
Supplementary information

Analysis of 173,303 exomes and genomes in the Pakistan Genome Resource

In the format provided by the
authors and unedited

Supplementary Contents

Supplemental Text	1
In vitro profiling of RXFP1 variants from PGR	1
End-truncating pLoF variants	2
Supplementary Figures	4
Supplementary Figure 1. Full RXFP1 Western blot scan	4
Supplementary Figure 2. LOFTEE END_TRUNC filter misclassifies variants	5
Supplementary Table Legends	5
Supplementary Table 1. PGR and study characteristics	5
Supplementary Table 2. Pairwise population F_{ST} values	6
Supplementary Table 3. pLoF and damaging missense biomarker burden PheWAS	6
Supplementary Table 4. HBB burden PheWAS	6
Supplementary Table 5. Associations between F_{ROH} and quantitative traits in PGR	6
Supplementary Table 6. Associations between F_{ROH} and binary traits in PGR	7
Supplementary Table 7. Number of homLoF-carrying individuals per gene identified in PGR	7
Supplementary Table 8. Presence or absence of homLoF variants in PGR and gnomAD populations for each gene	7
Supplementary Table 9. Depletion/enrichment of homLoF variants and recessive constraint in PGR for relevant gene sets	7
Supplementary Table 10. homLoF variants for clinically important genes and corresponding phenotypes	7
Supplementary Table 11. RXFP1 PGR burden PheWAS	8
Supplementary Table 12. RXFP1 recall-by-genotype echocardiogram results	8
Supplementary Table 13. RXFP1 recall-by-genotype homLoF variant carrier echocardiogram parameters	8
Supplementary Table 14. CIDEA burden PheWAS	8
Supplementary Table 15. TRPM8 recall-by-genotype study variants and associated number of carriers that were profiled in the cold pressor test	9
References	9
PGR Data Access GANTT Chart.....	9

Supplemental Text

In vitro profiling of RXFP1 variants from PGR

A subset of missense variants present in conserved positions and nonsense variants were selected to evaluate impact on protein expression and activity in transiently transfected HEK293 cells, a cell type that does not express detectable levels of RXFP1 protein. Because RXFP1 is a G_s -coupled GPCR, functionality was assessed by the ability of RXFP1's ligand, relaxin-2, to induce an increase in intracellular cAMP levels. Profiled variants are shown in Extended Data Figure 12.

Many of the variants were still able to generate protein that could be found in the membrane fraction of transfected HEK293 cells (Extended Data Figure 12a). V121G, L197F, and T751M had substantially lower levels of protein expression than other tested variants but were nonetheless detectable.

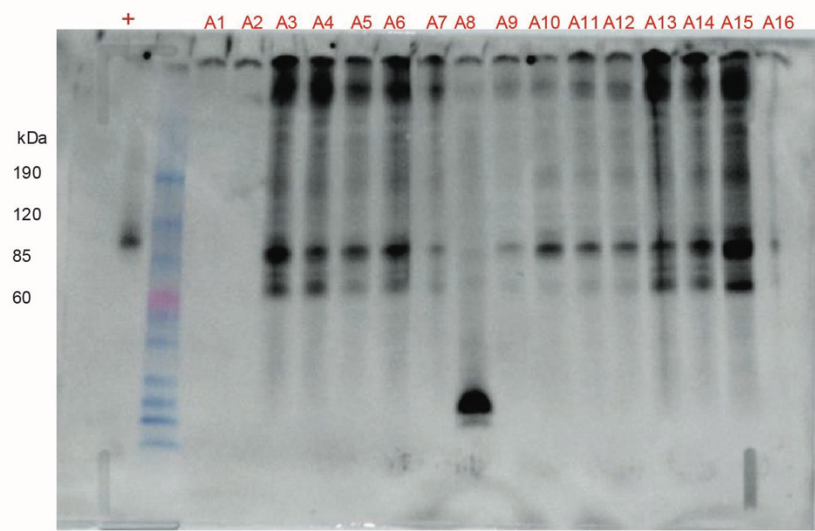
Evaluating the ability of variants to respond to RXFP1's native ligand, relaxin-2, revealed a wide divergence of activity amongst the variants (Extended Data Figure 12b), with some demonstrating activity comparable to reference RXFP1 allele, some with hypomorphic activity profiles (e.g., Y557C and T751M), and some with no detectable activity (e.g., V121G and S132Ter). One end-truncated variant, E743Ter, retained significant activity, illustrating the utility of functional confirmation of the impact of end-truncating variants. A summary of the maximum cAMP responses expressed as a percentage of the reference allele is shown in Extended Data Figure 12c. Variants with no significant activity (<5% of WT E_{max}) were used to seed recall by genotype studies and were also included in burden analyses. The positions of these variants are shown in Extended Data Figure 12d.

End-truncating pLoF variants

In PGR and gnomAD, frameshift and stop gain variants occur at higher frequencies towards the ends of transcripts, consistent with previous reports¹ and reflecting reduced overall evolutionary constraint (Supplementary Figure 2a). This trend is more pronounced for variants with homozygous carriers (Supplementary Figure 2b). To attempt to account for this, the current version of LOFTEE calculates, for each stop gain and frameshift variant, the number of base pairs deleted weighted by the GERP ("Genomic Evolutionary Rate Profiling") scores of the deleted positions. If this cumulative GERP score is less than -58, the variant is filtered as "END_TRUNC". There are several problems with this approach. First, for variants close to the end of the transcript, the cumulative GERP score will rarely have the distance to accumulate to a value of -58. Specifically, while 11.3% of all gnomAD and PGR stop gain and frameshift variants are labeled by LOFTEE as END_TRUNC, of variants occurring in the last percentile of their respective transcripts, only 3.0% are labeled as END_TRUNC, the lowest rate observed across all percentiles (Supplementary Figure 2a). As a result, 16,061 genes have at least one frameshift or stop gain variant that is not annotated as END_TRUNC despite falling in the last 5% of the transcript. For 1,162 of these genes, a homozygous variant of the same description occurs. Secondly, in genes subject to low overall conservation as defined by GERP, variants early in the transcript may be annotated as

END_TRUNC. In PGR and gnomAD, END_TRUNC variants are found in the first 1/3rd of 1,879 genes. Thirdly, at the individual gene level, many genes have END_TRUNC variants that are upstream of non-END_TRUNC variants. Specifically, this occurs for 4,870 (90%) of the 5,400 genes with at least one END_TRUNC variant and one non-END_TRUNC variant (Supplementary Figure 2c). As the cumulative GERP score is a function of the absolute number of bases from the end of a transcript, the artifacts described above are most pronounced for small genes (Supplementary Figure 2d).

Supplementary Figures

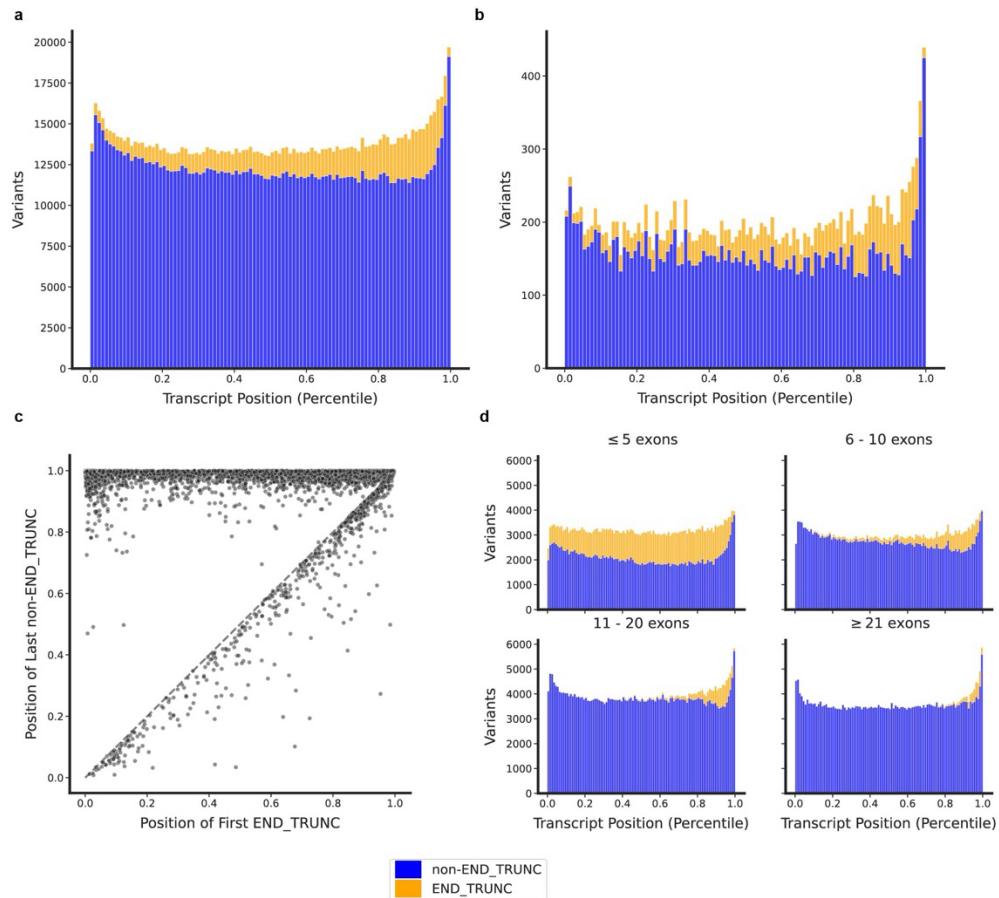


To produce Extended Data Figure 12A: we excised a stable cell membrane extract control left of the molecular weight ladder and cropped out the bottom of the image

A1	no transfection
A2	pcDNA 3.1
A3	WT RXFP1
A4	G30V
A5	N48Y
A6	Y80C
A7	V121G
A8	S132*
A9	V197F
A10	T235I
A11	Y557C
A12	F635Y
A13	C640F
A14	P735L
A15	E743*
A16	T751M

2E6 HEK293, 2ug DNA, 48 hours
+: hRXFP1 HEK293 stable cells
1: mouse anti – RXFP1 15ug/ml ON
2: goat anti-mouse HRP 2 hours RT
detection: KPL Lumi-GLO substrate, 1min

Supplementary Figure 1. Full RXFP1 Western blot scan



Supplementary Figure 2. LofTEE *END_TRUNC* filter misclassifies variants

(A) Counts of frameshift and stop gain variants found in PGR or gnomAD in each transcript position (percentile), colored by whether LofTEE labels them as “*END_TRUNC*”. **(B)** Variant counts as in **(A)** but limited to variants with at least one homozygous carrier in PGR or gnomAD. **(C)** For 5,400 genes with at least one *END_TRUNC* and one non-*END_TRUNC* frameshift or stop gain, the percentile position of the first *END_TRUNC* variant (x-axis) is plotted against the percentile position of the last non-*END_TRUNC* variant (y-axis). **(D)** Variant counts as in **(A)** subset by number of exons in the transcript.

Supplementary Table Legends

Supplementary Table 1. PGR and study characteristics

Baseline characteristics of PGR participants including self-reported ethnicity, case/control breakdown, and availability of a set of biomarkers and binary outcomes analyzed in this study. For each analysis, details about which individuals were included.

Supplementary Table 2. Pairwise population F_{ST} values

F_{ST} values calculated for pairs of PGR ethnicities, 1000 genomes (1KG) populations, and populations from the Human Genome Diversity Project (HGDP). CSA, Central/South Asian; MID, Middle Eastern; SW USA, individuals of African descent residing in the southwestern USA

Supplementary Table 3. pLoF and damaging missense biomarker burden PheWAS

Results of a burden PheWAS of 46 biomarkers. Two sets of variants (“masks”) were considered: (1) the set of high confidence (HC) pLoF variants, and (2) the set of HC pLoF variants plus missense variants predicted to be damaging (REVEL score > 0.7). Associations that met a strict Bonferroni multiple testing p-value threshold ($P < 9.9e-08$) are included in the table. Summary statistics from the AstraZeneca PheWAS Portal are included for comparison. ptv5pcnt, protein truncating variants with minor allele frequency ≤ 0.05 within the cohort and within gnomAD; ptvaredmg, protein truncating variants ≤ 0.001 within the cohort and within gnomAD and non-synonymous variants predicted to be damaging (REVEL score ≥ 0.25) with minor allele frequency ≤ 0.0005 within the cohort and within gnomAD; No with QV, number of participants with the qualifying variant (QV); No without QV, number of participants without the qualifying variant (QV). *For AstraZeneca PheWAS data, PLA2G7 levels were used rather than activity.

Supplementary Table 4. HBB burden PheWAS

Burden PheWAS of HBB pLoF variants in PGR. NAFLD, non-alcoholic fatty liver disease; BMI, body mass index; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; Apo, apolipoprotein; eGFR, estimated glomerular filtration rate; ALT, alanine aminotransferase; AST, aspartate aminotransferase; HetRA, heterozygous reference/alternative; HomAA, homozygous alternative; AF, cumulative allele frequency; OR, odds ratio; pLoF, putative loss-of-function.

Supplementary Table 5. Associations between F_{ROH} and quantitative traits in PGR

Summary stats for associations between F_{ROH} and quantitative phenotypes under different testing scenarios. All phenotypes were analyzed using linear regression with age, age², gender, sequencing platform, and the first 10 genetic PCs as covariates. Phenotypes were transformed using a rank-based inverse normal transformation (RINT) except where otherwise specified. Cochran’s Q test was used to assess heterogeneity of effect sizes by ethnicity, age bin, and sex. BMI, body mass index; WHR, waist-to-

hip ratio; LDL-C, low density lipoprotein cholesterol; HDL-C, high density lipoprotein cholesterol; BP, blood pressure; AST, aspartate aminotransferase; ALT, alanine aminotransferase; eGFR, estimated glomerular filtration rate

Supplementary Table 6. Associations between F_{ROH} and binary traits in PGR

Summary stats for associations between F_{ROH} and binary outcomes under different testing scenarios. All phenotypes were analyzed using logistic regression with age, age², gender, sequencing platform, and the first 10 genetic PCs as covariates. Cochran's Q test was used to assess heterogeneity of effect sizes by ethnicity, age bin, and sex. AFib, atrial fibrillation; NAFLD, non-alcoholic fatty liver disease; COPD, chronic obstructive pulmonary disease

Supplementary Table 7. Number of homLoF-carrying individuals per gene identified in PGR

Homozygous pLoF variant counts are shown in the format

"chromosome:position:reference_allele:alternative_allele" (GRCh38). In cases where individuals are homozygous for multiple pLoF variants in a gene, the individual variants are separated with a '/'.

Supplementary Table 8. Presence or absence of homLoF variants in PGR and gnomAD populations for each gene

Each row indicates whether a homLoF variant in the specified gene was observed in PGR and for each gnomAD population.

Supplementary Table 9. Depletion/enrichment of homLoF variants and recessive constraint in PGR for relevant gene sets

Gene homLoF odds ratios, homozygote depletion, and autozygous allele depletion stats for various gene sets. GSEA², Gene Set Enrichment Analysis; HPA, Human Protein Atlas³; AAR, weighted average Autozygous Allele Ratio

Supplementary Table 10. homLoF variants for clinically important genes and corresponding phenotypes

Examples of genes of clinical importance for which there are homLoF variant carriers in PGR or discovered through recall-by-genotype studies. For individuals who were identified in PGR and enrolled in a recall-by-genotype study, phenotypic and demographic parameters reported here are from the recall

study. Age of onset, when available, is shown in parentheses. BMI, body mass index; LDL-C, low-density lipoprotein cholesterol; AST, aspartate aminotransferase; ALT, alanine aminotransferase; eGFR, estimated glomerular filtration rate; MI, myocardial infarction; NAFLD, non-alcoholic fatty liver disease. Coding (HGVS_c) and protein (HGVS_p) sequence changes are relative to Ensembl canonical transcripts version 107. HGVS, Human Genome Variation Society; n/a, not available; kg, kilograms; m, meters; mg, milligrams; dl, deciliters; mL, milliliters; L, liter

Supplementary Table 11. RXFP1 PGR burden PheWAS

Burden PheWAS of RXFP1 pLoF variants and *in vitro* confirmed pLoF variants. NAFLD, non-alcoholic fatty liver disease; BMI, body mass index; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; Apo, apolipoprotein; eGFR, estimated glomerular filtration rate; ALT, alanine aminotransferase; AST, aspartate aminotransferase; HetRA, heterozygous reference/alternative; HomAA, homozygous alternative; AF, cumulative allele frequency; OR, odds ratio; pLoF, putative loss-of-function.

Supplementary Table 12. RXFP1 recall-by-genotype echocardiogram results

Summary of results from PheWAS of RXFP1 loss-of-function with categorical and continuous traits computed from echocardiograms (ECHOs) in the RXFP1 recall-by-genotype study. Results from linear mixed effects models and logistic regression are reported. Only additive models were considered.

Supplementary Table 13. RXFP1 recall-by-genotype homLoF variant carrier echocardiogram parameters

Table of quantitative and categorical measurements of cardiac structure and function computed from echocardiograms of individuals with RXFP1 complete loss-of-function in the recall-by-genotype study.

Supplementary Table 14. CIDEA burden PheWAS

Burden PheWAS of CIDEA pLoF variants and damaging missense variants (REVEL score >0.7) in PGR. NAFLD, non-alcoholic fatty liver disease; BMI, body mass index; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; Apo, apolipoprotein; eGFR, estimated glomerular filtration rate; ALT, alanine aminotransferase; AST, aspartate aminotransferase; HetRA, heterozygous reference/alternative; HomAA, homozygous alternative; AF, cumulative allele frequency; OR, odds ratio; pLoF, putative loss-of-function.

Supplementary Table 15. TRPM8 recall-by-genotype study variants and associated number of carriers that were profiled in the cold pressor test

HomRR, homozygous reference; HetRA, heterozygous; HomAA, homozygous alternative

References

- 1 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828 (2012). <https://doi.org/10.1126/science.1215040>
- 2 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005). <https://doi.org/10.1073/pnas.0506580102>
- 3 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015). <https://doi.org/10.1126/science.1260419>

PGR Data Access GANTT Chart

- **Submission (Day 0)**
 - Applicant emails a research proposal (<1 page) to info@cncd.org delineating what research questions they have and what data (genotype, phenotype) they would need access to address this question
- **Initial review: feasibility + IRB compatibility (Days 1–45)**
 - Screen for general feasibility, alignment with existing/approved IRB coverage, basic governance fit, and anticipated costs
 - *Outcome:* proceed / request revisions / decline
- **Scientific & data access review by data scientists (Days 45–90)**
 - Assess data suitability for research query, technical feasibility of data transfer and any contractual obligations, privacy-preserving approach (high level), and resource needs
 - *Outcome:* proceed / request revisions / decline
- **Committee decision (Days 90–180)**
 - Consolidate reviews, final determination, conditions (if any)
- **Applicant notification & agreement execution (Days 180-365)**
 - Communicate decision; if approved, complete required data access agreement / terms and any institutional sign-offs
- **Provisioning & access start (Days 180-365+)**
 - Account/workspace setup, dataset provisioning, confirmation of access start date