Rare Variant Pathogenicity Triage and Inclusion of Synonymous Variants Improves Analysis of Disease Associations

RIDGE DERSHEM,[1*], RAGHU P.R. METPALLY,[1*§], KIRK JEFFREYS,[1], SARATHBABU KRISHNAMURTHY,[1], DAVID J. CAREY,[1], MICHAL HERSHFINKEL,[2], JANET D. ROBISHAW,[3], GERDA E. BREITWIESER,[1§]

[1]*Functional and Molecular Genomics, Geisinger Health System, Danville, PA,*

[2]*Faculty of Health Sciences, Ben-Gurion University of the Negev, Israel*

[3]*College of Medicine, Florida Atlantic University, Boca Raton, FL*

Keywords: zinc receptor, GPR39, sequence kernel association test, whole exome sequencing, synonymous variants, codon bias, orphan G protein-coupled receptors

*contributed equally

[§]Corresponding authors:

Raghu P.R. Metpally, Weis Center for Research, Geisinger Clinic, Functional and Molecular Genomics, 100 N. Academy Avenue, Danville, PA 17822-2608, rmetpally@geisinger.edu; phone 570 271-8669.

Gerda E. Breitwieser, Weis Center for Research, Geisinger Clinic, Functional and Molecular Genomics, 100 N. Academy Avenue, Danville, PA 17822-2604; gebreitwieser@geisinger.edu; phone 570 271-6675.

Abstract

Many G protein-coupled receptors (GPCRs) lack common variants that lead to reproducible genome-wide disease associations. Here we used rare variant approaches to assess the disease associations of 85 orphan / understudied GPCRs in an unselected cohort of 51,289 individuals. Rare loss-of-function plus likely pathogenic / pathogenic missense variants and a subset of rare synonymous variants were used as independent data sets for sequence kernel association testing (SKAT). Strong, phenome-wide disease associations common to at least two variant categories were found for ~40% of the GPCRs. Functional analysis of rare missense and synonymous variants of GPR39, a Family A GPCR activated by Zn^{2+}, validated the bioinformatics and SKAT analyses by demonstrating altered expression and/or Zn^{2+} activation for both variant classes. Results support the utility of rare variant analyses for determining disease associations for genes without impactful common variants, and the importance of including rare synonymous variants.

Background and Rationale

The superfamily of G protein-coupled receptors (GPCRs) translates extracellular signals from light, metabolites and hormones into cellular changes, which makes them the targets of a significant fraction of drugs currently on the market[1]. Genome-wide association studies (GWAS) on common variants in GPCRs have begun to identify their contributions to various disease processes[2,3]. However, many GPCRs lack common variants and alternate strategies are needed to understand their roles. Recently, sequence kernel association testing (SKAT) on rare variants in GPCRs have been developed to assess disease associations. While the traditional method relies on binning all rare variants within a genomic region or gene[4,5], more recent methods are designed to group rare variants most likely to contribute to disease associations, or aggregate variants based on domain or family structures[6,7].

In this study, we performed SKAT analysis on rare variants in 85 orphan or understudied GPCRs that have not been amenable to GWAS studies. We binned according to the following functional classes: loss-of-function (LOF, truncation and frameshift) variants; missense variants with predicted pathogenicity; or synonymous variants showing altered codon bias. We performed independent SKAT analyses on the various functional classes to determine their disease associations. Remarkably, for those orphan/understudied GPCRs with sufficient numbers of patients for statistical analyses, we found the top disease associations were common to all functional classes.

Next, we focused on *GPR39* to assess the validity of the disease association results. Among its particular advantages, *GPR39* contained sufficient numbers of variants in all three functional classes to be amenable to rare variant approaches. Second, *GPR39* is a small Family A GPCR, allowing rapid generation of mutants. Third, GPR39 is activated by extracellular Zn^{2+} and is coupled to inositol phosphate production[8,9], permitting straightforward functional analyses. Finally, we have only a rudimentary understanding of GPR39 function despite its broad expression and multiple potential role(s) in human physiology[10,11]. Our results demonstrate the validity of a combined computational and functional approach to provide important insights in orphan/understudied GPCR function and clinical implications. They also focus attention on the importance of synonymous variants having altered site-specific codon usage on

disease associations. This strategy can be readily applied to other classes of genes without functionally important common variants.

Results

Large-scale studies of orphan/understudied GPCRs to characterize natural genetic variation in the human population can provide insights into the biological function and/or potential causal contributions to disease processes[12]. The DiscovEHR collaboration represents a tremendous resource[13], which has to date provided whole exome sequences and clinical information on 51,289 individuals. In this cohort, we identified sequence variants (common (MAF≥1) and rare (MAF<1%)) for 85 orphan/understudied GPCR genes. To predict their functional impact, variants were annotated and sorted into three classes in order of predicted severity, i.e., Loss-of-Function (LOF), missense and synonymous variants. The LOF class includes variants having a premature stop codon, loss of a start or stop codon, or disruption of a canonical splice site. The missense class contained predicted pathogenic (pP) or likely pathogenic (pLP) variants identified with the RMPath pipeline (Supplemental Methods). Finally, the synonymous class had variants with significantly different codon frequency relative to the reference codon (termed SYN_ΔCB, i.e., synonymous variants with altered codon bias).

Determining disease associations for genes having only low frequency variants requires binning of variants to increase statistical power. Binning methods have focused on gene or genetic region, and have recently been expanded to incorporate regulatory regions and/or pathways by incorporating biological information from curated knowledge databases[14]. For this study, we were specifically looking for clinically impactful rare coding variants which could be validated by functional studies of the relevant GPCRs. We used sequence kernel association tests (SKAT) and binned variants according to functional classes described above, and focused on the disease associations which were common to more than one class of variant. **Table 1** shows the top disease associations for the 15 GPCRs that had sufficient variants in the LOF class. Supporting validity, associations between GPR37L1 and epilepsy[15], and LGR5 and alopecia[16,17] have been previously identified by other means. Attesting to the discovery potential, other disease

associations were novel, including GPR161 and hyperhidrosis, LGR4 and Sicca syndrome, GPR153 and hyperpotassemia, GPR84 and anemia, and GPR39 and peripheral nerve damage. Notably, the phenome-wide disease associations for the predicted LOF class showed congruence across the other variant classes for the subset of GPCRs having sufficient missense or SYN_ΔCB variants. While predicted LOF variants are easily identified and likely to have the greatest functional impact, missense variants classed as pLP + pP can also have significant effects on GPCR function. For the subset of GPCRs without sufficient LOF variants, we found significant disease associations for the missense and SYN_ΔCB classes of variants, **Table 1**. Some of the associations validate results found in previous studies, including GPR183 and disorders of liver[18], and GPR85 with acute myocardial infarction[19]. Other associations were novel, and represent potential targets for future study, including GPR132 and disturbances in sulphur-bearing amino acid metabolism, GPR176 with asthma, GPR12 with epilepsy, and GPRC5D with renal osteodystrophy. Parallel SKAT analyses of the three distinct classes of rare variants produced two important results. First, independent analysis of distinct classes of variants yielded concordant disease associations. Second, the most consistent source of phenome-wide disease associations was SYN_ΔCB. Altogether, these results provide a strong rationale for including all three functional variant classes in disease association analyses.

Disease associations provide a rich source of hypotheses regarding the potential contributions of GPCRs. The first step in defining a causal role for these rare variants, however, is functional validation of the SKAT results. Here we focused in particular on rare synonymous variants with altered codon frequency (SYN_ΔCB), whose functional consequences are not well understood. Many of the GPCRs have no known agonist, and their dominant signaling pathways have not been characterized (**Table 1**). We therefore chose to focus on *GPR39*, the Zn^{2+} receptor, which has been characterized in both cellular and knockout mouse models, and both agonist and dominant signaling pathway are known[10,11].

All common and rare variants within *GPR39* were identified from the DiscovEHR[13] cohort (**Figure 1)**. None of the *GPR39* variants were present in the ClinVar database. Accordingly, all LOF, missense, and synonymous variants were analyzed and classified as described for **Table 1**. The common variants were

used for Phenome-wide Association Studies (PheWAS). The three common variants included a missense variant (2:133174764, Ala50Val) predicted to be benign, a synonymous variant (2:133174999; Thr128Thr), which changed a common to the rarest codon (ACA, (28%, to ACG, 12%), and a non-coding variant (2:133402607). Notably, no phenome-wide disease associations were detected (Supplemental **Figure S1**).

The rare variants used for SKAT, i.e., LOF, missense (pLP + pP), and SYN_ΔCB variants, were distributed throughout the coding region of *GPR39* (**Figure 2**). Examination of the 39 missense variants that were categorized as pLP and pP confirmed that the RMPath pipeline identified variants which are likely to impact GPR39 function (**Figure 2**). The three pP variants (S78L, C108R, R133G) included a cysteine participating in a disulfide bond at the extracellular face of GPR39[20]. The pLP variants included one which has been shown to participate in binding Zn^{2+} (H19R)[21], two variants in a disulfide bond cysteine (C191G, C191W)[20], a residue in the highly conserved NPXXY motif of Family A GPCRs (N340S)[22], and a variant distal to the palmitoylation site in the proximal carboxyl terminus (R352Q). Other residues in the pLP category introduce charges or proline residues within the transmembrane helices or extracellular loops (S49N, M87I, T122M, A167S, P179T, R193C, S222Y, T291I, A293P, R302W, R302Q, A307V, P310A, Y314R, R320Y, S336R, R352Q, R362C, H367R, R386C, and R390C). We also identified the subset of 21 variants which introduced at least a two-fold change in site-specific codon frequency (SYN_ΔCB; denoted by arrowheads in **Figure 2**).

The initial SKAT screen compared the top associations of all three independent classes of rare variants. To more fully explore the results for *GPR39*, we identified all significant disease associations. We increased power by combining all individuals heterozygous for rare LOF, pLP and pP variants (termed LOF/MISS). Analysis was restricted to those ICD9 codes with at least 200 individuals in the DiscovEHR cohort having 3 or more independent instances of the code in their electronic health record (EHR)[23], and the results are plotted against ICD9 codes in **Figure 3A**. A second, independent SKAT analysis was run on the rare SYN_ΔCB variants, plotted in **Figure 3B**. Both the LOF/MISS and SYN_ΔCB analyses had a significant number of phenome-wide associations in common. All results common to both groups, adjusted

for false discovery rate, are listed in **Table 2**. Top associations for *GPR39* included lesion of the ulnar nerve and benign prostate hyperplasia with urinary obstruction. Note that individuals who had both rare LOF/MISS and SYN_ΔCB variants were excluded from the analyses. To validate the importance of rare variant pathogenicity triage for binning prior to SKAT analysis, we combined all rare variants (LOF, missense, and synonymous variants; 209 variants, 177 used in analysis) and ran SKAT. The only significant disease association was upper quadrant abdominal pain (ICD9 789.02; Q-value 0.0058). As a separate dataset, we ran SKAT on those rare missense variants predicted to be benign (**Figure 1**, pB; 61 variants); no statistically significant disease associations were obtained (Q-values=1.0). Likewise, we ran SKAT analysis on all rare synonymous variants (47 variants, 45 included in analysis), and found top associations similar to those found for SYN_ΔCB, but with lower Q-values, indicating dilution of impactful variants, i.e., lesion of ulnar nerve (Q-value 1.14E-5), and benign prostate hyperplasia with urinary tract obstruction and other lower urinary symptoms (Q-value 1.71E-5). We conclude that pathogenicity triage to refine rare variant binning was crucial to identifying disease associations for the LOF and missense variants, and binning of the subset of rare synonymous variants having large changes in local codon frequency bias strengthened disease associations. Further, refining the rare variant input into the SKAT analyses yielded disease associations common to multiple independent classes of variants, strengthening confidence in the associations.

To confirm the functional basis for the observed disease associations, we generated the rare missense and SYN_ΔCB variants in the reference human *GPR39* background. The cDNA construct was engineered to include an amino terminal 3X FLAG epitope followed by a minimal bungarotoxin binding site for ease of expression analysis and subcellular localization, respectively. Expression in HEK293 cells of equivalent amounts of cDNA and subsequent western blotting of lysates showed significant differences in net expression levels of the rare missense and SYN_ΔCB variants. Ten missense variants had significantly reduced expression and one had significantly increased expression relative to WT GPR39 (Supplemental **Figure S2**). Rare missense variants with altered expression were distributed throughout

GPR39 structural domains, as color coded in **Figure 4A**. Likewise, SYN_ΔCB variant expression was significantly greater than WT GPR39 for three variants, while fourteen variants, including the three pP variants, had significantly reduced expression (Supplemental **Figure S2**), quantified in **Figure 4B**. Surprisingly, despite the change from rare to common codon frequency for the majority of SYN_ΔCB variants, only four showed increased expression relative to WT. Four variants went from common to rare codons, denoted by * on the x-axis in **Figure 4B**. Contrary to expectations, Q265 was expressed at significantly higher levels than WT (~150%), despite the large decrease in codon frequency. Overall, results argue that factors in addition to site-specific codon usage bias determine net expression levels for rare synonymous variants.

GPR39 signaling is complex, with multiple signaling outputs. However, in both endogenous and heterologous expression systems[8,9,24], GPR39 activates inositol trisphosphate (IP_3)-dependent release of intracellular Ca^{2+}. The first step in the signaling pathway is G_q-mediated activation of phospholipase C, which produces diacylglycerol and IP_3. Here we exposed HEK293 cells transiently expressing WT or variant GPR39 to its agonist, $ZnCl_2$, and used a FRET assay to quantify accumulation of intracellular inositol monophosphate (IP_1) in the presence of LiCl. **Figure 5A** illustrates relative levels of IP_1 for WT and representative missense variants having decreased (R386C) or increased (W314R) EC_{50} for activation by $ZnCl_2$. **Figure 5B** plots the EC_{50}s for WT and all other missense variants, color-coded for their locations in GPR39 domains. Four variants, S78L, M87I, T122M and R133G, had no measurable activity, likely accounted for by low levels of expression. Three additional variants, C108R, R193C and R302Q, had dose-response curves that were significantly right-shifted, precluding fitting to determine EC_{50}s or V_{MAX}s, despite expression levels not significantly different from WT. Likewise, **Figure 5C** illustrates the dose-response relations for WT GPR39 and representative SYN_ΔCB variants which decrease (V354) or increase (I29) the EC_{50} for $ZnCl_2$ activation, and **Figure 5D** plots the EC_{50}s of WT GPR39 and all SYN_ΔCB variants analyzed. In contrast to the missense variants, all SYN_ΔCB variants had $ZnCl_2$-stimulated IP_1 activity. Six variants had increased EC_{50}s and three variants having significantly reduced EC_{50}s relative to WT.

Surface expression of all rare missense and SYN_ΔCB variants was assessed by BgTx-Alexa488 binding on non-permeabilized cells. Surface expression was observed for all missense and SYN_ΔCB variants (**Figures 5E, 5F**), and generally correlated with the extrapolated V_{max} values for the IP_1 assays (**Figures 5G, 5H**). Notable exceptions include the missense variants C108R, R193C and R320Q, which had surface expression comparable to WT but little IP_1 activity, precluding estimates of EC_{50} or V_{MAX}. Likewise, the most striking SYN_ΔCB variant, Q265, had an overall increase in net expression (**Figure 4B**), increased EC_{50} relative to WT (**Figure 5D**), and increased surface localization and V_{MAX} relative to WT (**Figure 5F, 5H**). Overall, these results argue for a striking similarity in functional impact on expression and IP_1 signaling for both rare missense and SYN_ΔCB variants of GPR39, rationalizing the disease associations identified for both classes of rare variants. The congruence of top disease associations between rare LOF, pathogenic missense and SYN_ΔCB variants for a significant subset of GPCRs (recall **Table 1**) argues strongly for the utility of these analytical approaches for identification of functionally impactful rare variants, and the imperative to consider rare synonymous variants in genomic analysis.

Discussion

The related GWAS and PheWAS approaches, which utilize common variants, have produced strong, validated targets for drug development[2]. However, many genes do not have common variants, and additional approaches are required. The advantages of rare coding variant approaches are two-fold. First, rare variants can have stronger effects on protein function, and second, assessment of the impact of rare coding variants on protein expression and function is straightforward. However, due to their low frequency, it is necessary to combine rare coding variants to increase the statistical power for detecting disease associations. In this study, we used pathogenicity triage as a basis for rare variant binning, based on the hypothesis that aggregating variants with similar directions of effect on function would amplify disease associations. We considered, in order of the strength of their potential effects on protein function, predicted LOF variants, missense variants predicted to be pathogenic or likely pathogenic, and synonymous variants

having altered codon frequency bias. For GPR39, we compared agnostic binning of all rare variants with the disease associations identified for the three sorted classes of variants. Of note, binning of all rare variants yielded no significant disease associations, while curated classes of those same rare variants had concordant top phenome-wide disease associations, obtained with non-overlapping individuals. Likewise, SKAT analysis of the subset of rare missense variants predicted to be benign yielded no significant disease associations. Results argue strongly that binning all rare variants in a gene or genetic region dilutes the effects of impactful variants, and appropriate pathogenicity triage of rare variants is an integral first step in successful application of disease association analyses.

Over the past 10 years, rare synonymous variants have been shown to impact protein expression and function and therefore disease associations with a frequency comparable to non-synonymous variants, although the mechanism(s) by which they exert their biological function are diverse[25-36]. Rare synonymous variants have not, however, been systematically utilized for disease association analysis[31,37-39]. Here we demonstrate that a subset of GPCRs (38%) have impactful rare synonymous variants which yield disease associations comparable to LOF/missense variants in the same gene, which may provide impetus for exploratory drug development despite the lack of an identified endogenous agonist. Our data also highlight the importance of improved approaches for pathogenicity triage of rare synonymous variants. For GPR39, SKAT analysis on all rare synonymous variants for the most part yielded disease associations similar to those obtained with the SYN_ΔCB subset, but having reduced Q-values, arguing for dilution of impactful variants. However, a few disease associations, including psoriatic arthropathy, showed higher Q-values in the combined synonymous variant analysis (1.79E-10) versus SYN_ΔCB (0.00013), implying that simply isolating those synonymous variants having altered local codon usage bias missed some variants with significant functional impacts (see **Table 3**). Using both LOF/missense and synonymous variants as independent data sets in the SKAT analyses increases confidence in the identified disease associations and provide the impetus for further functional studies.

A disease association does not constitute proof of causation, but implies an impact on protein expression or function, which must be validated experimentally. As proof-of-concept, *GPR39* was chosen because its agonist and dominant signaling pathway were known, and rare variant classes produced strongly congruent disease associations. We identified 168 rare (MAF<1%) variants in *GPR39* in the 51,289 WES DiscovEHR cohort. Most individuals were heterozygous for the rare variants studied, and only 1/327 individuals had more than one rare variant. Of the 110 missense variants, the RMPath pipeline predicted that 28 were likely to be pathogenic (pLP + pP). Among these, rare variants previously shown to be important for GPR39 function were identified, including the zinc binding site[21], a critical extracellular disulfide bond[20], the pH sensor[9], and the NPXXY motif common to all Family A GPCRs[22]. When expressed in HEK293 cells, variants had the expected effects on EC_{50}s and maximal activities. Of the 47 synonymous variants, the codon usage table identified 21 variants that resulted in altered codon usage frequency; most had significant effects on expression, surface localization, and Zn^{2+}-induced IP_1 production. Surprisingly, the structural distributions of rare missense and SYN_ΔCB variants were distinct. The majority of missense variants were localized to the transmembrane domain (77%), half in the extracellular loops (10/26), one an intracellular loop and the remainder in the transmembrane helices (9/26). In contrast, ~50% of SYN_ΔCB variants were in the carboxyl terminus; the remainder were distributed among the transmembrane helices (3/20) and intracellular (2/20) and extracellular (4/20) loops. The carboxyl terminal variants in both classes had significant effects on expression and signaling, primarily reducing V_{MAX}. The carboxyl terminus of GPR39 is relative large (108 amino acids) and has 22 putative phosphorylation sites. GPR39 signaling outputs are regulated by agonist-toggled binding of protein kinase inhibitor β[40]. In addition, phosphorylation by Rho kinase mediates desensitization of GPR39[41]. Thus, carboxyl terminal missense or synonymous variants may alter secondary structure and/or phosphorylation to impact signaling. Results focus attention the carboxyl terminus as a critical domain mediating distinct aspects of GPR39 signaling. It should be noted that GPR39 signaling is complex, including G_s-mediated cAMP production, G_q-mediated IP_3 accumulation, and SRF-SRE-mediated transcriptional regulation via $G_{12/13}$[41]. The current study

examined only G_q-mediated IP_3 generation, leaving open the possible impact of missense and/or synonymous variants on signaling bias.

Based on ClinVar and related databases, *GPR39* has not been causally linked to disease. Notably, both LOF/MISS and SYN_ΔCB variants yielded disease associations by SKAT which suggest a role for GPR39 in nerve function, benign hyperplasia of the prostate, diseases of hair and hair follicles, and benign essential hypertension. These associations are in agreement with cellular and animal studies which suggest that GPR39 regulates neuronal transport of potassium and chloride, attenuating post-synaptic excitability[10,42-44]. Animal models suggest that GPR39-mediated up-regulation of the neuronal K^+/Cl^- cotransporter attenuates seizure activity[44], and knockout of GPR39 promotes depression[45]. GPR39 is up-regulated by some anti-depressants[46] in mouse models, and in one study in humans, shown to be significantly reduced in the hippocampus and cortex of suicide victims[47]. The present study focused on the general patient population, and all of the patients in the present study were heterozygous for *GPR39* rare variants. The highest levels of Zn^{2+} in the body are in the prostate and seminal fluid[48], and GPR39 is differentially expressed in normal versus malignant prostate cells, where it regulates cell growth and proliferation[8,10,11], potentially contributing to the association with benign prostate hyperplasia observed in the present work. Finally, the observation that individuals heterozygous for rare variants in GPR39 have a significant association with diseases of hair and hair follicles should be considered in light of a recent study that found that GPR39 marks a class of stem cells in the sebaceous gland and contributes to wound healing in mice[49]. Overall, the identification of a subset of clinical phenotypes that echo cell culture and animal model studies supports disparate roles for GPR39 in human physiology, and validates the SKAT approaches used herein to identify disease associations. Additional phenotypes, not previously associated with GPR39 including psoriatic arthropathy and macular puckering of the retina, deserve further attention. In conclusion, we have identified strong disease associations for GPR39 using either LOF/missense or rare synonymous variants, which confirm and extend cellular and animals studies. Rare variants from both the missense and synonymous classes altered GPR39 expression and function, providing a rationale for disease associations. The approach applied to a larger set of orphan/understudied GPCRs also yielded robust

associations with phenome-wide significance, providing potential novel targets for further investigation. The methods described in this report are easily applied to any set of genes of potential clinical importance with available genomic and integrated longitudinal electronic health records. Finally, this study highlights the importance of rare synonymous variants in human physiology, and argues for their routine inclusion in any comprehensive analysis of genomic variants as potential causes of disease.

Materials and Methods

Whole Exome Sequence (WES) Data. The MyCode® Community Health Initiative recruits Geisinger Health System (Geisinger) patients from a broad range of inpatient and outpatient clinics to integrate genetic information with their clinical Electronic Health Record (EHR) data to foster discoveries in health and disease[23]. For this study, we used WES from 51,289 participants[13], which includes patients enrolled through primary care and specialty outpatient clinics. Participants were 59% female, with a median age of 61 years, and predominantly Caucasian (98%) and non-Hispanic/Latino (95%). Patient demographics are in Supplemental Table 1. Whole exome sequencing was performed using 75 bp paired-end sequencing on an Illumina v4 HiSeq 2500 to a coverage depth sufficient to provide greater than 20x haploid read depth of over 85% of targeted bases for at least 96% of samples. Sample preparation, sequencing, sequence alignment, variant identification, genotype assignment and quality control steps were carried out as previously described[23].

Analysis Methods

Site specific rare codon change. Codon frequency tables provide the relative fraction of different codons used in the genes of an organism, and are available in Codon Usage tables[50,51] and https://www.genscript.com/tools/codon-frequency-table. For all rare synonymous codons, we determined the ratio of the genome-wide codon fraction of the *GPR39* reference codon over the synonymous codon at that position, e.g., for Ser147Ser (tcG/tcA), 0.06/0.15 = 0.4, i.e., a 2.5-fold change. All synonymous variant with at least two-fold change in site specific fraction of codon usage change were used in the SKAT analysis.

Phenome-wide association analysis (PheWAS). De-identified EHR data was obtained from an approved data broker. All unique ICD9 codes with ≥200 patients (regardless of genotype) with at least 3 independent incidences of the ICD9 code were extracted from the EHR. Individuals having 1-2 calls were excluded, and those having no calls of a particular ICD9 code were considered as controls. All non-Europeans were excluded from the analysis, as was one sample from pairs of closely related individuals up to first cousins. All models were adjusted for sex, age, age^2 and first 4 principal components. Plink2 [https://www.cog-genomics.org/plink2] and/or PheWAS R package were used for association analyses, which were performed for three GPR39 common variants (2:133174764, Ala50Val; 2:133174999; Thr128Thr; and 2:133402607, noncoding). All analyses/plotting were carried out using R (3.2) and/or GraphPad Prism (V.6).

Sequence Kernel Association Testing (SKAT). To test whether *GPR39* rare variants (MAF < 0.01) that might be associated with clinical phenotypes, we performed sequence kernel association test (using default weights; SKAT R package) a gene-based binned statistical test comparing the burden of rare variants in cases and controls. Cases were defined as individuals with at least 3 independent ICD-9 calls derived from the EHR. We excluded all individuals with 1 or 2 ICD-9 calls and all ICD-9 codes with less than 200 cases. Individuals with no ICD-9 calls were treated as controls. We created SNP set ID files that defined various variant sets, including (1) only potential loss of function (splicing, stop gain, stop loss, or frameshift) plus potential pathogenic missense (pLP and pP) variants, and (2) synonymous variants having large codon usage frequency changes (at least a two-fold change in frequency between human reference codon and altered codon, SYN_ΔCB). Age, age^2, sex, BMI, and first four principal components were used as co-variates. All non-European samples were excluded from the analyses. False discovery rate estimation was measured by adjusted P-values (Q-values ≤ 0.01) for a given ICD-9 association set of P-values[52].

Generation of *GPR39* variants. All missense and SYN_ΔCB variants were generated by gene synthesis (Thermo Fisher Scientific, Gene Art) based on the NP_001499.1 reference sequence (GeneID 2863, 453 amino acids), and included tandem tags at the amino terminus to facilitate expression and functional studies,

i.e., after the initiation methionine codon, the 3X FLAG epitope, followed by a minimal bungarotoxin binding site were inserted. The complete inserted sequence is: GACTACAAAGACCATGACGGTGATTATAAAGATCATGATATCGATTACAAGGATGACGATG ACAAGGGTTGGAGATACTACGAGAGCTCCCTGGAGCCCTACCCTGACGGT.

Expression of GPR39 variants. Wild type (WT) and variant constructs were transiently transfected into HEK293 cells, lysed after 48 hours, and then 25 μg of lysates were separated on 4-15% SDS PAGE gels (BioRad), blotted to nitrocellulose, blocked with 5% milk in TBS-T, and then exposed for 1 hour at room temperature to anti-FLAG monoclonal antibody conjugated to HRP (Sigma). Blots were exposed to SuperSignal West Pico Chemiluminescence Substrate, (Thermo Fisher), and images recorded on a FUJIFILM LAS-4000mini luminescence analyzer and processed with Image-Gauge version 3.0. GPR39 expression was quantified using Multi-Gauge software, normalizing to WT expression and corrected for HEK293 background on the same blot.

Functional analysis of GPR39 variants. WT and variant constructs were transiently transfected into HEK293 cells, cultured for 24 hours, then split and equivalent numbers of cells plated into poly-D-lysine-coated 96 well plates for a further 24 hours for IP_1 analysis. Replicates were treated with various concentrations of $ZnCl_2$ for 10 min at 37 °C. Levels of IP_1 were determined as recommended by manufacturer (IP-One ELISA kit, Cisbio USA, Inc.). Plates were read on a POLARStar Omega (BMG Labtech) plate reader. EC_{50}s were determined by fitting normalized dose-response data with the Michaelis-Menten equation, and V_{MAX} were separately calculated on raw data (GraphPad Prism, V6). Parallel plates were exposed to Alexa-594-conjugated bungarotoxin to assess surface expression of GPR39 variants (3 μg/ml, 5 min at 37 °C, 5% CO_2). After removal of bungarotoxin and washes with PBS, fluorescence was read on the POLARStar Omega (BMG Labtech) reader. Each variant was assayed in ≥3 independent transfections; WT, standards and solution blanks were included in each assay.

Acknowledgements

We are grateful to Geisinger Health System and the Regeneron Genetics Center for access to the DiscovEHR cohort.

Author contributions

RD, KJ performed experiments, RD contributed to data analysis and edited the manuscript, SK analyzed genomic and clinical data, DJC, MH and JDR made helpful suggestions and edited the manuscript, RPRM developed and implemented genomic analysis and edited the manuscript, RPRM and GEB conceived the project, GEB analyzed data, generated figures, and wrote the manuscript.

References

1. Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schioth, H. B., & Gloriam, D. E. Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829-842 (2017).
2. Roth, B. L., Kroeze, W. K. Integrated approaches for genome-wide interrogation of the druggable non-olfactory G protein-coupled receptor family. *J. Biol. Chem.* **290**, 19471-19577 (2015).
3. Kovacs, P., Schonberg, T. The relevance of genomic signatures at adhesion GPCR loci in humans. *Handb. Exp. Pharmacol.* **234**, 179-217 (2016).
4. Wu, M. C., 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.* **89**, 82-93 (2011).
5. Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani, D. C., Wurfel, M. M., Lin, X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 225-237 (2012).
6. Richardson, T. G., Shihab, H. A., Rivas, M. A., McCarthy, M. I., Campbell, C., Timpson, N. J., Gaut, T. R. A protein domain and family based approach to rare variant association analysis. *PLoS One* **11**, e0153803 (2016).
7. Friedrichs, S., Malzahn, D., Pugh, E. W., Almeida, M., Liu, X. Q., Bailey, J. N. Filtering genetic variants and placing informative priors based on putative biological function. *BCM Genet.* **71 Suppl. 2**, 8 (2016).
8. Asraf, H., Salomon, S., Nevo, A., Sekler, I., Mayer, D., Hershfinkel, M. The ZnR/GPR39 interacts with the CaSR to enhance signaling in prostate and salivary epithelia. *J. Cell Physiol.* **229**, 868-877 (2014).
9. Cohen, L., Asraf, H., Sekler, I., Hershfinkel, M. Extracellular pH regulates zinc signaling via an Asp reside of the zinc-sensing receptor (ZnR/GPR39). *J. Biol. Chem.* **287**, 33339-33350 (2012).
10. Sunuwar, L., Gilad, D., Hershfinkel, M. The zinc sensing receptor, ZnR/GPR39, in health and disease. *Front. Biosci (Landmark Ed)* **22**, 1469-1492 (2017).
11. Hershfinkel, M. The Zinc Sensing Receptor, ZnR/GPR39, in health and disease. *Int. J. Mol. Sci.* **19**, 437-458 (2018).

12. Sriram, K., Insel, P. A. GPCRs as targets for approved drugs: How many targets and how many drugs? *Mol. Pharmacol.* Jan.3 DOI: 10.1124/mol.117.111062 (2018).

13. Dewey, F. E., Murray, M. F., Overton, J. D., Habegger, L., Leader, J. B., Fetterolf, S. M., O-Dushlaine, C., Van Hout, C. V., Staples, J., Fonzaga-Jauregui, C., Metpally, R.,Pendergrass, S. A., Giovanni, M. A., Kirchner, H. L., Balasubramanian, S., Abul-Husn, N. S., Hartzel, D. N., Lavage, D. R., Kost, K. A., Packer, J. S., Lopez, A. E., Penn, J., Mukherjee, S., Gosalia, N., Kanagaraj, M., Li, A. H., Mitnaul, L. J., Adams, L. J., Person, T. N., Praveen, K., Marcketta, A., Lebo, M. S., Austin-Tse, C. A., Mason-Suares, H. M., Bruse, S., Mellis, S., Phillips, R., Stahl, N., Murphy, A., Economides, A., Skelding, K. A., Still, C. D., Elmore, J. R., Borecki, I. B., Yancopoulos, G. D., Davis, F. D., Faucett, W. A., Gottesman, O., Ritchie, M. D., Shuldiner, A. R., Reid, J. G., Ledbetter, D. H., Baras, A., Carey, D. J. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, pii aaf6814 (2016).

14. Basile, A. O., Byrska-Bishop, M., Wallace, J., Frase, A. T., Ritchie, M. D. Novel features and enhancements in BioBin, a tool for biologically inspired binning and association analysis of rare variants. *Bioinformatics* **34**, 527-529 (2018).

15. Giddens, M. M., Wong, J. C., Schroeder, J. P., Farrow, E. G., Smith, B. M., Owino, S., Soden, S. E., Meyer, R. C., Saunders, C., LePichon, J. B., Weinshenker, D., Escayg, A., Hall, R. A. *GPR37L1* modulates seizure susceptibility: Evidence from mouse studies and analyses of a human *GPR37L1* variant. *Neurobiology of Disease* **106**, 181-190 (2017).

16. Hoeck, J. D., Biehs, B., Kurtova, A. V., Kljavin, N. M., de Sousa, E. Melo F., Alicke, B., Koeppen, H., Modrusan, Z., Piskol, R., de Sauvage, F. J. Stem cell plasticity enables hair regeneration following LGR5+ cell loss. *Nat. Cell Biol.* **19**, 666-675 (2017).

17. Michel, L., Reygagne, P., Benech, P., Jean-Louis, F., Scalvino, S., Ly Ka So, S., Hamidou, Z., Bianovici, S., Pouch, J., Ducos, B., Bonnet, M., Bensussan, A., Patatian, A., Lati, E., Wdzieczak-Bakala, J., Choulot, J.-C., Loing, E., Hocquaux, M. Study of gene expression alteration in male androgenic

alopecia: evidence of predominant molecular signaling pathways. *Brit. J. Dermatol.* **177**, 1322-1336 (2017).

18. Guillemot-Legris, O., Mutemberezi, V., Cani, P. D., Muccioli, G. G. Obesity is associated with changes in oxysterol metabolism and levels in mice liver, hypothalamus, adipose tissue and plasma. *Sci. Reports 6*, 19694 (2016).

19. Li, Y., Dong, J., Xuan, Y.-H., Liu, S.-S., Luo, J.-Y., Xiao, Y.-J., Tian, Z.-P., Sun, Z.-J. Identification of microRNA expression in a rat model of post-infarction heart failure. *Int. J. Clin. Exp. Pathol.* **10**, 4729-4738 (2017).

20. Storjohann, L., Holst, B., Schwartz, T. W. A second disulfide bridge from the N-terminal domain to extracellular loop 2 dampens receptor activity in GPR39. *Biochem.* **47**, 9198-9207 (2008).

21. Storjohann, L., Holst, B., Schwartz, T. W. Molecular mechanism of Zn^{2+} agonism in the extracellular domain of GPR39. *FEBS Lett.* **582**, 2583-2588 (2008).

22. Yuan, S., Filipek, S., Palxzewski, K., Vogel, H. Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway. *Nat. Commun.* **5**, 4733 (2014).

23. Carey, D. J., Fetterolf, S. N., Davis, F. D., Faucett, W. A., Kirchner, H. L., Mirshahi, U., Murray, M. F., Smelser, D. T., Gerhard, G. S., Ledbetter, D. H. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet. Med.* **18**, 906-913 (2016).

24. Verhulst, P. J., Lintermans, A., Janssen, S., Loeckx, D., Hummelreich, U., Buyse, J., Tack, J., Depoortere, I. GPR39, a receptor of the ghrelin receptor family, plays a role in the regulation of glucose homeostasis in a mouse model of early onset diet-induced obesity. *J. Neuroendocrinol.* **23**, 490-500 (2011).

25. Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E., Kimchi-Sarfaty, C. Exposing synonymous mutations. *Trends Genet.* **30**, 308-321 (2014).

26. Bali, V., Bebok, Z. Decoding mechanisms by which silent codon changes influence protein biogenesis and function. *Int. J. Biochem. Cell. Bol.* **64**, 58-74 (2015).

27. Brar, G. A. Beyond the triplet code: context cues transform translation. *Cell* **167**, 1681-1692 (2016).

28. Supek, F., Minana, B., Valcarcel, J., Gabaldon, T., Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324-1335 (2014).

29. Fernandez-Calero, T., Cabrera-Cabrera, F., Ehrlich, R., Marin, M. Silent polymorphisms: can the tRNA population explain changes in protein properties? Life, 6:9; doi:10.3390/life6010009 (2016).

30. Jung, H., Lee, D., Park, D., Kim, Y. J., Park, W. Y., Hong, D., Park, P. J., Lee, E. Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47**, 1242-1248 (2015).

31. Soussi, T., Taschner, P. E., Samuels, Y. Synonymous somatic variants in human cancer are not infamous: a plea for full disclosure in databases and publications. *Hum. Mutat.* **38**, 339-342 (2017).

32. Kukreti, R., Tripathi, S., Bhatnagar, P., Gupta, S., Chauhan, C., Kubendran, S., Janardhan Reddy, Y. C., Jain, S., Brahmachari, S. K. Association of DRD2 gene variant with schizophrenia. *Neurosci. Lett.* **392**, 68-71 (2006).

33. Grymek, K., Lukasiewica, S., Faron-Goreckaa, A., Tworzydlo, M., Polit, A., Dziedzicka-Wasylewska, M. Role of silent polymorphrisms within the dopamine D1 receptor associated with schizophrenia on D1-D2 receptor heterodimerization. *Pharmacol. Rep.* **61**, 1024-1033 (2009).

34. Duan, J., Wainwright, M.S., Comeron, J.M., Saitou, N., Sanders, A.R., Gelernter, J., Gejman, P. V. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **12**, 205-216 (2003).

35. Kimchi-Sarfaty, C., Oh, J.M., Kim, I.-W., Sauna, Z E., Calcagno, A. M., Ambudkar, S. V., Gottesman, M. M. A “silent” polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**, 525-528 (2007).

36. Shah, K., Cheng, Y., Hahn, B., Bridges, R., Bradbury, N.A., Mueller, D.M. Synonymous codon usage affects the expression of wild type and F508del CFTR. *J. Mol. Biol.* **427**, 1464-1479 (2015).

37. Zhang, T., Wu, Y., Lan, Z., Shi, Q., Yang, Y., Guo, J. Syntool: a novel region-based intolerance score to single nucleotide substitution for synonymous mutations predictions based on 12,136 individuals. *BioMed Res. Int.* **2017**, 5096208 (2017).

38. Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E., Kimchi-Sarfaty, C. Exposing synonymous mutations. *Trends Genet.* **30**, 308-321 (2014).

39. Bali, V., Bebok, Z. Decoding mechanisms by which silent codon changes influence protein biogenesis and function. *Int. J. Biochem. Cell. Bol.* **64**, 58-74 (2015).

40. Kovacs, Z., Schacht, T., Hermann, A. K., Albrecht, P., Lefkimmiatis, K., Methner, A. Pr otein kinase inhibitor β enhances the constitutive activity of G-protein-coupled zinc receptor GPR39. *Biochem. J.* **462**, 125-132 (2014).

41. Shimizu, Y., Koyama, R., Kawamoto, T. Rho kinase-dependent desensitization of GPR39; a unique mechanism of GPCR downregulation. *Biochem. Pharmacol.* **140**, 105-114 (2017).

42. Chorin, E., Vinograd, O., Fleidervish, I., Gilad, D., Hermann, S., Sekler, I., Aizenman, E., Hershfinkel, M. Upregulation of KCC2 activity by zinc-mediated neurotransmission via the mZnR/GPR39 receptor. *J. Neurosci*. **31**, 12916-12926 (2011).

43. Saadi, R. A., He, K., Hartnett, K. A., Kandler, K, Hershfinkel, M., Aizenman, E. SNARE-dependent upregulation of potassium chloride co-transporter 2 activity after metabotropic zinc receptor activation in rat cortical neurons in vitro. *Neurosci.* **210**, 38-46 (2012).

44. Gilad, D., Shorer, S., Ketzef, M., Friedman, A., Sekler, I., Aizenman, E., Hershfinkel, M. Homeostatic regulation of KCC2 activity by the zinc receptor mZnR/GPR39 during seizures. *Neurobiol. Dis.* **81**, 4-13 (2015).

45. Mlyniec, K., Budziszewska, B., Holst, B., Ostachowicz, B., Nowak, G. GPR39 (zinc receptor) knockout mice exhibit depression-like behavior and CREB/BDNF down-regulation in the hippocampus. *Int. J. Neuropsychopharmacol.* **18**, pii pyu002 (2014).

46. Mlyniec, K., Nowak, G. Up-regulation of the GPR39 Zn^{2+}-sensing receptor and CREB/BDNF/TrkB pathway after chronic but not acute antidepressant treatment in the frontal cortex of zinc-deficient mice. *Pharmacol. Rep.* **67**, 1135-1140 (2015).

47. Mlyniec, K., Doboszewska, U., Szewczyk, B., Sowa-Kucma, M., Misztak, P., Piekoszewski, W., Trela, F., Ostachowicz, B., Nowak, G. The involvement of the GPR39-Zn^{2+}-sensing receptor in the

pathophysiology of depression. Studies in rodent models and suicide victims. *Neuropharmacol.* **79,** 290-297 (2014).

48. Costello, L. C., Franklin, R. B. A comprehensive review of the role of zinc in normal prostate function and metabolism; and its implications in prostate cancer. *Arch. Biochem. Biophys.* **611**, 100-112 (2016).

49. Zhao, H., Qiao, J., Zhang, S., Zhang, H., Lei, H., Lei, X., Deng, Z., Ning, L., Cao, Y., Guo, Y., Liu, S., Duan, E. GPR39 marks specific cells within the sebaceous gland and contributes to skin wound healing. *Sci. Rep.* **4**, 7913 (2015).

50. Nakamura, Y., Gojobori, T., Ikemura, T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*. **28**, 292 (2000).

51. Athey, J., Alexaki, A., Osipova, E., Rostovtsev, A., Santana-Quintero, L. V., Katneni, U., Simonyan, V., Kimchi-Sarfaty, C. A new and updated resource for codon usage tables. *BMC Bioinformatics* **18**, 391 (2017).

52. Storey, J. D., Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci.* **100**, 9440-9445 (2003).

Table 1. Orphan GPCRs with overlapping disease associations for multiple classes of rare variants. All orphan GPCRs with sufficient numbers of rare LOF, MISS or SYN_ΔCB variants in at least two categories were analyzed by SKAT. Absence of results in a category (marked by X) indicates insufficient number of variants and/or individuals having the variant for well-powered SKAT. Q-values provide an estimate of the false discovery rate, with significance set at Q-value < 0.001, i.e., 0.1% of tests will result in false positive associations[52].

			LOF	MISS	SYN_ΔCB
GPCR	ICD9	Description	Q-value	Q-value	Q-value
LGR4	710.2	Sicca syndrome	5.03E-29	1.17E-11	6.45E-12
GPR150	372.14	Other chronic allergic conjunctivitis	1.86E-28	9.94E-30	0.00037
GPR153	276.7	Hyperpotassemia	1.65E-27	7.76E-14	7.51E-20
GPR149	374.3	Ptosis of eyelid, unspecified	4.23E-27	1.71E-06	4.98E-06
GPR161	780.8	Generalized hyperhidrosis	2.37E-23	X	5.20E-24
GPR37L1	345.1	Generalized convulsive epilepsy, w/o intract.epilepsy	1.46E-15	6.29E-17	1.60E-17
GPR84	280.8	Other specified iron deficiency anemias	1.66E-12	6.07E-06	1.47E-12
GPR39	**354.2**	**Lesion of ulnar nerve**	**5.01E-12**	**9.63E-07**	**2.55E-12**
LGR6	583.81	Nephritis and nephropathy, acute or chronic not specified	2.13E-11	0.000221	8.35E-17
LGR5	704	Alopecia NOS	4.11E-10	X	1.73E-10
GPR1	287.5	Thrombocytopenia NOS	9.98E-07	2.50E-05	1.04E-06
GPR18	583.81	Nephritis and nephropathy, acute or chronic not specified	5.53E-06	6.93E-06	1.28E-06
HCAR1	214.1	Lipoma of other skin and subcutaneous tissue	6.16E-05	X	6.16E-05
GPR156	739.1	Nonallopathic lesions of cervical region NOS	6.55E-05	X	7.86E-05
GPR179	270.4	Disturbances of sulphur-bearing amino-acid metabolism	9.78E-05	X	7.44E-05
GPR176	493.91	Asthma, unspecified type, with status asthmaticus	X	8.05E-31	1.44E-23
GPR132	270.4	Disturbances of sulphur-bearing amino-acid metabolism	X	2.97E-27	3.25E-21
GPR85	410.1	Acute myocardial infarction, of other anterior wall	X	4.65E-24	4.09E-33
GPR183	573.8	Other specified disorders of liver	X	9.60E-17	1.08E-16
GPR12	345.1	Generalized convulsive epilepsy, w/o intract.epilepsy	X	3.65E-11	3.55E-11
GPR135	350.1	Trigeminal neuralgia	X	1.28E-09	1.32E-09
GPR26	372.14	Other chronic allergic conjunctivitis	X	2.95E-09	2.82E-09
MRGPRE	401.1	Benign essential hypertension	X	7.23E-08	3.22E-05
GPR15	455.6	Hemorrhoids NOS	X	1.10E-07	0.000528
GPR83	250.8	Diabetes mellitus type II (non-insulin dependent)	X	2.02E-07	3.56E-09
GPRC5D	588	Renal osteodystrophy	X	9.36E-07	7.18E-18
GPRC5B	381.81	Dysfunction of eustachian tube	X	8.98E-05	7.86E-05
GPR141	786.07	Wheezing	X	2.26E-06	1.98E-07
GPR6	214.1	Lipoma of other skin and subcutaneous tissue	X	0.000123	1.31E-06
GPR31	41.02	Group B Streptococcus infection	X	0.000136	0.000136
MRGPRF	305	Alcohol abuse, unspecified drinking behavior	X	0.000188	0.000302
GPR4	585.2	Chronic kidney disease, Stage II (mild)	X	0.00043	0.000403
GPR142	704	Alopecia NOS	X	0.000757	0.000526

Table 2. SKAT results for rare GPR39 variants in 51,289 DiscovEHR cohort. All LOF plus pathogenic missense (pLP + pP) were binned and analyzed by SKAT (39 markers; 177 individuals). The SYN_ΔCB were separately binned and analyzed (21 markers; 198 individuals). Shown are all significant associations that were found for both the LOF/MISS and SYN_ΔCB rare variant groups. Associations are listed in order of Q-values of SYN_ΔCB results.

		LOF + pLP + pP	SYN_CUB
CODE	DESCRIPTION	Qvalue	Qvalue
354.2	Lesion of ulnar nerve	9.55E-09	2.55E-12
600.21	BPH, localized, with lower urinary tract symptoms (LUTS)	3.50E-07	2.43E-07
362.56	Macular puckering of retina	0.0002029	9.79E-05
704.8	Other specified diseases of hair and hair follicles	0.004823329	0.00010162
696	Psoriatic arthropathy	0.000193128	0.00013058
401.1	Benign essential hypertension	0.000833645	0.00071272
591	Hydronephrosis	0.001252239	0.00109841
276.9	Electrolyte and fluid disorders not elsewhere classified	0.005113828	0.00415868

Table 3. SKAT analysis of all rare synonymous GPR39 variants in 51,289 DiscovEHR cohort. All rare synonymous variants were binned and analyzed by SKAT (47 markers; 940 individuals). Shown are all significant associations listed in order of Q-values with cutoff at Q<0.001.

CODE	DESCRIPTION	Qvalue
696	Psoriatic arthropathy	1.79E-10
354.2	Lesion of ulnar nerve	1.14E-05
600.21	BPH, localized, with lower urinary tract symptoms (LUTS)	1.71E-05
110.4	Dermatophytosis of foot	0.00018
995.3	Allergy unspecified not elsewhere classified	0.00049
372.3	Conjunctivitis NOS	0.00087

Figure Legends

Figure 1. Pathogenicity triage of 159 coding variants in *GPR39* identified in 51,289 WES. Common variants with MAF≥1% were categorized as likely benign (MAF between 1-5%) and benign (MAF≥5%). All variants having MAF<1% sorted to synonymous, loss of function (LOF; frameshift or premature stop codon), and missense. Missense variants were thus Variants of Unknown Significance (VUS) and subjected to bioinformatics pathogenicity triage using RMPath pipeline, and classified as predicted benign, pB, likely benign, pLB, likely pathogenic, pLP or pathogenic pP.

Figure 2. Snake plot of GPR39 topology indicating rare variants identified in 51,289 WES. The GPR39 sequence was color-coded to indicate reference codon usage frequency. For all amino acids having multiple potential codons, red indicates the most common codon, with progression from common to rare colored as gradations to light pink (rare). Arrowheads indicate the locations of rare synonymous variants exhibiting the largest changes in codon bias, either from common to rare or rare to common. Asterisks indicate rare LOF variants (truncations or frameshifts) and arrows indicate locations of rare missense variants combined for SKAT analysis.

Figure 3. Manhattan plots of SKAT results for *GPR39* rare variants. A. Rare LOF and MISS variants were combined (39 total variants) and analysis run using SKAT. Top phenome-wide associations are labelled. B. Rare SYN_ΔCB variants (21 total variants) were analyzed with SKAT; top phenome-wide associations are labelled. Full details of associations common to both variant categories are presented in Table 2.

Figure 4. Expression of rare missense and synonymous GPR39 variants. A. Quantitation of expression levels of *GPR39* rare variants relative to WT. All blots had WT and HEK293 only lanes for background subtraction and normalization. Data presented as % of WT expression ± S.D. (n ≥ 3 independent transfections). Representative blots in **Figure S2**. **B.** Quantitation of expression levels of GPR39 rare

SYN_ΔCB variants relative to WT. Details same as in **A**. Statistical analysis for both figures was two-tailed t-test relative to WT, with *p<0.05, ** p<0.01, ***p<0.001 and ****p<0.0001.

Figure 5. Functional analysis of rare missense and synonymous GPR39 variants. A. IP1 responses to various concentration of $ZnCl_2$ for all variants that expressed ≥ 50% of WT. Representative curves and fits for WT, W314R and R386C are illustrated. Plotted are mean ± S.D. of three independent experiments, lines show fits to Michaelis-Menten equation. **B.** EC_{50}s for all dose-response relations (as in **A**), plotted as mean ± S.E. for n=3 independent experiments. **C.** IP1 responses of representative SYN_ΔCB variants showing altered EC_{50}s compared to WT, I29 and V354. Details as in **A**. **D.** EC_{50}s for all dose-response relations (as in **A**) for SYN_ΔCB variants. **E.-F.** Bungarotoxin binding (see Methods) was used to assess surface localization of rare missense (**E**) and SYN_ΔCB (**F**) variants. **G.-H.** Extrapolated V_{MAX}s of Michaelis-Menten fits to IP1 dose-response relations (see Methods). For all plots, *p<0.05, **p<0.01, ***p<0.001, and ****p<0.0001, by two-tailed t-test relative to WT, assuming unequal variances.

Figure 1

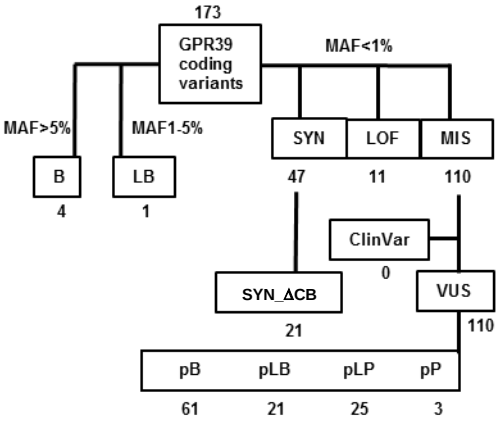

173
GPR39 coding variants
MAF<1%
MAF>5%
MAF1-5%
SYN
LOF
MIS
B
LB
47
11
110
4
1
ClinVar
0
SYN_ΔCB
VUS
21
110
pB
pLB
pLP
pP
61
21
25
3

Figure 2

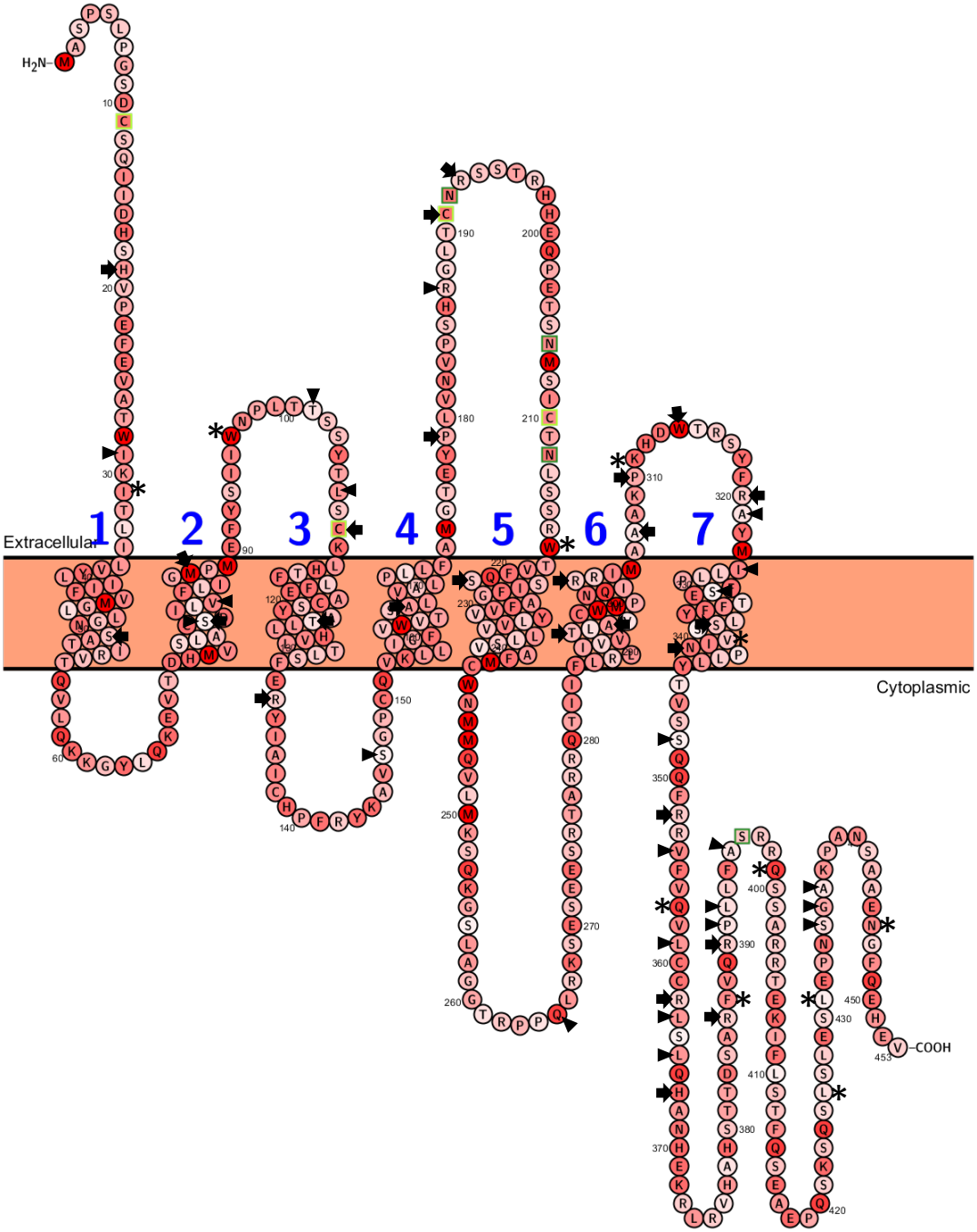


Figure 3

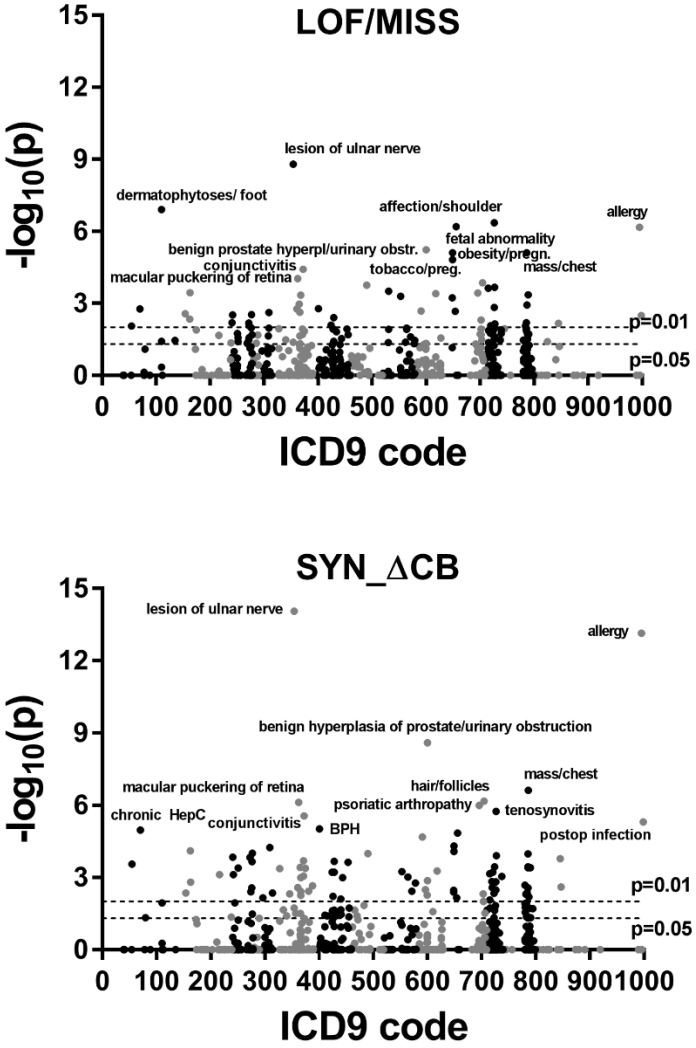

LOF/MISS
-log10(p)
ICD9 code
lesion of ulnar nerve
dermatophytoses/ foot
affection/shoulder
allergy
fetal abnormality
benign prostate hyperpl/urinary obstr.
obesity/pregn.
conjunctivitis
tobacco/preg.
mass/chest
macular puckering of retina
p=0.01
p=0.05
SYN_ΔCB
-log10(p)
ICD9 code
lesion of ulnar nerve
allergy
benign hyperplasia of prostate/urinary obstruction
mass/chest
macular puckering of retina
hair/follicles
psoriatic arthropathy
tenosynovitis
chronic
HepC
conjunctivitis
BPH
postop infection
p=0.01
p=0.05

Figure 4

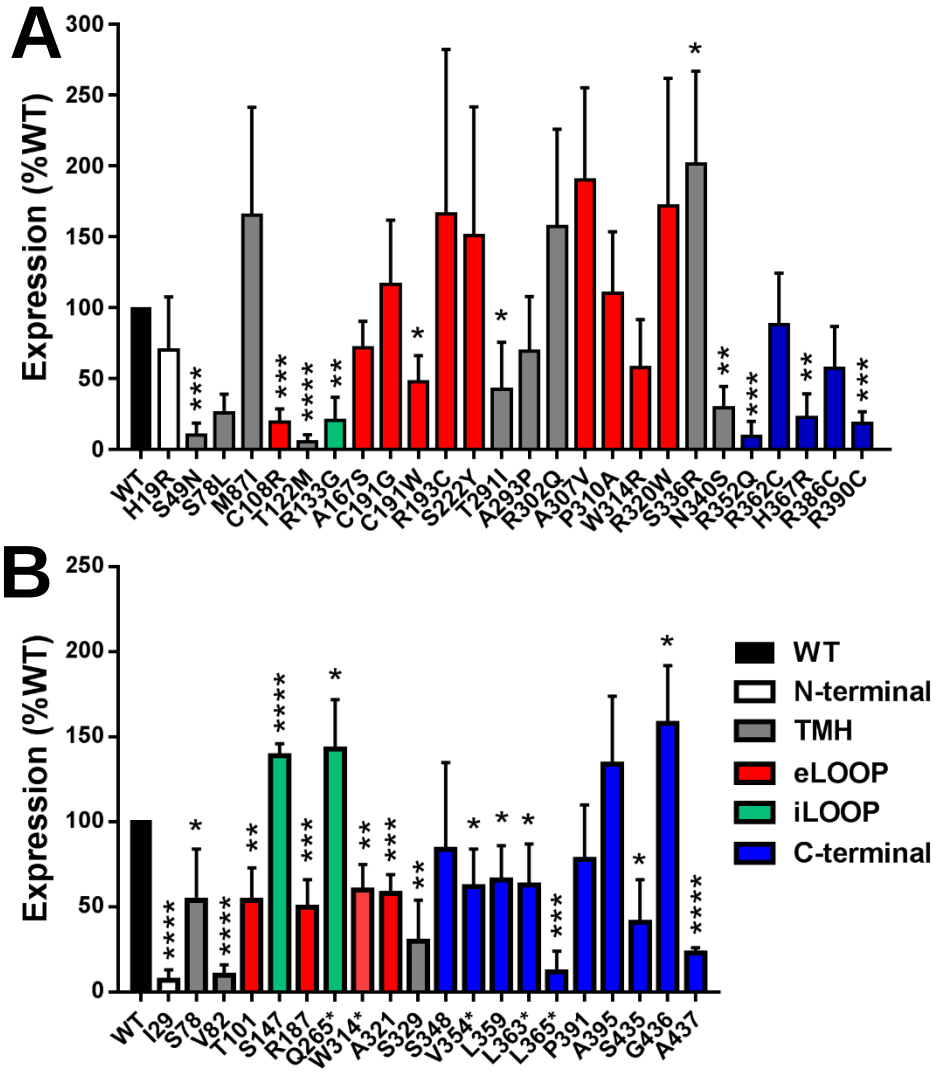
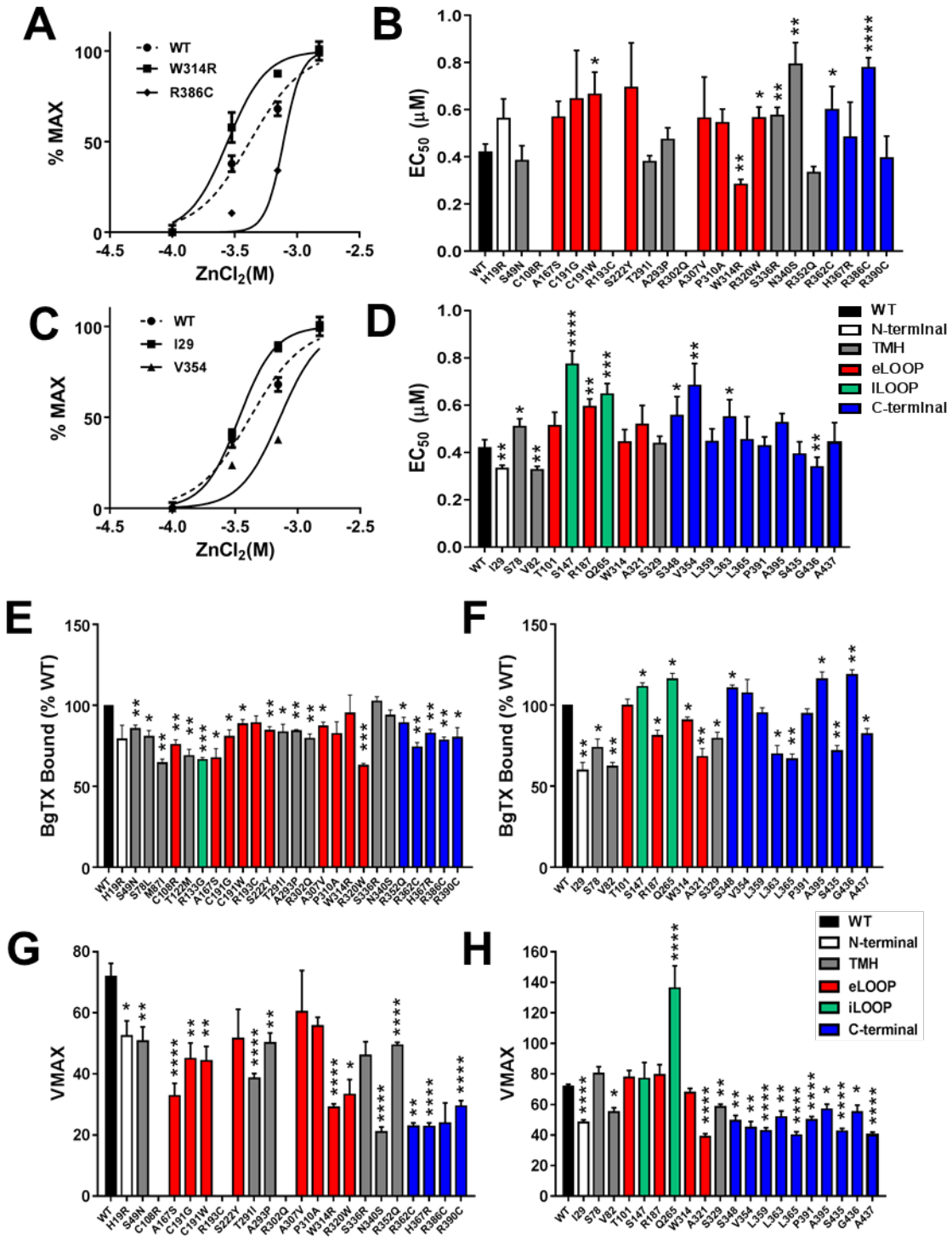

A
Expression (%WT)
300
250
200
150
100
50
0
WT
H19R
S49N
S78L
M87I
C108R
T122M
R133G
A167S
C191G
C191W
R193C
S222Y
T291I
A293P
R302Q
A307V
P310A
W314R
R320W
S336R
N340S
R352Q
R362C
H367R
R386C
R390C
B
Expression (%WT)
250
200
150
100
50
0
WT
I29
S78
V82
T101
S147
R187
Q265*
W314*
A321
S329
S348
V354*
L359
L363*
L365*
P391
A395
S435
G436
A437
WT
N-terminal
TMH
eLOOP
iLOOP
C-terminal

Figure 5


A
WT
W314R
R386C
% MAX
ZnCl2(M)
B
EC50 (μM)
C
WT
I29
V354
% MAX
ZnCl2(M)
D
EC50 (μM)
WT
N-terminal
TMH
eLOOP
ILOOP
C-terminal
E
BgTX Bound (% WT)
F
BgTX Bound (% WT)
G
VMAX
H
VMAX
WT
N-terminal
TMH
eLOOP
iLOOP
C-terminal

Supplemental Methods.

In-silico Radical Mutations Pathogenicity (RMPath) prediction pipeline.

Rare variants are categorized into two groups: (1) nonsense and frame-shift variants, and (2) missense variants, and are scored by giving numerical values to outcome predictions, as described below:

(1) Scoring of nonsense and frame-shift variants.

SnpEff fields LOF[*].PERC ≥ 0.5 or NMD[*].NUMTR > 3 are considered potential Loss of function variants.

(2) Scoring of missense variants.

1. SIFT prediction: D (damaging) → score 2; All others → score 0.

2. PolyPhen2 prediction: i) HDIV: Probably damaging (D) → score 2; Possibly damaging (P) → score 1 ; All others → score 0. ii) HVAR: Probably damaging (D) → score 2; Possibly damaging (P) → score 1; All others → score 0.

3. MutationTaster prediction: Disease causing (D) → score 2; All others → score 0.

4. MutationAssessor prediction: High (H) → score 2; Medium (M) → score 1; All others → score 0.

5. FATHMM prediction: D (damaging) → score 2; All others → score 0.

6. LRT prediction: D(deleterious) → score 2; All others → score 0.

7. PROVEAN prediction: D(damaging) → score 2; All others→ score 0.

8. GRANTHAM Score: ≥140 → score 4; ≥120 → score 3; ≥90 → score 2; ≥70 → score 1; All others → score 0.

9. Genomic Evolutionary Rate Profiling GERP++ rejected substitutions (RS) score: ≥6 → score 4; ≥5 → score 3; ≥3.5 → score 2; ≥2 → score 1; All others → score 0.

Using cumulative score (represented a RMPath score) generated by the listed tools (maximum possible score, 24), the rare variants were sorted into 4 classes: predictive Benign (up to 25% of max score), predictive Likely Benign (26 to 44%), predictive Likely Pathogenic (45 to 69%), predictive Pathogenic

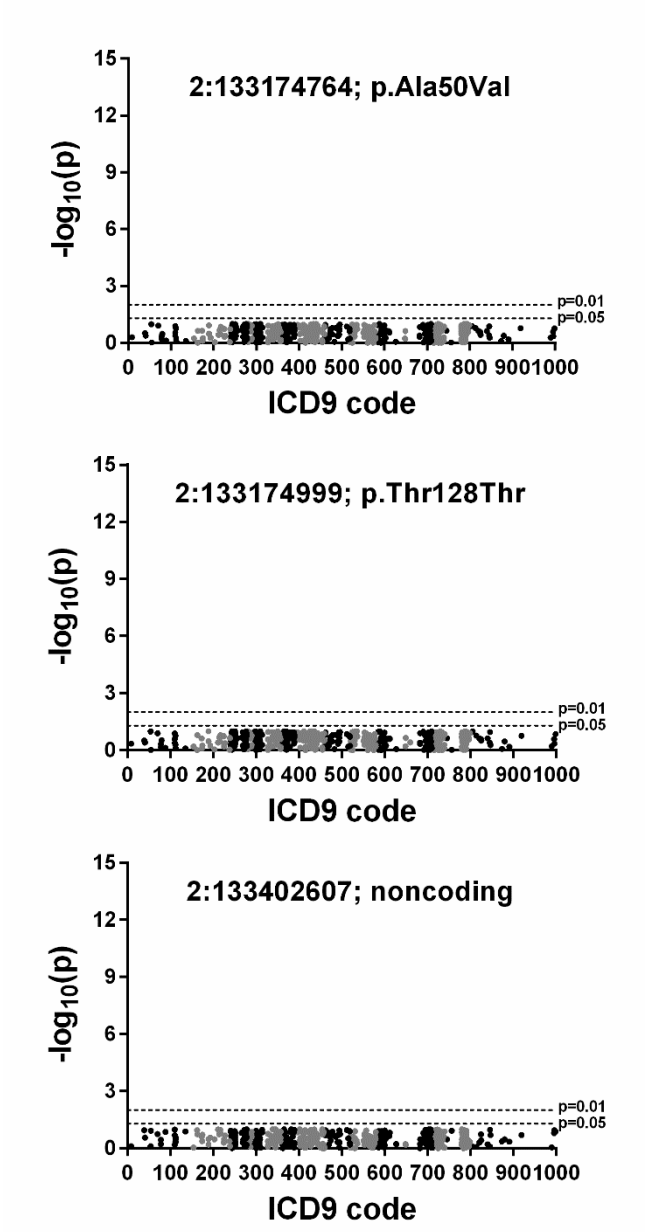
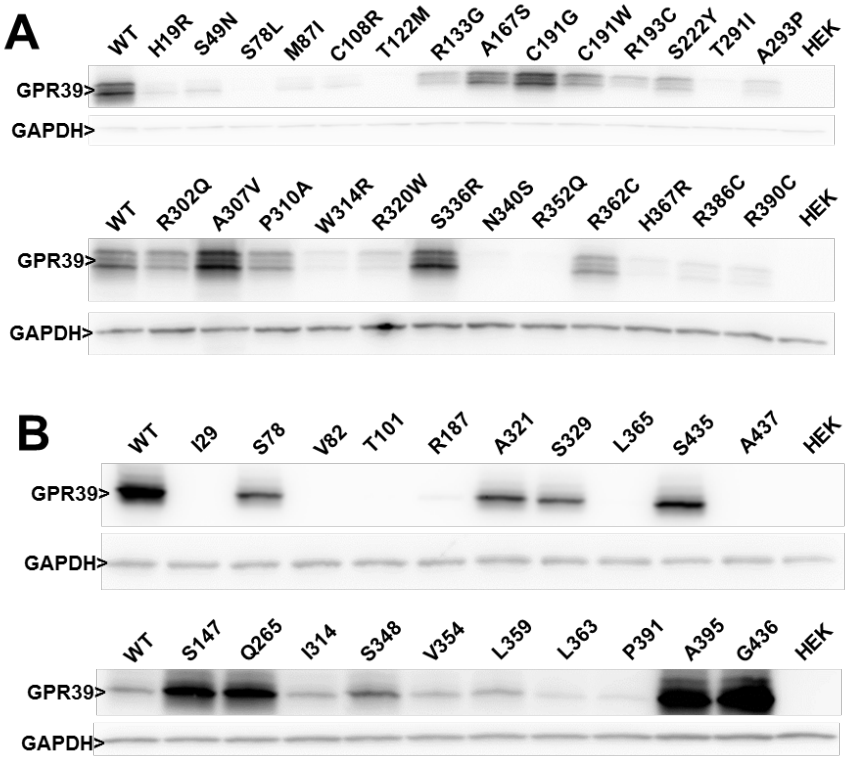
(≥70% of max score). For clinical evaluation, all predicted variants must be treated as Variant of Unknown Significance (VUS) unless family co-segregation and/or functional analyses are required to definitively establish a variant as pathogenic.

Supplemental Table 1. Demographics of the DiscovEHR cohort used in the present study.

Abbreviations, EHR, electronic health record, BMI, body mass index.

Basic demographics	**DiscoveEHR Sequenced Patients**
Total patients (N)	51,289
Female, N (%)	30,290 (59%)
Median age, yrs	61 (range 48-72)
Median BMI, kg/m^2	30 (range 26-36)
Median years of EHR data	14 (range 11-17)
Median medication orders/patient	129 (range 37-221)

Supplemental Figure S1. Manhattan plots of results of PheWAS analyses of common GPR39 variants. No association results exceeded the minimal P-value threshold of 0.05, uncorrected for multiple testing.



Supplemental Figure S2. Representative western blots of rare missense and synonymous variants identified in the 51,289 WES cohort. Western blots of HEK293 cells lysates expressing WT or rare variant missense or SYN_ΔCB cDNAs were run as described in Methods. **A.** Missense variants run in order from amino to carboxyl terminal location; each gel also had WT and untransfected HEK samples for normalization; GAPDH was used as loading control. **B.** Expression of rare synonymous variants was more variable, and rare variants were sorted for high expression (top row) or low expression (bottom row) for better quantitation; WT and HEK were run on each gel for normalization; GAPDH was used as loading control.