METHOD

PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins

Nadav Brandes[1*] , Nathan Linial[1] and Michal Linial[2*]

* Correspondence: nadav.brandes@
mail.huji.ac.il; michall@cc.huji.ac.il
[1]School of Computer Science and
Engineering, The Hebrew University
of Jerusalem, Jerusalem, Israel
[2]Department of Biological
Chemistry, The Alexander Silberman
Institute of Life Sciences, The
Hebrew University of Jerusalem,
Jerusalem, Israel

Abstract

We introduce Proteome-Wide Association Study (PWAS), a new method for detecting gene-phenotype associations mediated by protein function alterations. PWAS aggregates the signal of all variants jointly affecting a protein-coding gene and assesses their overall impact on the protein's function using machine learning and probabilistic models. Subsequently, it tests whether the gene exhibits functional variability between individuals that correlates with the phenotype of interest. PWAS can capture complex modes of heritability, including recessive inheritance. A comparison with GWAS and other existing methods proves its capacity to recover causal protein-coding genes and highlight new associations. PWAS is available as a command-line tool.

Keywords: GWAS, Machine learning, Protein function, Protein annotations, UK Biobank, Recessive heritability

Background

Genome-wide association studies (GWAS) seek to robustly link genetic loci with diseases and other heritable traits [1–4]. In the past decade, the method has implicated numerous variant-phenotype associations [5] and driven important scientific discovery [6, 7]. Nowadays, thanks to the rapid development of large-scale biobanks with well-genotyped and well-phenotyped cohorts, conducting case-control studies has become easier than ever. The UK Biobank (UKBB) is a flagship project of these efforts, having recruited a cohort of over 500,000 individuals, each with a full genotype and thousands of curated phenotypes (including medical history, lab tests, a variety of physical measures and comprehensive lifestyle questionnaires) [8, 9].

Despite the enormous impact of GWAS, inherent difficulties still limit its success [2, 10]. Among the key factors is its limited statistical power, partly caused by the large number of tested variants across the genome. This limiting factor is especially crucial when dealing with rare variants of small effect sizes [10]. Due to linkage disequilibrium

Brandes *et al. Genome Biology* (2020) 21:173

Page 2 of 22

(LD) and population stratification, even when a genomic locus is robustly implicated with a phenotype, pinning the exact causal variants is a convoluted task [6].

Three major strategies are commonly used for prioritizing the most likely entities (e.g., variants or genes) causally affecting the phenotype. The most common strategy is the fine-mapping of raw GWAS results [11–13]. Fine-mapping of GWAS summary statistics often relies on functional annotations of the genome, under the assumption that functional entities are more likely to be causal. However, even following fine-mapping, many of the significant GWAS associations remain without any known biological mechanistic interpretation.

To arrive at more interpretable, actionable discoveries, another commonly used strategy is to prioritize genes (or other functional entities) rather than variants. There are numerous methods that aggregate GWAS summary statistics at the level of genes, often by combining them with data from expression quantitative trait locus (eQTL) studies or functional annotations of genes and pathways [14–17].

The third strategy seeks to implicate genes directly, by carrying the association tests at the level of annotated functional elements in the first place. The most commonly used gene-level method is SKAT, which aggregates the signal across an entire genomic region, be it a gene or any other functional entity (or just a collection of SNPs) [18, 19]. Another approach, recently explored by methods such as PrediXcan [20] and TWAS [21], tests whether the studied phenotypes correlate with gene expression levels predicted from genetic variants. Under this paradigm, the association test comprised two stages. First, an independent reference panel is used to train a prediction model of gene expression (in a particular tissue) as a function of the genetic makeup of a sample. The learned model is then applied on the phenotyped dataset, and the predicted gene expression levels are tested against phenotypes of interest. The advantages of this approach include a reduced burden of multiple testing and more concrete and interpretable discoveries.

A natural enhancement to these approaches would be a protein-centric method that considers the effects of genetic variants on the functionality of genes, rather than affecting their abundance (be it at the transcript or protein level).

We present Proteome Wide Association Study (PWAS) (Fig. 1). PWAS is based on the premise that causal variants in coding regions affect phenotypes by altering the biochemical functions of the genes' protein products (Fig. 1a). Such functional alterations could be, for example, changes to a protein's enzymatic activity or binding capacity (e.g., of a ligand, DNA/RNA molecule, or another protein). To capture these effects, PWAS quantifies the extent to which proteins are damaged given an individual's genotype. Specifically, PWAS considers any variant that affects the coding regions of genes (e.g., missense, nonsense, frameshift). It quantifies the impact of these variants on the function of the affected proteins using FIRM, a machine learning model that considers the rich proteomic context each affecting variant [22]. These predicted effects are then combined with the genotyping data of the cohort and aggregated into per-gene functional predictions, assigning each protein-coding gene functional effect scores (Fig. 1b). For each gene (in the context of a specific individual), PWAS assigns two scores, to cover the two elementary modes of heritability: dominant and recessive inheritance (other modes of heritability, including additive effects, can be represented as a composition of the two). Intuitively, the dominant effect score is intended to express the
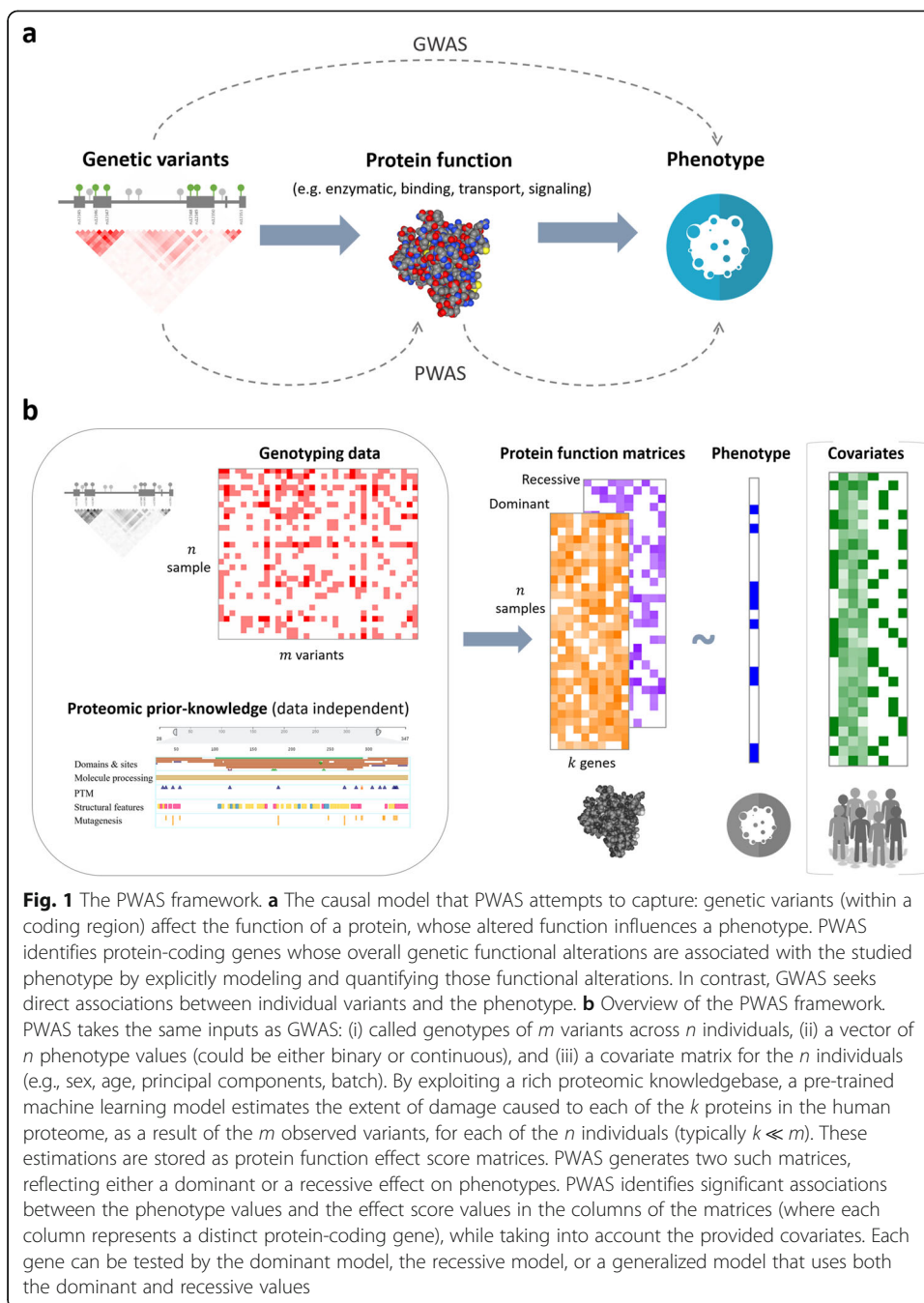
Fig. 1 The PWAS framework. **a** The causal model that PWAS attempts to capture: genetic variants (within a coding region) affect the function of a protein, whose altered function influences a phenotype. PWAS identifies protein-coding genes whose overall genetic functional alterations are associated with the studied phenotype by explicitly modeling and quantifying those functional alterations. In contrast, GWAS seeks direct associations between individual variants and the phenotype. **b** Overview of the PWAS framework. PWAS takes the same inputs as GWAS: (i) called genotypes of m variants across n individuals, (ii) a vector of n phenotype values (could be either binary or continuous), and (iii) a covariate matrix for the n individuals (e.g., sex, age, principal components, batch). By exploiting a rich proteomic knowledgebase, a pre-trained machine learning model estimates the extent of damage caused to each of the k proteins in the human proteome, as a result of the m observed variants, for each of the n individuals (typically $k \ll m$). These estimations are stored as protein function effect score matrices. PWAS generates two such matrices, reflecting either a dominant or a recessive effect on phenotypes. PWAS identifies significant associations between the phenotype values and the effect score values in the columns of the matrices (where each column represents a distinct protein-coding gene), while taking into account the provided covariates. Each gene can be tested by the dominant model, the recessive model, or a generalized model that uses both the dominant and recessive values

probability of at least one hit damaging the protein function, while the recessive score attempts to express the probability of at least two damaging hits. PWAS then tests, using routine statistical analysis, if a gene's effect scores are associated with the phenotype. In the case of a binary phenotype, a significant correlation would mean that the effect scores of cases are different than those of controls, namely that the protein is more (or less) damaged in affected individuals.

Like other gene-based approaches, PWAS enjoys a reduced burden of multiple testing correction. In addition, it provides concrete functional interpretations for the protein-coding genes it discovers (Fig. 1a). By aggregating the signal spread across all

the variants affecting the same gene, it can uncover associations that would remain un-detectable at the per-variant resolution, especially when rare variants are involved.

To examine the properties of PWAS, we first test it on simulated data, analyzing its statistical power across different settings. We then test it on real data derived from the UKBB, to demonstrate its wide applicability across a diverse set of phenotypes. We further compare the results of PWAS to established methods, specifically to standard GWAS and SKAT. Finally, we highlight the associations uniquely discovered by PWAS.

PWAS is available at https://github.com/nadavbra/pwas.

Results

Functional effect scores

We analyzed a cohort derived from the UKBB. Of ~ 18K analyzed protein-coding genes, 17,843 were affected by at least one non-synonymous variant reported in the UKBB. On average, each of these genes was affected by 35.9 such variants (Fig. 2a).
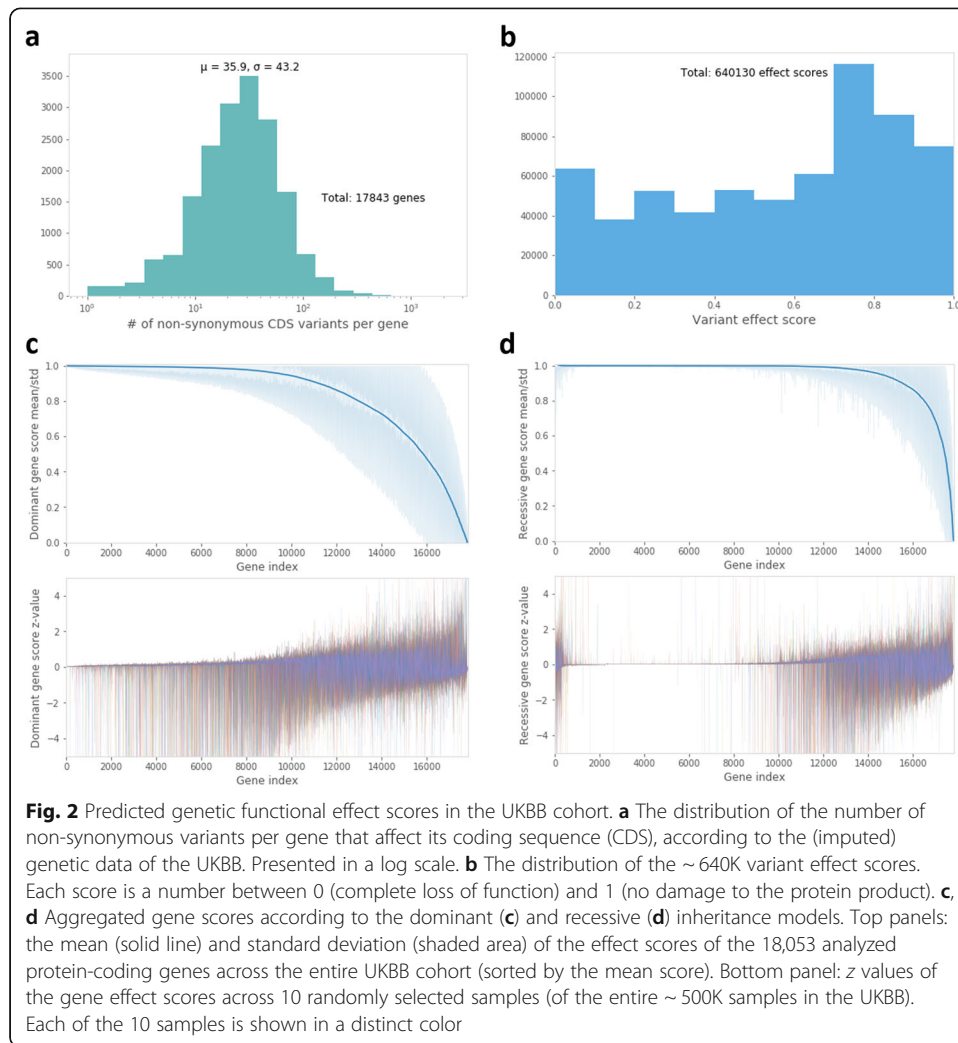
The derivation of the gene effect score matrices is comprised of two steps. First, FIRM is used to predict an effect score for each protein-affecting variant (Fig. 2b). In-tuitively, these predicted effect scores can be interpreted as the probability of the variant-affected protein to retain its function. The variant scores are then integrated with the cohort genotypes and aggregated together to derive per-sample dominant and recessive effect scores at the gene level (Fig. 2c, d). As expected, dominant genetic ef-fects (capturing single hits) are more prevalent than recessive effects (of double hits). The derived gene scores capture genetic variability in the UKBB population observed even within a small number of samples. The objective of PWAS is to test whether this functional genetic variability correlates with phenotypes.

Simulation analysis

To examine the discovery potential of PWAS compared to GWAS and SKAT, we con-ducted a simulation analysis (Fig. 3). The simulation was carried on real genetic data (from the UKBB cohort), with phenotypes simulated by mixing genetic signal and noise. To test the sensitivity of PWAS to the inevitable inaccuracies of FIRM, we exam-ined the effect of a noise parameter (ϵ) influencing its predictions. Specifically, we dis-torted the variant effect scores predicted by FIRM (in the range between 0 and 1) with additive Gaussian noise of standard deviation ϵ. It appears that under the modeling as-sumptions of the simulation, PWAS is not very sensitive to limited inaccuracies of the underlying machine learning predictor.

Based on the simulation results, we expect the advantage of PWAS to be the most substantial when dealing with recessive inheritance. We find that with small effect size ($\beta = 0.01$), at least 100K samples are required to obtain sufficient statistical power (given 173 covariates). When the effect size is higher ($\beta = 0.05$), cohorts of 10K samples could be sufficient.

It is important to state that phenotypes were simulated from the genetic data by a modeling scheme compatible with the assumptions of PWAS. Therefore, these results should not be seen as evidence for the dominance of PWAS over GWAS or SKAT in the real world. Rather, these simulations simply examine the method's range of applic-ability and assess the amount of data required for sufficient statistical power under the

Fig. 2 Predicted genetic functional effect scores in the UKBB cohort. **a** The distribution of the number of non-synonymous variants per gene that affect its coding sequence (CDS), according to the (imputed) genetic data of the UKBB. Presented in a log scale. **b** The distribution of the ~ 640K variant effect scores. Each score is a number between 0 (complete loss of function) and 1 (no damage to the protein product). **c**, **d** Aggregated gene scores according to the dominant (**c**) and recessive (**d**) inheritance models. Top panels: the mean (solid line) and standard deviation (shaded area) of the effect scores of the 18,053 analyzed protein-coding genes across the entire UKBB cohort (sorted by the mean score). Bottom panel: *z* values of the gene effect scores across 10 randomly selected samples (of the entire ~ 500K samples in the UKBB). Each of the 10 samples is shown in a distinct color
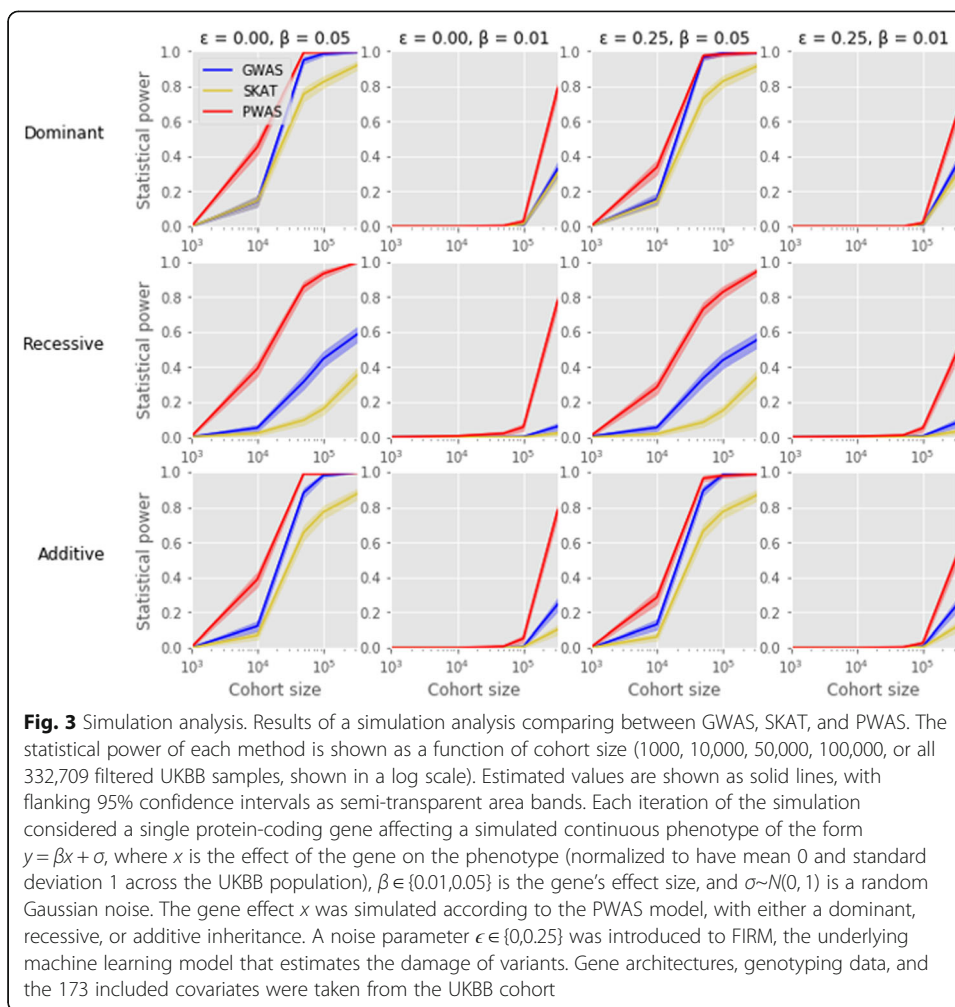
settings for which it was designed. In addition to this protein-centric modeling scheme, we also examined phenotypes simulated under a standard linear model, as well as binary phenotypes (Additional file 1: Fig. S1).

Case study: colorectal cancer

To examine PWAS on real phenotypes, we begin with a case study of colorectal cancer. A cohort of 260,127 controls and 2822 cases was derived from the UKBB to detect predisposition genes leading to increased risk of colorectal cancer through germline variants.

To exemplify how PWAS works, we begin with a demonstration of the analysis over a specific gene—*MUTYH* (Fig. 4a), a well-known predisposition gene for colorectal cancer [23]. In the studied cohort, there are 47 non-synonymous variants affecting the gene's protein sequence. When considered by standard per-variant GWAS, the most significant of these variants yields a *p* value of 1.2E−03. Even if the entire flanking region of the gene is considered (up to 500,000 bp from each side of its open reading frame), the strongest significance obtained is still only *p* = 6.3E−04, far from the exome-

Fig. 3 Simulation analysis. Results of a simulation analysis comparing between GWAS, SKAT, and PWAS. The statistical power of each method is shown as a function of cohort size (1000, 10,000, 50,000, 100,000, or all 332,709 filtered UKBB samples, shown in a log scale). Estimated values are shown as solid lines, with flanking 95% confidence intervals as semi-transparent area bands. Each iteration of the simulation considered a single protein-coding gene affecting a simulated continuous phenotype of the form $y = \beta x + \sigma$, where x is the effect of the gene on the phenotype (normalized to have mean 0 and standard deviation 1 across the UKBB population), $\beta \in \{0.01, 0.05\}$ is the gene's effect size, and $\sigma \sim N(0, 1)$ is a random Gaussian noise. The gene effect x was simulated according to the PWAS model, with either a dominant, recessive, or additive inheritance. A noise parameter $\epsilon \in \{0, 0.25\}$ was introduced to FIRM, the underlying machine learning model that estimates the damage of variants. Gene architectures, genotyping data, and the 173 included covariates were taken from the UKBB cohort

wide significance threshold (5E−07). When analyzed by PWAS, on the other hand, this association exhibits overwhelming significance (FDR q value = 2.3E−06), far beyond the commonly used FDR significance threshold ($q < 0.05$).

PWAS was able to uncover the association by aggregating signal spread across a large number of different variants, with 5 of the 47 protein-affecting variants showing mild associations ($p < 0.05$). As these 5 variants show consistent directionality (all risk increasing), and as most of them are predicted to be likely damaging, they were effectively aggregated into gene scores that significantly differ between cases and controls. Specifically, the *MUTYH* gene is significantly more damaged in cases than in controls according to the PWAS framework. The association is only significant according to the recessive model, with an estimated effect size of $d = -0.079$ (standardized mean difference in the gene effect scores between cases and controls). This observation is consistent with previous reports about *MUTYH*, claiming a recessive inheritance mode [23].

To recover all protein-coding genes associated with colorectal cancer according to PWAS, we analyzed 18,053 genes (Fig. 4b), discovering 6 significant associations (Table 1). Of these 6 associations, 5 are supported by some literature evidence, 3 of which with level of evidence we consider strong. In 4 of the 5 supported associations, the directionality of the association reported in the literature (i.e., protective or risk
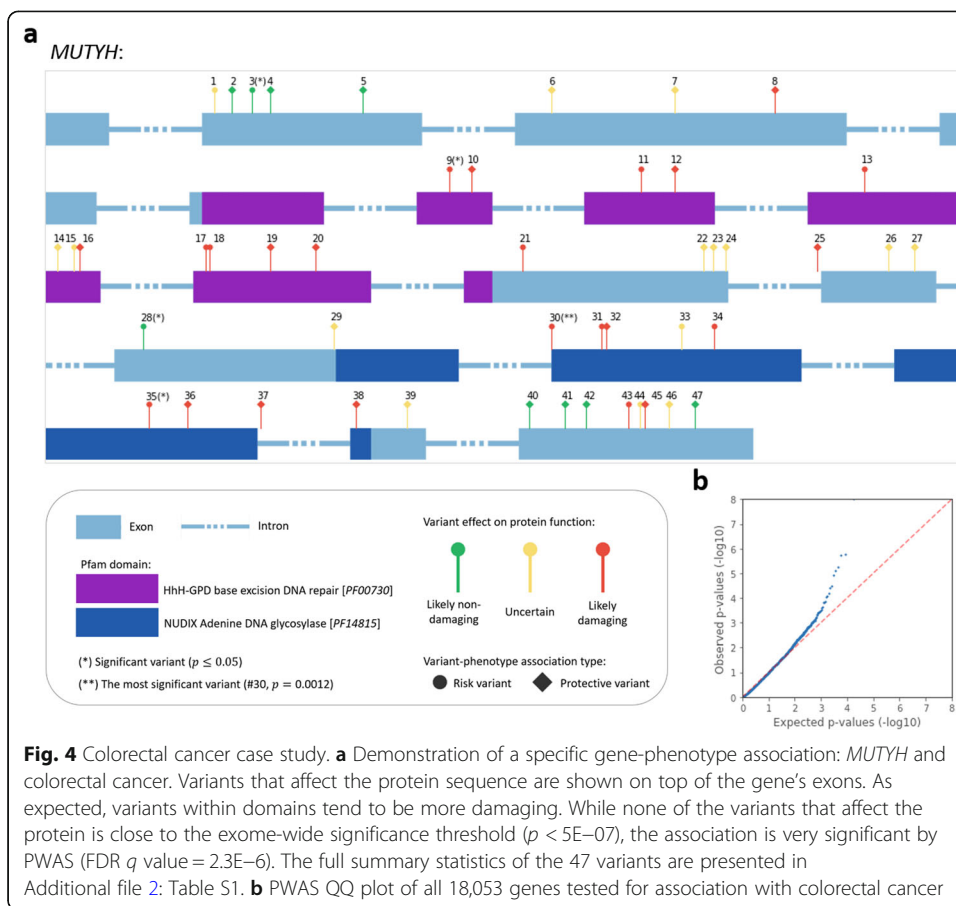
Fig. 4 Colorectal cancer case study. **a** Demonstration of a specific gene-phenotype association: *MUTYH* and colorectal cancer. Variants that affect the protein sequence are shown on top of the gene's exons. As expected, variants within domains tend to be more damaging. While none of the variants that affect the protein is close to the exome-wide significance threshold ($p < 5E–07$), the association is very significant by PWAS (FDR q value = 2.3E–6). The full summary statistics of the 47 variants are presented in Additional file 2: Table S1. **b** PWAS QQ plot of all 18,053 genes tested for association with colorectal cancer

gene) agrees with the effect size (Cohen's *d*) detected by PWAS (only in *POU5F1B* it is inversed). Of the 6 genes, only *POU5F1B* is affected by a variant exceeding the exome-wide significance (rs6998061, $p = 1.4E–07$). The 5 other genes are not discovered by GWAS, even when considering all the variants in the gene's region (up to 500,000 bp away from the gene). Notably, while GWAS determines significance by the Bonferroni-corrected significance level ($p < 5E–07$ for coding regions), PWAS determines significance by FDR ($q < 0.05$), like other gene-based methods.

Applicability of PWAS across 49 different phenotypes

Having case studied PWAS for a specific phenotype, we turn to consider its applicability for a diverse set of 49 prominent phenotypes (Fig. 5a). We applied both standard GWAS and PWAS across the 49 phenotypes on the same UKBB cohort (~ 330K samples), obtaining a rich collection of associations (Fig. 5b, c). Altogether, PWAS discovered 12,444 gene-phenotype associations, only 5294 of which (43%) contain a GWAS-significant non-synonymous variant in the gene's coding region (Fig. 5b). In other words, although PWAS considers the exact same set of variants, in 57% of the associations, it is able to recover an aggregated signal that is overlooked by GWAS when considering each of the variants individually. Even when considering all the variants in the proximity of the gene to account for LD (up to

500,000 bp to each side of the coding region), 2743 of the 12,444 PWAS associations (22%) are still missed by GWAS (Fig. 5c, d).

Full summary of all 49 tested phenotypes, with complete per-gene summary statistics, is available in Additional file 3: Table S2 (for all the significant PWAS associations) and Additional file 4: Table S3 (with all 18,053 tested protein-coding genes). QQ plots of all 49 phenotypes are available in Additional file 1: Fig. S2.

To confirm the importance of the predicted functional effect scores assigned to variants, we tested the performance of a version of PWAS where the effect scores of non-synonymous variants were shuffled prior to their aggregation into gene scores. Indeed, we find that the original version of PWAS (capturing gene function) outperforms the shuffled version (Additional file 1: Fig. S3).

Comparison with SKAT

Having established the discovery power of PWAS beyond standard GWAS, we also compare it to SKAT [18], the most commonly used method for detecting genetic associations at the gene level. Importantly, whereas SKAT attempts to recover all existing genetic associations, PWAS focuses specifically on protein-coding genes that are associated with a phenotype through protein function.

We find that PWAS is superior to SKAT in the number of discovered associations for most phenotypes (Fig. 6a). We also examined the extent of overlap between the results reported by each of the two methods (see the "consensus" bars in Fig. 6a). It appears that PWAS and SKAT tend to recover distinct sets of genes, so the two methods can be considered as largely complementary.

To assess the quality of discoveries, we appeal to Open Targets Platform (OTP) [32], an exhaustive resource curating established gene-disease associations based on multiple layers of evidence, and OMIM [33], the most prominent catalog of human genes implicated in genetic disorders. We compared the quality of associations discovered by the two methods, according to OTP-derived evidence scores, across the 24 tested diseases that are recorded in OTP (Fig. 6b). According to this metric, the results of PWAS and SKAT appear to be largely comparable, with consensus genes showing stronger evidence.

We further investigate how the two methods (PWAS and SKAT) recover externally validated associations provided by OTP (Fig. 6c) and OMIM (Fig. 6d). Of 4944 associations with strong support by OTP, 9 were recovered by SKAT compared to 6 recovered by PWAS. In the case of OMIM, which provides an even more restricted list of 202 high-quality gene-disease associations with known molecular basis, PWAS was somewhat superior (12 compared to 7 recovered associations, with the 7 being a subset of the 12). We observe no obvious trend between the types of phenotypes (e.g., cancer or other diseases) and the significance of associations obtained by the two methods (see colors in Fig. 6c, d).

Based on this comparative analysis, we conclude that PWAS and SKAT are complementary and that it may be advantageous to use both in association studies. We stress that the two methods are very distinct in the type of associations they seek and how they model them.

Highly significant associations not dominated by single variants

Among all the discovered associations, we seek to highlight those that are particularly characteristic to our new method, namely results that are uniquely discovered by

Table 1 Significant colorectal cancer genes detected by PWAS

Symbol	Name	Chrom	Most significant variant in the region	Most significant protein affecting variant	Generalized PWAS q value	Dominant PWAS q value	Dominant PWAS Cohen's d	Recessive PWAS q value	Recessive PWAS Cohen's d	Literature evidence
MUTYH	mutY DNA glycosylase	1	rs12139364 p = 6.3E−4	rs36053993 p = 1.2E−3	2.3E−6 (***)	0.34	n.s.	1.2E−4 (***)	− 0.079	**Strong**, biallelic mutations increase colorectal cancer risk by a factor of 17–44 [23]
FHL3	Four and a half LIM domains 3	1	rs147339918 p = 1.5E−3	rs145496383 p = 0.016	0.01 (*)	0.92	n.s.	0.024 (*)	− 0.037	**Moderate**, acts as tumor suppressor in breast and other cancer types [24]
OSTC	Oligosaccharyltransferase complex non-catalytic subunit	4	rs17038839 p = 9.1E−4	rs202168879 p = 0.016	0.01 (*)	0.96	n.s.	0.024 (*)	0.026	**Weak**, subunit of the OST complex which has been associated with lung and ovarian cancer [25, 26]
CDK2AP2/ DOC-1R	Cyclin-dependent kinase 2-associated protein 2	11	rs147242558 p = 2.7E−3	rs530762126 p = 0.4	0.024 (*)	1	n.s.	7.8E−3 (*)	− 5.4E−3	**Strong**, inhibits CDK2 and G1/S phase transition, a paralog of p14 and CDK2AP1, binds CDK2AP1 [27, 28]
POU5F1B/ BRN4	POU class 5 homeobox 1B	8	*rs6983267 p = 5.9E−9	*rs6998061 p = 1.4E−7	0.027 (*)	0.02 (*)	0.09	1	n.s.	**Strong**, promotes proliferation in several cancer types, GWAS hit in colorectal cancer [29–31]
CCDC172	Coiled-coil domain containing 172	10	rs200485970 p = 1.6E−4	rs532636333 p = 0.055	0.035 (*)	0.96	n.s.	0.059	n.s.	**None**

n.s. non-significant

Fig. 5 PWAS enriches GWAS discoveries across phenotypes. **a** We analyzed 23 binary phenotypes, 25 continuous phenotypes, and 1 categorical phenotype (male-balding patterns) derived from ~ 330K UK Biobank samples. Within binary phenotypes, the number of cases spans orders of magnitude (from only 127 in systemic sclerosis to 62K in hypertension). **b**, **c** Partition of the significant protein-coding genes, across the different phenotypes, that were detected by GWAS, PWAS, or both. The total number of significant genes is shown in brackets. In **b**, a gene was considered significant by GWAS if a non-synonymous variant within the coding region of the gene passed the exome-wide significance threshold ($p < 5E-07$). In **c**, a relaxed criterion was taken, considering all variants within 500,000 bp to each side of the coding region of the gene (here showing only the PWAS significant genes). **d** The number of significant genes per phenotype found by PWAS alone, according to the relaxed criterion of GWAS, as defined in **c** (i.e., without any significant variant within 500,000 bp)

PWAS and show strong evidence of being causal. To this end, we filtered associations by highly strict criteria: (i) strong significance (FDR q value < 0.01), (ii) no other significant genes in the region, and (iii) no single dominating variant association. Of the 2743 gene-phenotype associations uniquely found by PWAS (Fig. 5d), 48 meet these criteria
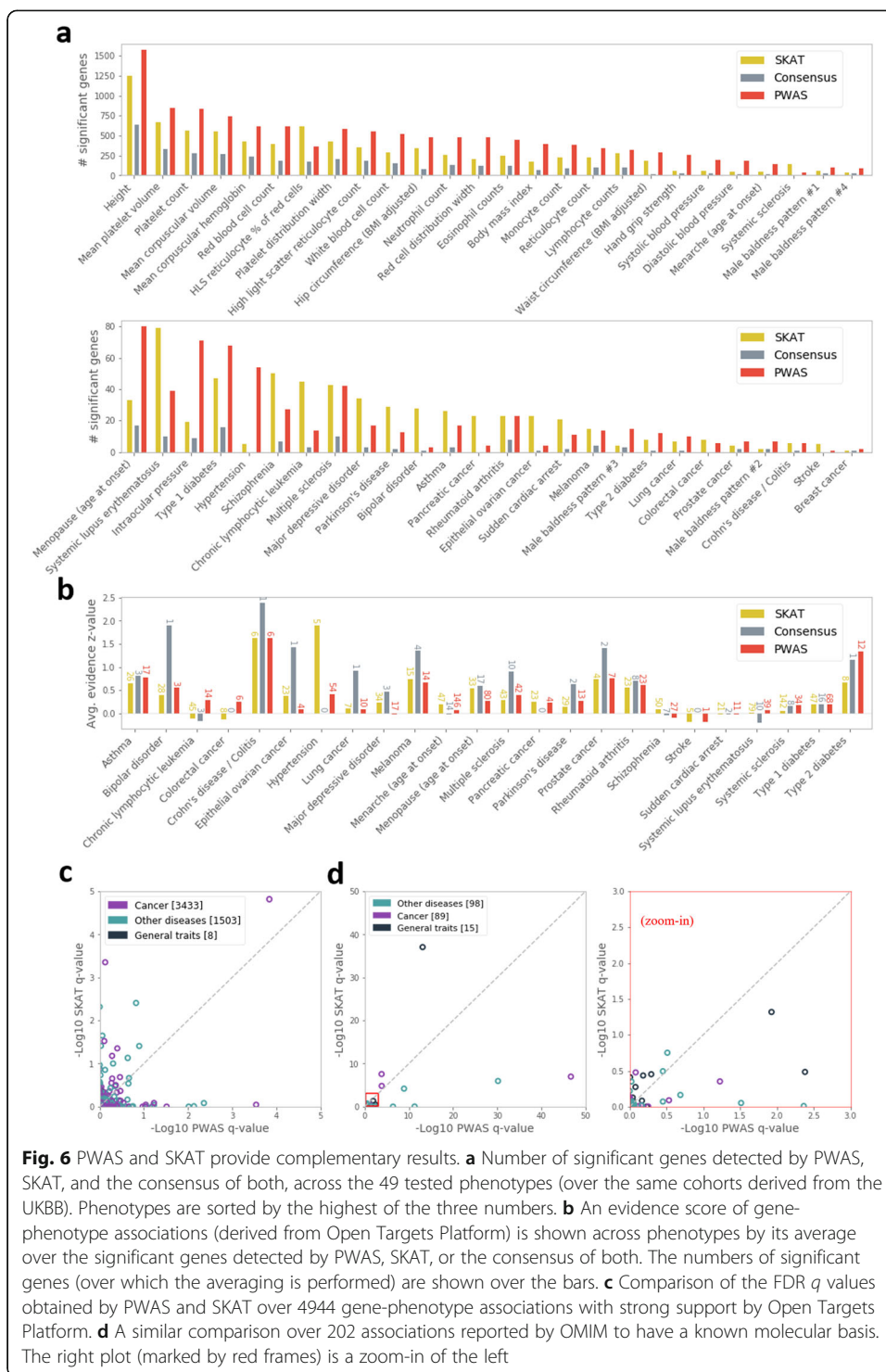
Fig. 6 PWAS and SKAT provide complementary results. **a** Number of significant genes detected by PWAS, SKAT, and the consensus of both, across the 49 tested phenotypes (over the same cohorts derived from the UKBB). Phenotypes are sorted by the highest of the three numbers. **b** An evidence score of gene-phenotype associations (derived from Open Targets Platform) is shown across phenotypes by its average over the significant genes detected by PWAS, SKAT, or the consensus of both. The numbers of significant genes (over which the averaging is performed) are shown over the bars. **c** Comparison of the FDR *q* values obtained by PWAS and SKAT over 4944 gene-phenotype associations with strong support by Open Targets Platform. **d** A similar comparison over 202 associations reported by OMIM to have a known molecular basis. The right plot (marked by red frames) is a zoom-in of the left

and are referred to as "PWAS-exclusive" associations (Table 2; the full list is provided in Additional file 5: Table S4).

As expected, the PWAS-exclusive genes show no GWAS signal at all, and the PWAS associations are constrained to the associated genes (Fig. 7a). When considered by SKAT, none of the 48 associations comes up as significant (Fig. 7b), even though SKAT was not included in
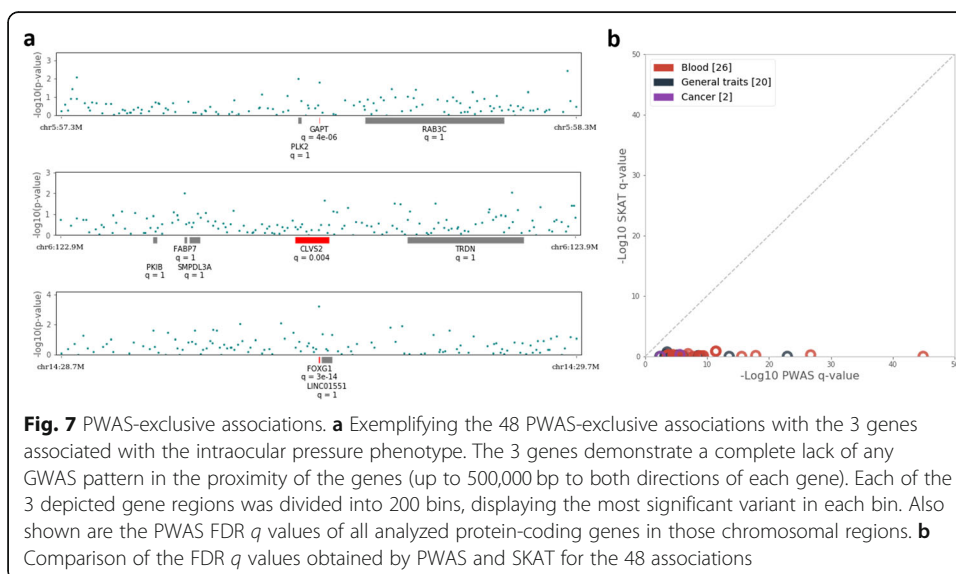
Table 2 Ten selected PWAS-exclusive associations

Phenotype	Gene symbol	Chrom	Generalized *q* value	Dominant *q* value	Dominant *r*/ Cohen's *d*	Recessive *q* value	Recessive *r*/ Cohen's *d*
HLS reticulocyte % of red cells	IL6	7	1.8E−126	1	n.s.	7.3E−128	− 0.043
Hip circumference (BMI adjusted)	SLC39A8	4	1.1E−23	1.6E−22	− 0.019	0.00081	− 0.0083
Eosinophil counts	FOXP1	3	2.6E−16	0.89	n.s.	9.8E−17	− 0.016
Intraocular pressure	FOXG1	14	2.7E−14	1	n.s.	2.6E−15	0.031
Lymphocyte counts	INPP1	2	3.4E−12	0.5	n.s.	1.9E−12	− 0.014
Eosinophil counts	CD80	3	3.1E−06	0.97	n.s.	1.1E−06	− 0.01
Intraocular pressure	GAPT	5	4.4E−06	1	n.s.	6.5E−07	− 0.021
Red cell distribution width	MLLT3	9	1.2E−05	0.76	n.s.	8.5E−06	− 0.01
Chronic lymphocytic leukemia	FAM160B1	10	0.0033	0.99	n.s.	0.048	0.06
Intraocular pressure	CLVS2	6	0.0043	0.98	n.s.	0.0018	0.016

n.s. non-significant

the criteria for defining those associations. Interestingly, most of the PWAS-exclusive associations are driven by recessive inheritance. Among the ten genes listed in Table 2, only one (*SLC39A8*) shows a dominant inheritance pattern. This suggests that the modeling of recessive inheritance is a unique advantage of PWAS over GWAS.

Some of the listed associations are strongly supported by the literature. For example, interleukin 6 (*IL6*), here implicated with high light scatter (HLS) reticulocyte percentage of red blood cells with overwhelming significance (PWAS FDR *q* value = 1.8E−126), is known for its capacity to impair hemoglobin production and erythroid maturation. A connection of *IL6* to erythroid maturation, anemia, and inflammation through impairment of mitochondrial function was also established [34]. Moreover, *IL6* plays a role in the development of anemia of chronic kidney disease in children (CKD anemia). This *IL6*-dependent pathology is induced

Fig. 7 PWAS-exclusive associations. **a** Exemplifying the 48 PWAS-exclusive associations with the 3 genes associated with the intraocular pressure phenotype. The 3 genes demonstrate a complete lack of any GWAS pattern in the proximity of the genes (up to 500,000 bp to both directions of each gene). Each of the 3 depicted gene regions was divided into 200 bins, displaying the most significant variant in each bin. Also shown are the PWAS FDR *q* values of all analyzed protein-coding genes in those chromosomal regions. **b** Comparison of the FDR *q* values obtained by PWAS and SKAT for the 48 associations

by the destruction of red blood cells through its effects on the erythropoietin (Epo) axis, confirming a direct link of IL6 to the percentage of red blood cells [35].

Similarly, MLLT3, which appears to be associated with red blood cell distribution width through recessive inheritance according to PWAS (FDR q value = 8.5E−06, r = − 0.01), was indeed reported to be a key regulatory gene in the bone marrow [36]. Among the 49 phenotypes tested in this work, we found the gene to also be significant in numerous other blood cell traits, as well as hand grip strength (Additional file 3: Table S2). Likewise, CD80, which PWAS associates with eosinophil counts through recessive inheritance (FDR q value = 1.1E−06, r = − 0.01), indeed has an important role in antigen presentation by eosinophils [37]. FOXP1 is another gene associated with eosinophil counts through recessive inheritance according to PWAS (FDR q value = 9.8E−17, r = − 0.016). While no direct evidence for this association is reported, FOXP1 is known to affect monocyte differentiation and macrophage function [38].

In other examples, while there is no clear indication for the reported association, there does exist strong molecular plausibility. Another transcription factor belonging to the forked head family is *FOXG1*, which plays a key role in the development of the retina (a function conserved in all vertebrates) [39]. The gene was shown to be associated with visual impairment in both mouse and human cohorts [40]. However, it has never been directly linked to intraocular pressure, an association that we observe here with outstanding significance according to the recessive model of PWAS (FDR q value = 2.6E−15). Specifically, the normal function of the gene (i.e., lack of damaging variants) appears to be positively correlated (r = 0.031) with intraocular pressure.

Another example is *INPP1*, which encodes the enzyme inositol polyphosphate-1-phosphatase. In the existing literature, it is mainly reported in the context of autism and mood disorders [41], while reported genetic associations in the Open Targets Platform [32] focus mainly on autoimmune disorders and blood characteristics. Nonetheless, it does not appear to be linked to lymphocyte counts, an association we observe here (recessive FDR q value = 1.9E−12, r = − 0.014). In general, genetic study of blood phenotypes appears to be somewhat neglected, and it is often uncertain how such associations relate to clinical outcomes.

In some instances, we find little to no literature evidence for reported PWAS-exclusive associations. For example, *GAPT* and *CLVS2* are found to be associated with intraocular pressure. *GAPT* (growth factor receptor-bound protein 2-binding adapter protein, transmembrane) plays a role in regulating B cell activation and proper maintenance of the marginal zone [42]. *CLVS2* (clathrin vesicle-associated Sec14 protein 2) is involved in cell membrane trafficking [43]. In both cases, a link to intraocular pressure is not yet reported. Another significant PWAS association lacking literature support is *FAM160B1* with respect to leukemia. Despite the lack of existing literature support for those connections, the strong associations established by PWAS provide strong evidence for potential links which deserve further exploration.

Discussion

In this work, we have introduced a new functional protein-centric approach to association studies. We have demonstrated its applicability to a broad range of prominent human phenotypes and established its utility in supplementing existing methods and highlighting novel associations.

Due to its explicit gene-based functional model, PWAS provides more interpretable results than other methods. Like other gene-based approaches seeking to establish associations of concrete genes, it requires no post-analysis fine-mapping. Furthermore, as PWAS relies on an explicit functional model, it is better posed to suggest causal relationships. Specifically, a significant PWAS association would suggest that variants disrupting the function of the implicated protein might influence the studied phenotype (in the case of a disease, increase or decrease one's risk). Furthermore, PWAS can determine whether the proposed causal effect appears to be dominant, recessive, or some mixture of the two (e.g., additive). Yet, while PWAS is more suggestive of causality than other methods, significant results are still susceptible to spurious correlations. In particular, the problem of LD [7] is still far from resolved, and significant PWAS associations, like any genetic associations, should be interpreted with caution. It should also be kept in mind that most GWAS associations occur in non-coding regions [44], which PWAS completely ignores in its current form.

By aggregating all variants affecting the same gene into unified statistics, PWAS is able to detect a signal that is too weak and spread to appear in per-variant GWAS (Fig. 7, Table 2). It is particularly important in the case of rare variants, which account for much of the heritability [45]. In fact, PWAS can successfully handle even variants that occur only once in the cohort (including, in principle, de novo variants). As long as the observed variants fit the overall trend observed in the studied gene (e.g., that they are more damaging in cases compared to controls), even singletons can increase the statistical power of the test. In this work, however, we relied on imputed genotypes which cannot capture variants that are too rare. As a result, some biological signal was probably missed (e.g., damaged genes that were mistaken to be intact due to ungenotyped variants). We therefore anticipate that PWAS can substantially benefit from exome sequencing (as opposed to SNP array genotypes). It should be noted that while more accurate genotyping should indeed enhance its statistical power, the reliance of PWAS on rigorous statistics keeps it robust to false discoveries as other GWAS methods, even with imperfect genotyping.

A rather unique feature of PWAS is its separate dominant and recessive inheritance models. Although there are strong indications that the commonly used additive model can capture most of the heritability of complex human traits [46], non-additive and epistatic effects play a key role in many phenotypes [47]. While there have been efforts to address epistatic effects in GWAS [48], the special case of recessive inheritance in complex traits has been largely neglected. Our results show that recessive inheritance is indeed substantial in a variety of phenotypes. Twenty-four percent of the recovered PWAS associations are significant by the recessive but not the dominant model. PWAS is uniquely posed, among present methods, to handle recessive inheritance, as per-gene recessive inheritance is much more sensible than per-variant. Specifically, PWAS is able to capture the prevalent instances of compound heterozygosity (due to its per-gene aggregation), whereas per-variant GWAS would fail to detect such recessive effects [49].

Another important advantage of PWAS over existing methods is its reduced computational burden in multi-phenotype datasets (such as the UKBB). PWAS aggregates all the genetic data into compact gene score matrices, whose size is much smaller than the raw genotyping data (as there are typically substantially fewer genes than variants). These matrices store all of the relevant genetic information (encompassing the assessed functional state of the proteome in each of the cohort samples) and can be independently tested against each phenotype.

PWAS belongs to the growing family of methods that seek genetic associations through modeling of functional genomic properties. While methods such as PrediXcan [20] and TWAS [21] model gene expression, PWAS models protein function, which, in principle, is completely orthogonal to the signal of gene abundance. We purposefully employ a very abstract definition of the term "protein function" to encapsulate anything the protein is supposed to do in the cell such that disturbing it (by variants altering the protein sequence) could lead to phenotypic effects (e.g., missense variants affecting a membrane receptor protein could interfere with its signal transduction function and result in predisposition to cancer). We consider PWAS complementary to methods that model other functional aspects of the genome.

Contrary to expression-based methods, PWAS assigns gene scores that depend only on genotypes. Gene expression is highly volatile, with substantial variability between tissues, epigenetic conditions, and many other non-genetic factors. TWAS, for example, requires the researcher to choose a target tissue for the association tests (e.g., based on prior assumptions about the phenotype under study) [50]. Damage to protein products, on the other hand, is mostly a direct result of genetic makeup. This benefits PWAS in two major ways. First, it offers reduced computational complexity, since it is sufficient to compute the gene score matrices only once, which can later be used for any phenotype. More importantly, it relieves the researcher from the need to select a specific tissue or expression profile for the analysis. Indeed, most human traits are not confined to specific tissues, let alone specific cellular conditions, making the selection of a relevant reference panel for expression-based methods non-trivial.

A potential limitation of PWAS is its reliance on complete individual-level data (including genotype and phenotype information). Unlike other methods, it is unable to analyze summary GWAS statistics alone. This reliance on raw data is due to the non-linear nature of the aggregation algorithm used to derive gene effect scores from variant effect scores (see the "Methods" section). It remains open whether a simplified linear version of PWAS could be derived or at least a version simple enough that can be applied on summary statistics. On the positive side, with modern biobanks and genetic cohorts (e.g., UKBB, SFARI [51]), large-scale datasets are becoming increasingly available for direct modeling and analysis.

Conclusions

We have presented PWAS as a novel protein-centric method for genetic association studies providing functionally interpretable gene results. We have demonstrated the validity of PWAS through comparison to multiple external resources and shown its added value to commonly used methods across a wide range of prominent phenotypes, including numerous new discoveries. We argue that integrating rich machine learning models based on prior knowledge, as exemplified in this work, is a promising avenue to novel insight and discovery in human genetics.

Methods

UK Biobank cohort

Throughout this work, we used genetic and phenotypic data from the UK Biobank (UKBB) resource [8, 9] (application ID 26664).

From the entire UKBB cohort of 502,539 samples, we filtered 409,584 labeled as Whites/Caucasians according to both self-reported ethnicity and their genetics. We

removed 312 samples with mismatching self-reported and genetics-derived sex. Finally, we removed 75,848 samples to keep only one representative of each kinship group of related individuals, obtaining a final cohort of 333,424 samples.

Specification of the 49 phenotypes used in this work is available in Additional file 6: Table S5. The table specifies how each phenotype was defined (based on either a UKBB field or ICD-10 codes) and whether it was restricted to a specific gender. The set of all ICD-10 codes associated with a sample was derived from the following UKBB fields: 41202, 41204, 40006, 40001, 40002, and 41201.

When testing a specific phenotype, we also filtered out samples with missing values in that phenotype (e.g., for height, we filtered out 690 samples, obtaining a cohort of 332,734 samples). When testing phenotypes defined by ICD-10 codes, we filtered out all samples without any recorded ICD-10 code. This further removed 70,475 samples from the cohort, leaving 262,949 samples in those phenotypes. The final cohort size used for testing each phenotype is listed in Additional file 3: Table S2. In the rare cases where samples had multiple records of the same continuous phenotype (e.g., from different visits to the UKBB assessment centers), we took the maximum value.

All the association tests carried out in this work (with either of the three used methods, i.e., GWAS, PWAS, or SKAT) included the following covariates: sex (binary), year of birth (numeric), the 40 principal components of the genetic data provided by the UKBB (numeric), the UKBB genotyping batch (one-hot-encoding with 105 categories), and the UKBB assessment centers associated with each sample (binary, with 25 categories). Altogether, 173 covariates (including a constant intercept) were included. For specific phenotypes, additional covariates were included as part of the phenotype's definition (e.g., "hip circumference adjusted for BMI" included BMI as an additional covariate; see Additional file 6: Table S5).

The computational pipeline for processing the UK Biobank data is open source and available at https://github.com/nadavbra/ukbb_parser (through both Python and command-line interface).

Variant functional effect scores

The gene effect scores used by PWAS are derived from the aggregation of per-variant effect scores (Fig. 2b). Each non-synonymous variant in the coding region of a gene which affects the resulted protein sequence is assigned a functional effect score that aims to capture its propensity to damage the protein product of the gene. Specifically, PWAS considers the following types of variants as affecting protein sequence: missense, nonsense, frameshift, in-frame indel, and canonical splice-site variants. The predicted effect score of a variant is a number between 0 (complete loss of function) and 1 (no functional effect). Intuitively, it reflects the probability that the affected gene retains its function given the variant.

To predict the effect of missense variants, PWAS employs a machine learning model. Specifically, we used the FIRM predictor [22] (https://github.com/nadavbra/firm). Unlike commonly used prediction tools assessing mutation pathogenicity (e.g., CADD [52], SIFT [53], Polyphen2 [54], MutationTaster2 [55]), FIRM is designed to assess the damage of variants at the molecular level (rather than the clinical outcome at the organism level). This distinction is particularly important when PWAS is used for phenotypes

without clinical significance (e.g., height). FIRM attempts to capture gene function in its broadest sense (e.g., enzymatic reaction, molecular interaction, cellular pathways), thereby allowing PWAS, in principle, to model any protein-phenotype effect.

To assess the impact of a missense variant on gene function, FIRM considers its rich proteomic context, which it encodes as a set of 1109 numerical features (which are used by the underlying machine learning model to predict its effect score). The full specification of the features used by FIRM is described elsewhere [22]. They include (i) the position of the variant within the protein sequence, (ii) the identity of the reference and alternative amino acids and the amino acid composition of the protein in various regions of the protein with respect to the variant, (iii) abundance of annotations extracted from UniProt [56] (e.g., phosphorylation and other post-translational modifications, active sites, secondary structure), and (iv) details of Pfam domains [57] in the proximity of the variant.

Missense variants comprise the vast majority of non-synonymous variants [58]. For other variant types, effect scores were derived through rougher, rule-based formulas. Specifically, nonsense, frameshift, and canonical splice site variants (i.e., variants affecting the first or last two letters of an intron) were assumed to be loss-of-function variants and assigned a score of 0. In-frame indels were assigned an effect score based on the numbers of substituted, inserted, and deleted amino acids (see Additional file 1: Supplementary Methods).

Gene functional effect scores

To calculate gene effect scores (Figs. 1b and 2c–d, PWAS aggregates variant effect scores (see previous section) integrated with the cohort's genotyping data. Unlike variant-level scores, gene scores are sample specific (i.e., depending on each sample's genotype). PWAS supports two aggregation schemes, resulting in "dominant" and "recessive" gene scores. Intuitively, dominant scores reflect the probability of at least one damaging event, whereas recessive scores reflect the probability for at least two.

Let $s_1, \ldots, s_k$ be the functional effect scores of the k variants potentially affecting a protein-coding gene (assigned by the scheme detailed in the previous section). For a given variant $i \in [k]$ (in the context of a given sample), let $0 \leq p_0^{(i)}, p_1^{(i)}, p_2^{(i)} \leq 1$ (satisfying $\sum_{j=0}^{2} p_j^{(i)} = 1$) indicate the genotyping probabilities of the variant (i.e., $p_j^{(i)}$ is the probability of variant i to occur j times in the given sample). Recall that we intuitively interpret s_i as the probability that the gene retains its function following the variant effect.

A question arises how to estimate the probability of the gene to retain its function if variant i occurs twice (an event of probability $p_2^{(i)}$). A possible approach would be to treat the two occurrences of the variant as independent, so the probability would be s_i^2. Another approach is to treat the two occurrences as fully dependent (i.e., either the variant is damaging or it is not), taking the probability to be simply s_i like in the heterozygous case. To accommodate this uncertainty, we chose to introduce a parameter $\mu \in [0, 1]$ and take the effect to be $\mu s_i + (1 - \mu)s_i^2$. The parameter μ can be thought of as the probability of the homozygous effect to be dependent (i.e., when $\mu = 0$, it is completely independent, and when $\mu = 1$, it is fully dependent). Overall, the probability that the gene retains its function considering variant i (in the context of that sample) would

be $x_i := p_0^{(i)} \cdot 1 + p_1^{(i)} \cdot s_i + p_2^{(i)} \cdot (\mu s_i + (1 - \mu)s_i^2)$.

Note that in reality, the scores s_i are not purely probabilistic entities. More likely, they capture both the probability of gene damage and its extent (so, s_i and x_i can be more realistically interpreted as damage expectations rather than probabilities). That is another reason why the independent case ($\mu = 0$) might be more appropriate than the dependent case ($\mu = 1$), as two hits of a variant often cause more damage than a single hit. Taking the same expression (s_i) to estimate the outcome of these two events would miss this effect.

We want the dominant effect score of the gene to reflect the probability that it retains its function (given the sample's genotyping of the k variants and their effect scores). If we simplistically assume that the k variants independently affect the gene, then it retains its function with probability $x_1 \cdots x_k$. Here too, some degree of dependence might better reflect the dominant effect of the gene. Under full dependence, we would take the score to be $\min\{x_1, ..., x_k\}$ (i.e., the overall effect on the gene is the effect of the most damaging variant). To allow a more refined dependence model, let us write the multiplication $\prod_{i=1}^{k} x_i$ as $\exp(-\sum_{i=1}^{k} \log \frac{1}{x_i})$. The term $\sum_{i=1}^{k} \log \frac{1}{x_i}$ is the ℓ_1 norm of the vector $(\log \frac{1}{x_1}, ..., \log \frac{1}{x_k})$. We introduce another parameter $1 \le p \le \infty$ and take the dominant score to be $D := \exp(-\|(\log \frac{1}{x_1}, ..., \log \frac{1}{x_k})\|_p)$. Note that when $p = \infty$, we get the full independence score $\min\{x_1, ..., x_k\}$.

For deriving the recessive effect score of the gene, we would like to express the probability of at most one damaging event (so, its complementary event would represent the probability of at least two damaging events). Assuming independence, that probability would be $x_1 \cdots x_k + \sum_{i=1}^{k} x_1 \cdots x_{i-1} y_i x_{i+1} \cdots x_k$, where y_i expresses the probability of variant i damaging exactly one copy of the gene. Specifically, we define $y_i := p_1^{(i)} \cdot (1 - s_i) + p_2^{(i)} \cdot (1 - \mu) \cdot 2s_i(1 - s_i)$. The second coefficient is explained by $2s_i(1 - s_i)$ being the probability of variant i introducing exactly one hit, given that each of its two copies is independent; when they are fully dependent, that is not possible for the two copies to introduce exactly one hit. When all $x_i \ne 0$, we can rewrite that expression as $(x_1 \cdots x_k)(1 + \sum_{i=1}^{k} \frac{y_i}{x_i})$. Like with the dominant score, we parameterize $(x_1 \cdots x_k)$ into $D_p := \exp(-\|(\log \frac{1}{x_1}, ..., \log \frac{1}{x_k})\|_p)$, and $\sum_{i=1}^{k} \frac{y_i}{x_i}$ into $\zeta_q := \|(\frac{y_1}{x_1}, ..., \frac{y_k}{x_k})\|_q$ for some parameter values $1 \le p, q \le \infty$. The recessive effect score is then taken to be $(1 + \zeta_q)D_p$. However, this term is not well-defined when there is $x_i = 0$. To derive the recessive score in that case, we can calculate $\lim_{x_i \to 0}(1 + \zeta_q)D_p$ (see Additional file 1: Supplementary Methods for details) and obtain:

$$R := \begin{cases} 0 & \exists i \ne j, x_i = x_j = 0 \\ y_i & \exists! i, x_i = 0, p > 1 \\ y_i \prod_{j \ne i} x_j & \exists! i, x_i = 0, p = 1 \\ (1 + \zeta_q)D_p & \forall i, x_i \ne 0 \end{cases}$$

To summarize, the aggregation scheme takes as input the individual variant scores $s_1, ..., s_k$ (which are sample independent) and the genotyping probabilities of the k variants within the given sample $p_0^{(i)}, p_1^{(i)}, p_2^{(i)}$ ($i \in [k]$), to produce the dominant and recessive gene scores of the

gene. The dominant score D relies on two parameters (μ and p), whereas the recessive score R depends on three parameters (μ, p, and q). Note that the parameters μ and p used by the two scoring schemes need not take the same values in the two contexts (despite sharing a similar purpose). For clarity, we denote the parameters of D by μ_D and p_D, and the parameters of R by μ_R, p_R, and q_R. Overall, the effect score aggregation scheme of PWAS is parameterized by 5 distinct parameters.

To find optimal parameter values, we fit the aggregation scheme on known gene-phenotype associations derived from OMIM [33], taking the combination of 5 parameters that optimize the recovered significance of these associations (see Additional file 1: Supplementary Methods). Importantly, the gene-phenotype associations used to find the optimal parameters do not overlap with the associations used to evaluate PWAS throughout this work (e.g., Fig. 6c). In particular, they involve other phenotypes that were not studied in the primary analysis. The obtained parameter values used throughout the analyses presented in this work are $\mu_D = 1$, $p_D = 1.25$, $\mu_R = 0.5$, $p_R = \infty$, and $q_R = 3$.

Non-modeled genomic properties

In its current form, PWAS does not consider structural and copy number variations, as they do not naturally fit into the framework of dominant and recessive heritability modes. Non-canonical splicing effects are also not considered at present, as they are not supported by FIRM. In general, the effects of splicing events are considered to be hard to model [59]. Furthermore, weak splicing events are often associated with alternative splicing of non-canonical protein isoforms. To allow simple modeling and interpretation of the results, PWAS considers only canonical protein isoforms (see Additional file 1: Supplementary Methods).

It should also be noted that the recessive model assumes standard autosomal inheritance, and PWAS does not properly address recessive inheritance in sex and mitochondrial chromosomes. Another current limitation of the recessive model has to do with the absence of phased genotypes in the UKBB resource. For a recessive genetic effect to take place, both copies of a gene (on the two copies of the relevant chromosome) should be affected. Due to the lack of phased genotypes, PWAS is unable to determine if different variants affect the same or different copies of the gene. Therefore, our modeling choice was to assume that different variants affect different gene copies (see previous section).

Importantly, these non-modeled genomic properties can only affect the statistical power of PWAS, but should not lead to false discoveries (see next section).

Statistical analysis

To test whether a gene is associated with a phenotype, PWAS conducts linear or logistic regression (depending on whether the phenotype is continuous or binary, respectively). A categorical phenotype is split into multiple binary phenotypes (each isolating one of the categories in a one-vs.-rest manner). The regression model includes all 173 covariates (see the "UK Biobank cohort" section) and the relevant gene scores (dominant, recessive, or both). Specifically, when testing for a dominant inheritance, the term $\beta_D \cdot D$ is included in the regression model, where D is the dominant score of the gene, and β_D is the corresponding regression coefficient. The null hypothesis of the

Brandes *et al. Genome Biology*   (2020) 21:173

Page 20 of 22

regression under dominant inheritance is $H_0 : \beta_D = 0$. Similarly, when testing for a recessive inheritance, the term $\beta_R \cdot R$ is included, and the null hypothesis is $H_0 : \beta_R = 0$. When the test is carried according to the generalized model, both terms are included in the regression, and the tested null hypothesis is $H_0 : \beta_D = \beta_R = 0$. Unless stated explicitly that the dominant or recessive model is used, all the p values reported in this work refer to the generalized model.

As PWAS relies on routine statistical analysis to calculate significance, its results are valid (in terms of avoiding false discoveries) regardless of how accurately the calculated gene scores reflect the true underlying biology. While better scoring schemes are expected to provide increased statistical power, protection against type I errors is guaranteed irrespectively.

To provide a fair comparison to PWAS, the results of GWAS and SKAT reported in this work were performed using identical statistical analysis over the same data (see Additional file 1: Supplementary Methods).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02089-x.

Additional file 1: Supplementary materials: Figures S1-S7 and Supplementary Methods.

Additional file 2: Table S1: The 47 imputed variants affecting the protein sequence of the *MUTYH* gene.

Additional file 3: Table S2: Full per-gene summary statistics of all significant PWAS associations across all 49 tested phenotypes.

Additional file 4: Table S3: Full per-gene summary statistics of all 18,053 tested protein-coding genes across all 49 tested phenotypes.

Additional file 5: Table S4: The 48 gene-phenotype associations defined as "PWAS-exclusive".

Additional file 6: Table S5: Specification of the 49 tested phenotypes.

Additional file 7: Table S6: Specification of the 6 OMIM diseases used to fit the effect score aggregation models.

Additional file 8: Table S7: Enrichment factor between our GWAS results to the GWAS Catalog across all 49 tested phenotypes and parameter values.

Additional file 9: Review history.

Availability of data and materials
PWAS is available through command-line interface as part of an open-source project (MIT license) at https://github.com/nadavbra/pwas [60]. The specific version used in this study (PWAS 1.0.4) is available at https://doi.org/10.5281/zenodo.3902592 [61].
The dataset used in this study was derived from the UK Biobank [8, 9], as detailed in the "Methods" section.

Ethics approval and consent to participate
Approval to use the UK Biobank dataset was obtained as part of our project titled "Novel machine learning framework for improved inference in GWAS" (application ID 26664). Ethical approval for this study was obtained from the

Brandes *et al. Genome Biology*   (2020) 21:173

Page 21 of 22

committee for ethics in research involving human subjects at the Faculty of Medicine, The Hebrew University of Jerusalem, Israel (approval number 13082019).

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References
1. Lewis CM, Knight J. Introduction to genetic association studies. Cold Spring Harb Protoc. 2012;2012(3):pdb--top068163.
2. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013 Jan [cited 2014 Jul 11];9(1):29. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3750305&tool=pmcentrez&rendertype=abstract.
3. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47(3):284.
4. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):7.
5. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2018;47(D1):D1005–12.
6. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7–24.
7. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101(1):5–22.
8. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779.
9. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. Genome-wide genetic data on ~500,000 UK Biobank participants. BioRxiv. 2017;166298.
10. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747.
11. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Hum Mol Genet. 2015;24(R1):R111–9.
12. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19(8):491.
13. de Bunt M, Cortes A, Brown MA, Morris AP, McCarthy MI, Consortium I, et al. Evaluating the performance of fine-mapping strategies at common variant GWAS loci. PLoS Genet. 2015;11(9):e1005535.
14. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48:481 Available from: https://doi.org/10.1038/ng.3538.
15. Hormozdiari F, Van De Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL signals detects target genes. Am J Hum Genet. 2016;99(6):1245–60.
16. Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun. 2015;6:5890 Available from: https://doi.org/10.1038/ncomms6890.
17. Hormozdiari F, Kichaev G, Yang W-Y, Pasaniuc B, Eskin E. Identification of causal genes for complex traits. Bioinformatics. 2015;31(12):i206–13 Available from: https://doi.org/10.1093/bioinformatics/btv240.
18. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82–93.
19. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91(2):224–37.
20. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47:1091 Available from: https://doi.org/10.1038/ng.3367.
21. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48:245 Available from: https://doi.org/10.1038/ng.3506.
22. Brandes N, Linial N, Linial M. Quantifying gene selection in cancer through protein functional alteration bias. Nucleic Acids Res. 2019;27(24). https://ascopubs.org/doi/full/10.1200/JCO.2008.21.6853?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed&.
23. Lubbe SJ, Di Bernardo MC, Chandler IP, Houlston RS, others. Clinical implications of the colorectal cancer risk associated with MUTYH mutation. J Clin Oncol 2009;27(24):3975–3980.
24. Niu C, Yan Z, Cheng L, Zhu J, Zhang H, Xu X, et al. Downregulation and antiproliferative role of FHL3 in breast cancer. IUBMB Life. 2011;63(9):764–71.
25. Sambrooks CL, Baro M, Quijano A, Narayan A, Cui W, Greninger P, et al. Oligosaccharyltransferase inhibition overcomes therapeutic resistance to EGFR tyrosine kinase inhibitors. Cancer Res. 2018;78(17):5094–106.
26. Vaňhara P, Horak P, Pils D, Anees M, Petz M, Gregor W, et al. Loss of the oligosaccharyl transferase subunit TUSC3 promotes proliferation and migration of ovarian cancer cells. Int J Oncol. 2013;42(4):1383–9.
27. Liu Q, Liu X, Gao J, Shi X, Hu X, Wang S, et al. Overexpression of DOC-1R inhibits cell cycle G1/S transition by repressing CDK2 expression and activation. Int J Biol Sci. 2013;9(6):541.

28. Ertekin A, Aramini JM, Rossi P, Leonard PG, Janjua H, Xiao R, et al. Human cyclin-dependent kinase 2-associated protein 1 (CDK2AP1) is dimeric in its disulfide-reduced state, with natively disordered N-terminal region. J Biol Chem. 2012; 287(20):16541–9.
29. Hayashi H, Arao T, Togashi Y, Kato H, Fujita Y, De Velasco MA, et al. The OCT4 pseudogene POU5F1B is amplified and promotes an aggressive phenotype in gastric cancer. Oncogene. 2015;34(2):199.
30. Pan Y, Zhan L, Chen L, Zhang H, Sun C, Xing C. POU5F1B promotes hepatocellular carcinoma proliferation by activating AKT. Biomed Pharmacother. 2018;100:374–80.
31. Yao L, Tak YG, Berman BP, Farnham PJ. Functional annotation of colon cancer risk SNPs. Nat Commun. 2014;5:5114.
32. Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N, et al. Open targets platform: new developments and updates two years on. Nucleic Acids Res. 2018;47(D1):D1056–65.
33. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(suppl_1):D514–7.
34. McCranor BJ, Kim MJ, Cruz NM, Xue Q-L, Berger AE, Walston JD, et al. Interleukin-6 directly impairs the erythroid development of human TF-1 erythroleukemic cells. Blood Cells Mol Dis. 2014;52(2–3):126–33.
35. Ulich TR, Yin SM, Egrie JC, others. The erythropoietic effects of interleukin 6 and erythropoietin in vivo. Exp Hematol 1991;19(1):29–34.
36. Pina C, May G, Soneji S, Hong D, Enver T. MLLT3 regulates early human erythroid and megakaryocytic cell fate. Cell Stem Cell. 2008;2(3):264–73.
37. Tamura N, Ishii N, Nakazawa M, Nagoya M, Yoshinari M, Amano T, et al. Requirement of CD80 and CD86 molecules for antigen presentation by eosinophils. Scand J Immunol. 1996;44(3):229–38.
38. Shi C, Sakuma M, Mooroka T, Liscoe A, Gao H, Croce KJ, et al. Down-regulation of the forkhead transcription factor Foxp1 is required for monocyte differentiation and macrophage function. Blood. 2008;112(12):4699–711.
39. Picker A, Cavodeassi F, Machate A, Bernauer S, Hans S, Abe G, et al. Dynamic coupling of pattern formation and morphogenesis in the developing vertebrate retina. PLoS Biol. 2009;7(10):e1000214.
40. Boggio EM, Pancrazi L, Gennaro M, Lo RC, Mari F, Meloni I, et al. Visual impairment in FOXG1-mutated individuals and mice. Neuroscience. 2016;324:496–508.
41. Serajee FJ, Nabi R, Zhong H, Huq AHMM. Association of INPP1, PIK3CG, and TSC2 gene variants with autistic disorder: implications for phosphatidylinositol signalling in autism. J Med Genet. 2003;40(11):e119.
42. Liu Y, Zhang W. Identification of a new transmembrane adaptor protein that constitutively binds Grb2 in B cells. J Leukoc Biol. 2008;84(3):842–51.
43. Saito K, Tautz L, Mustelin T. The lipid-binding SEC14 domain. Biochim Biophys Acta (BBA)-Molecular Cell Biol Lipids. 2007;1771(6):719–26.
44. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2013;42(D1):D1001–6.
45. Speed D, Cai N, Johnson MR, Nejentsev S, Balding DJ, Consortium U, et al. Reevaluation of SNP heritability in complex human traits. Nat Genet. 2017;49(7):986.
46. Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 2008;4(2):e1000008.
47. Moore JH, Williams SM. Epistasis and its implications for personal genetics. Am J Hum Genet. 2009;85(3):309–20.
48. Niel C, Sinoquet C, Dina C, Rocheleau G. A survey about methods dedicated to epistasis detection. Front Genet. 2015;6: 285.
49. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012;13(2):135.
50. Mancuso N, Gayther S, Gusev A, Zheng W, Penney KL, Kote-Jarai Z, et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. Nat Commun. 2018;9(1):1–11.
51. Feliciano P, Daniels AM, Snyder LG, Beaumont A, Camba A, Esler A, et al. SPARK: a US cohort of 50,000 families to accelerate autism research. Neuron. 2018;97(3):488–93.
52. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310.
53. Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1 [cited 2016 Jan 21];31(13):3812–4. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkg509.
54. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;76(1):7–20.
55. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11(4):361.
56. Consortium U. UniProt: a hub for protein information. Nucleic Acids Res. 2014;43(D1):D204–12.
57. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):D222-30. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965110&tool=pmcentrez&rendertype=abstract.
58. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv. 2005;344:1-20. https://www.sciencedirect.com/science/article/abs/pii/S037811190400650X?via%3Dihub.
59. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, et al. Function of alternative splicing. Gene. 2005;344:1–20.
60. Brandes N. pwas. Github. 2020. https://github.com/nadavbra/pwas. Accessed 11 Apr 2020.
61. Brandes N. pwas 1.0.4. zenodo. 2020. https://doi.org/105281/zenodo3902592. Accessed 11 Apr 2020.

Publisher's Note