European Journal of Human Genetics (2015) 23, 523–529

www.nature.com/ejhg

ARTICLE

Phenome-wide association studies (PheWASs) for functional variants

Zhan Ye[1], John Mayer[1], Lynn Ivacic[2], Zhiyi Zhou[3], Min He[1,2], Steven J Schrodi[2], David Page[4], Murray H Brilliant[2] and Scott J Hebbring*[,2,4]

The genome-wide association study (GWAS) is a powerful approach for studying the genetic complexities of human disease. Unfortunately, GWASs often fail to identify clinically significant associations and describing function can be a challenge. GWAS is a phenotype-to-genotype approach. It is now possible to conduct a converse genotype-to-phenotype approach using extensive electronic medical records to define a phenome. This approach associates a single genetic variant with many phenotypes across the phenome and is called a phenome-wide association study (PheWAS). The majority of PheWASs conducted have focused on variants identified previously by GWASs. This approach has been efficient for rediscovering gene–disease associations while also identifying pleiotropic effects for some single-nucleotide polymorphisms (SNPs). However, the use of SNPs identified by GWAS in a PheWAS is limited by the inherent properties of the GWAS SNPs, including weak effect sizes and difficulty when translating discoveries to function. To address these challenges, we conducted a PheWAS on 105 presumed functional stop-gain and stop-loss variants genotyped on 4235 Marshfield Clinic patients. Associations were validated on an additional 10 640 Marshfield Clinic patients. PheWAS results indicate that a nonsense variant in *ARMS2* (rs2736911) is associated with age-related macular degeneration (AMD). These results demonstrate that focusing on functional variants may be an effective approach when conducting a PheWAS.
European Journal of Human Genetics (2015) **23,** 523–529; doi:10.1038/ejhg.2014.123; published online 30 July 2014

INTRODUCTION

The phenome-wide association study (PheWAS) design is emerging as a complementary/alternative approach to the genome-wide association study (GWAS).[1] This is driven in part by challenges in interpreting GWAS results. GWASs often fail to identify clinically significant associations and it can be equally challenging to characterize biological function when a significant fraction of single-nucleotide polymorphisms (SNPs) identified by GWASs are tag SNPs in intergenic regions with no known function.[2] The PheWAS methodology introduces a paradigm shift by considering the phenome to be as useful as the genome when discovering gene–disease associations. Whereas GWAS associates a phenotype with genotypes across the genome, PheWAS associates a genotype with phenotypes across the phenome. For the majority of PheWASs, phenomes have been defined by structured data in an electronic medical record (EMR) system; specifically by International Classification of Disease version 9 (ICD9) codes. In the United States, ICD9 codes are used as diagnostic codes for billing purposes. ICD9 codes are limited by variable positive and negative predictive values for describing disease phenotypes, but offer an extremely efficient mechanism to broadly capture patient histories for thousands of diseases at varying levels of phenotypic resolution. Nearly 17 000 possible phenotypes can be differentiated by the ICD9 coding system,[3] and SNPs can be associated with each ICD9 code across the phenome.

The first PheWAS was published by Denny *et al*[4] in 2010 and focused on five disease-associated SNPs identified by GWASs. In this proof-of-principle study, each SNP was associated with hundreds of phenotypes (ICD9 codes) across the phenome and was capable of identifying expected associations when using a genotype-to-phenotype approach. Since this proof-of-principle study, other groups have assessed previously reported GWAS SNPs.[5–8] The use of SNPs identified by GWAS for PheWAS has the advantage of leveraging known association data when interpreting PheWAS results. For example, PheWAS results demonstrate that multiple sclerosis (MS) and erythematous conditions, including rosacea, may share a common genetic etiology (ie, *HLA-DRB1*1501*),[4,6] demonstrating the capacity of PheWAS to identify pleiotropic effects for GWAS SNPs.

A challenge with using GWAS SNPs for PheWAS is the lack of biological/functional data associated with those SNPs. An alternative PheWAS approach may be to focus on variants with expected function, such as stop-gain and stop-loss variants. A stop-gain variant, or nonsense variant, introduces a premature stop codon in the mRNA coding sequence, whereas a stop-loss variant disrupts a current stop codon. Stop-gain variants may deactivate a protein by altering protein stability, result in the deletion of important protein domains, and/or cause reduced mRNA expression because of nonsense-mediated mRNA decay mechanisms.[9] In genetic association studies, nonsense SNPs have a higher probability of being associated with a phenotype with often higher effect sizes, compared with other classes of variation (eg, missense SNPs).[10] Furthermore, nonsense mutations represent a class of variation that explains a significant proportion of Mendelian diseases,[11] and nonsense

[1]Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI, USA; [2]Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI, USA; [3]Parkland Center for Clinical Innovation, Parkland Health and Hospital System, Dallas, TX, USA; [4]Computation and Informatics in Biology and Medicine, University of Wisconsin Madison, Madison, WI, USA
*Correspondence: Dr SJ Hebbring, Center for Human Genetics, Marshfield Clinic Research Foundation, 1000 N Oak Avenue, Marshfield, WI 54449, USA. Tel/Fax: +1 715 389 3112; E-mail: Hebbring.scott@mcrf.mfldclin.edu
Received 3 January 2014; revised 27 May 2014; accepted 30 May 2014; published online 30 July 2014

variants of unknown clinical significance may have similar evolutionary selective pressures as known disease-causing mutations.[12] The biological significance of stop-loss variants is less well characterized. These variants may cause loss of function by disrupting protein stability or could result in the gain of function by the addition of extra amino acids.

In this study, we test the hypothesis that presumed functional variants may be associated with human disease and that these associations can be identified by PheWASs. Focusing on functional variants in PheWAS, primarily loss-of-function variants, is analogous to the type of reverse genetics experiment normally reserved for an animal model system (eg, transgenic knockout mouse). Whereas the use of GWAS SNPs in PheWAS leverages known phenotypic data, concentrating on functional variants leverages known biological insights.

MATERIALS AND METHODS

Ethics statement

This study was approved by the Marshfield Clinic Institutional Review Board in Marshfield, Wisconsin, and written and informed consent was acquired for all participants.

Patient population

Genotyped samples have been described elsewhere[13,14] and have been applied to a previously reported PheWAS.[6] Briefly, all individuals genotyped are self-identified white/non-Hispanic Marshfield Clinic patients recruited into the Personalized Medicine Research Project (PMRP). PMRP represents a homogenous population with 77% of participants claiming German ancestry.[13] In this PheWAS, 4235 patients were included in a discovery set and 10 640 patients were included in a validation set. The 4235 patients included in the discovery set were all over age 50 years (mean 74 years), have on average over 30 years of data in the EMR, and were originally selected as a subpopulation of PMRP to examine genetic associations with high-density lipoprotein levels or cataract disease.[15] Importantly, all 4235 samples have been genotyped with Illumina 660W SNP chip. The 10 640 patient validation set represents all additional PMRP patients over the age of 40 years. The validation subset is younger (mean 59 years of age), but with comparable years of EMR data.

Phenome definition

The phenome was defined by patient EMR data as described previously.[6] Briefly, ICD9 codes were used to define cases and controls at varying levels of phenotypic resolution (eg, ICD9 695, 695.1, 695.11). Patients coded for any one specific code became a 'case' for that code, whereas those not coded for any one specific code became a 'control.' Based on prevalence of ICD9 coding in the population, 'rule-of-two' was used to define cases for common conditions (>300 cases). Rule-of-two requires a patient to be coded two or more times to be considered a case. Because of privacy concerns, phenotypes observed less than eight times in the cohort were excluded. A total of 4841 phenotypes/ICD9 codes defined the phenome.

Statistical analysis

A total of 105 stop-gain/loss SNPs and 5 PheWAS control SNPs were genotyped in the discovery set (see Supplementary Table 1 for a list of candidate SNPs and a list of changes observed for all rs-numbers described in the study). Candidate SNPs were selected based on potential clinical relevance or availability of pre-existing genotype data. For candidate SNPs where direct genotype data were unavailable, an appropriate tag SNP was used ($r^2>0.9$). A total of 31 stop-gain/loss SNPs with the strongest biological and/or statistical PheWAS results were further genotyped in the validation set. Additional details on the SNP selection and genotyping methods are available in Supplementary Methods. Each SNP was associated across the phenome.

For common ICD9 codes, logistic regression analyses were used. Both adjusted and unadjusted models were investigated. Covariates included sex and years of EMR data. Age was not included as a covariate because of potential confounding effects that may be unique to this population (Supplementary Material and Supplementary Figures S1 and S2). For rare ICD9 codes where cell counts for a genotype fell below five in a contingency table, Fisher's exact test was used, similar to methods described in previous PheWASs.[4,6] Q–Q plots were generated for every SNP (data not shown) and at the study-wide level (Supplementary Material and Supplementary Figure S3) to measure systematic confounding or bias in the SNP–phenotype associations. The PheWAS results from the discovery set and validation set were combined by meta-analysis. Meta-analysis was implemented using a random effect model and pooling of studies using the Mantel–Haenszel method. Heterogeneity was measured by Q and I^2. All analyses were conducted using the R statistical software package (www.r-project.org, Vienna, Austria).

RESULTS

Phenomes

The discovery set phenome included 4841 ICD9 codes documented for 4235 patients with extensive genetic data available through the PMRP. PMRP patients are primarily Caucasian adult patients receiving care in the Marshfield Clinic system,[13] and the discovery set represents PMRP patients over the age of 50 years. A validation set was also drawn from the PMRP cohort and represented all additional PMRP patients over the age of 40 years. Despite being drawn from PMRP, the discovery and validation sets differed by case size and age. The mean and median case size for each phenotype in the discovery set was 217 and 59, respectively, compared with 312 and 100, respectively, in the larger validation set. The average current age in the discovery set was $\sim$74 years. In addition, the average current age in the validation set was $\sim$59 years. The average age difference between cases and unaffected controls for all ICD9 codes was larger for the older discovery set compared with the younger validation set (Supplementary Figure S1). The average age difference between cases (age at diagnosis) and controls (current age) for the discovery set and validation set was 11.5 and 8.7 years, respectively. Furthermore, this age difference resulted in 3.3 times as many cases per control in the discovery set compared with the validation set (Supplementary Figure S2).

Control SNPs

For all five PheWAS control SNPs, the ICD9 codes that defined the expected phenotypes were associated with their respective SNPs (Figure 1 and Table 1). Because the PheWAS for rs3135388 was reported previously (Figure 1a),[6] focus will be placed on the remaining four PheWAS control SNPs. In almost all cases, the expected ICD9 codes were ranked at or near the top of their respective PheWAS (Table 1), including rs9501572. Rs9501572 tags for *HLA-B27* and is known to be associated with ankylosing spondylitis.[16] Surprisingly, even with only 10 cases coded for 'other inflammatory spondylopathies' (ICD9 720.8), this code was the top PheWAS result for rs9501572 ($P=3.3\times10^{-4}$; Figure 1b).

PheWAS control SNP rs12678919 is in linkage disequilibrium (LD) with the nonsense SNP rs328 ($r^2=1$). Rs328 induces a premature stop codon in the gene for lipoprotein lipase (LPL). LPL is involved in lipid metabolism and rs328 is associated with both triglyceride and high-density lipoprotein levels.[2] The second most significant PheWAS association for rs12678919 was pure hyperglyceridemia (ICD9 272.1, $P=1.3\times10^{-4}$; Figure 1c). Interestingly, the ICD9 code for chronic inflammation of the gall bladder ('chronic cholecystitis,' ICD9 575.11) was the fourth most significant PheWAS result for this SNP ($P=5.0\times10^{-4}$). Elevated triglyceride levels are a risk factor for chronic cholecystitis and may lead to an increased risk for gallbladder cancer.[17]

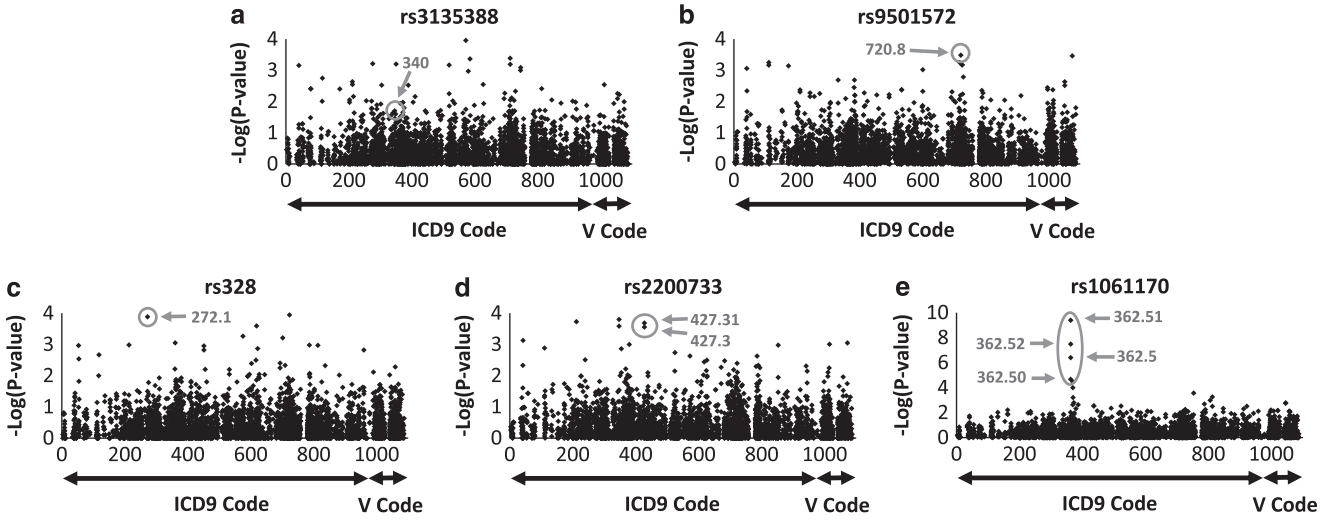

Figure 1 Manhattan plots of unadjusted −log10 (*P*-values) for the 4841 ICD9 and V codes that define the phenome. Highlighted are association results for (**a**) multiple sclerosis (ICD9 340) for rs3135388, (**b**) other inflammatory spondylopathies (ICD9 720.8) for rs9501572, (**c**) pure hyperglyceridemia (ICD9 272.1) for rs328, (**d**) atrial fibrillation (ICD9 427.31 and 427.3) for rs2200733, and (**e**) age-related macular degeneration (AMD) (ICD9 362.50, 362.51, 362.52, and 362.5) for rs1061170.

Table 1 PheWAS results for the five positive control SNPs genotyped in the discovery set

SNP	*Type*	*Gene*	*ICD9*	*Description*	*Cases (MAF)*	*Controls (MAF)*	P-*value*	*OR (95% CI)*	*Rank*
rs2200733	Intergenic	*4q25*	427.31	Atrial fibrillation	892 (0.15)	3340 (0.12)	2.1E−4	1.3 (1.1–1.5)	3
rs1061170	Missense	*CFH*	362.51	Nonexudative senile macular degeneration	1000 (0.44)	3224 (0.36)	4.1E−10	1.4 (1.2–1.5)	1
rs328	Stop-gain	*LPL*	272.1	Pure hyperglyceridemia	295 (0.059)	3772 (0.11)	1.3E−4	0.53 (0.37–0.75)	2
rs9501572	B27	*HLA-B*	720.8	Other inflammatory spondylopathies	10 (0.60)	4183 (0.25)	3.3E−4	4.5 (1.8–11)	1
rs3135388	DRB1*1501	*HLA-DRB1*	340	Multiple sclerosis	20 (0.30)	4210 (0.14)	2.4E−2	2.6 (1.3–5.1)	83

Abbreviations: CI, confidence interval; ICD9, International Classification of Disease, version 9; MAF, minor allelic frequency; OR, allelic odds ratio; SNP, single-nucleotide polymorphism.

PheWAS control SNP rs2200733 has been previously shown to be associated with atrial fibrillation by GWAS.[18,19] The third and fifth most significant phenotypes in this PheWAS were atrial fibrillation codes (ICD9 427.31, $P=2.1\times10^{-4}$; ICD9 427.3, $P=2.8\times10^{-4}$; Figure 1d). This SNP has also been reported to be associated with ischemic stroke,[20] but the associations for the ICD9 codes describing cerebrovascular disease (ICD9 430–438) were not significant.

The strongest PheWAS associations identified were observed for the control SNP rs1061170, a nonsynomous SNP in the gene for complement factor H (CFH). Rs1061170 has been associated with age-related macular degeneration (AMD) and represents one of the most significant SNPs identified by GWAS.[2] The four top PheWAS findings for this SNP included nonexudant macular degeneration (ICD9 362.51, $P=3.5\times10^{-10}$), exudative macular degeneration (ICD9 362.52, $P=9.2\times10^{-8}$), macular degeneration unspecified (ICD9 362.50, $P=2.8\times10^{-5}$), and the code that defines AMD more broadly (ICD9 362.5, $P=4.0\times10^{-7}$; Figure 1e). Furthermore, ICD9 codes describing other sight loss phenotypes were also associated (eg, visual loss not otherwise specified, ICD9 369.9, $P=6.2\times10^{-4}$). These results may indicate high correlations across some ICD9 codings.

Stop-gain/loss SNPs

Discovery set. A total of 105 stop-gain/loss variants were selected for examination by PheWAS. Of the 105 SNPs, 22 are in Online Mendelian Inheritance in Men (OMIM) genes (Supplementary Table S1). In the discovery set, 35 SNPs had at least one ICD9 code with an association of $P<1\times10^{-4}$. The strongest PheWAS signal for the stop-gain/loss SNPs was for rs3731608, an apparent nonsense SNP in GABA receptor modulator DBI, with ICD9 V58.6 that defines current long-term drug use ($P=3.2\times10^{-6}$; Table 2). Of the 35 SNPs with at least one significant association, 6 had top ICD9 codes with unlikely genetic origins. For example, rs5758511, a nonsense SNP in *CENPM*, was associated with acquired deformities of the toe (ICD9 735). These six SNPs were not considered for replication studies. Conversely, functional SNPs in OMIM genes with available biological/clinical background information were given extra scrutiny.

SNP rs2736911 introduces a premature stop codon (R38X) in the *ARMS2* gene. Multiple SNPs in multiple genes, all in LD with one another, including a missense SNP 3′ downstream of rs2736911 in *ARMS2* (rs10490924), have been associated with AMD in multiple GWASs.[2] Importantly, this region is believed to be relevant to the less common, 'wet' form of AMD.[21] According to 1000Genomes and HapMap results, rs2736911 and rs10490924 are not in LD in Caucasian populations ($r^2=0.001$–0.067). The importance of *ARMS2* is still uncertain because of unknown function for the missense SNP, LD patterns spanning multiple genes, and little knowledge regarding the function of those genes in relation to AMD.[22] Characterizing a presumed functional variant in *ARMS2*, independent of GWAS SNPs, may assist when understanding the importance of this gene for AMD. Based on PheWAS results, the nonsense SNP rs2736911 was weakly associated with the ICD9 code defining wet AMD (ICD9 362.52, $P=0.030$; Figure 2a); no other AMD-related ICD9 codes were associated.

Table 2 Association results for top PheWAS signals ($P < 1.0 \times 10^4$) in the discovery set and validation set

					Discovery set				*Validation set*			
SNP	*Type*	*Gene*	*ICD9*	*Description*	*P-value*	*OR (95% CI)*	*Cases (MAF)*	*Controls (MAF)*	*P-value*	*OR (95% CI)*	*Cases (MAF)*	*Controls (MAF)*
rs10491178	G	*ABCA10*	388	Other disorders of ear	4.0E−5	0.50 (0.36–0.69)	786 (0.028)	2717 (0.054)	0.42	0.92 (0.76–1.1)	1353 (0.046)	7733 (0.05)
rs10838851	G	*OR4×1*	617.3	Endometriosis of pelvic peritoneum	8.1E−5	2.4 (1.5–3.6)	44 (0.42)	4191 (0.24)	0.64	1.1 (0.85–1.3)	236 (0.24)	10 404 (0.24)
rs12520799	G	*C5orf20*	295.62	Chronic schizophrenic disorders	6.7E−5	11 (2.5–47)	10 (0.90)	4224 (0.45)	0.80	1.0 (0.60–1.7)	28 (0.46)	10 612 (0.46)
rs12568784	G	*FLG2*	722.7	Intervertebral disc disorder with myelopathy	5.2E−5	4.9 (2.3–11)	13 (0.46)	4222 (0.15)	0.08	1.2 (0.54–2.8)	19 (0.18)	10 621 (0.16)
rs1667366	L	*ZNF568*	716.95	Unspecified arthropathy of pelvic region and thigh	7.5E−5	1.8 (1.4–2.5)	94 (0.64)	4141 (0.49)	0.30	1.2 (0.86–1.6)	81 (0.53)	10 559 (0.49)
rs1790218	G	*SLC22 A10*	447.7	Aortic ectasia	7.9E−5	1.4 (0.68–3.0)	14 (0.50)	4221 (0.41)	0.50	0.74 (0.44–1.2)	32 (0.34)	10 608 (0.41)
rs1861050	G	*CC2D2A*	656	Fetal and placental problems affecting management of mother	8.4E−5	0.78 (0.29–2.1)	61 (0.033)	4167 (0.042)	0.13	1.2 (0.78–1.8)	260 (0.048)	9856 (0.041)
rs2176186	G	*C2orf83*	719.16	Hemarthrosis, lower leg	4.0E−5	2.1 (1.2–3.4)	30 (0.47)	4205 (0.30)	0.11	1.4 (0.91–2.3)	40 (0.39)	10 599 (0.31)
rs2176186	G	*C2orf83*	211.3	Benign neoplasm of colon	6.2E−5	2.3 (1.5–3.4)	49 (0.49)	4186 (0.30)	0.53	0.76 (0.38–1.5)	22 (0.25)	10 617 (0.31)
rs3731608	G	*DBI*	V58.6	Current long-term drug use	3.2E−6	0.70 (0.60–0.82)	3725 (0.20)	510 (0.26)	0.75	0.99 (0.93–1.1)	6671 (0.20)	3949 (0.20)
rs3743503	L	*CENPN*	389.22	Hearing loss-bilateral mixed	7.1E−5	0.91 (0.45–1.8)	38 (0.12)	4197 (0.13)	0.12	1.093 (0.63–1.9)	56 (0.13)	10 582 (0.12)
rs41463245	G	*CCL26*	530	Diseases of esophagus	8.8E−5	9.6 (4.0–23)	30 (0.10)	4200 (0.011)	1.0	0.50 (0.07–3.6)	68 (0.0070)	10 569 (0.015)
rs4148974	G	*NDUFV3*	438.84	Other late effects of cerebrovascular disease, ataxia	9.4E−5	5.3 (2.3–13)	14 (0.25)	4221 (0.059)	1.0	0.88 (0.12–6.6)	10 (0.050)	10 626 (0.057)
rs5065	L	*NPPA*	781.8	Neurologic neglect syndrome	2.4E−5	4.6 (1.8–12)	9 (0.44)	4226 (0.15)	0.26	1.5 (0.63–3.8)	15 (0.20)	10 625 (0.14)
rs5744168	G	*TLR5*	V13.09	Personal history of other specified urinary system disorders	7.7E−5	3.0 (1.7–5.3)	47 (0.16)	4177 (0.059)	0.16	1.6 (0.88–2.9)	61 (0.098)	10 575 (0.064)
rs6024911	L	*GCNT7*	363.4	Choroidal degenerations	4.4E−5	1.7 (1.3–2.1)	161 (0.31)	4070 (0.22)	0.63	0.92 (0.70–1.2)	173 (0.20)	10 466 (0.21)
rs650825	G	*OPRM1*	553	Other hernia of abdominal cavity without mention of obstruction or gangrene	5.3E−5	1.3 (1.1–1.5)	826 (0.29)	3013 (0.24)	0.42	0.96 (0.86–1.1)	1097 (0.24)	8930 (0.25)
rs677830	G	*OPRM1*	553	Other hernia of abdominal cavity without mention of obstruction or gangrene	5.3E−5	1.3 (1.1–1.5)	826 (0.29)	3013 (0.24)	0.42	0.95 (0.86–1.1)	1097 (0.24)	8930 (0.25)
rs745961	L	*CEP89*	V05.3	Need for prophylactic vaccination and inoculation against viral hepatitis	7.0E−5	1.3 (1.1–1.5)	490 (0.41)	3744 (0.35)	0.67	1.0 (0.94–1.1)	1409 (0.36)	9230 (0.36)
rs7485773	G	*ACSM4*	562.12	Diverticulitis of colon with hemorrhage	3.1E−5	4.6 (2.5–8.4)	39 (0.17)	4194 (0.042)	0.67	1.3 (0.31–5.4)	19 (0.053)	10 621 (0.041)
rs79448530	G	*C11orf40*	483	Pneumonia due to other specified organism	7.5E−5	5.3 (2.6–11)	21 (0.24)	4213 (0.056)	0.27	1.7 (0.68–4.3)	28 (0.089)	10 612 (0.054)
rs850763	G	*SLC5A9*	V15.09	Other allergy, other than to medicinal agents	1.7E−5	1.8 (1.4–2.3)	153 (0.27)	4081 (0.17)	0.95	0.99 (0.81–1.2)	312 (0.17)	10 328 (0.17)
rs935706	L	*ZNF568*	716.95	Unspecified arthropathy of pelvic region and thigh	7.5E−5	1.8 (1.4–2.5)	94 (0.64)	4141 (0.49)	0.30	1.2 (0.86–1.6)	81 (0.53)	10 558 (0.49)
rs9427397	G	*FCGR2A*	531.7	Chronic gastric ulcer without mention of hemorrhage or perforation	5.7E−5	2.8 (1.8–4.5)	43 (0.29)	4190 (0.13)	0.84	1.0 (0.52–2.0)	40 (0.13)	10 600 (0.12)
rs9973206	G	*USP29*	562.11	Diverticulitis of colon (without mention of hemorrhage)	5.7E−5	1.5 (1.2–1.7)	359 (0.24)	3875 (0.18)	0.61	0.95 (0.76–1.2)	278 (0.17)	10 223 (0.18)

Abbreviations: CI, confidence interval; G, stop-gain; ICD9, International Classification of Disease, version 9; L, stop-loss; MAF, mean allelic frequency; OR, allelic odds ratio; SNP, single-nucleotide polymorphism.

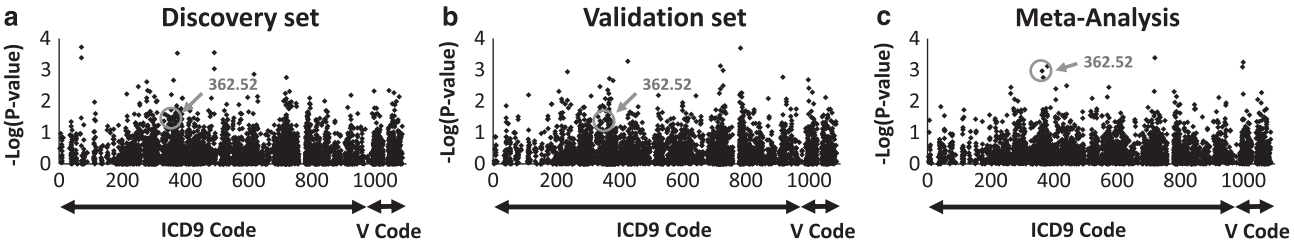

Figure 2 Manhattan plots of unadjusted −log10 (*P*-values) for the nonsense variant rs2736911 in *ARMS2*. Highlighted is the ICD9 code for wet age-related macular degeneration (AMD) in the (**a**) discovery set, (**b**) validation set, and (**c**) meta-analysis.

Validation set. A total of 31 stop-gain/loss SNPs were genotyped in the validation set for replication of discovery set results, including 24 with a clinically relevant ICD9 code ($P<1.0\times10^{-4}$) and 7 other SNPs with biologically/clinically interesting PheWAS results (eg, rs2736911 in *ARMS2*). Replication consisted of an independent PheWAS for the 31 SNPs. At the phenotypic level, none of the top ICD9 codes were replicated in the independent analysis (Table 2). However, the ICD9 code defining wet AMD was consistent with the direction of the discovery set (ICD9 362.52, $P=0.081$; Figure 2b).

To better characterize the SNPs selected for validation, and because the discovery and validation sets represented unique populations based on age differences described previously (Supplementary Figures S1 and S2), a meta-analysis was conducted using a random effect model with pooling of studies by the Mantel–Haenszel method. As expected, the ICD9 codes with $P<1.0\times10^{-4}$ in the discovery set were not significant in the meta-analysis when discovery set *P*-value biases were considered. Conversely, the fifth most significant association from the PheWAS meta-analysis of rs2736911 was the ICD9 code that defines wet AMD (ICD9 362.52, $P=0.0011$, OR = 0.69; Figure 2c). When considering other GWAS results,[2] in combination with the results of the PheWAS described here, it may be more likely that *ARMS2* is the candidate involved in wet AMD compared with other genes in the region.

In addition to the disease-specific meta-analyses for the associations identified in the discovery set, a meta-analysis across the phenome of all 31 SNPs genotyped in both the discovery and validation sets was assessed to generate new hypotheses. The most significant PheWAS meta-analysis was between rs1861050 and the ICD9 code that defines 'encounter for other and unspecified procedures and aftercare' (ICD9 V58, $P=1.8\times10^{-6}$). The importance of this association, and others reported (Supplementary Table S2), is unclear and will require further study.

DISCUSSION

All genetic-based PheWASs published thus far have focused on genetic variants with known phenotypic associations driven primarily by GWAS results.[1] The advantage of focusing on GWAS SNPs in PheWAS includes the potential to identify variants with pleiotropic effects, as emphasized by previously published results and associations observed with PheWAS control SNPs in this study. For example, nonsense SNP rs328 in *LPL* is not only associated with triglyceride levels, but may also be a risk factor for chronic cholecystitis. The use of GWAS SNPs in PheWAS capitalizes on known association data. Unfortunately, the majority of GWAS SNPs, except for the *LPL* example given above, are intergenic SNPs with unknown function, making translation of association results into biological insight a challenge. An alternative PheWAS approach may focus instead on known functional variants.

Like GWAS, PheWAS is a hypothesis-generating approach that is challenged by multiple comparison testing. The common approach to account for multiple testing in PheWAS is the use of a Bonferroni correction.[1,4,6,23–25] With 4841 phenotypes, 110 SNPs, and a study-wide α of 0.05, an association with $P<9.4\times10^{-8}$ would be required to identify a statistically significant association assuming independence. The only association that meets this criterion is that between the PheWAS control SNP rs1061170 in *CFH* and ICD9 codes that define AMD (Figure 1e). Although a Bonferroni threshold has been commonly applied in PheWASs, a Bonferroni correction may be overly conservative when correlations exist between ICD9 codes.[1] Regardless, interesting associations were identified in this study that may provide insight into the genetic etiologies of complex conditions.

We demonstrated that stop-gain/loss variants may be used to identify important genes/polymorphisms involved in human disease. Furthermore, by focusing on functional variants, biological insights may be inferred. For example, several SNPs in the chromosomal region 10q26.13 have been associated with AMD, including wet AMD.[2] This region contains multiple genes, including *ARMS2* and *HTRA1*. One of the GWAS-significant SNPs associated with AMD is a missense SNP in *ARMS2* (rs10490924). Because of LD across the region and the lack of biological insights for the genes and SNPs in the region, the importance of *ARMS2* in AMD is uncertain. Our PheWAS results demonstrate that a nonsense SNP in *ARMS2* (rs2736911), independent of the missense SNP, is associated with AMD (Figure 2c). The loss-of-function allele for rs2736911 had a protective effect (OR = 0.69) that infers that the missense SNP rs10490924 could result in a gain of function. When considering previously reported GWASs, this PheWAS result strengthens the hypothesis that *ARMS2* is involved in the pathophysiology of AMD and that two or more variants in *ARMS2* may affect risk. Although unlikely, we cannot rule out the potential that these independent coding variants may be in LD with other functional variants outside of *ARMS2*.

The example of *ARMS2* described above demonstrates that a variant with presumed function can be used to identify gene–disease associations by PheWAS. *ARMS2* is one of 22 OMIM genes of focus in the present study, due in part to presumed clinical importance (Supplementary Table S1). However, many of the other SNPs in OMIM genes did not result in expected associations. The lack of associations could be the result of inherent differences between the discovery set and validation set. The discovery set was older than the validation set. As a result, the older discovery set had more cases for every control compared with the validation set (Supplementary Figure S2). This could indicate that a proportion of controls in the younger validation set may end up developing a disease at an older age. The impact of this is uncertain given that the number of controls is far greater than the number of cases for any given phenotype in a PheWAS.

The lack of associations identified for SNPs in OMIM genes could also be the result of incomplete phenomes, uncertainties in disease manifestations, and/or challenges in describing true function for any given stop-gain/loss SNP. For example, rs1861050 is a nonsense variant in *CC2D2A*. Mutations in *CC2D2A* are believed to cause COACH syndrome, although rs1861050 has not been implicated. COACH syndrome is a rare autosomal recessive disease related to Joubert syndrome.[26] Based on PheWAS results, there were a total of 25 individuals in the discovery and validation set homozygous for the nonsense variant. No ICD9 codes indicative of COACH syndrome were associated with this SNP. This could be explained by a variety of reasons. First, the condition may be too rare to be captured within the phenome. Joubert syndrome, and several other congenital diseases, can be coded as ICD9 759.89. This code does not exist in our phenome. Because of privacy concerns, only those ICD9 codes that existed eight or more times in the population were analyzed.[6] Second, COACH syndrome is frequently characterized by mental disabilities, potentially limiting the capacity of affected subjects to consent into PMRP. Alternatively, unaffected individuals homozygous for the nonsense SNP may be an indication of incomplete penetrance. Lastly, and perhaps most likely, challenges remain when interpreting function for some apparent nonsense SNPs.

Stop-gain/loss SNPs were selected because they fall in a class of variation that may be more likely to perturb function and result in a pathogenic effect. MacArthur *et al*[27] were one of the first to systematically survey loss-of-function variants for disease associations. Using the Wellcome Trust Case Control Consortium, 417 loss-of-function variants were associated with seven complex phenotypes in nearly 16 000 patients. With a limited number of phenotypes, this association study was mostly negative, but was able to rediscover an association between a frame shift variant in *NOD2* and Crohn's disease. This study also highlighted challenges when predicting function in that nonsense SNPs are enriched toward the 3′ end of affected genes, suggesting that partial truncation is tolerable.[27] Furthermore, SNP function may also be mitigated by unappreciated alternative splicing. In the *CC2D2A* example mentioned above, *CC2D2A* encodes three splice isoforms. Rs1861050 codes for a nonsense SNP in mRNA transcript NM_020785, but codes for a synonymous variant in the two other mRNA transcripts (NM_001080522 and NM_001164720). Disease-causing mutations identified in *CC2D2A* have been detected in the NM_001080522 transcript.[28] Interestingly, this SNP has been reported to be associated with conduct disorder by GWAS, although the SNP did not reach GWAS significance ($P = 8 \times 10^{-6}$).[29] The ICD9 code for conduct disorder (ICD9 314) was modestly associated with the rs1861050 genotype in the meta-analysis ($P = 0.037$). Further investigation into the importance of this SNP in conduct disorder is necessary.

As evident by the five PheWAS controls SNPs (Figure 1 and Table 1), and GWAS SNPs previously assessed by PheWAS,[1] it is relatively straightforward to identify true associations when knowledge of expected associations is incorporated into the interpretation of PheWAS results. When there is little information on expected phenotypes, identifying true associations is a challenge. This is especially true when no association meets a conservative Bonferroni correction. Regardless, this PheWAS demonstrates the potential to identify gene–disease associations when focusing on functional variants. This study also highlights the limitations of PheWAS. For example, many conditions in a phenome may be rare. With a small number of cases, the power to detect an association may be difficult. Focusing on deleterious variants may increase power if these variants have higher effect sizes compared with other types of variation,[10] such as those identified by GWAS. In the future, the challenge of small sample sizes may be exacerbated when ICD10 coding is implemented. Whereas ICD9 offers 17 000 possible codes, ICD10 has nearly 1 55 000 possible codes.[30] The limitation of sample size is undoubtedly temporary. As biobanks continue to grow, and genetics/genomics is incorporated into standard medical practice, sample size may not be a limiting factor in future PheWASs. This expectation may open the door to PheWASs that focus on presumed functional variants of unknown significance identified by clinical next-generation sequencing.

As demonstrated by this and other reports, PheWAS has the capacity to rediscover SNP–disease associations when known phenotype data are incorporated into the interpretation of PheWAS results. This study also demonstrates that focusing on variants with presumed function and incorporating biological insights may be an effective PheWAS strategy to identify SNP–disease associations. Importantly, this approach may offer additional biological insights that GWAS SNPs tend not to provide as demonstrated by the *ARMS2* nonsense SNP described above. As the genome and the phenome become better defined, PheWAS may provide an effective role when evaluating the use of human genetics for the application of individualized medicine.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support from the National Center for Research Resources (1UL1RR025011), National Center for Advancing Translational Sciences (9U54TR000021), U.S. National Library of Medicine (5T15LM007359 and 1K22LM011938), and the Marshfield Clinic Research Foundation. We also thank Rachel Stankowski for her assistance in editing this manuscript.

1 Hebbring SJ: The challenges, advantages, and future of phenome-wide association studies. *Immunology* 2013; **141**: 157–165.
2 Hindorff LA, MacArthur J, Morales J *et al*: A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed 24 July 2013.
3 U.S. Department of Health and Human Services: improving the health, safety, and well-being of America 2008. Available at: http://www.hhs.gov/news/press/2008pres/08/20080815a.html. Accessed 9 September 2013.
4 Denny JC, Ritchie MD, Basford MA *et al*: PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010; **26**: 1205–1210.
5 Denny JC, Bastarache L, Ritchie MD *et al*: Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013; **31**: 1102–1110.
6 Hebbring SJ, Schrodi SJ, Ye Z *et al*: A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun* 2013; **14**: 187–191.
7 Pendergrass SA, Brown-Gentry K, Dudek S *et al*: Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* 2013; **9**: e1003087.
8 Liao KP, Kurreeman F, Li G *et al*: Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non-rheumatoid arthritis controls. *Arthritis Rheum* 2013; **65**: 571–581.
9 Kervestin S, Jacobson A: NMD: a multifaceted response to premature translational termination. *Nature Rev Mol Cell Biol* 2012; **13**: 700–712.
10 Chen R, Davydov EV, Sirota M, Butte AJ: Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS One* 2010; **5**: e13574.
11 Stenson PD, Mort M, Ball EV *et al*: The Human Gene Mutation Database: 2008 update*Genome Med* 2009; **1**: 13.
12 Abecasis GR, Altshuler D, Auton A *et al*: A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
13 McCarty CA, Wilke RA, Giampetro PF, Wesbrook SD, Caldwell MD: Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Personalized Med* 2005; **2**: 49–79.
14 McCarty CA, Chisholm RL, Chute CG *et al*: The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011; **4**: 13.

15 Turner SD, Berg RL, Linneman JG *et al*: Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. *PLoS One* 2011; **6**: e19586.
16 Thomas GP, Brown MA: Genetics and genomics of ankylosing spondylitis. *Immunol Rev* 2010; **233**: 162–180.
17 Wistuba II, Gazdar AF: Gallbladder cancer: lessons from a rare tumour. *Nat Rev Cancer* 2004; **4**: 695–706.
18 Gudbjartsson DF, Holm H, Gretarsdottir S *et al*: A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat Genet* 2009; **41**: 876–878.
19 Gudbjartsson DF, Arnar DO, Helgadottir A *et al*: Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* 2007; **448**: 353–357.
20 Gretarsdottir S, Thorleifsson G, Manolescu A *et al*: Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Ann Neurol* 2008; **64**: 402–409.
21 Dewan A, Liu M, Hartman S *et al*: HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* 2006; **314**: 989–992.
22 Yu W, Dong S, Zhao C *et al*: Cumulative association between age-related macular degeneration and less studied genetic variants in PLEKHA1/ARMS2/HTRA1: a meta and gene-cluster analysis. *Mol Biol Rep* 2013; **40**: 5551–5561.
23 Denny JC, Crawford DC, Ritchie MD *et al*: Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet* 2011; **89**: 529–542.
24 Ritchie MD, Denny JC, Zuvich RL *et al*: Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 2013; **127**: 1377–1385.
25 Shameer K, Denny JC, Ding K *et al*: A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum Genet* 2013; **133**: 95–109.
26 Online Mendelian Inheritance in Man, OMIM®: *McKusick-Nathans Institute of Genetic Medicine*. Baltimore, MD: Johns Hopkins University, 2013. Available at: https://omim.org/. Accessed 1 September 2013.
27 MacArthur DG, Balasubramanian S, Frankish A *et al*: A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 2012; **335**: 823–828.
28 Gorden NT, Arts HH, Parisi MA *et al*: CC2D2A is mutated in Joubert syndrome and interacts with the ciliopathy-associated basal body protein CEP290. *Am J Hum Genet* 2008; **83**: 559–571.
29 Dick DM, Aliev F, Krueger RF *et al*: Genome-wide association study of conduct disorder symptomatology. *Mol Psychiatry* 2011; **16**: 800–808.
30 Centers for Medicare and Medicaid Services: ICD-10. Baltimore (MD): CMS.gov. 2010. Available at: http://www.cms.gov/Medicare/Coding/ICD10/index.html?redirect=/icd10. Accessed 9 September 2013.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)