dependent penetrance estimates, interpreting incidental or secondary findings from clinical testing, and informing direct-to-consumer genetic testing. At this point, we are left with some fundamental conceptual questions about the nature of "monogenic" disease. When should variants exhibiting reduced penetrance—a term frequently used in the diagnosis of rare genetic disease—be called risk or susceptibility factors, terms generally used in the study of common disease? When should a gene-disease relationship be termed variable expressivity rather than normal variation? Should "pathogenic" be reserved only for highly penetrant variants that cause a tightly defined disease entity, or can it apply to any variant associated, however weakly, with a clinically relevant phenotype? As genome-wide sequencing becomes widely used in routine clinical practice, research cohorts, and direct-to-consumer testing, understanding this spectrum will become both increasingly important and tractable.

Supplemental Data

Supplemental Data include three figures and three tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.12.015.

Acknowledgements

This research has been conducted using the UK Biobank Resource. This work was carried out under UK Biobank project number 871. M.N.W. and S.E.J. are funded by the Medical Research Council (grant: MR/M005070/1). A.R.W. and T.M.F. are supported by the European Research Council (grants: SZ-245 50371-GLUCOSEGENES-FP7-IDEAS-ERC and 323195). R.B. is funded by the Wellcome Trust and Royal Society (grant: 104150/Z/14/Z). J.T. is funded by a Diabetes Research and Wellness Foundation fellowship. K.A.P. has a postdoctoral fellowship funded by the Wellcome Trust (110082/Z/15/Z). T.W.L. is funded by the Wellcome Trust and the Royal Society (105636/Z/14/Z). A.T.H. is a Wellcome Trust Senior Investigator (WT098395/Z/12/Z), and A.T.H. is a National Institute for Health Research (NIHR) Senior Investigator.

Declaration of Interests

The authors declare no competing interests.

Received: September 5, 2018
Accepted: December 20, 2018
Published: January 17, 2019

Web Resources

DiabetesGenes, http://www.diabetesgenes.org
Evoker Lite, https://github.com/dlrice/evoker-lite
Gene2Phenotype, https://www.ebi.ac.uk/gene2phenotype
Online Mendelian Inheritance in Man, http://www.omim.org
UK Biobank Dataset, https://biobank.ctsu.ox.ac.uk/

References

1. Kraft, P., Zeggini, E., and Ioannidis, J.P.A. (2009). Replication in genome-wide association studies. Stat. Sci. *24*, 561–573.
2. Park, S., Lee, S., Lee, Y., Herold, C., Hooli, B., Mullin, K., Park, T., Park, C., Bertram, L., Lange, C., et al. (2015). Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families. BMC Med. Genet. *16*, 62.
3. Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. *15*, 1496–1502.
4. Gratten, J., and Visscher, P.M. (2016). Genetic pleiotropy in complex traits and diseases: Implications for genomic medicine. Genome Med. *8*, 78.
5. Visscher, P.M., and Yang, J. (2016). A plethora of pleiotropy across complex traits. Nat. Genet. *48*, 707–708.
6. Boycott, K.M., and Innes, A.M. (2017). When one diagnosis is not enough. N. Engl. J. Med. *376*, 83–85.
7. Theunissen, T.E.J., Sallevelt, S.C.E.H., Hellebrekers, D.M.E.I., de Koning, B., Hendrickx, A.T.M., van den Bosch, B.J.C., Kamps, R., Schoonderwoerd, K., Szklarczyk, R., Mulder-Den Hartog, E.N.M., et al. (2017). Rapid resolution of blended or composite multigenic disease in infants by whole-exome sequencing. J. Pediatr. *182*, 371–374.e2.
8. Ritchie, M.D., and Van Steen, K. (2018). The search for gene-gene interactions in genome-wide association studies: Challenges in abundance of methods, practical considerations, and biological interpretation. Ann. Transl. Med. *6*, 157.
9. Cooper, D.N., Krawczak, M., Polychronakos, C., Tyler-Smith, C., and Kehrer-Sawatzki, H. (2013). Where genotype is not predictive of phenotype: Towards an understanding of the

molecular basis of reduced penetrance in human inherited disease. Hum. Genet. *132*, 1077–1130.

10. Gillentine, M.A., Lupo, P.J., Stankiewicz, P., and Schaaf, C.P. (2018). An estimation of the prevalence of genomic disorders using chromosomal microarray data. J. Hum. Genet. *63*, 795–801.
11. Wright, C.F., FitzPatrick, D.R., and Firth, H.V. (2018). Paediatric genomics: Diagnosing rare disease in children. Nat. Rev. Genet. *19*, 253–268.
12. Hormozdiari, F., Zhu, A., Kichaev, G., Ju, C.J.-T., Segrè, A.V., Joo, J.W.J., Won, H., Sankararaman, S., Pasaniuc, B., Shifman, S., and Eskin, E. (2017). Widespread allelic heterogeneity in complex traits. Am. J. Hum. Genet. *100*, 789–802.
13. McClellan, J., and King, M.-C. (2010). Genetic heterogeneity in human disease. Cell *141*, 210–217.
14. Wright, C.F., Middleton, A., Burton, H., Cunningham, F., Humphries, S.E., Hurst, J., Birney, E., and Firth, H.V. (2013). Policy challenges of clinical genome sequencing. BMJ *347*, f6845.
15. Katsanis, N. (2016). The continuum of causality in human genetic disorders. Genome Biol. *17*, 233.
16. Minikel, E.V., Zerr, I., Collins, S.J., Ponto, C., Boyd, A., Klug, G., Karch, A., Kenny, J., Collinge, J., Takada, L.T., et al. (2014). Ascertainment bias causes false signal of anticipation in genetic prion disease. Am. J. Hum. Genet. *95*, 371–382.
17. Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N.E. (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. Am. J. Epidemiol. *186*, 1026–1034.
18. Ganguli, M., Lytle, M.E., Reynolds, M.D., and Dodge, H.H. (1998). Random versus volunteer selection for a community-based study. J. Gerontol. A Biol. Sci. Med. Sci. *53*, M39–M46.
19. DeBoever, C., Tanigawa, Y., Lindholm, M.E., McInnes, G., Lavertu, A., Ingelsson, E., Chang, C., Ashley, E.A., Bustamante, C.D., Daly, M.J., and Rivas, M.A. (2018). Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. Nat. Commun. *9*, 1612.
20. Bastarache, L., Hughey, J.J., Hebbring, S., Marlo, J., Zhao, W., Ho, W.T., Van Driest, S.L., McGregor, T.L., Mosley, J.D., Wells, Q.S., et al. (2018). Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science *359*, 1233–1239.
21. Tuke, M.A., Ruth, K.S., Wood, A.R., Beaumont, R.N., Tyrrell, J., Jones, S.E., Yaghootkar, H., Turner, C.L.S., Donohoe, M.E., Brooke, A.M., et al. (2017). Mosaic Turner syndrome shows reduced phenotypic penetrance in an adult population study compared to clinically ascertained case. bioRxiv. https://doi.org/10.1101/177659.
22. Minikel, E.V., Vallabh, S.M., Lek, M., Estrada, K., Samocha, K.E., Sathirapongsasuti, J.F., McLean, C.Y., Tung, J.Y., Yu, L.P.C., Gambetti, P., et al.; Exome Aggregation Consortium (ExAC) (2016). Quantifying prion disease penetrance using large population control cohorts. Sci. Transl. Med. *8*, 322ra9.
23. Shah, N., Hou, Y.-C.C., Yu, H.-C., Sainger, R., Caskey, C.T., Venter, J.C., and Telenti, A. (2018). Identification of misclassified clinvar variants via disease population prevalence. Am. J. Hum. Genet. *102*, 609–619.
24. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). 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.
25. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2017). Genome-wide genetic data on ∼500,000 UK Biobank participants. bioRxiv. https://doi.org/10.1101/166298.
26. Auer, P.L., and Lettre, G. (2015). Rare variant association studies: Considerations, challenges and opportunities. Genome Med. *7*, 16.
27. Harrison, S.M., Riggs, E.R., Maglott, D.R., Lee, J.M., Azzariti, D.R., Niehaus, A., Ramos, E.M., Martin, C.L., Landrum, M.J., and Rehm, H.L. (2016). Using clinvar as a resource to support variant interpretation. Curr Protoc Hum Genet *89*, 8.16.1–8.16.23.
28. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
29. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
30. Ellard, S., Lango Allen, H., De Franco, E., Flanagan, S.E., Hysenaj, G., Colclough, K., Houghton, J.A.L., Shepherd, M., Hattersley, A.T., Weedon, M.N., and Caswell, R. (2013). Improved genetic testing for monogenic diabetes using targeted next-generation sequencing. Diabetologia *56*, 1958–1963.
31. Hattersley, A.T., and Patel, K.A. (2017). Precision diabetes: Learning from monogenic diabetes. Diabetologia *60*, 769–777.
32. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al.; DDD study (2015). Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data. Lancet *385*, 1305–1314.
33. Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433–438.
34. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.
35. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.
36. Bragin, E., Chatzimichali, E.A., Wright, C.F., Hurles, M.E., Firth, H.V., Bevan, A.P., and Swaminathan, G.J. (2014). DECIPHER: Database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res. *42*, D993–D1000.
37. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789–D798.
38. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

39. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

40. Ewing, C.M., Ray, A.M., Lange, E.M., Zuhlke, K.A., Robbins, C.M., Tembe, W.D., Wiley, K.E., Isaacs, S.D., Johng, D., Wang, Y., et al. (2012). Germline mutations in HOXB13 and prostate-cancer risk. N. Engl. J. Med. *366*, 141–149.

41. Antoniou, A.C., Casadei, S., Heikkinen, T., Barrowdale, D., Pylkäs, K., Roberts, J., Lee, A., Subramanian, D., De Leeneer, K., Fostira, F., et al. (2014). Breast-cancer risk in families with mutations in PALB2. N. Engl. J. Med. *371*, 497–506.

42. Karlsson, R., Aly, M., Clements, M., Zheng, L., Adolfsson, J., Xu, J., Grönberg, H., and Wiklund, F. (2014). A population-based assessment of germline HOXB13 G84E mutation and prostate cancer risk. Eur. Urol. *65*, 169–176.

43. Erkko, H., Xia, B., Nikkilä, J., Schleutker, J., Syrjäkoski, K., Mannermaa, A., Kallioniemi, A., Pylkäs, K., Karppinen, S.-M., Rapakko, K., et al. (2007). A recurrent mutation in PALB2 in Finnish cancer families. Nature *446*, 316–319.

44. Smith, F.J.D., Irvine, A.D., Terron-Kwiatkowski, A., Sandilands, A., Campbell, L.E., Zhao, Y., Liao, H., Evans, A.T., Goudie, D.R., Lewis-Jones, S., et al. (2006). Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris. Nat. Genet. *38*, 337–342.

45. Palmer, C.N.A., Irvine, A.D., Terron-Kwiatkowski, A., Zhao, Y., Liao, H., Lee, S.P., Goudie, D.R., Sandilands, A., Campbell, L.E., Smith, F.J.D., et al. (2006). Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. Nat. Genet. *38*, 441–446.

46. Jordan, N., Williams, N., Gregory, J.W., Evans, C., Owen, M., and Ludgate, M. (2003). The W546X mutation of the thyrotropin receptor gene: Potential major contributor to thyroid dysfunction in a Caucasian population. J. Clin. Endocrinol. Metab. *88*, 1002–1005.

47. Lee, A.J., Wang, Y., Alcalay, R.N., Mejia-Santana, H., Saunders-Pullman, R., Bressman, S., Corvol, J.-C., Brice, A., Lesage, S., Mangone, G., et al.; Michael J. Fox LRRK2 Cohort Consortium (2017). Penetrance estimate of LRRK2 p.G2019S mutation in individuals of non-Ashkenazi Jewish ancestry. Mov. Disord. *32*, 1432–1438.

48. Goldwurm, S., Tunesi, S., Tesei, S., Zini, M., Sironi, F., Primignani, P., Magnani, C., and Pezzoli, G. (2011). Kin-cohort analysis of LRRK2-G2019S penetrance in Parkinson's disease. Mov. Disord. *26*, 2144–2145.

49. Wu, X., Tang, K.-F., Li, Y., Xiong, Y.-Y., Shen, L., Wei, Z.-Y., Zhou, K.-J., Niu, J.-M., Han, X., Yang, L., et al. (2012). Quantitative assessment of the effect of LRRK2 exonic variants on the risk of Parkinson's disease: A meta-analysis. Parkinsonism Relat. Disord. *18*, 722–730.

50. Zhang, L., Hirano, A., Hsu, P.-K., Jones, C.R., Sakai, N., Okuro, M., McMahon, T., Yamazaki, M., Xu, Y., Saigoh, N., et al. (2016). A PERIOD3 variant causes a circadian phenotype and is associated with a seasonal mood trait. Proc. Natl. Acad. Sci. USA *113*, E1536–E1544.

51. Jones, S.E., Lane, J.M., Wood, A.R., van Hees, V.T., Tyrrell, J., Beaumont, R.N., Jeffries, A.R., Dashti, H.S., Hillsdon, M., Ruth, K.S., et al. (2018). Genome-wide association analyses of chronotype in 697,828 individuals provides new insights into circadian rhythms in humans and links to disease. bioRxiv. https://doi.org/10.1101/303941.

52. Bevan, C.L., Brown, B.B., Davies, H.R., Evans, B.A., Hughes, I.A., and Patterson, M.N. (1996). Functional analysis of six androgen receptor mutations identified in patients with partial androgen insensitivity syndrome. Hum. Mol. Genet. *5*, 265–273.

53. Kelly, M.A., Caleshu, C., Morales, A., Buchan, J., Wolf, Z., Harrison, S.M., Cook, S., Dillon, M.W., Garcia, J., Haverfield, E., et al. (2018). Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: Recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. Genet. Med. *20*, 351–359.

54. Laver, T.W., Colclough, K., Shepherd, M., Patel, K., Houghton, J.A.L., Dusatkova, P., Pruhova, S., Morris, A.D., Palmer, C.N., McCarthy, M.I., et al. (2016). The common p.R114W HNF4A mutation causes a distinct clinical bubtype of monogenic diabetes. Diabetes *65*, 3212–3217.

55. Osbak, K.K., Colclough, K., Saint-Martin, C., Beer, N.L., Bellanné-Chantelot, C., Ellard, S., and Gloyn, A.L. (2009). Update on mutations in glucokinase (GCK), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. Hum. Mutat. *30*, 1512–1526.

56. Chakera, A.J., Steele, A.M., Gloyn, A.L., Shepherd, M.H., Shields, B., Ellard, S., and Hattersley, A.T. (2015). Recognition and management of individuals with hyperglycemia because of a heterozygous glucokinase mutation. Diabetes Care *38*, 1383–1392.

57. Kashiyama, K., Nakazawa, Y., Pilz, D.T., Guo, C., Shimada, M., Sasaki, K., Fawcett, H., Wing, J.F., Lewin, S.O., Carr, L., et al. (2013). Malfunction of nuclease ERCC1-XPF results in diverse clinical manifestations and causes Cockayne syndrome, xeroderma pigmentosum, and Fanconi anemia. Am. J. Hum. Genet. *92*, 807–819.

58. King, R.A., Willaert, R.K., Schmidt, R.M., Pietsch, J., Savage, S., Brott, M.J., Fryer, J.P., Summers, C.G., and Oetting, W.S. (2003). MC1R mutations modify the classic phenotype of oculocutaneous albinism type 2 (OCA2). Am. J. Hum. Genet. *73*, 638–645.

59. Preising, M.N., Forster, H., Tan, H., Lorenz, B., de Jong, P.T.V.M., and Plomp, A.S. (2007). Mutation analysis in a family with oculocutaneous albinism manifesting in the same generation of three branches. Mol. Vis. *13*, 1851–1855.

60. Lunetta, K.L., Day, F.R., Sulem, P., Ruth, K.S., Tung, J.Y., Hinds, D.A., Esko, T., Elks, C.E., Altmaier, E., He, C., et al.; EPIC-InterAct Consortium; and Generation Scotland (2015). Rare coding variants and X-linked loci associated with age at menarche. Nat. Commun. *6*, 7756.

61. Topaloglu, A.K., Reimann, F., Guclu, M., Yalin, A.S., Kotan, L.D., Porter, K.M., Serin, A., Mungan, N.O., Cook, J.R., Imamoglu, S., et al. (2009). TAC3 and TACR3 mutations in familial hypogonadotropic hypogonadism reveal a key role for Neurokinin B in the central control of reproduction. Nat. Genet. *41*, 354–358.

62. Origa, R. (1993). Beta-Thalassemia. In GeneReviews, R.A. Pagon, M.P. Adam, H.H. Ardinger, S.E. Wallace, A. Amemiya, L.J. Bean, T.D. Bird, C.-T. Fong, H.C. Mefford, and R.J. Smith, et al., eds. (University of Washington, Seattle). https://www.ncbi.nlm.nih.gov/pubmed/20301599.

63. Turan, S., and Bastepe, M. (2015). GNAS spectrum of disorders. Curr. Osteoporos. Rep. *13*, 146–158.

64. Zhu, J., Li, S., Ramelot, T.A., Kennedy, M.A., Liu, M., and Yang, Y. (2018). Structural insights into the impact of two holoprosencephaly-related mutations on human TGIF1 homeodomain. Biochem. Biophys. Res. Commun. *496*, 575–581.

65. Tatton-Brown, K., Loveday, C., Yost, S., Clarke, M., Ramsay, E., Zachariou, A., Elliott, A., Wylie, H., Ardissone, A., Rittinger, O., et al.; Childhood Overgrowth Collaboration (2017). Mutations in epigenetic regulation genes are a major cause of overgrowth with intellectual disability. Am. J. Hum. Genet. *100*, 725–736.

66. Jefferson, J.A., Lemmink, H.H., Hughes, A.E., Hill, C.M., Smeets, H.J., Doherty, C.C., and Maxwell, A.P. (1997). Autosomal dominant Alport syndrome linked to the type IV collage alpha 3 and alpha 4 genes (COL4A3 and COL4A4). Nephrol. Dial. Transplant. *12*, 1595–1599.

67. Heidet, L., Arrondel, C., Forestier, L., Cohen-Solal, L., Mollet, G., Gutierrez, B., Stavrou, C., Gubler, M.C., and Antignac, C. (2001). Structure of the human type IV collagen gene COL4A3 and mutations in autosomal Alport syndrome. J. Am. Soc. Nephrol. *12*, 97–106.

68. Douglas, J., Cilliers, D., Coleman, K., Tatton-Brown, K., Barker, K., Bernhard, B., Burn, J., Huson, S., Josifova, D., Lacombe, D., et al.; Childhood Overgrowth Collaboration (2007). Mutations in RNF135, a gene within the NF1 microdeletion region, cause phenotypic abnormalities including overgrowth. Nat. Genet. *39*, 963–965.

69. Horner-Johnson, W., and Bailey, D. (2013). Assessing understanding and obtaining consent from adults with intellectual disabilities for a health promotion study. J. Policy Pract. Intell. Disabil. *10*, 260–265.