bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TITLE

Assessing the pathogenicity, penetrance and expressivity of putative disease-causing variants in a population setting

AUTHORS

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

[1]Genetics of Complex Traits, Institute of Biomedical and Clinical Science, University of Exeter Medical School, RILD, Royal Devon & Exeter Hospital, Barrack Road, Exeter, EX2 5DW, UK

ABSTRACT

Over 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 may be over-estimated by clinical ascertainment. As more people undergo genome sequencing there is an increasing need to assess the true penetrance of alleles. Until recently, this was not possible in a population-based setting. Here, we use data from 388,714 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 1,244 of 4,585 (27%) putatively clinically relevant rare variants genotyped on the UKB microarray as high-quality. We defined “rare” as variants with a minor allele frequency of <0.01, and “clinically relevant” as variants that were either classified as pathogenic/likely pathogenic in ClinVar or are in genes known to cause two specific monogenic diseases in which we have some expertise: Maturity-Onset Diabetes of the Young (MODY) and severe developmental disorders (DD). We assessed the penetrance and pathogenicity of these high-quality variants by testing their association with 401 clinically-relevant traits available in UKB.

We identified 27 putatively clinically relevant rare variants associated with a UKB trait but that exhibited reduced penetrance or variable expressivity compared with their associated disease. For example, the P415A *PER3* variant that has been reported to cause familial advanced sleep phase syndrome is present at 0.5% frequency in the population and associated with an odds ratio of 1.38 for being a morning person (P=2x10^{-18}). We also observed novel associations with relevant traits for heterozygous carriers of some rare recessive conditions, e.g. heterozygous carriers of the R799W *ERCC4* variant that causes Xeroderma pigmentosum were more susceptible to sunburn (one extra sunburn episode reported, P=2x10^{-8}). Within our two disease subsets, we were able to refine the penetrance estimate for the R114W *HNF4A* variant in diabetes (only ~10% by age 40yrs) and refute the previous disease-association of *RNF135* in developmental disorders.

In conclusion, this study shows that very large population-based studies will help refine the penetrance estimates of rare variants. This information will be important for anyone receiving information about their health based on putatively pathogenic variants.

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 problems with study design, 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 there is likely to be a continuum of causality that exists for many human disorders[15]. Fortunately, rare and common disease studies suffer from opposing ascertainment biases. Clinically-ascertained cohorts are enriched for individuals with a specific clinical presentation, and will therefore tend to over-estimate the penetrance of any disease-causing variants identified[16]. In contrast, population cohorts tend to be enriched for healthy individuals (so-called “healthy volunteer” selection bias) who have both the time and ability to volunteer for a study[17,18], and will therefore tend to under-estimate penetrance. Population cohorts with 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), or one particular disease, or were too small to statistically test phenotypic penetrance. With its wealth of linked phenotypic and clinical information on >450,000 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 to assay 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, which can then be used to address biological and clinical questions.

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 variants in ClinVar that have been classified as “pathogenic/likely pathogenic” by at least one submitter[27], as well as variants in genes known to cause two specific monogenic diseases in which we have some expertise: maturity-onset diabetes of the young (MODY) and developmental disorders (DD).

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

METHODS

UKB cohort

UKB recruited over 500,000 individuals aged 37-73 years between 2006-2010 from across the UK. Participants provided a range of information via questionnaires and interviews (e.g. demographics, health status, lifestyle) and anthropometric measurements, blood pressure readings, blood, urine and saliva samples were taken for future analysis. Genotypes for single nucleotide variants (SNVs) and insertions/deletions (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 (http://UKB.ctsu.ox.ac.uk)[25]. We limited our analysis to 388,714 QC-passed white Europeans.

Variant prioritisation

Variants were annotated using Annovar[28] and MAFs were calculated using PLINK[29]. To prioritise variants of potential clinical importance, we selected those with at least one classification of pathogenicity (pathogenic or likely pathogenic) in the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/)[27], including those with conflicting classifications. In addition, irrespective of their presence in ClinVar, we selected predicted protein truncating variants (PTV; stopgain SNVs and frameshift indels) and known pathogenic functional variants (nonsynonymous SNVs and inframe indels) in genes known to cause MODY[30,31] (https://www.diabetesgenes.org) and dominant DD[32,33] (https://www.ebi.ac.uk/gene2phenotype) for detailed evaluation. These diseases and genes were selected due to our own prior experience, the availability of well-curated gene lists that include mode of inheritance and mechanism of action, and the different priors associated with finding diabetes (a common disease) and severe DD (a rare disease) in UKB. We excluded common variants (MAF>0.01), as these have already been thoroughly investigated through genome-wide association studies[34,35], and further refined the list of variants to include only those where the Hardy–Weinberg equilibrium (HWE) P>0.05 and the proportion of missing genotypes across all samples <0.01 (n=4,585).

Assessing analytical validity

To assess the analytical validity of these variants, we used Evoker Lite (https://github.com/dlrice/evoker-lite) to generate cluster plots of intensities, and 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/uncertain; 4=good quality, clearly separable clusters but noisy data; and 5=good quality, clear separation between clusters (**Supplementary Figure 1**). In an initial subset of 750 variants that was independently evaluated by two scientists (**Supplementary Figure 2**), 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 $\geq$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 genotypes calls at these variants were entirely due to calls from a single

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

batch. Across the 1,244 variants, the highest proportion of rare genotype calls in a single batch was 4 from a total of 13 for Affx-89007317. A plot of total allele count for each variant against maximum allele count across each individual batch demonstrated a linear association with no clear outlying variants.

Assessing clinical relevance

We ran a phenome-wide association for all of our 1,244 high-quality rare variants against a curated list of 401 clinically-relevant traits in UKB (**Supplementary Table 1**) in 388,714 QC-passed white Europeans using PLINK[29], and those with a Bonferroni-corrected $p < 1x10^{-7}$ (0.05/(401*1244)) were prioritised for detailed evaluation. For continuous traits, we used linear regression adjusting for age, sex (unless a sex-specific trait), centre, chip and ten ancestry principal components. For binary traits, we used Fisher's exact test as the primary association method and performed logistic regression adjusting for the same covariates as for continuous traits as a sensitivity analysis. We excluded variants now considered by recent reclassifications in ClinVar to be benign. 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 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 (1Mb 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 0.00001 frequency are not reliably genotyped

Across all the variants evaluated for analytical validity using combined cluster plots (n unique=4,585, see Methods), we categorised 27% as high quality (average score$\geq$4), 64% as low quality likely false positives (average score$\leq$2.5), and 9% as unclear (**Table 1**). There was a strong correlation between the analytical validity quality score and MAF (**Table 1** and **Figure 1**), as well as 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% CI = 0.943-0.956); the false positive rate (FPR) at MAF>0.00005 was ~20%, while FPR~60% at MAF>0.00001.

Reduced penetrance estimates for known pathogenic variants

The 1,244 high-quality putative pathogenic rare variants, with their ClinVar-associated disease and the allele frequencies in UKB and gnomAD, are shown in **Supplementary Table 2**. Of these variants only 27 were associated with one of the 401 traits we tested against in UKB with $p<1x10^{-7}$ (**Table 2**). Of these, 13 have previously been linked with a dominant disease. For two variants, where penetrance had previously been estimated from large clinical cohorts[40,41], we found substantially reduced penetrance in our population-based study. Specifically, we observed well-established associations between variants in *PALB2*[40] and *HOXB13*[41] and breast cancer and prostate cancer respectively, where the odds ratios were around half the previous estimates from family-based disease studies (in both cases, ~4.5 in UKB

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

versus ~9.5 in the family-based studies)[40,41]. The other 11 variants were causally linked to disease, but penetrance estimates were not available from the literature for comparison. However, we observed that these variants were associated with a related trait in our population-based cohort (**Table 2**), suggesting reduced penetrance versus their presumed monogenic forms. Two PTVs in *FLG* that cause ichthyosis vulgaris[42] were associated with a 2-fold increased odds of Eczema. A PTV in *TSHR* that causes nonautoimmune hyperthyroidism[43] was associated with a 3-fold increased odds of hypothyroidism. A nonsynonymous variant in *LRRK2* that causes Parkinson's disease[44] was associated with a 5-fold odds of having a parent with Parkinson's disease. A nonsynonymous variant in *PER3* previously classified as pathogenic for advanced sleep phase syndrome had an odds ratio of only 1.38 for being a morning person[45,46] compared to a reported 2 hour shift in midpoint sleep. Height, skeletal weight and male pattern baldness were negatively associated with two nonsynonymous variants in *AR* that cause partial androgen insensitivity syndrome[47]. Finally, a nonsynonymous variant in *MYH7*, which has been classified by a ClinGen Expert Panel as pathogenic for hypertrophic cardiomyopathy[48] was associated with a reduced pulse rate of 5 beats per minute.

We specifically investigated known pathogenic variants and PTVs in MODY genes, where we found two rare variants that were high quality, definitely pathogenic and strongly associated with diabetes (**Table 2**): a very rare stop-gain variant in *GCK* (OR=68 95% CI: 14, 328, $P=2x10^{-8}$), and a nonsynonymous variant (p.R114W) in *HNF4A* (OR=2.9 95% CI: 1.7, 5.0, $P=3x10^{-4}$). Both associated with diabetes in UKB, in-line with previous findings[49–51]. However, the penetrance of the *HNF4A* variant was previously estimated to be up to 75% at age 40-years based on a large MODY diabetes cohort[49], while 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 the interpretation of genetic test results: if the R114W variant was found in an affected individual following clinical testing, it may still be the primary cause of their diabetes, while 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 variant in *ERCC4*, which causes recessive xeroderma pigmentosum[52], and two nonsynonymous variants in *OCA*, which causes oculocutaneous albinism[53,54] were associated with ease of sunburn. A stopgain variant in *TACR3*, which causes recessive hypogonadotropic hypogonadism[55,56] was associated with an 8 month increase in age at menarche. In addition, variants in six genes known to cause different recessive blood-related disorders were associated with decreased mean corpuscular volume and/or increased red blood cell distribution width (including *HBB* which causes β-thalassemia but where the carrier state is already known to cause the much milder β-thalassemia minor[57]).

Refuting previous disease associations

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 in genes that are

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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* is the only one of the five genes with a high probability of being loss-of-function intolerant (pLI)[38] based on the frequency of loss-of-function variants in ExAC[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 variants in *GNAS* that cause Albright hereditary osteodystrophy are located in later, highly constrained exons[58]. Similarly, the stop-gain variant in *TGIF1* is located in the first exon of the gene, where multiple PTVs in gnomAD[38] are also located, while *TGIF1* pathogenic variants causing holoprosencephaly are located in the final exons of the gene where they affect DNA binding affinity[59]. Finally, a frameshift deletion in *HIST1H1E* is located near the start of the single exon of this gene; however, in contrast, 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[60]. Hence, we believe that these three rare PTVs are benign due to their location, despite being PTVs in genes that cause dominant DD via haploinsufficiency.

For the other three 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* – where pathogenic variants are thought to cause a rare dominant form of Alport syndrome (as well as benign familial hematuria)[61,62] – 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* – where haploinsufficiency is thought to cause macrocephaly, macrosomia and facial dysmorphism syndrome[63] – 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, coupled 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.

CONCLUSIONS

Previous studies have been unable to analyse 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 minimised the impact of UKB in the rare disease community. We have established a method for evaluating the analytical validity of rare variants genotyped by microarray, using combined intensity plots for individual variants across all genotyping batches. Although we initially tried to examine variant cluster plots for each batch separately, as recommended by UKB, this proved impossible due to the rarity of most clinically important variants. MAF was an extremely good predictor of the likelihood of a variant being genotyped well by the UKB arrays (**Figure 1**). At MAF>0.00005 (~50 heterozygous individuals) FPR~7% and most variants were well genotyped, while FPR~60% at MAF>0.00001

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(~10 heterozygous individuals), and we classified all variants at MAF< 0.000005 (~5 heterozygous individuals) as being low quality. This has important implications for epidemiological research carried out uncritically using 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 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**, plus additional well-established associations for variants where MAF>0.01). 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 associated with appropriate traits in UKB (**Table 2**) and 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, due to the rarity of the condition, the relatively senior age of the UKB population and the inherent challenges of consenting individuals with severe DD to population biobanks[64]. We are therefore confident that the PTV variants identified in dominant DD genes in UKB are benign (**Table 3**), and in refuting previous associations between haploinsufficiency in *RNF135* and *COL4A3* and dominant DD (which 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 to estimate lower bounds for the penetrance of pathogenic disease-causing variants, and refine our understanding of the links between rare variants (MAF<0.01) and monogenic diseases. Performing a similar analysis on ultra-rare variants (MAF<0.00001) 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 interpreting incidental or secondary findings from clinical testing, and for 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.

ACKNOWLEDGEMENTS

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This research has been conducted using the UK Biobank Resource. This work was carried out under UK Biobank project number 871.

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TABLES

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

Table 2. Pathogenic variants. Reduced penetrance, variable expressivity and carrier phenotypes for rare (MAF<0.01) ClinVar pathogenic variants with genome-wide significant associations in UKB.

Table 3. Benign variants. Classification of likely pathogenic variants in maturity-onset diabetes of the young (MODY) and developmental disorders (DD) from UKB.

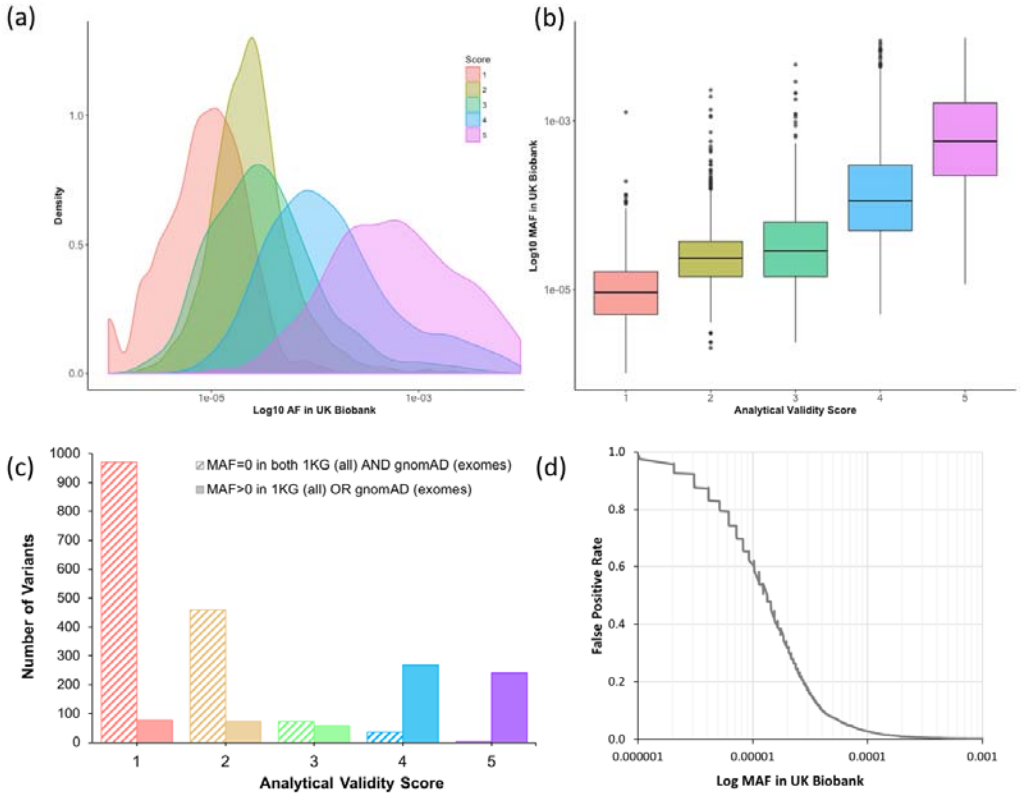
Supplementary Table 1. Curated traits included from UKB.

Supplementary Table 2. 1,244 high quality putative pathogenic variants analysed.

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

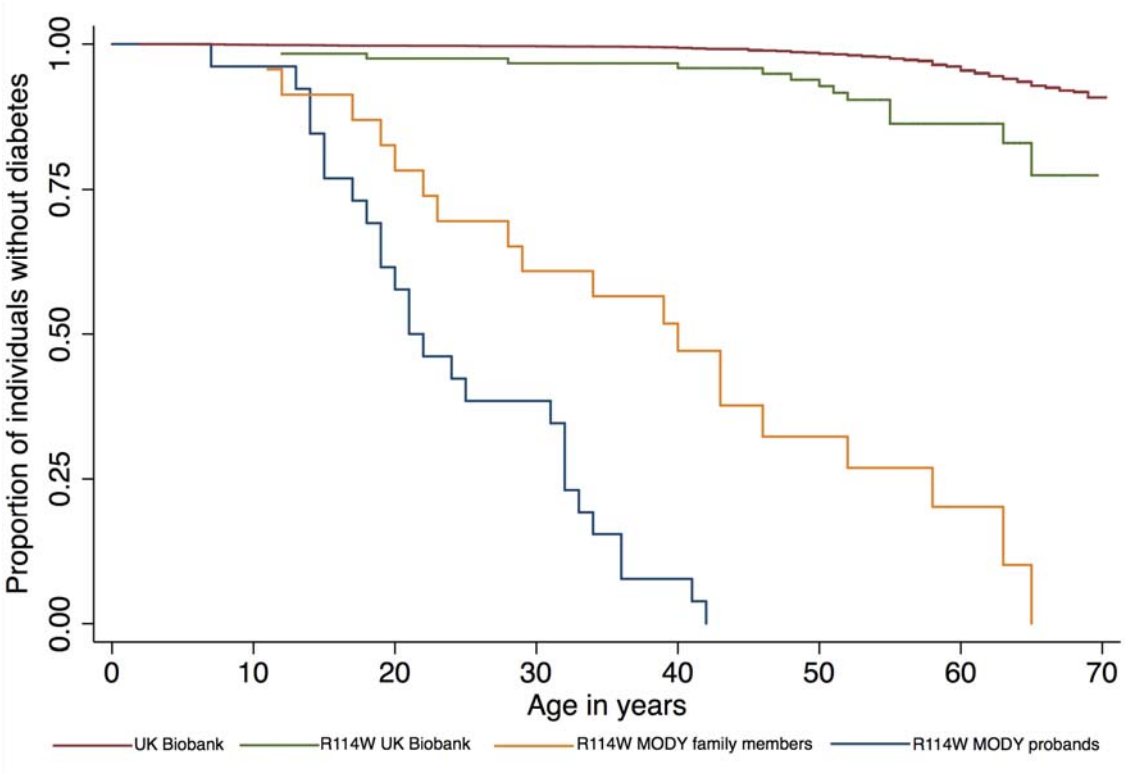
FIGURES

Figure 1. Correlation between MAF and analytical validity quality score. **(a)** Density plot and **(b)** boxplot of manual quality scores (from 1-5, see Supplementary Figure 1) of genotype data in UKB versus minor allele frequency (MAF) for 4,585 putatively clinically important variants, where MAF<0.01, 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; **(d)** Estimation of the false positive rate (FPR) versus MAF for variants assayed using the UKB genotyping arrays, calculated by grouping quality scores into low (score=1 or 2) and high (score=4 or 5) and using the rocreg command in Stata to fit a ROC curve. Red=score 1; gold=score 2; green = score 3; blue = score 4; purple = score 5.



bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2. Penetrance estimate for *HNF4A* p.R114W in UK Biobank compared to previously published estimates from MODY cohort studies. A Kaplan-Meier plot of proportion of individuals that are diabetes-free against age for 388,174 individuals from UK Biobank (red line); 122 UK Biobank individuals that are heterozygous for *HNF4A* p.R114W (green line) and 26 MODY referral probands (blue line) and 24 family members of the probands (yellow line) from Laver *et al.* [49]



bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SUPPLENTARY FIGURES

Supplementary Figure 1. Combined cluster intensity plots.
Intensity plots combined across all batches are shown for five variants, all with a UKB MAF = 0.00004. The clustering quality of heterozygous variants was manually assessed and ranked from 1-5. **(a)** Score 1 = poor quality, no discernible separate clusters; **(b)** Score 2 = poor quality, no discernible separate clusters but noisy data; **(c)** Score 3 = unclear/uncertain; **(d)** Score 4 = good quality, clearly separable clusters but noisy data; **(e)** Score 5 = good quality, clear separation between clusters.

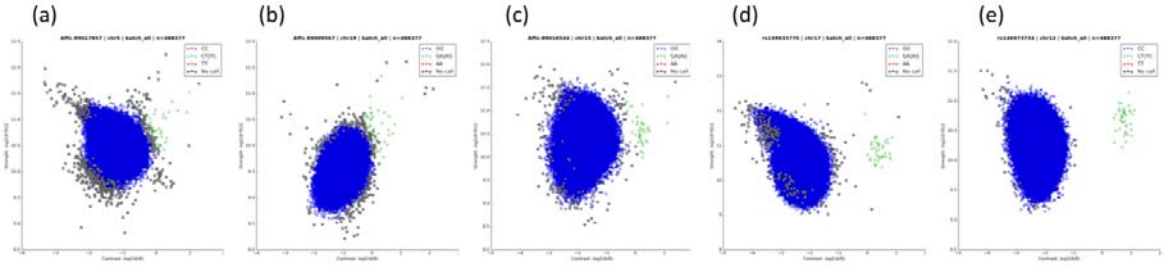
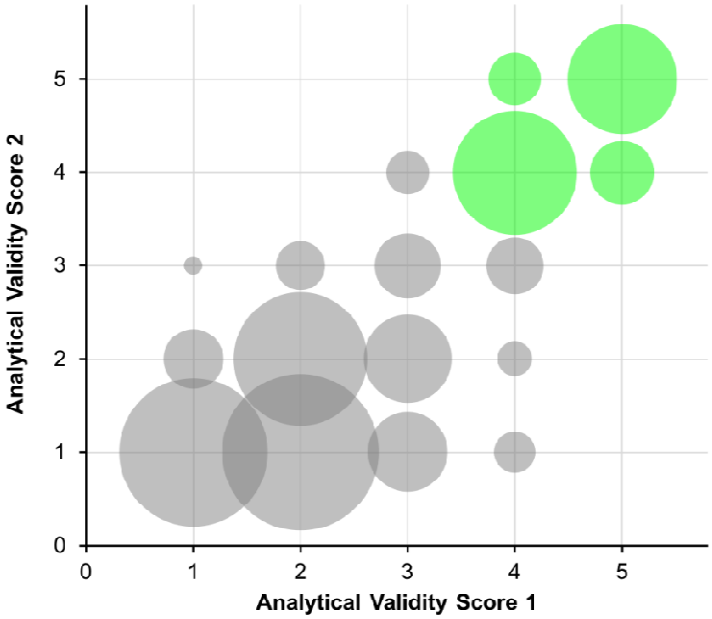


Figure 2. Comparison of quality scores between two independent scorers.
Two scientists independently scored the quality (from 1-5, see Figure 1) of combined cluster plots for 750 variants. The R^2 between their scores was 0.8, and there was a 95% agreement in low quality (score=1 or 2) versus high quality (score=4 or 5) variants.



bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.
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., et al. (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*,

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

M39-46.
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., et al. (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.
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. (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.
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., et al. (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., et al. (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. (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.

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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*, D789-98.
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. (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. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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*, E1536-44.
46. 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.
47. 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.
48. 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.
49. Laver, T.W., Colclough, K., Shepherd, M., Patel, K., Houghton, J.A.L., Dusatkova, P.,

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Pruhova, S., Morris, A.D., Palmer, C.N., McCarthy, M.I., et al. (2016). The Common p.R114W HNF4A Mutation Causes a Distinct Clinical Subtype of Monogenic Diabetes. Diabetes *65*, 3212–3217.
50. 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.
51. 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.
52. 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.
53. 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.
54. 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.
55. 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. (2015). Rare coding variants and X-linked loci associated with age at menarche. Nat. Commun. *6*, 7756.
56. 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.
57. 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, R.J. Smith, et al., eds. (Seattle (WA): University of Washington, Seattle), p.
58. Turan, S., and Bastepe, M. (2015). GNAS spectrum of disorders. Curr. Osteoporos. Rep. *13*, 146–158.
59. 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.
60. Tatton-Brown, K., Loveday, C., Yost, S., Clarke, M., Ramsay, E., Zachariou, A., Elliott, A., Wylie, H., Ardissone, A., Rittinger, O., et al. (2017). Mutations in Epigenetic Regulation Genes Are a Major Cause of Overgrowth with Intellectual Disability. Am. J. Hum. Genet. *100*, 725–736.
61. 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.
62. 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.
63. Douglas, J., Cilliers, D., Coleman, K., Tatton-Brown, K., Barker, K., Bernhard, B., Burn, J., Huson, S., Josifova, D., Lacombe, D., et al. (2007). Mutations in RNF135, a gene within the NF1 microdeletion region, cause phenotypic abnormalities including overgrowth. Nat. Genet. *39*, 963–965.

bioRxiv preprint doi: https://doi.org/10.1101/407981; this version posted September 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

64. 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. Intellect. Disabil. *10*, 260–265.