HGG Advances

A power-based sliding window approach to evaluate the clinical impact of rare genetic variants in the nucleotide sequence or the spatial position of the folded protein

Elizabeth T. Cirulli,[1,6,7,*] Kelly M. Schiabor Barrett,[1] Alexandre Bolze,[1] Daniel P. Judge,[2] Pamala A. Pawloski,[3] Joseph J. Grzymski,[4,5] William Lee,[1] and Nicole L. Washington[1]

Summary

Systematic determination of novel variant pathogenicity remains a major challenge, even when there is an established association between a gene and phenotype. Here we present Power Window (PW), a sliding window technique that identifies the impactful regions of a gene using population-scale clinico-genomic datasets. By sizing analysis windows on the number of variant carriers, rather than the number of variants or nucleotides, statistical power is held constant, enabling the localization of clinical phenotypes and removal of unassociated gene regions. The windows can be built by sliding across either the nucleotide sequence of the gene (through 1D space) or the positions of the amino acids in the folded protein (through 3D space). Using a training set of 350k exomes from the UK Biobank (UKB), we developed PW models for well-established gene-disease associations and tested their accuracy in two independent cohorts (117k UKB exomes and 65k exomes sequenced at Helix in the Healthy Nevada Project, myGenetics, or In Our DNA SC studies). The significant models retained a median of 49% of the qualifying variant carriers in each gene (range 2%–98%), with quantitative traits showing a median effect size improvement of 66% compared with aggregating variants across the entire gene, and binary traits' odds ratios improving by a median of 2.2-fold. PW showcases that electronic health record-based statistical analyses can accurately distinguish between novel coding variants in established genes that will have high phenotypic penetrance and those that will not, unlocking new potential for human genomics research, drug development, variant interpretation, and precision medicine.

Introduction

Statistical analyses of rare genetic variants in large populations present unique challenges. Variants that are only observed in a handful of people, or even one person, lack statistical power to identify whether they are associated with a trait.[1] Gene-based collapsing methods navigate this problem by grouping together similar rare variants, often by predictions of functional consequence, to improve power.[2] However, not all nonsynonymous variants in a gene, even when they are predicted to be damaging by various *in silico* tools, can be expected to have the same effect on a phenotype, and grouping them together in this way dilutes the overall signal.

In addition to utilizing cellular and model organism functional assays, many analysis approaches have been developed to identify and prioritize the types of rare variants and gene regions that are most important for an association between a gene and a particular phenotype. First, analyses of LoF variants (loss of function: nonsense, frameshifts, and essential splice sites) and coding variants (damaging missense and in-frame indels) are often performed separately, to distinguish their effects.[3–5] Algorithms like SKAT (Sequence Kernel Association Test) allow genetic variants in a single gene to have different effect sizes and directions of effect, giving an overall signal for the gene even when not all variants behave the same.[6] Other studies have focused on just analyzing the intolerant regions of the gene or specific gene domains to identify the source of a gene's signal.[7,8] Sliding window and clustering methods to localize the most important regions of the gene have also been tried.[9–13] However, there is still a need to develop a flexible and unbiased statistical analysis method that effectively selects the parts of a gene to include in association studies and apply it to improving the overall statistical associations for rare coding variants, especially when it comes to parsing out the portions of the gene that are *not* associated with the trait.

Here, we present Power Window (PW), a novel technique that leverages paired clinical phenotypes with genetic sequence to identify regions of a gene where rare nonsynonymous variants of any type—for example, LoF or coding—are statistically significantly associated with a trait. We use PW to build regional LoF and coding models for well-established gene-disease associations in a large training set and test these refinements in two additional cohorts. Not only do these models replicate in cohorts with a different composition of variants, but many drive

[1]Helix, 101 S Ellsworth Ave Suite 350, San Mateo, CA 94401, USA; [2]Division of Cardiology, Medical University of South Carolina, 30 Courtenay Drive, MSC 592, Charleston, SC 29425, USA; [3]HealthPartners, Minneapolis, MN 55125, USA; [4]University of Nevada, 2215 Raggio Pkwy, Reno, NV 89512, USA; [5]Renown Institute for Health Innovation, Reno, NV 89512, USA
[6]X (formerly Twitter): @ecirulli
[7]Lead contact
*Correspondence: liz.cirulli@helix.com

https://doi.org/10.1016/j.xhgg.2024.100284.

dramatic improvements to the effect size or odds ratio (OR), especially for coding variants. PW showcases that even in the absence of family data, prior clinical evidence for that variant, or functional tests, EHR (electronic health record)-based statistical analyses alone can be used to refine gene signals to determine the types and locations of variants that will have high penetrance in population cohorts. This highly accurate method for prioritizing genetic variants associated with health outcomes unlocks new potential for common disease genomic risk screening.

Subjects and methods

Genetic data and phenotypes

We utilized the UK Biobank (UKB) population level exome OQFE pVCFs for 470k individuals (field 23157, with genotypes set to missing when DP (read depth) < 7 for SNVs and <10 for indels, and variants excluded if there were no homozygotes or the max allelic balance was <0.15 for SNVs or <0.2 for indels as per Backman et al.[14]) as well as the imputed genotypes from genome-wide association study genotyping (field 22801–22823). We also utilized 62,406 samples that were sequenced and analyzed at Helix using the Exome+® assay as previously described, recruited from the Healthy Nevada Project (n = 37,989, sequenced January 2018 to March 2023); myGenetics (n = 15,104, sequenced May 2022 to March 2023); and In Our DNA SC (n = 9,313, sequenced December 2021 to March 2023) (Table S1).[8] Standard QC for all samples included removing individuals with a difference between genetic and self-reported sex and removing contaminated samples. For the UKB cohort (n with exomes = 470k), participants range in age as of 2022 from 51 to 88 and are 55% female, while the Helix age range is from 18 to 89+ and is 69% female. The UKB is 83% composed of individuals who are genetically similar to British Europeans, with another 10% with genetic similarity to other Europeans and 7% genetically similar to other ancestry groups, and the Helix cohorts are 77% composed of individuals with genetic similarity to Europeans, 14% with genetic similarity to those from the Americas, and 9% with genetic similarity to those of other ancestries. No filtering was applied to the cohorts based on genetic similarity.

The Helix cohorts were reviewed by Salus IRB (Institutional Review Board; reliance on Salus for all sites) and approved (approval number 21143), the WCG IRB (Western Institutional Review Board, WIRB-Copernicus Group) and approved (approval number 20224919), the Medical University of South Carolina Institutional Review Board for Human Research and approved (approval number Pro00129083), and the University of Nevada, Reno Institutional Review Board and approved (approval number 7701703417). The UKB study was approved by the North West Multicenter Research Ethics Committee, United Kingdom. All participants gave their informed, written consent before participation. All data used for research were de-identified.

Helix cohort phenotypes were processed from Epic/Clarity EHR data as previously described and updated as of January 2023.[8] International Classification of Diseases, Ninth and Tenth Revision codes and associated dates (ICD-9 and ICD-10-CM) were collected from available diagnosis tables (from problem lists, medical histories, admissions data, surgical case data, account data, claims, and invoices). The data were sourced from EHR data formatted using the OMOP CDM v5.4. Each ICD source code was mapped to a source concept id and the source concept id was used to extract the relevant diagnoses. Quantitative phenotypes were transformed via rank-based inverse normal transformation.

UKB data were provided from the UKB resource (accessed September 2022). ICD codes and associated dates (both ICD-9 and ICD-10) were collected from inpatient data (category 2000), cancer register (category 100092) and the first occurrences (category 1712), which records the earliest instance of a diagnosis from the Primary Care data, Hospital inpatient data, Death Register records, and self-reported medical conditions mapped to an ICD-10 code at a three-character resolution (i.e., E11 instead of E11.0). UKB quantitative phenotypes were processed using the Neale lab modified version of PHESANT as previously described, which rank-transforms quantitative traits to normally distributed data and divides categorical traits into binary sets.[8,15,16]

ICD codes were translated to phecodes as previously described[17–20] (Table S1).

Genetic analysis

Variant annotation was performed with VEP (Variant Effect Predictor) 104.[21] Coding regions were defined according to Gencode version GENCODE 33, and the MANE (Matched Annotation from NCBI and EMBL-EBI) transcript (or Ensembl canonical transcript if there was no MANE transcript) was used to determine variant consequence. [22,23] Variants were restricted to CDS (coding sequence) regions and essential splice sites. We did not restrict variants according to missingness or Hardy-Weinberg equilibrium. Genotype processing was performed in Hail 0.2.115–10932c754edb.[24]

The collapsing analysis was performed as previously described.[17] Qualifying variants included coding (missense_variant, inframe_deletion, or inframe_insertion) that was not PolyPhen and SIFT benign, or LoF (stop_lost, start_lost, splice_donor_variant, frameshift_variant, splice_acceptor_variant, or stop_gained).[25,26] A MAF (minor allele frequency) cutoff of 0.1% was used in all gnomAD populations as well as locally within each genetic similarity group. [27] For visualization, UniProt was used to identify gene domains.[28]

We used regenie for genetic analyses, which builds a whole genome regression model to account for relatedness and population stratification and also accounts for case-control imbalances.[29] The covariates we included were age, sex, age*sex, age*age, sex*age*age, and bioinformatics pipeline version, and we used a set of 184,445 common variants to build the whole genome regression model. Ancestries were analyzed together, as we have previously established that this method works well for analyzing rare causal variants grouped together.[8,17,30]

Training PW models

PW uses a sliding window methodology to create windows with equal numbers of rare variant carriers across the gene (Figure 1A). To train, we built windows in 350k UKB samples; this otherwise randomly selected set (of the 470k UKB exome release) included all individuals who were first- or second-degree relatives or twins, so that the subsequent testing set would only include unrelated individuals. When building the windows, we required them to each contain 20 qualifying variant carriers out of these 350k individuals (see Figure S1 for stats about the windows). In the case of homozygotes, the number of rare variant alleles was used. PW analyses were run for 65 established gene/disease relationships (see the results section). Specifically, we built the windows separately for LoF and coding models, and we only built

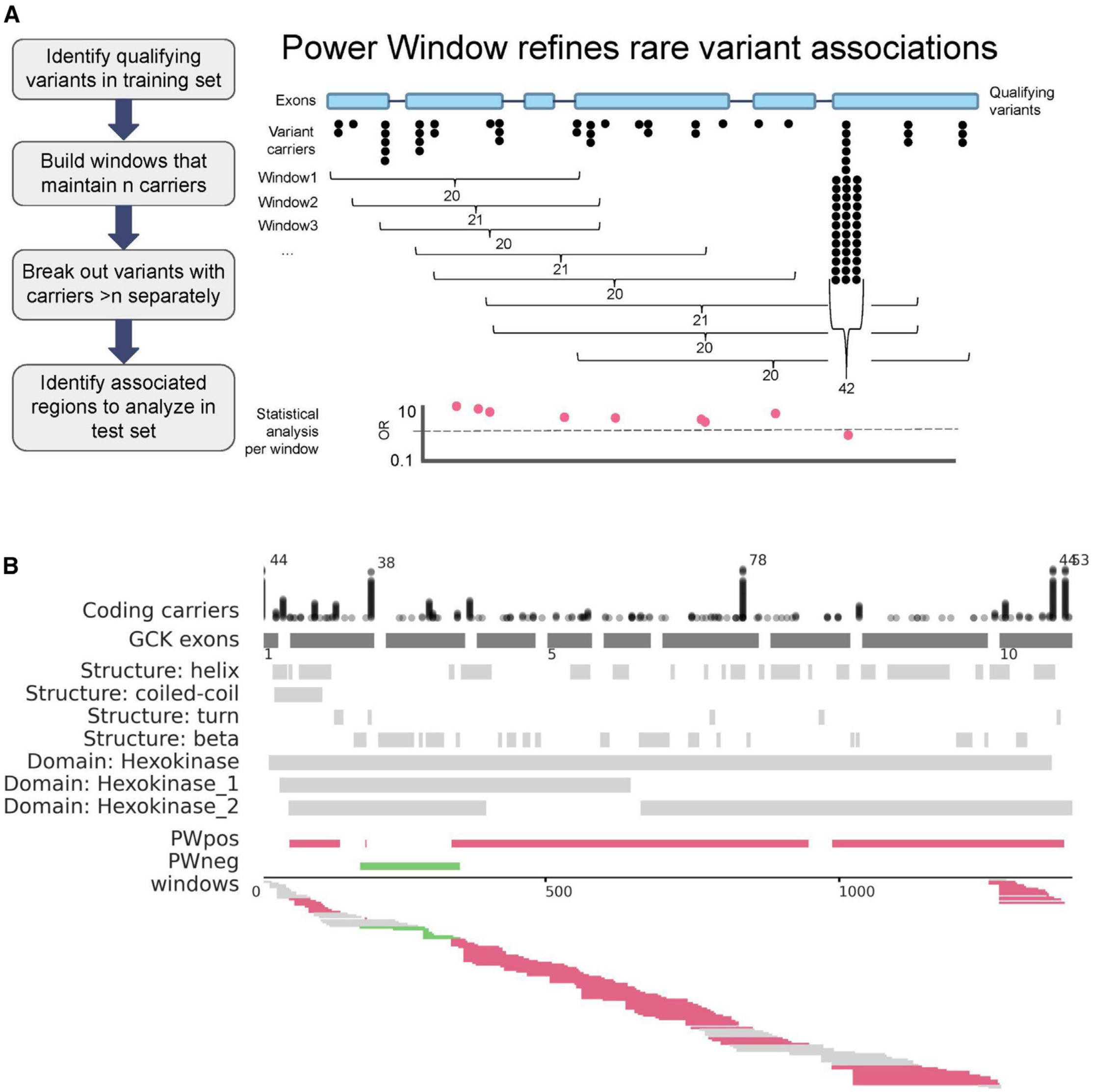


Figure 1. Power Window methodology and as applied to *GCK* in UKB training data
(A) Diagram of the methodology.
(B) Power Window (PW) applied to *GCK* and glucose levels in the UKB350k discovery cohort. Tracks are drawn against the *GCK* canonical coding transcript ENST00000403799.8. Coding position is shown at the bottom scale. Coding carriers: each dot represents an individual carrier and dots are stacked for each carrier at a given position. For brevity, carriers are trimmed to 35 and total number of carriers is indicated when total carriers >40. *GCK* exons: exons (dark gray to scale; introns not to scale). Exon number indicated below exon track. Secondary structure and major structural domains are shown according to UniProt. PW: bedtools merge of all significant $\text{PW}_{\text{coding}}$ windows with a positive direction of effect (beta >0.5; pink), as indicated in the "windows" track below. opp-PW: merger of all significant windows with a negative direction of effect (beta <−0.5; green). Windows: each window that was generated through applying the PW algorithm is shown, with a window size of 20 carriers per window. Significant association with glucose levels is indicated when beta <−0.5 or beta >0.5 (97.5% confidence that the true beta is not 0 in a sample of 20 individuals with a normalized phenotype). Windows are shown in pink if significant under the positive model and green if significant under the negative model, or gray if not significant (beta between −0.5 and 0.5).

windows for gene/model combinations that had at least 40 qualifying rare variant carriers in the training set (this excluded seven genes: *SF3B1* [MIM: 605590)], *GP9* [MIM: 173515], *CDKN2A* [MIM: 600160], *SLC4A1* [MIM: 109270], *GCK* [MIM: 138079], *FCGRT* [MIM: 601437], and *GFI1B* [MIM: 604383] from the LoF model and no genes from the coding model). We then ran a regression analysis of the relevant phenotype in regenie on this training set for each window generated for each gene.

For PW utilizing the 3D space of the folded protein to build windows (3DPW), predicted coordinates for the atoms in the protein were obtained from Alphafold, using the MANE Overlap dataset when available (mane_overlap_v4.tar) and otherwise the compressed *Homo sapiens* proteome (UP000005640_9606_HUMAN_v4.tar) (proteins ≥2,700 amino acids in length).[31,32] The predicted coordinates for proteins with at least 2,700 amino acids were split over multiple files, and their coordinates were harmonized using Kabsch's algorithm implemented with the scipy.spatial.procrustes function in Python. The position for each amino acid was chosen based on the alpha carbon x, y, and z coordinates. Distances between amino acids were calculated according to

Euclidean distance. A window was built centered around each amino acid change, with inclusion in the window determined by the Euclidean distance to other amino acid changes, spreading out in 3D space according to the structure of the protein.

We next decided which windows to retain in our final model for each gene using a confidence interval approach. While frameworks for testing the significance of a mixture of high- and low-effect windows of rare variants in a gene exist from prior works, here we chose a simple approach that uses an effect size cutoff from the widely used association analysis software regenie, with the final goal of testing the output of that model for significance in an independent cohort.[9–13] For a sample size of 20 carriers measured for a rank-based inverse normal transformed phenotype with a true mean of 0, then 97.5% of the time, the observed mean in the sample will be between −0.5 and 0.5. An effect size >|0.5| was therefore chosen as the cutoff for a window to be considered associated with the quantitative trait. Simulations for the gene *GCK* and the phenotype glucose in which causal regions were defined and glucose levels reassigned according to different parameters can be found in Figure S2 and provide more information on how the parameters of percent of gene implicated, number of distinct implicated regions in the gene, percent of implicated variants that are causal, effect size of causal variants, and different cutoffs for building models affect the outcomes. For binary traits, the cutoff was an OR higher than the maximum expected 99% of the time if the true OR was 1 in a sample of 20, given binomial probability and the case frequency for that phenotype. For example, if the true OR was 1, then in a sample of 20 carriers, 99% of the time a binary trait occurring in 1 out of 100 people would be expected to have fewer than three case carriers and thus an OR <17.5 when compared with the 349,980 non-carriers. If the binary trait occurred in 1 out of 20 people, then the OR cutoff would be 6.3 (5+ case carriers). The OR cutoff was thus different for each binary trait and tailored to its frequency, including tailored to sex-specific analyses when necessary, such as for breast cancer.

The final PW model for each gene separated the gene into regions that had statistical evidence for an association with the trait (PW) and those that did not (non-PW). For quantitative traits, sometimes two directions of effect were observed within different regions of the same gene, in which case an additional model was produced for the regions that had the opposite direction of effect from the main signal for the gene (opp-PW). Models were built separately for coding (PW_{coding}) and LoF (PW_{LoF}) annotations.

Testing PW in independent samples

We next tested the PW models, non-PW models, and whole-gene models in an independent set of 117k unrelated UKB samples. Individuals were analyzed separately according to whether they had variants within the PW regions, within the non-PW regions, or, for quantitative traits, within the opp-PW regions. We compared the ORs and betas in this independent testing cohort between the whole-gene model, the PW model, and the non-PW model, where little signal is expected to remain if the method is appropriately picking out the associated portions of the gene. We additionally assessed significance in the test set by removing individuals without a variant in the gene and directly comparing the PW with non-PW individuals. Models were considered to be significant if they had a p value less than 0.0005, which corrects for 96 tests (the 96 gene-phenotype-model combinations that produced PW models that did not just include 0% or 100% of the gene).

For the significant PW models, we additionally performed a replication study in an independent set of 65k samples sequenced at Helix. Phenotypes were available to check in this secondary cohort for 29 of the significant models.

Results

A statistical power-based sliding window method to localize rare variant association signals within a gene

The basic concept of a sliding window analysis is to group variants located near each other, in either 1D or 3D space, into one unit and analyze them together to improve power, much like a gene-based collapsing analysis but at a more localized scale. Rather than size our sliding window by the number of variants, bases, or amino acids covered, our sliding window moves to maintain roughly the same number of people within the cohort with a rare qualifying variant, and thus the statistical power, within each window (Figure 1A). When a single variant is well powered on its own, it is removed out into its own separate analysis, and the window slides past it, continuing to group surrounding variants as appropriate. We call this technique the Power Window (PW) method.

Parameters and functions

The carrier count used to define windows can be adjusted to fit different scenarios. There is a balance to strike, rooted in the currently available cohort sample size, between being able to home in on a specific region (small window size) and having adequate power to observe statistical associations (large window size) (Figure S3). We use a window size of 20 qualifying variant carriers in a training set of 350k UKB exomes because this powers us to identify associations with quantitative traits with an effect size approximately >|1| for a normalized trait (see Subjects and methods and Figure S2). This breaks each gene into a mean of 357 coding windows and 67 LoF windows (Figure S1, where details on *TTN*—which is excluded from stats given here due to its size—are also available). With the same number of qualifying variant carriers within each window, the statistical power for discovery is the same for each window as our analysis slides across the gene or protein. Nucleotides or amino acids that fall within the region of an associated window are assigned the same value in new datasets as that window has in the training set. Thus, new mutations of the same predicted impact (LoF, missense, etc.) that do not occur in the training dataset can still be assigned a value based on their location in the new dataset. This also means that gene or protein regions with no variation in the training set are assigned a value based on the values of variants in the surrounding regions, which defines whether or not they are within the boundaries of an associated window (for example, the third exon in Figure 1A). The ability of this method to home in on more and more specific regions of the gene increasingly improves as sample sizes grow and the distance covered by each window shrinks.

PW models refine gene-disease association signals

We evaluate the PW method using 65 genes with which we previously demonstrated rare variants to be significantly associated with well-documented phenotypes in the UKB using a gene-based collapsing analysis of LoF or coding variant models.[8,17] We chose these associations because they have strong, well-established effects that can be identified in smaller cohorts than the sample used here. Because these gene-phenotype combinations already have a strong statistical signal at the gene level, they are good candidates to test whether the signal is truly gene-wide or can be further localized. There are 37 quantitative traits and 28 binary traits included in this set. For each gene, we build and analyze both coding and LoF windows, irrespective of the original whole-gene association model, for the relevant phenotype in a training set of 350k UKB exomes and build a final model that separates the gene into PW regions and non-PW regions (see Subjects and methods).

We start with a PW algorithm that builds windows across the linear sequence of the coding nucleotides (through 1D space). We first examine if there are differences in how a PW analysis of coding variants (PW_{coding}) or LoF variants (PW_{LoF}) separates regions within a gene (Figure 2). We anticipate that PW_{LoF} models will usually implicate the entire gene instead of specific regions, as generally a LoF variant anywhere in the gene is likely to have a similar effect (with tissue-specific splicing causing some exceptions, see Schiabor Barrett et al.[30]).

Overall for the 65 genes, we find that applying PW_{coding} retains a mean of 34% of the carriers in each gene (range 0%–98%), while PW_{LoF} retains 74% (range 0%–100%) of carriers. In terms of the percent of the coding sequence (CDS) retained in the PW model in each gene, this is 40% (range <0%–99%) for PW_{coding} and 79% (range 0%–100%) for PW_{LoF} (Figure S4).[8,33] While a small percentage of the gene being retained indicates that the association signal is very specific, PW also identifies the gene-phenotype associations where a rare variant anywhere in the gene appears to have a similar effect, which occurs frequently for PW_{LoF} but also for some PW_{coding} models. We find that 48% of the PW_{LoF} models retain >90% of the CDS of the gene, compared with 9% for PW_{coding}.

To better understand the ways in which PW is able to refine the signal for different gene-phenotype relationships, we group the same 65 genes based on whether their primary association from the gene-based collapsing analysis was for LoF only or for both coding and LoF models to see if there are differences in the resulting PW models. We observe that the PW_{LoF} model retains most of its carriers, and thus CDS, regardless of the original whole-gene signal (Figure 2). In contrast, for PW_{coding}, the portion of the gene that is kept depends heavily on the type of association seen at the whole-gene level (Figure 2): if the whole-gene association signal is mainly for the LoF model ($n = 29$), then a mean of 17% (range 0%–69%) of the carriers are retained by the PW_{coding} model, translating to a mean of 27% (range 0%–90%) of the CDS. When the whole-gene association can be identified using both coding and LoF models ($n = 36$), then a mean of 48% (range 2%–98%) of the carriers are retained, corresponding to a mean of 50% (range 3%–99%) of the CDS (Figure 2).

PW models dramatically improve ORs and effect sizes in new datasets

To confirm the refinements established in the training set of 350k UKB exomes, we evaluate the predictive power of the PW models in an independent testing set of 117k UKB exomes. For each gene, we identify test set individuals with qualifying variants in the regions included in the PW models (PW) or excluded from them (non-PW). This analysis includes a mean of 51/11 new coding/LoF variants per gene, adding to the mean of 340/66 coding/LoF variants already included from the training set (for *TTN*, which we exclude from the summary stats due to its size, the values are 1,373/212 new and 9,275/1,182 existing coding/LoF variants). We analyze the relevant phenotypes in these groups compared with non-carriers and compare the resulting ORs (binary traits) and effect sizes (betas; quantitative traits) (Figures 3, S6, and S7). We then identify models that show statistical support for PW as follows.

For binary traits with PW_{coding}, we find that 24 out of 28 gene-disease associations show a higher OR in the PW compared with whole-gene model (mean fold improvement = 3.4, range = 0.45–18.4), and six (21%) of these differences are statistically significant compared with the non-PW portions of the gene (Bonferroni correction at $p < 0.0005$); Figures 3A and S6A). For example, the association between coding variants in *OCA2* (MIM: 611409) and blonde hair color has an OR of 2.8 (95% confidence interval [CI] 2.3–3.4) for the whole gene, but PW_{coding} splits this into 33% of carriers with an OR of 6.3 (95% CI 4.8–8.4) in the test set, while the 67% that are non-PW carriers have an OR of 1.59 (95% CI 1.2–2.9; p value 5.1e−20 for a comparison between PW and non-PW in the test set) (Table S1). The regions implicated in this model are in the extracellular and citrate transporter domains of this gene, which can be hypothesized to impact its ability to maintain melanosome pH (Table S2; Figure S5). [34]

For quantitative traits with PW_{coding}, the resulting models are even more reliable for the test set: we find that 36 out of 37 genes show a more extreme effect size in the PW compared with whole-gene model (mean percent improvement = 145%, range = 0% to 1,331%), and 22 (59%) of these PW effects are statistically significantly different from the non-PW effects in the test set (Figures 3B and S6B). For example, the association between coding variants in *GFI1B* (MIM: 604383) and normalized mean platelet volume has an effect size of 0.52 (95% CI 0.41–0.62) for the whole gene, but PW_{coding} splits this

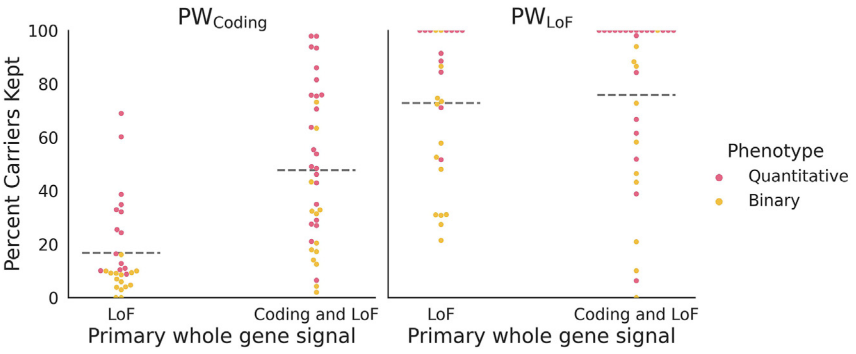


Figure 2. Percent carriers kept in Power Window models
For each gene (n = 65), the percent of the rare variant carriers from the whole-gene model who were retained by the PW model was evaluated for PW_{Coding} and PW_{LoF}. Within each model, the phenotypes were grouped into quantitative (pink) or binary (yellow) traits, and the genes were grouped based on the original genome-wide association as follows: LoF: original whole-gene associations had an absolute beta at least 3x as high in the LoF model as the coding model; coding and LoF: original whole-gene associations had a beta <3x as high for LoF as for coding (no gene-phenotype combinations had a whole-gene coding model absolute beta that was at least 3x higher than the whole-gene LoF beta). Two genes had 0 windows included in the final model for PW_{coding} (*IFT140* and *NF1*), and 1 for PW_{LoF} (*JAK2*). Dotted horizontal line indicates the mean percent carriers kept in that category. The percent CDS kept can be found in (Figure S4).

into 64% of carriers with an effect size of 0.71 (95% CI 0.57–0.84) in the test set, while the 36% that are non-PW carriers have an effect size of 0.18 (95% CI 0.01–0.35; p value 2.3e−6 for a comparison between PW and non-PW in the test set) (Table S1). The associated regions overlap zinc finger domains, which have previously been demonstrated to be involved with this trait (Figure S5).[8,35,36]

Compared with the original whole-gene model, the level of improvement from PW is often quite dramatic: for example, 55% of the significant PW_{coding} models for quantitative traits show a more than 60% improvement in the normalized effect size. Overall, the median fold improvement for significant PW_{coding} models is 2.2 for ORs (range 1.5–12.3; mean 4.1; binary traits) and 65% for percent change in effect size (range 2%–1,331%; mean 154%; quantitative traits) (Figure 4).

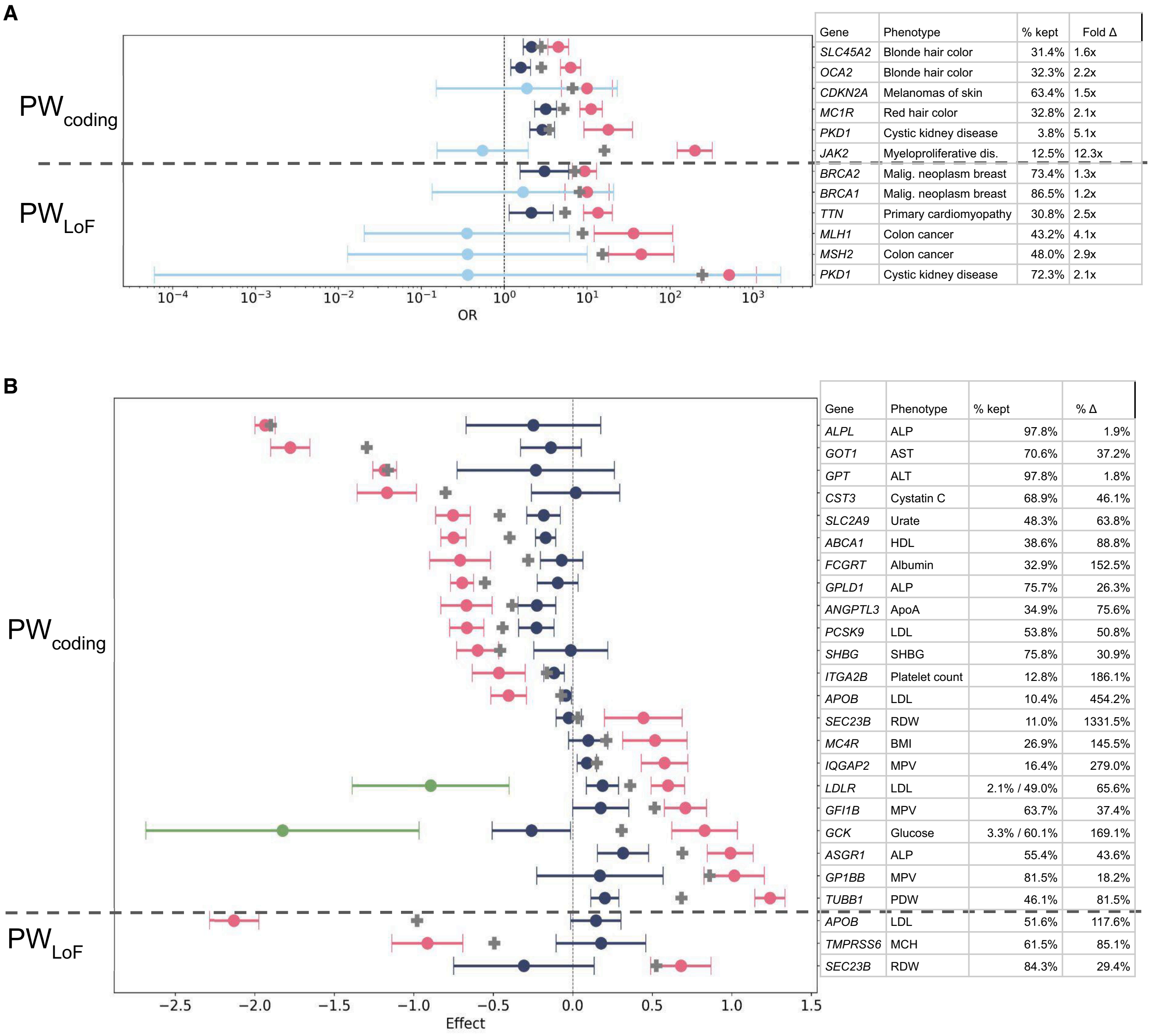
In contrast, the PW_{LoF} models show little to modest improvement over using the whole gene for either quantitative or binary traits and generally do not show a significant difference between PW and non-PW regions (Figures 3A and 3B bottom, S6B, and S7B). For most genes, when there is a LoF association, the statistical signal for LoF variants is spread across the entire gene, resulting in nearly the entire gene being included in PW_{LoF} models. For example, LoF variants anywhere in *LDLR* (MIM: 606945) are generally considered pathogenic, and indeed PW_{LoF} implicates the entire gene (Figure S7B). However, in specific situations, PW_{LoF} models perform well. For example, the PW_{LoF} model for *TTN* (MIM: 188840) with cardiomyopathy improves the OR by 2.5x in the test set by mostly restricting to cardiac-expressed exons, and the PW_{LoF} model for *APOB* with LDL (low-density lipoprotein) levels produces an effect size improvement of 118% by removing the last 1,082 bases of the gene (two-thirds of the final exon) as well as three single variants that did not show associations (Tables S1 and S2).[30] We also test using LOFTEE (Loss-Of-Function Transcript Effect Estimator) to remove low confidence LoF variants from the whole-gene models, which improves the effect size for some genes, but does not consistently result in a positive change (Figures S6B and S7B).[37] In contrast, our method to use phenotypic evidence to tailor which pLoF variants can be safely excluded from the LoF category could help curate potentially pathogenic variants.

Finally, we test the significant PW models in an independent cohort of 65k individuals sequenced at Helix. In the 65 genes, there are a mean of 78/21 new coding/LoF variants tested per gene in this set, as well as 106/14 coding/LoF variants per gene that were already used in the UKB analysis (for *TTN*, which we summarize separately, these numbers are 2,431/331 new and 3,219/152 already used). We find that 25 of the 27 significant models that have cases in both PW and non-PW regions in this replication cohort show the same direction of effect as the UKB. However, due to the smaller sample sizes, the difference between the PW and non-PW models is only statistically significant ($p < 0.002$, correcting for 27 tests) in 10 of the models in the Helix cohorts (Figure S8).

Overall, our results demonstrate that the PW methodology successfully segments gene regions that contribute to gene-based collapsing analysis signals, dramatically improving the effects seen compared with whole-gene models.

Using protein structures to build analysis windows through 3D space

We next built windows for analysis by focusing on the 3D protein structure (3DPW). This allows variants to be grouped together even when they are distant from each other in the coding sequence but are near each other in the final 3D location in the folded protein. Using predictions from Alphafold, we calculated the Euclidean distance between all pairs of amino acid changes in each analyzed protein.[31,32] Just as with the PW version that focused on the linear coding sequence (through 1D space), this analysis built windows out from each amino acid change until 20 carriers were identified in the training set (Figure 5). Here, we only focused on coding models, as the precise location of LoF variants within the 3D structure of the protein is less



Gene	Phenotype	% kept	Fold Δ
SLC45A2	Blonde hair color	31.4%	1.6x
OCA2	Blonde hair color	32.3%	2.2x
CDKN2A	Melanomas of skin	63.4%	1.5x
MC1R	Red hair color	32.8%	2.1x
PKD1	Cystic kidney disease	3.8%	5.1x
JAK2	Myeloproliferative dis.	12.5%	12.3x
BRCA2	Malig. neoplasm breast	73.4%	1.3x
BRCA1	Malig. neoplasm breast	86.5%	1.2x
TTN	Primary cardiomyopathy	30.8%	2.5x
MLH1	Colon cancer	43.2%	4.1x
MSH2	Colon cancer	48.0%	2.9x
PKD1	Cystic kidney disease	72.3%	2.1x

Gene	Phenotype	% kept	% Δ
ALPL	ALP	97.8%	1.9%
GOT1	AST	70.6%	37.2%
GPT	ALT	97.8%	1.8%
CST3	Cystatin C	68.9%	46.1%
SLC2A9	Urate	48.3%	63.8%
ABCA1	HDL	38.6%	88.8%
FCGRT	Albumin	32.9%	152.5%
GPLD1	ALP	75.7%	26.3%
ANGPTL3	ApoA	34.9%	75.6%
PCSK9	LDL	53.8%	50.8%
SHBG	SHBG	75.8%	30.9%
ITGA2B	Platelet count	12.8%	186.1%
APOB	LDL	10.4%	454.2%
SEC23B	RDW	11.0%	1331.5%
MC4R	BMI	26.9%	145.5%
IQGAP2	MPV	16.4%	279.0%
LDLR	LDL	2.1% / 49.0%	65.6%
GFI1B	MPV	63.7%	37.4%
GCK	Glucose	3.3% / 60.1%	169.1%
ASGR1	ALP	55.4%	43.6%
GP1BB	MPV	81.5%	18.2%
TUBB1	PDW	46.1%	81.5%
APOB	LDL	51.6%	117.6%
TMPRSS6	MCH	61.5%	85.1%
SEC23B	RDW	84.3%	29.4%

Figure 3. Performance of significant Power Window models in the 117k UKB test set
Performance is shown for (A) binary and (B) quantitative traits. Odds ratio (for binary traits in A) and effect size (for quantitative traits in B) values for gene regions are defined by inclusion in a PW model (PW, pink dot) or exclusion from a PW model (non-PW, blue dot) are plotted against the score for the whole gene (gray cross), together with 95% confidence intervals. Each row is an independently tested gene-phenotype association. Percent of rare variant carriers in the gene that were included in the PW model are indicated. When there is a significant opposite direction of effect within a gene, these are isolated to their own group (opp-PW, green dot) and independently tested, and percent shown is for negative effect/positive effect models. Models are only shown here if there is a significant difference between the PW and non-PW models in the independent test set. Remaining models tested, but failing this criterion, are shown in Figures S6 and S7. ALP, alkaline phosphatase; ALT, alkaline aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; MCH, mean corpuscular hemoglobin; MPV, mean platelet volume; PDW, platelet distribution width; RDW, red blood cell distribution width.

likely to be relevant. The 3DPW method split each protein into roughly the same number of windows as the 1D method, but the coding nucleotide distance covered by each window now ranged from 1 to 11,920 (mean of 1,124), whereas the range was 1–1,396 (mean of 61) for the 1D method. Overall, 59% of the windows included variants that were more than 150 coding nucleotides away from each other, which was only true for 6% of the $\text{PW}_{\text{coding}}$ windows from the 1D model.

After building $\text{3DPW}_{\text{coding}}$ models in the training set, we assessed their performance in the test set and compared them with the PW models built in 1D space. We found that these two methods produced nearly identical models, with 89% of the variants included in the significant $\text{PW}_{\text{coding}}$ model also included in the $\text{3DPW}_{\text{coding}}$ models. In terms of performance in the test set, the models were nearly identical, with similar predictive power (Figure 5). This similarity occurs because while distant amino acids

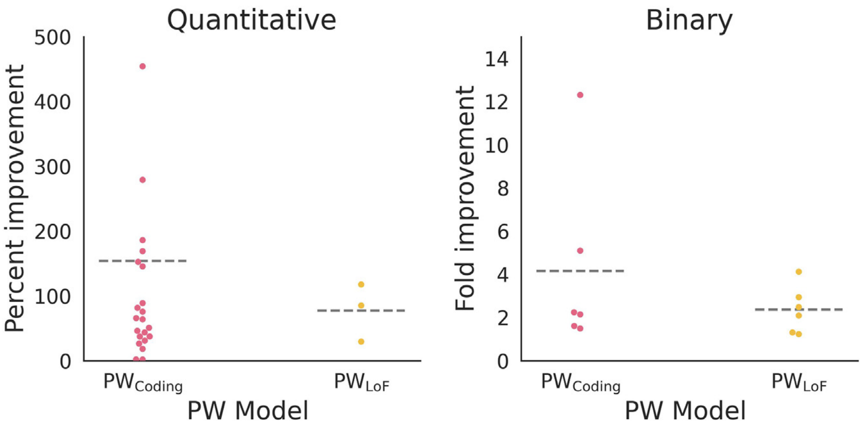


Figure 4. PW effect size improvement for significant models
PW-based improvement for significant gene-trait models measured by percent change in normalized effect size or fold improvement in OR. The results are grouped according to whether they were run with PW_{coding} or PW_{LoF} and whether the phenotype was binary or quantitative. For quantitative traits, the stats given are (PW effect size – whole-gene effect size)/whole-gene effect size. One model with >500% improvement is not shown (1,331% improvement for PW_{coding} for *SEC23B* with red blood cell distribution width). For binary traits, the stats given are the PW OR divided by the whole-gene OR. Gray dashed lines indicate the mean value for that category.

can be grouped together, a substantial proportion of the variants included in 3D windows are also close to each other in 1D space (Figure 5A). For example, 54% of the variants in 3D windows are within 50 coding nucleotides of the central variant, which was also true of 70% of the variants in the 1D windows. There is likely to be more distinction seen between the two methods when a signal is highly localized to a small portion of the protein.

PW successfully isolates gene regions with different directions of effect

Each of the associations that are refined with PW has an underlying story that is rooted in the biology of the gene and the phenotype. One unique refinement that PW can make over a whole-gene model is identifying regions of the gene with opposite directions of effect, especially for quantitative traits. PW models identify opposite directions of effect within the same gene for nine PW_{coding} models and one PW_{LoF} model (Figure S7). However, only two were part of significant PW models in the test set (Figure 3).

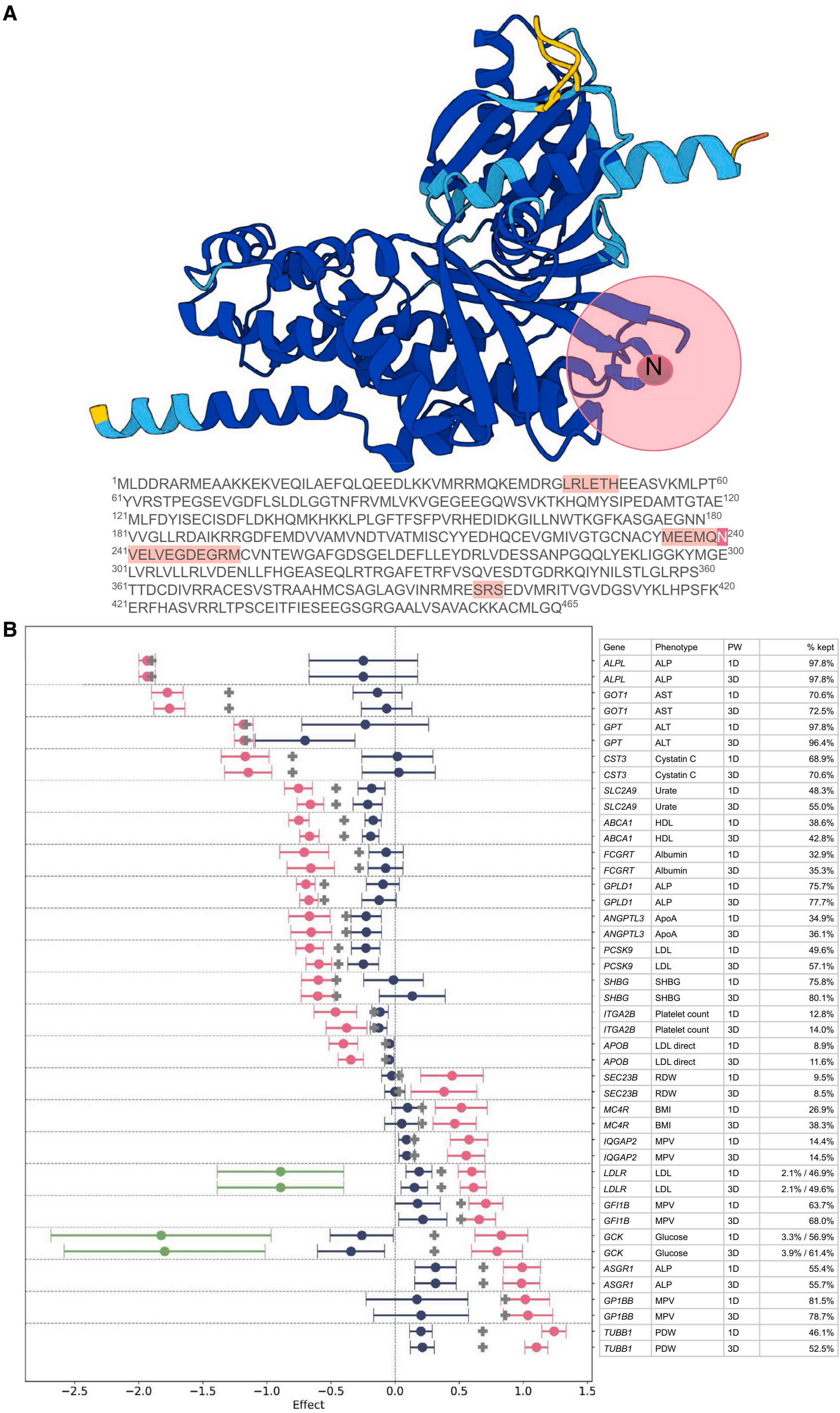
One significant PW_{coding} model with opposite directions of effect is *GCK*, where we identify that non-benign coding variants in half of the second and most of the third exon are associated with low glucose levels, while coding variants in most of the rest of the gene are associated with high glucose levels (Figure 1). Intriguingly, one variant in the low glucose region that is a single variant window (rs373418736), meaning that it is rare (MAF <0.1%) yet common enough to be analyzed separately (n > 20 carriers), shows an opposite direction of effect compared with the rest of the region. The PW methodology of breaking out single variant windows allows these opposing signals within the same regional location to be separated out and analyzed appropriately.

The other significant PW_{coding} model with opposite directions of effect is *LDLR* with LDL levels. Here, we identify that the variant rs377437226 in the cytoplasmic region is associated with lower LDL levels, while non-benign coding variants in most of the rest of the gene, which would be extracellular and thus where LDL binds to this gene product, are associated with higher LDL levels. This is consistent with the mechanism of disease, where the LDL receptor no longer interacts/binds LDL well, leaving cholesterol in the bloodstream, which results in high measured blood biomarker levels.

Discussion

Here, we present a method called Power Window (PW) that successfully identifies the variants and regions within genes that are responsible for the statistical associations found in gene-based collapsing analyses of rare variants. We tested this method on 65 established gene-phenotype associations of rare variants collapsed at the gene level. Our method distinguishes between the variants and gene regions that do and do not impact these traits, identifying statistically significant differences between distinct regions of the gene with a coding model 59% of the time for quantitative traits and 21% of the time for binary traits, as well as 16% of the time for LoF models. The method often dramatically improved the association signals, for example improving the normalized effect size by a median of 65% for significant coding models for quantitative traits when compared with an analysis that collapses across the whole gene. PW also identifies when the entire gene is implicated instead of specific regions, which was true for 41% of PW_{LoF} models and 0 PW_{coding} models. Finally, in some cases, PW even identifies variants and regions with opposite directions of effect in the same gene, in this case for *LDLR* and *GCK*.

PW uses the concept of a sliding window to distinguish gene regions that are associated with a phenotype from those that are not. However, unlike sliding window approaches that move according to a fixed number of bases or variants, PW slides according to the most important metric for rare variant analyses: the statistical power.[9–11] This focus on power, in this case dictated by the number of individuals carrying rare qualifying variants within a given region, is the critical feature for its success and is a key feature of its flexibility. Without this crucial innovation, it is difficult to tell whether regions without statistical



Gene	Phenotype	PW	% kept
ALPL	ALP	1D	97.8%
ALPL	ALP	3D	97.8%
GOT1	AST	1D	70.6%
GOT1	AST	3D	72.5%
GPT	ALT	1D	97.8%
GPT	ALT	3D	96.4%
CST3	Cystatin C	1D	68.9%
CST3	Cystatin C	3D	70.6%
SLC2A9	Urate	1D	48.3%
SLC2A9	Urate	3D	55.0%
ABCA1	HDL	1D	38.6%
ABCA1	HDL	3D	42.8%
FCGRT	Albumin	1D	32.9%
FCGRT	Albumin	3D	35.3%
GPLD1	ALP	1D	75.7%
GPLD1	ALP	3D	77.7%
ANGPTL3	ApoA	1D	34.9%
ANGPTL3	ApoA	3D	36.1%
PCSK9	LDL	1D	49.6%
PCSK9	LDL	3D	57.1%
SHBG	SHBG	1D	75.8%
SHBG	SHBG	3D	80.1%
ITGA2B	Platelet count	1D	12.8%
ITGA2B	Platelet count	3D	14.0%
APOB	LDL direct	1D	8.9%
APOB	LDL direct	3D	11.6%
SEC23B	RDW	1D	9.5%
SEC23B	RDW	3D	8.5%
MC4R	BMI	1D	26.9%
MC4R	BMI	3D	38.3%
IQGAP2	MPV	1D	14.4%
IQGAP2	MPV	3D	14.5%
LDLR	LDL	1D	2.1% / 46.9%
LDLR	LDL	3D	2.1% / 49.6%
GFI1B	MPV	1D	63.7%
GFI1B	MPV	3D	68.0%
GCK	Glucose	1D	3.3% / 56.9%
GCK	Glucose	3D	3.9% / 61.4%
ASGR1	ALP	1D	55.4%
ASGR1	ALP	3D	55.7%
GP1BB	MPV	1D	81.5%
GP1BB	MPV	3D	78.7%
TUBB1	PDW	1D	46.1%
TUBB1	PDW	3D	52.5%

Figure 5. 3D Power Window

(A) Protein structure of *GCK* as predicted by Alphafold. Colors on the 3D structure indicate level of Alphafold confidence, with dark blue indicating very high (pLDDT >90), light blue confident (90 > pLDDT >70), yellow low (70 > pLDDT >50), and red very low (pLDDT <50). For a hypothetical window, the highlighted amino acid 240N in pink is the center of the example Window. The window is built spreading out in 3D space according to the structure of the protein. The transparent pink circle indicates the boundaries for the analyzed

(legend continued on next page)

associations are truly not associated or simply lack power. Performing an analysis where all the windows of the gene have the same statistical power allows you to rank the parts of the gene according to their associations in a manner that is unbiased by variant frequency.

Performing analyses with PW is now feasible with the availability of very large sample sizes and well represented phenotypes from population studies and biobanks and will continue to improve as sample sizes increase. Being able to break a gene with an established association down into windows and still have a reasonable number of carriers of rare genetic variants—which are themselves rare events—in each window requires sample sizes that until now had been unreasonable, and which remain unreasonable for many genes. For example, *MIP* (MIM: 154050), which has a significant association with cataracts, had 238 carriers of qualifying coding variants in the set of 350k UKB training exomes. A window size of 20 carriers produces only 55 windows for this gene, making it much less practical to study than a gene like *TTN*, where 1,362 windows were made for the LoF model.

PW is able to extend known LoF associations to coding variants by identifying the regions of the gene in which coding variants have an effect similar to that of LoF variants. This is important because the complexity of interpreting novel coding variants has remained a difficult problem in human genetics. For example, ACMG guidelines allow novel LoF variants to be considered pathogenic in a gene where other LoF variants are already established as pathogenic, whereas each individual non-LoF variant requires, for example, evidence of association in other individuals and functional studies to support any pathogenicity assertations.[38] PW uses statistical evidence to distinguish coding variants, even novel ones, that have an impact similar to those of LoF variants. As an example, LoF variants in *LDLR* are considered pathogenic for high endogenous LDL.[39] Indeed, our $\mathrm{PW_{LoF}}$ model confirms that LoF variants anywhere in this gene are associated with higher LDL levels (effect size 0.56; Figure S7B). However, $\mathrm{PW_{coding}}$ identifies that 49% of the carriers of rare non-benign coding variants in this gene are also associated with higher LDL levels, with a similar effect size (effect size 0.6, Figure 3). In this way, PW can uncover latent non-LoF signals from regions of a gene that might otherwise go unobserved as well as help classify variants of unknown significance (VUS; Figure S10). This extension of known biology to additional types of variants in the gene increases the number of people who would benefit from genetic screening and also improves our understanding of gene function.

PW was able to identify when variants in different gene regions could produce significantly different directions of effect on a phenotype. Among these was the association between *GCK* and glucose levels. It is known that depending on the variant effect on overall glucose metabolism, mutations can lead to either hyper- or hypoglycemia.[40] We find both of these effects, concurrently, with the PW analysis, highlighting the regions of the gene structurally and functionally associated with each resulting phenotype. While hypoglycemic variants have been reported throughout various regions of *GCK*, we find most variants are concentrated toward the 5′ end, some of which are very well described while others may help to further our mechanistic understanding of *GCK* and thus improve the scope of variant interpretation possible at this locus.[40] This finding is also supported by a recent deep scan mutagenesis study, which found that 5′-end mutations were more likely to lead to lower glucose levels.[41]

PW works well both when models are built according to the linear coding sequence of the gene (through 1D space) and when they are built according to the final structure of the folded protein (through 3D space). In the gene-phenotype combinations studied here, the models built from these two approaches were largely identical despite the analyzing windows containing very different sets of variants. Most of the genes analyzed here are ones where large swaths of the gene show an association between rare variants and the phenotype, so that the same areas of the gene were able to be implicated in both types of models. However, it is reasonable to postulate that genes where only a very specific portion of the protein is involved may show different signals in 1D and 3D models.

While the theoretical basis for the PW method is a substantial advancement for identifying the specific regions of genes in which rare variants are established to be associated with traits, there are still many refinements to the methodology that will be beneficial for future studies. One aspect is that the method requires a balance between zooming in on specific regions of the gene (requiring a small carrier sample size cutoff) and obtaining statistical significance for the regions (requiring a larger carrier sample size cutoff). Future studies with even larger sample sizes may reclassify some portions of the genes tested here, identifying some regions we erroneously excluded or identifying new regions to include. The presented method also relies on the bioinformatic prediction of which missense variants are damaging to filter those that qualify for consideration in the model, although we found similar results with different bioinformatic tools (Figure S11).[42,43]

Window. In the linear amino acid sequence shown below, 240N is shown in pink, and the amino acids where variants could be included in this window are in light pink.
(B) As in Figure 3, PW models and 95% confidence intervals are shown in the test set of 117k individuals for significant $\mathrm{PW_{coding}}$ models. Each gene-phenotype pair is shown twice, for the 1D and 3DPW models. The model results are very similar in terms of percent of carriers kept in the model and final effect sizes. The percentage to the right of the phenotype is the percent carriers retained in the model; when there is both a negative and positive model, the percent carriers retained in the negative model is shown first. For binary traits, see Figure S9. ALP, alkaline phosphatase; ALT, alkaline aminotransferase; AST, aspartate aminotransferase; BMI, body mass index MCH, mean corpuscular hemoglobin; MPV, mean platelet volume; PDW, platelet distribution width; RDW, red blood cell distribution width.

Other improvements will involve focusing on certain classes of variants in the gene, such as those computationally predicted to be gain or loss of function, those with functional data from screenings, those in transcripts expressed in tissues of interest, and different protein conformations.[41,44,45]

The PW method is centered on statistics and can accommodate all classes of genetic variation to fine-tune rare variant signals and gene-based results. We believe the PW technique can have an immediate impact on human genomics research, drug development, variant interpretation, and precision medicine. As displayed in this work, PW can improve the specificity of variant interpretation for rare variants in established gene-level disease associations through regional refinement of the association. It can also expand variant classification capabilities by combining PW analyses derived from the same phenotype with carriers defined by both LoF and coding variant models. In both cases, with more precise or complete genomic definitions to identify relevant carriers, phenotypic penetrance estimates among carriers are likely to improve—a key metric for the practice of precision medicine. We also anticipate that PW will be a powerful tool for discovery, as it will be able to characterize variant patterns in a locus to better understand the mechanism of disease, improve knowledge of gene function to predict new drug targets, and in some cases may even identify sub-gene-disease associations that are drowned out when rare variants are collapsed at the gene level.

Data and code availability

Statistics relating to the Power Window analysis for all included genes, calculated using the UKB 350k discovery cohort, are available in Table S3. UKB data are available for download (https://www.ukbiobank.ac.uk/) to qualified researchers. The Helix data are available to qualified researchers upon reasonable request and with permission of the Helix Steering Committee and Helix. Researchers who would like to obtain the raw genotype data related to this study will be presented with a Data Use Agreement, which requires that participants will not be reidentified and no data will be shared between individuals, third parties, or uploaded onto public domains. Helix encourages collaboration with scientific researchers on an individual basis. Examples of restrictions that will be considered in requests to data access include but are not limited to (1) whether the request comes from an academic institution in good standing and will collaborate with our team to protect the privacy of the participants and the security of the data requested; (2) type and amount of data requested; (3) feasibility of the research suggested; and (4) amount of resource allocation for Helix and member institutions required to support a collaboration. The code to calculate regions to use in Power Window analysis is available at https://github.com/ecirulli/PowerWindow/.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2024.100284.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 40436. This work uses data provided by patients and collected by the NHS as part of their care and support, re-used with the permission of the UK Biobank under Application Number 40436. Copyright © (2023), NHS England. All rights reserved. This research also used data assets made available by National Safe Haven as part of the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (grant ref MC_PC_20058). Funding was provided to DRI by the Nevada Governor's Office of Economic Development. Funding was provided to the Renown Institute for Health Innovation by Renown Health and the Renown Health Foundation. Funding was provided to the myGenetics program by HealthPartners. We acknowledge the entire Helix bioinformatics and lab teams for their contributions to the production of exome sequencing pipeline as well as the research administration team for coordinating the project. We thank all of the participants of the Healthy Nevada Project, myGenetics, and In Our DNA SC.

Declaration of interests

K.M.S.B., E.T.C., A.B., W.L., and N.L.W. are all employees of Helix. A patent has been filed by Helix for the Power Window analysis technique with E.T.C., K.M.S.B., and N.L.W. as inventors, and its current status is unpublished (application number 17575894).

Received: November 16, 2023
Accepted: March 15, 2024

Web resources

https://www.ukbiobank.ac.uk/
https://github.com/ecirulli/PowerWindow/

References

1. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415–425.
2. Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A.S., and Goldstein, D.B. (2019). Rare-variant collapsing analyses for complex traits: guidelines and applications. Nat. Rev. Genet. *20*, 747–759.
3. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Aslibekyan, S., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat. Genet. *52*, 969–983.
4. Li, Z., Li, X., Zhou, H., Gaynor, S.M., Selvaraj, M.S., Arapoglou, T., Quick, C., Liu, Y., Chen, H., Sun, R., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. Nat. Methods *19*, 1599–1611.

5. Li, X., Quick, C., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Selvaraj, M.S., Sun, R., Dey, R., Arnett, D.K., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. Nat. Genet. *55*, 154–164.
6. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
7. Gussow, A.B., Petrovski, S., Wang, Q., Allen, A.S., and Goldstein, D.B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. Genome Biol. *17*, 9.
8. Cirulli, E.T., White, S., Read, R.W., Elhanan, G., Metcalf, W.J., Tanudjaja, F., Fath, D.M., Sandoval, E., Isaksson, M., Schlauch, K.A., et al. (2020). Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat. Commun. *11*, 542.
9. Li, Z., Li, X., Liu, Y., Shen, J., Chen, H., Zhou, H., Morrison, A.C., Boerwinkle, E., and Lin, X. (2019). Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. Am. J. Hum. Genet. *104*, 802–814.
10. Bocher, O., Ludwig, T.E., Oglobinsky, M.-S., Marenne, G., Deleuze, J.-F., Suryakant, S., Odeberg, J., Morange, P.-E., Trégouët, D.A., Perdry, H., and Génin, E. (2022). Testing for association with rare variants in the coding and non-coding genome: RAVA-FIRST, a new approach based on CADD deleteriousness score. PLoS Genet. *18*, e1009923.
11. (2022). STAARpipeline: an all-in-one rare-variant tool for biobank-scale whole-genome sequencing data. Nat. Methods *19*, 1532–1533.
12. Ionita-Laza, I., Makarov, V., ARRA Autism Sequencing Consortium, and Buxbaum, J.D. (2012). Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. Am. J. Hum. Genet. *90*, 1002–1013.
13. McCallum, K.J., and Ionita-Laza, I. (2015). Empirical Bayes scan statistics for detecting clusters of disease risk variants in genetic studies. Biometrics *71*, 1111–1120.
14. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature *599*, 628–634.
15. astheeggeggs astheeggeggs/PHESANT (2017). GitHub. https://github.com/astheeggeggs/PHESANT.
16. Millard, L., Davies, N.M., Gaunt, T., Smith, G.D., and Tilling, K. (2017). PHESANT: a tool for performing automated phenome scans in UK Biobank. https://doi.org/10.1101/111500.
17. Schiabor Barrett, K.M., Bolze, A., Ni, Y., White, S., Isaksson, M., Sharma, L., Levin, E., Lee, W., Grzymski, J.J., Lu, J.T., et al. (2021). Positive predictive value highlights four novel candidates for actionable genetic screening from analysis of 220,000 clinicogenomic records. Genet. Med. *23*, 2300–2308.
18. Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J.C., et al. (2019). Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Med. Inform. *7*, e14325.
19. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. *31*, 1102–1110.
20. Bastarache, L., Hughey, J.J., Hebbring, S., Marlo, J., Zhao, W., Ho, W.T., Van Driest, S.L., McGregor, T.L., Mosley, J.D., Wells, Q.S., et al. (2018). Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science *359*, 1233–1239.
21. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.
22. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.
23. Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res. *46*, D754–D761.
24. (2022). HailTeamHail 0.2.21-f16fd64e0d77 (Github). https://github.com/hail-is/hail.
25. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
26. Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. *40*, W452–W457.
27. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
28. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489.
29. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2020). Computationally efficient whole genome regression for quantitative and binary traits. Cold Spring Harbor Lab. *53*, 1097–1103. https://doi.org/10.1101/2020.06.19.162354.
30. Schiabor Barrett, K.M., Cirulli, E.T., Bolze, A., Rowan, C., Elhanan, G., Grzymski, J.J., Lee, W., and Washington, N.L. (2023). Cardiomyopathy prevalence exceeds 30% in individuals with TTN variants and early atrial fibrillation. Genet. Med. *25*, 100012.
31. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.
32. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444.
33. (2022). Entry - *147796 - JANUS KINASE 2; JAK2 - OMIM. https://omim.org/entry/147796.
34. Wiriyasermkul, P., Moriyama, S., and Nagamori, S. (2020). Membrane transport proteins in melanosomes: Regulation of ions for pigmentation. Biochim. Biophys. Acta. Biomembr. *1862*, 183318.

35. Möröy, T., Vassen, L., Wilkes, B., and Khandanpour, C. (2015). From cytopenia to leukemia: the role of Gfi1 and Gfi1b in blood formation. Blood *126*, 2561–2569.

36. Polfus, L.M., Khajuria, R.K., Schick, U.M., Pankratz, N., Pazoki, R., Brody, J.A., Chen, M.-H., Auer, P.L., Floyd, J.S., Huang, J., et al. (2016). Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative GFI1B Splice Variants in Human Hematopoiesis. Am. J. Hum. Genet. *99*, 785.

37. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

38. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.

39. Miller, D.T., Lee, K., Chung, W.K., Gordon, A.S., Herman, G.E., Klein, T.E., Stewart, D.R., Amendola, L.M., Adelman, K., Bale, S.J., et al. (2021). ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. *23*, 1381–1390.

40. Osbak, K.K., Colclough, K., Saint-Martin, C., Beer, N.L., Bellanné-Chantelot, C., Ellard, S., and Gloyn, A.L. (2009). Update on mutations in glucokinase (GCK), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. Hum. Mutat. *30*, 1512–1526.

41. Gersing, S., Cagiada, M., Gebbia, M., Gjesing, A.P., Coté, A.G., Seesankar, G., Li, R., Tabet, D., Stein, A., Gloyn, A.L., et al. (2023). A comprehensive map of human glucokinase variant activity. Genome Biol. *24*, 97. https://doi.org/10.1101/2022.05.04.490571.

42. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. *99*, 877–885.

43. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492.

44. Schiabor Barrett, K.M., Cirulli, E.T., Bolze, A., Rowan, C., Elhanan, G., Grzymski, J.J., Lee, W., and Washington, N.L. (2022). TTN truncating variants in hiPSI exons show high penetrance for cardiomyopathy in carriers with atrial fibrillation. Preprint at bioRxiv. https://doi.org/10.1101/2022.06.06.22276058.

45. Stein, D., Bayrak, Ç.S., Wu, Y., Stenson, P.D., Cooper, D.N., Schlessinger, A., and Itan, Y. (2022). Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of diverse feature set. Preprint at bioRxiv. https://doi.org/10.1101/2022.06.08.495288.

HGGA, Volume *5*

Supplemental information

A power-based sliding window approach to evaluate the clinical impact of rare genetic variants in the nucleotide sequence or the spatial position of the folded protein

Elizabeth T. Cirulli, Kelly M. Schiabor Barrett, Alexandre Bolze, Daniel P. Judge, Pamala A. Pawloski, Joseph J. Grzymski, William Lee, and Nicole L. Washington

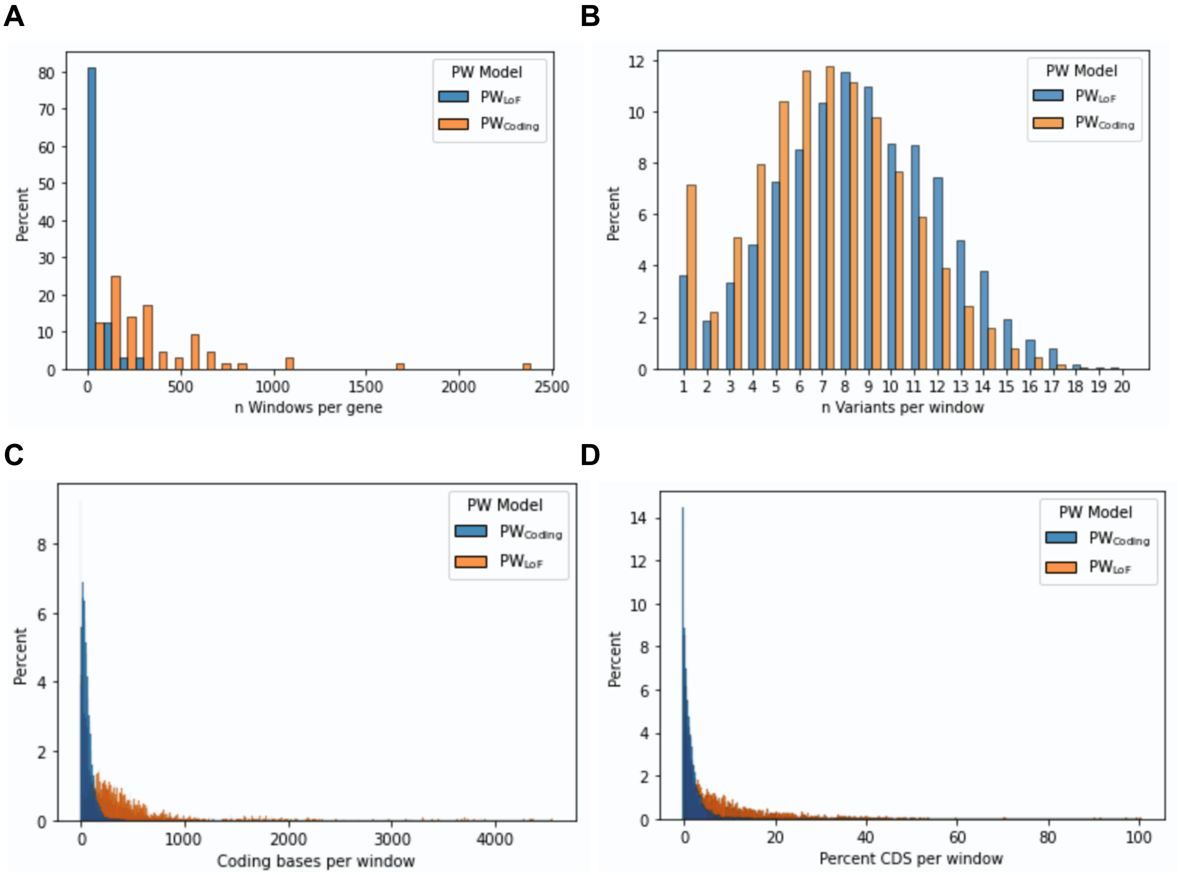
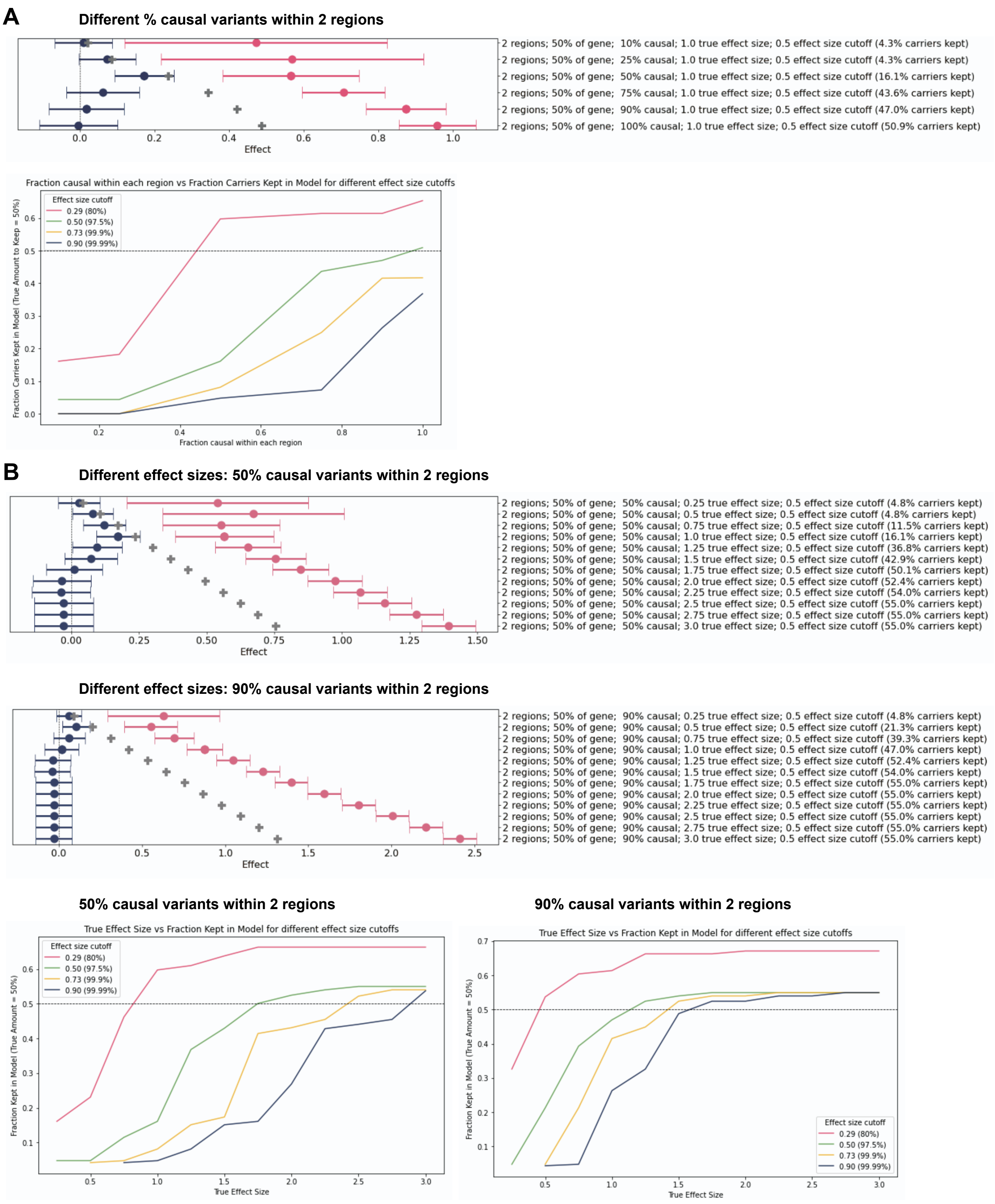
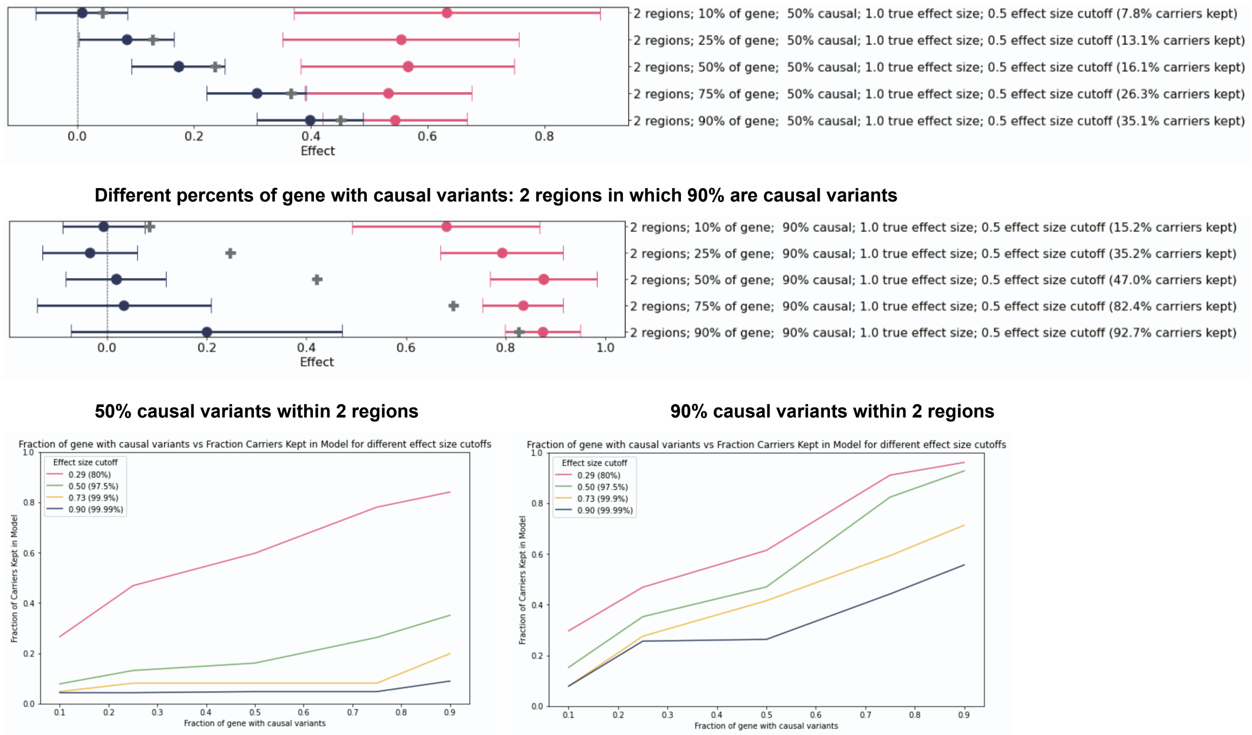


Figure S1. Stats for power windows. Because of the size of *TTN*, the stats for this gene are not displayed and instead are written separately in the legend. A) Histogram of number of windows analyzed per gene. The mean number of windows analyzed per gene was 357 / 67 for PW$_{Coding}$ / PW$_{LoF}$ (range 30-2385 and 2-302). *TTN* = 10,034 / 1362 for PW$_{Coding}$ / PW$_{LoF}$. B) Histogram of number of variants included in each window. The mean number of variants analyzed per window was 7 / 8 for PW$_{Coding}$ / PW$_{LoF}$ (range 1-20 and 1-20). *TTN* = 7 (1-18) / 11 (1-18) for PW$_{Coding}$ / PW$_{LoF}$. C) Histogram of number of coding bases included in each window; when a window only included one variant or only included variants at the same site, then the length is 1. The mean number of coding bases included per window was 61 / 404 for PW$_{Coding}$ / PW$_{LoF}$ (range 1-1396 and 1-4549). *TTN* = 74 (1-2847) / 943 (1-2632) for PW$_{Coding}$ / PW$_{LoF}$. D) Histogram of the percent of each CDS for the gene that was included in each window. The mean percent of the CDS of the gene that was analyzed in each window was 2% / 10% for PW$_{Coding}$ / PW$_{LoF}$ (range <1-36% and <1-100%). *TTN* = 0.07% (0.001%-2.6%) / 0.87% (0.001%-2.4%) for PW$_{Coding}$ / PW$_{LoF}$.


A
Different % causal variants within 2 regions
2 regions; 50% of gene; 10% causal; 1.0 true effect size; 0.5 effect size cutoff (4.3% carriers kept)
2 regions; 50% of gene; 25% causal; 1.0 true effect size; 0.5 effect size cutoff (4.3% carriers kept)
2 regions; 50% of gene; 50% causal; 1.0 true effect size; 0.5 effect size cutoff (16.1% carriers kept)
2 regions; 50% of gene; 75% causal; 1.0 true effect size; 0.5 effect size cutoff (43.6% carriers kept)
2 regions; 50% of gene; 90% causal; 1.0 true effect size; 0.5 effect size cutoff (47.0% carriers kept)
2 regions; 50% of gene; 100% causal; 1.0 true effect size; 0.5 effect size cutoff (50.9% carriers kept)
0.0
0.2
0.4
0.6
0.8
1.0
Effect
Fraction causal within each region vs Fraction Carriers Kept in Model for different effect size cutoffs
Effect size cutoff
0.29 (80%)
0.50 (97.5%)
0.73 (99.9%)
0.90 (99.99%)
Fraction Carriers Kept in Model (True Amount to Keep = 50%)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.2
0.4
0.6
0.8
1.0
Fraction causal within each region
B
Different effect sizes: 50% causal variants within 2 regions
2 regions; 50% of gene; 50% causal; 0.25 true effect size; 0.5 effect size cutoff (4.8% carriers kept)
2 regions; 50% of gene; 50% causal; 0.5 true effect size; 0.5 effect size cutoff (4.8% carriers kept)
2 regions; 50% of gene; 50% causal; 0.75 true effect size; 0.5 effect size cutoff (11.5% carriers kept)
2 regions; 50% of gene; 50% causal; 1.0 true effect size; 0.5 effect size cutoff (16.1% carriers kept)
2 regions; 50% of gene; 50% causal; 1.25 true effect size; 0.5 effect size cutoff (36.8% carriers kept)
2 regions; 50% of gene; 50% causal; 1.5 true effect size; 0.5 effect size cutoff (42.9% carriers kept)
2 regions; 50% of gene; 50% causal; 1.75 true effect size; 0.5 effect size cutoff (50.1% carriers kept)
2 regions; 50% of gene; 50% causal; 2.0 true effect size; 0.5 effect size cutoff (52.4% carriers kept)
2 regions; 50% of gene; 50% causal; 2.25 true effect size; 0.5 effect size cutoff (54.0% carriers kept)
2 regions; 50% of gene; 50% causal; 2.5 true effect size; 0.5 effect size cutoff (55.0% carriers kept)
2 regions; 50% of gene; 50% causal; 2.75 true effect size; 0.5 effect size cutoff (55.0% carriers kept)
2 regions; 50% of gene; 50% causal; 3.0 true effect size; 0.5 effect size cutoff (55.0% carriers kept)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
Effect
Different effect sizes: 90% causal variants within 2 regions
2 regions; 50% of gene; 90% causal; 0.25 true effect size; 0.5 effect size cutoff (4.8% carriers kept)
2 regions; 50% of gene; 90% causal; 0.5 true effect size; 0.5 effect size cutoff (21.3% carriers kept)
2 regions; 50% of gene; 90% causal; 0.75 true effect size; 0.5 effect size cutoff (39.3% carriers kept)
2 regions; 50% of gene; 90% causal; 1.0 true effect size; 0.5 effect size cutoff (47.0% carriers kept)
2 regions; 50% of gene; 90% causal; 1.25 true effect size; 0.5 effect size cutoff (52.4% carriers kept)
2 regions; 50% of gene; 90% causal; 1.5 true effect size; 0.5 effect size cutoff (54.0% carriers kept)
2 regions; 50% of gene; 90% causal; 1.75 true effect size; 0.5 effect size cutoff (55.0% carriers kept)
2 regions; 50% of gene; 90% causal; 2.0 true effect size; 0.5 effect size cutoff (55.0% carriers kept)
2 regions; 50% of gene; 90% causal; 2.25 true effect size; 0.5 effect size cutoff (55.0% carriers kept)
2 regions; 50% of gene; 90% causal; 2.5 true effect size; 0.5 effect size cutoff (55.0% carriers kept)
2 regions; 50% of gene; 90% causal; 2.75 true effect size; 0.5 effect size cutoff (55.0% carriers kept)
2 regions; 50% of gene; 90% causal; 3.0 true effect size; 0.5 effect size cutoff (55.0% carriers kept)
0.0
0.5
1.0
1.5
2.0
2.5
Effect
50% causal variants within 2 regions
True Effect Size vs Fraction Kept in Model for different effect size cutoffs
Effect size cutoff
0.29 (80%)
0.50 (97.5%)
0.73 (99.9%)
0.90 (99.99%)
Fraction Kept in Model (True Amount = 50%)
0.1
0.2
0.3
0.4
0.5
0.6
0.5
1.0
1.5
2.0
2.5
3.0
True Effect Size
90% causal variants within 2 regions
True Effect Size vs Fraction Kept in Model for different effect size cutoffs
Fraction Kept in Model (True Amount = 50%)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Effect size cutoff
0.29 (80%)
0.50 (97.5%)
0.73 (99.9%)
0.90 (99.99%)
0.5
1.0
1.5
2.0
2.5
3.0
True Effect Size

C

Different percents of gene with causal variants: 2 regions in which 50% are causal variant



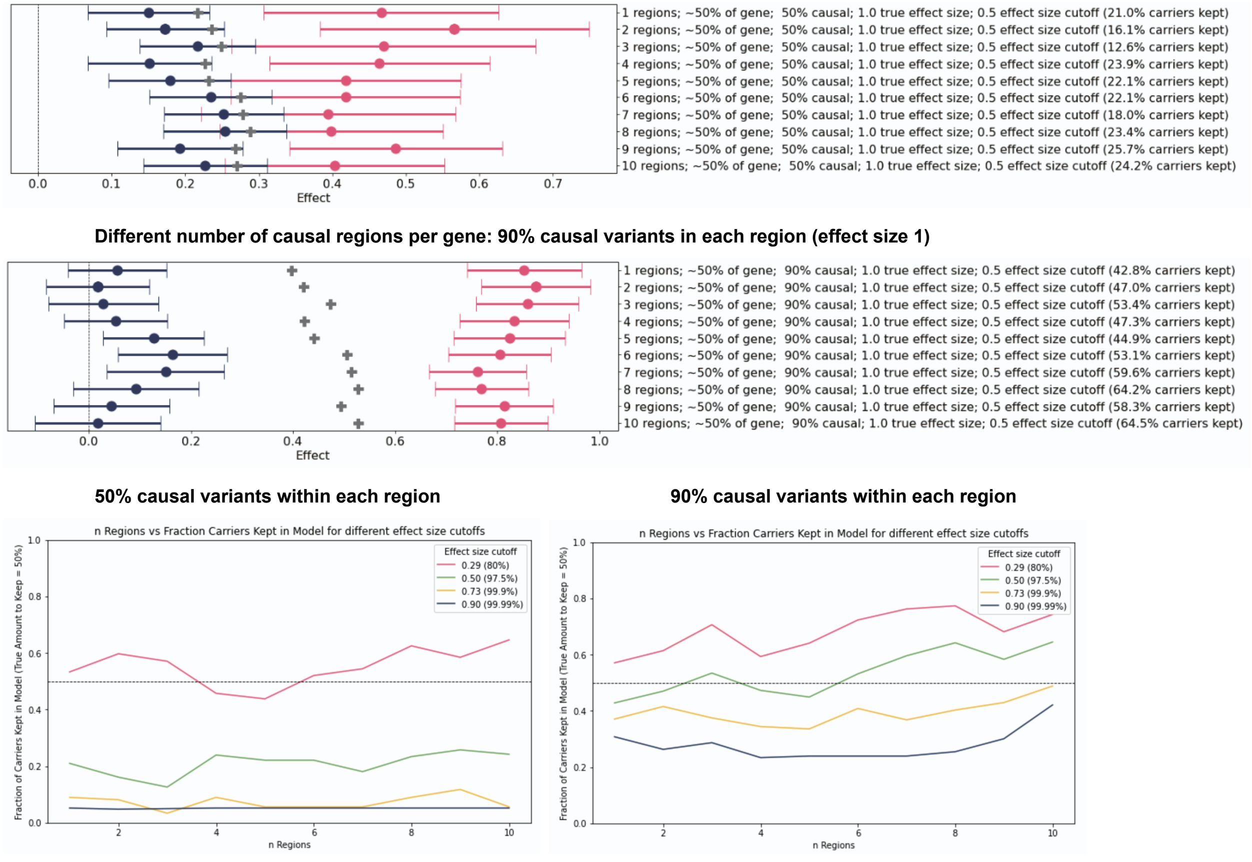
Different percents of gene with causal variants: 2 regions in which 90% are causal variants

50% causal variants within 2 regions

90% causal variants within 2 regions

D

Different number of causal regions per gene: 50% causal variants in each region (effect size 1)



Different number of causal regions per gene: 90% causal variants in each region (effect size 1)

50% causal variants within each region

90% causal variants within each region

Figure S2. Simulations showing the effect of tuning different parameters for Power Window in the example gene *GCK* and phenotype glucose with a coding model. For the simulations, causal regions were defined within *GCK* and glucose levels for each carrier were reassigned according to the parameters specified. The figures show the model performance in the independent replication set of 117k individuals, as in **Figure 3**. Effect size values for gene regions are defined by inclusion in a PW model (PW, pink dot) or exclusion from a PW model (non-PW, blue dot) are plotted against the score for the whole gene (grey cross), together with 95% confidence intervals. Each row has different parameters, as labeled, with the effect size cutoff for inclusion into the model set at 0.5 just like in the main manuscript. All models shown are significant at p<0.0005 in the test set unless otherwise specified. Each panel also includes a plot showing the relationship between the specified parameter and the percent of carriers kept in the model, with different effect size cutoffs for inclusion into the model (rejecting an effect size of 0 with 80% confidence, 97.5% (as in the main manuscript), 99.9%, and 99.99%). For all panels, "percent causal variants" is equivalent to "of the individuals with variants in the region, the percent whose variant is causal", i.e., putting asside that one variant may be seen in 1 individual and another in 5 individuals. **A)** Parameter changed: percent causal variants within associated regions of the gene. The simulation was set with 2 implicated regions taking up 50% of the gene in which causal variants had an effect size of 1, and the percent causal variants within those regions, ranged from 0% to 100%. To capture all of the associated regions with a cutoff of 0.5 when the effect size is 1, at least 75% of the variants in implicated regions must be causal. The only non-significant model was the one with 10% of the variants in the region being causal (p=0.01). **B)** Parameter changed: effect size of causal variants. The simulation was set with 2 implicated regions taking up 50% of the gene, in which either 50% or 90% of the variants were causal, and the effect size of the causal variants ranged from 0.25 to 3. To capture all of the associated regions with a cutoff of 0.5 when only 50% of the variants in implicated regions are truly associated, the true effect size of causal variants must be at least ~1.5; if 90% of the variants are truly associated, then the effect size of the causal variants must be at least ~1. The only non-significant models were the ones with an effect size of 0.25. **C)** Parameter changed: percent of gene implicated. The simulation was set with 2 implicated regions in which either 50% or 90% of the variants were causal with effect sizes of 1, with the percent of the gene implicated ranging from 10% to 90%. When 50% of the variants in the regions were causal, PW generally missed approximately half of the regions; when 90% of the variants in the regions were causal, PW identified the correct regions but also sometimes included some additional regions. **D)** Parameter changed: number of implicated regions. The simulation was set with the implicated regions taking up ~50% of the gene (note the actual percentages varied from 46-54% due to variations in numbers of variants carriers in different regions), in which either 50% or 90% of the variants were causal with an effect size of 1, and the number of regions ranged from 1 to 10. When 50% of the variants in the regions were causal, PW performed similarly regardless of the number of regions but missed more carriers when there were fewer regions; when 90% of the variants in the regions were causal, PW correctly captured them when they were split into 1 or 2 regions but allowed in more carriers than should have been included as the number of implicated regions in the gene grew.

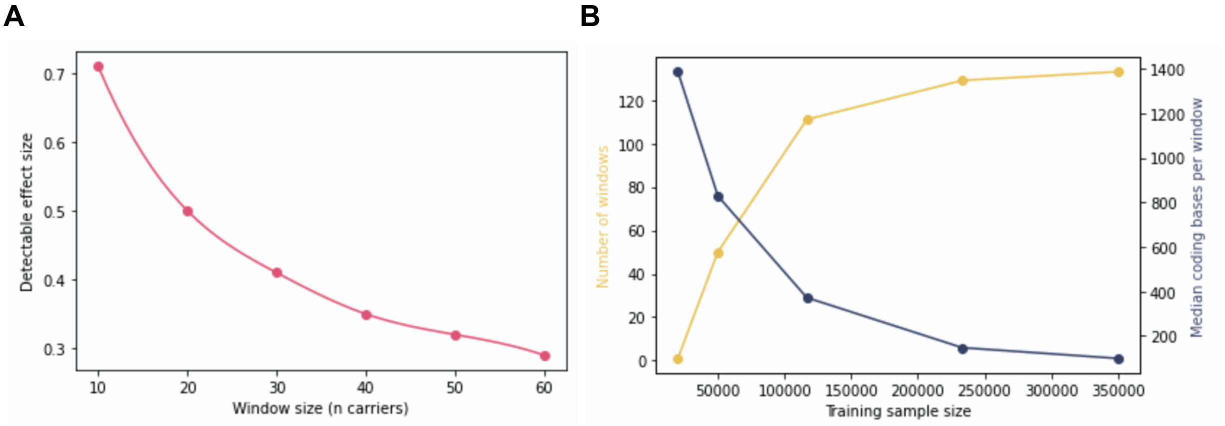


Figure S3. Properties of Power Window size. **A)** For a quantitative trait in a training set of 350k samples, the window size (number of carriers in the window) is compared to the effect size for which there is power to identify (with 97.5% confidence) that the beta is not 0. **B)** For the example gene *GCK* with a coding model (which in our main model had 133 windows, with a mean window length of 99 coding bases), the training sample size is compared to the physical length of each window in terms of coding position, for a set window size of 20 carriers per window. A larger sample size results in being able to home in on more specific regions for analysis.

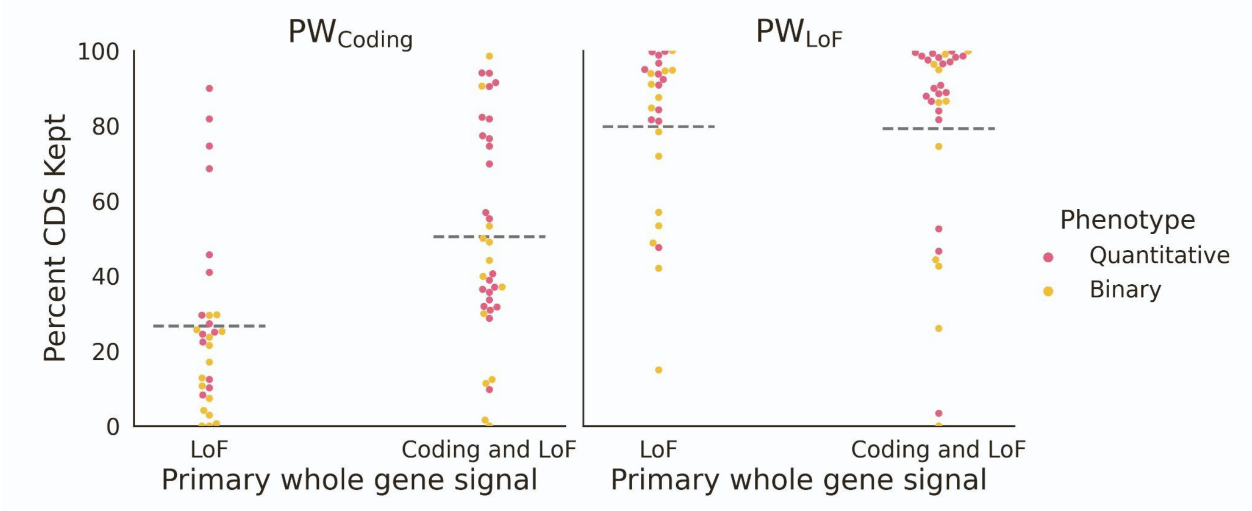
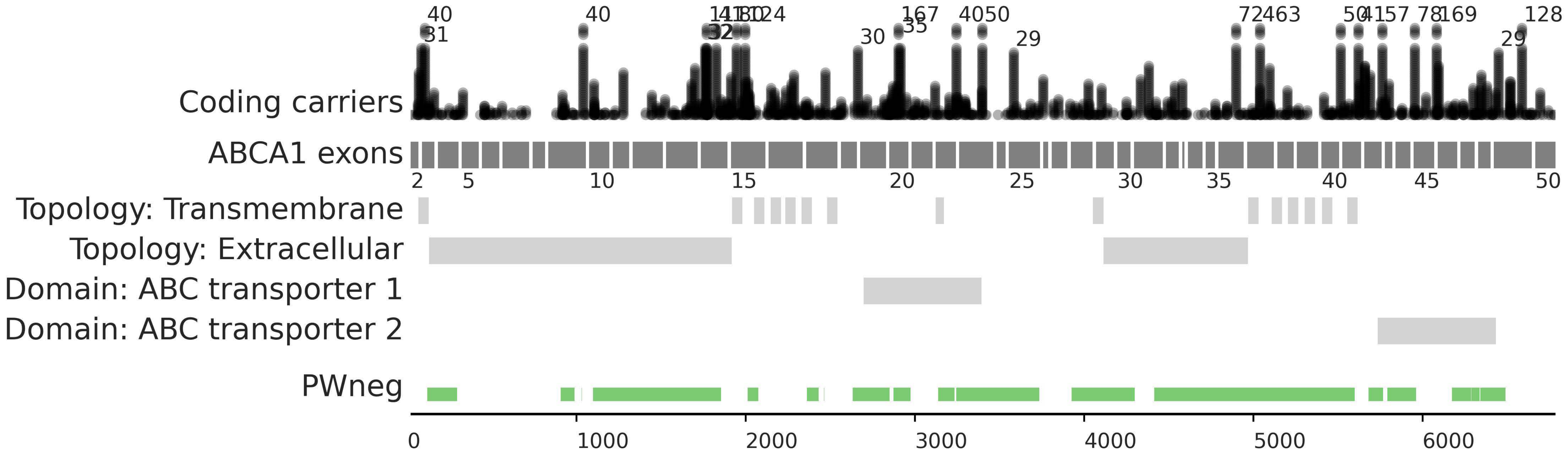
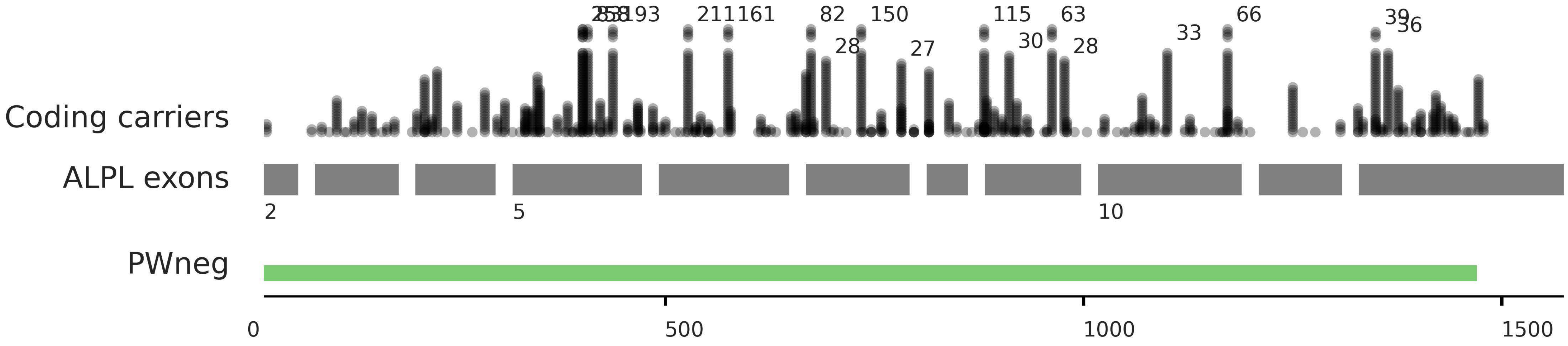
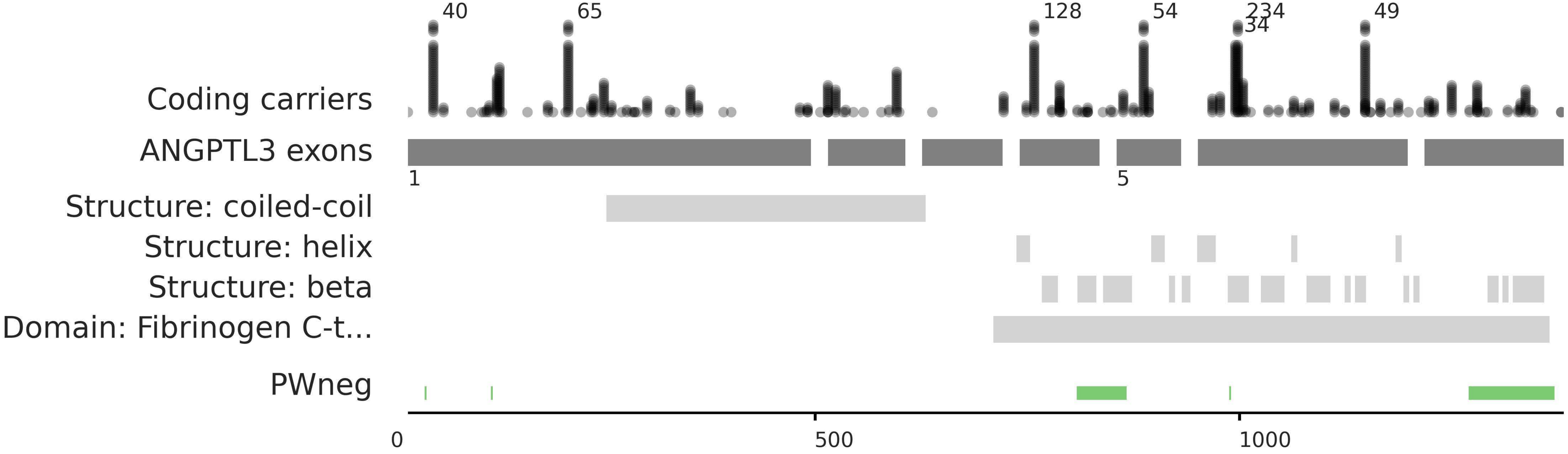
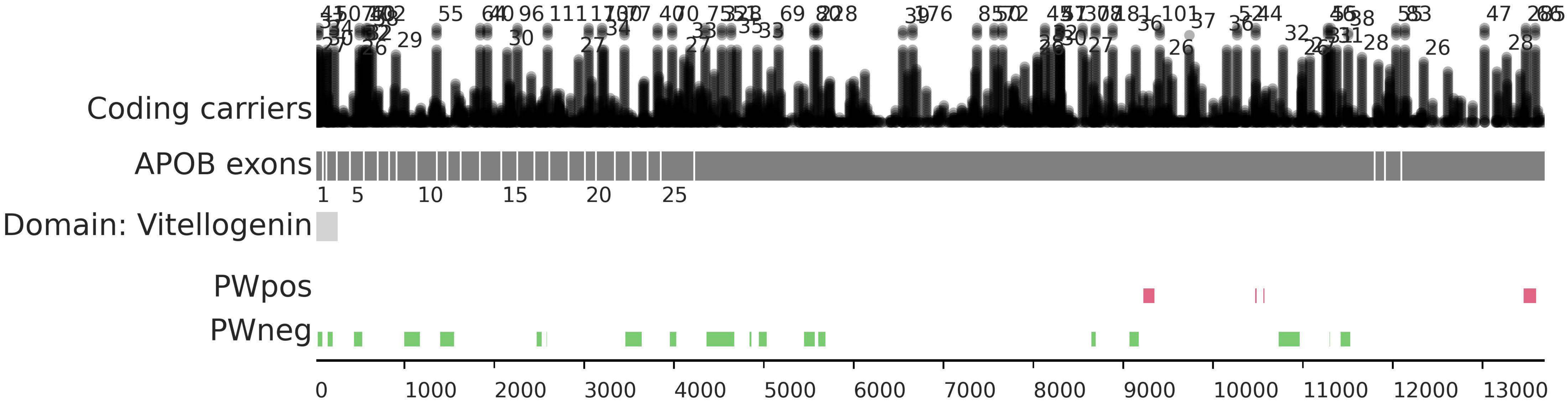


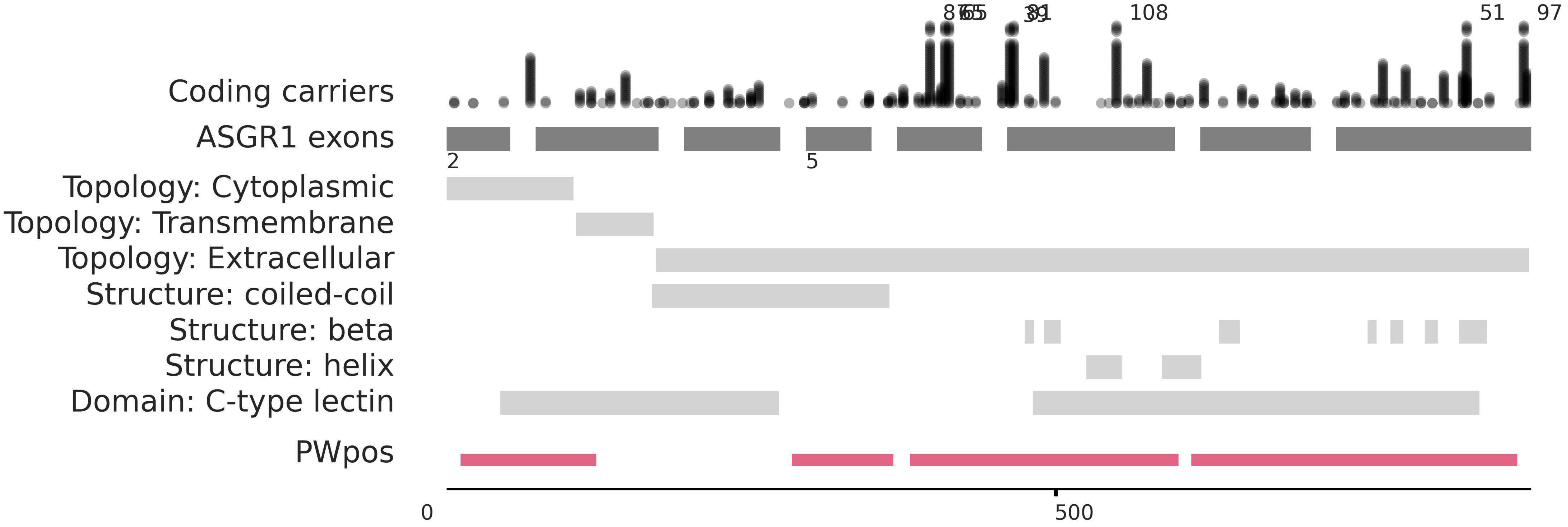
Figure S4. Percent CDS kept in Power Window models. For each gene (n=65), the percent of the CDS for the gene that was retained by the PW model was evaluated for (A) PW_{Coding} and (B) PW_{LoF}. Within each model, the phenotypes were grouped into quantitative (pink) or binary (yellow) traits, and the genes were grouped based on the original genome-wide association as follows: **LoF**: original whole-gene associations had an absolute beta at least 3x as high in the LoF model as the Coding model; **Coding and LoF**: original whole-gene associations had a beta <3x as high for LoF as for Coding (no gene-phenotype combinations had a whole-gene Coding model absolute beta that was at least 3x higher than the whole-gene LoF beta). Two genes had 0 windows included in the final model for PW_{Coding} (*ASXL1* and *NF1*), and 2 for PW_{LoF} (*GFI1B* and *JAK2*). Dotted horizontal line indicates the median percent carriers kept.


40
31
40
14110124
32
30
167
35
4050
29
72463
504157 78169
29
128
Coding carriers
ABCA1 exons
2
5
10
15
20
25
30
35
40
45
50
Topology: Transmembrane
Topology: Extracellular
Domain: ABC transporter 1
Domain: ABC transporter 2
PWneg
0
1000
2000
3000
4000
5000
6000

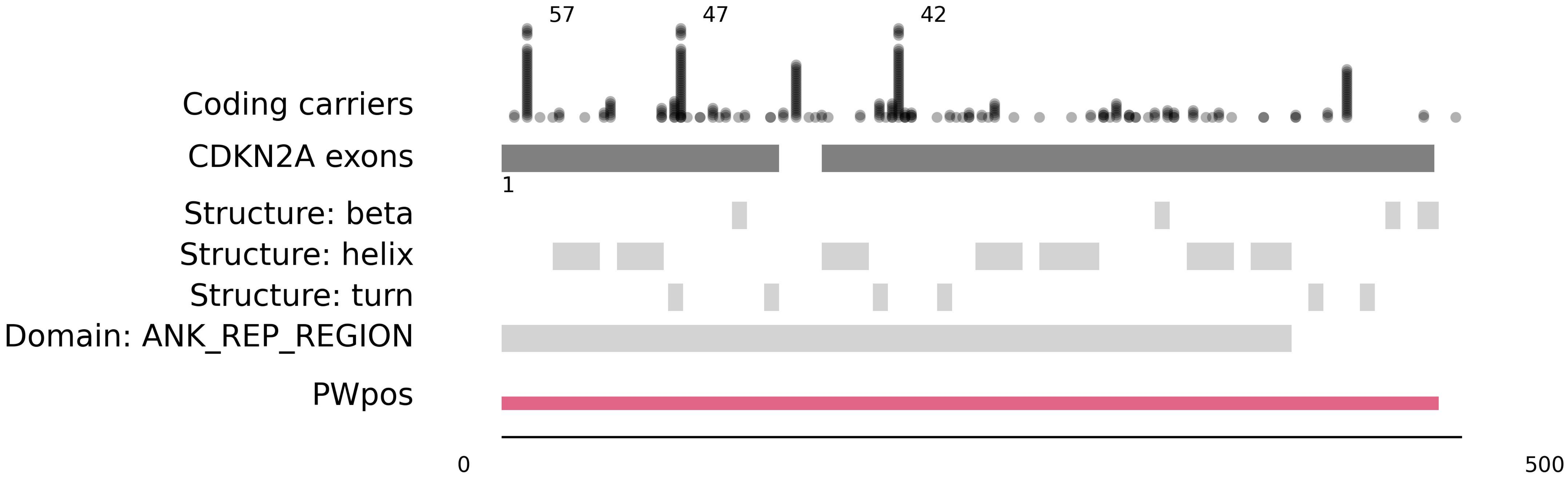

2853193
211
161
82
150
28
27
115
30
63
28
33
66
39
36
Coding carriers
ALPL exons
2
5
10
PWneg
0
500
1000
1500

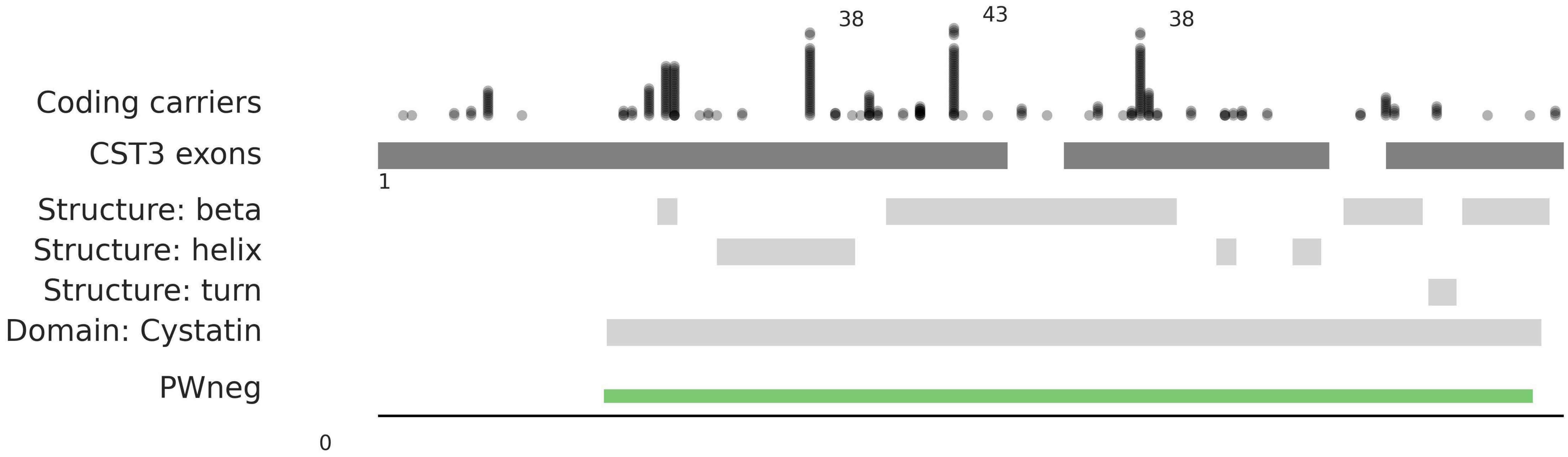

40
65
128
54
234
34
49
Coding carriers
ANGPTL3 exons
1
5
Structure: coiled-coil
Structure: helix
Structure: beta
Domain: Fibrinogen C-t...
PWneg
0
500
1000


Coding carriers
APOB exons
1
5
10
15
20
25
Domain: Vitellogenin
PWpos
PWneg
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000

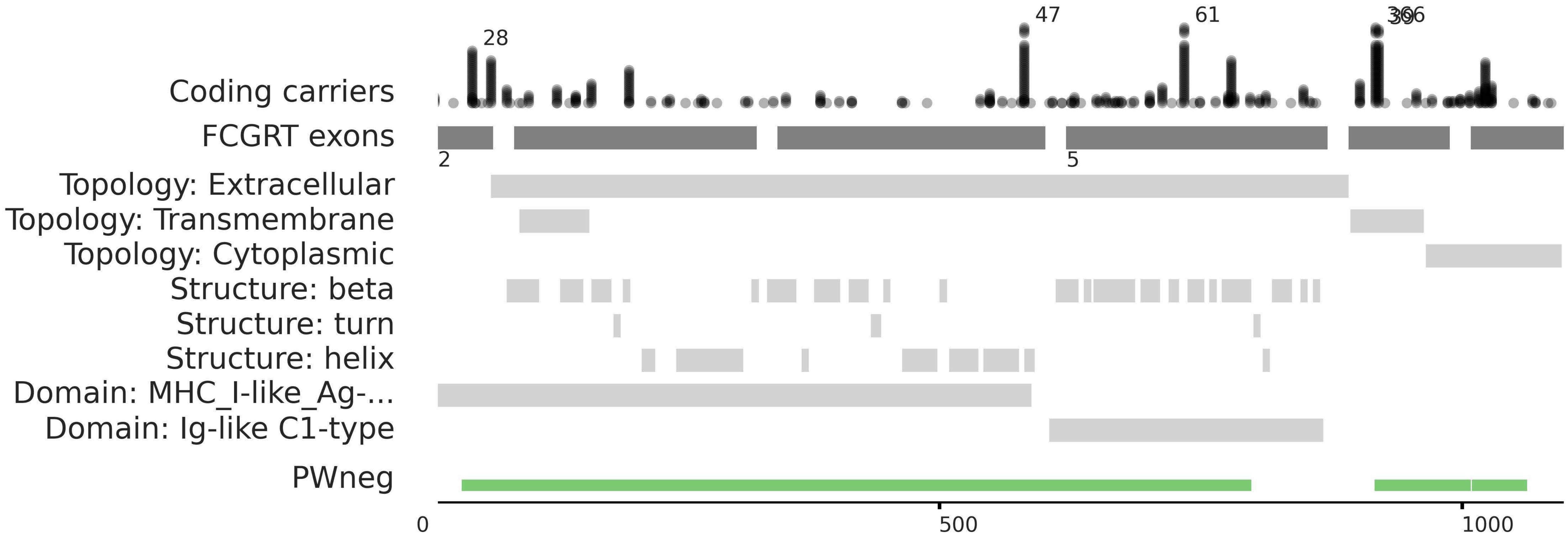


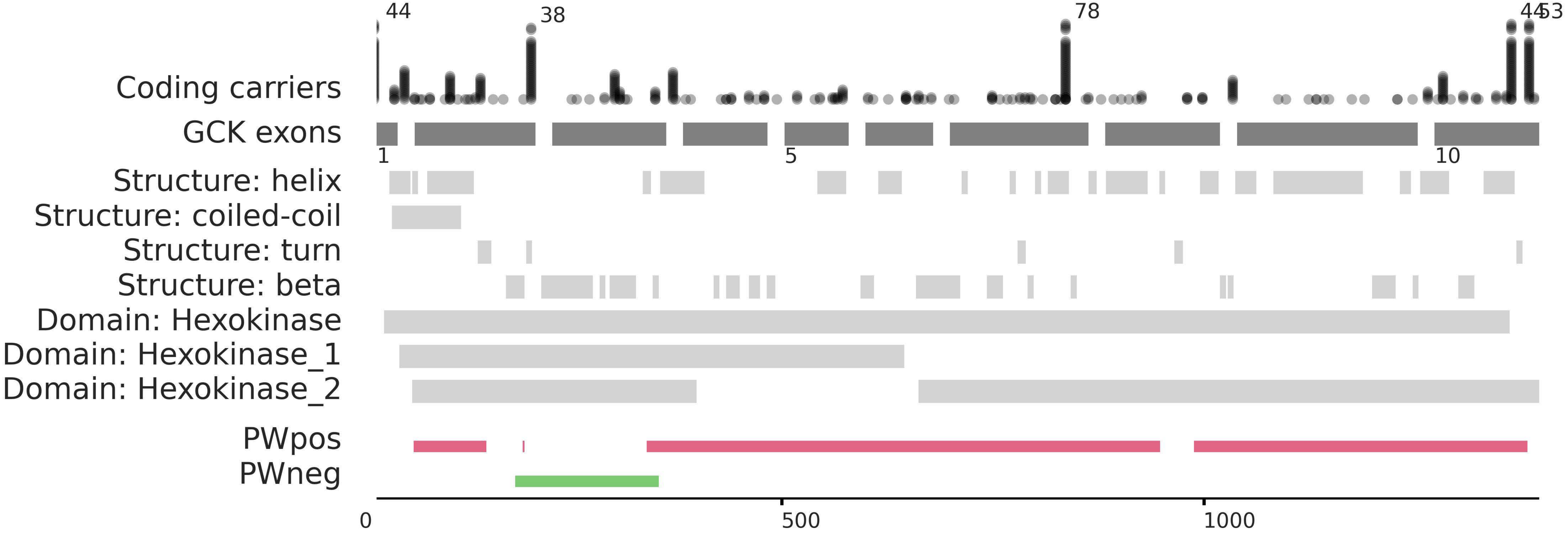
87
65
39
81
108
51
97
Coding carriers
ASGR1 exons
2
5
Topology: Cytoplasmic
Topology: Transmembrane
Topology: Extracellular
Structure: coiled-coil
Structure: beta
Structure: helix
Domain: C-type lectin
PWpos
0
500

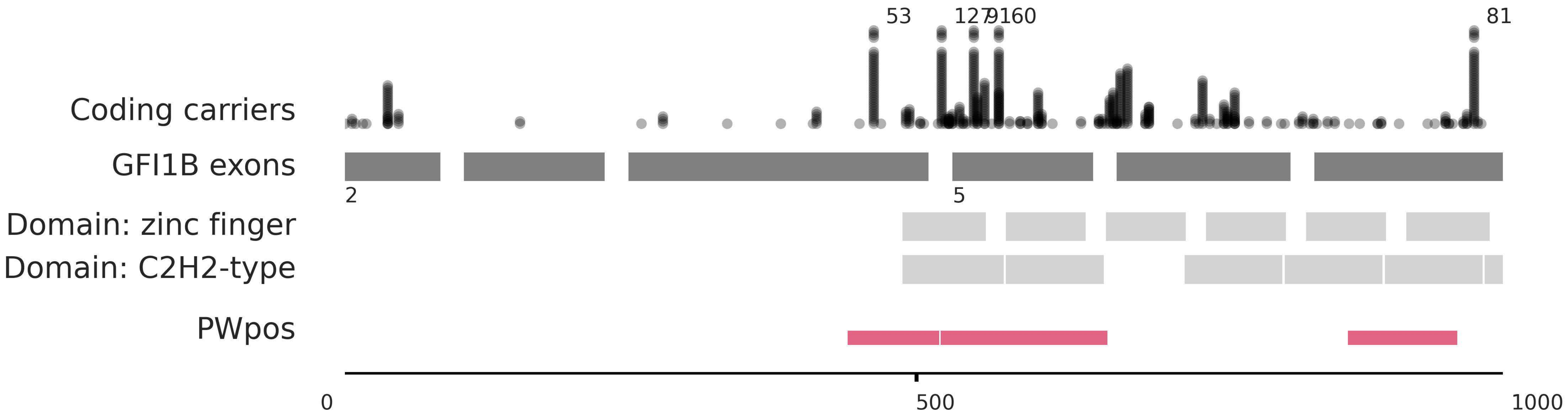

57
47
42
Coding carriers
CDKN2A exons
1
Structure: beta
Structure: helix
Structure: turn
Domain: ANK_REP_REGION
PWpos
0
500

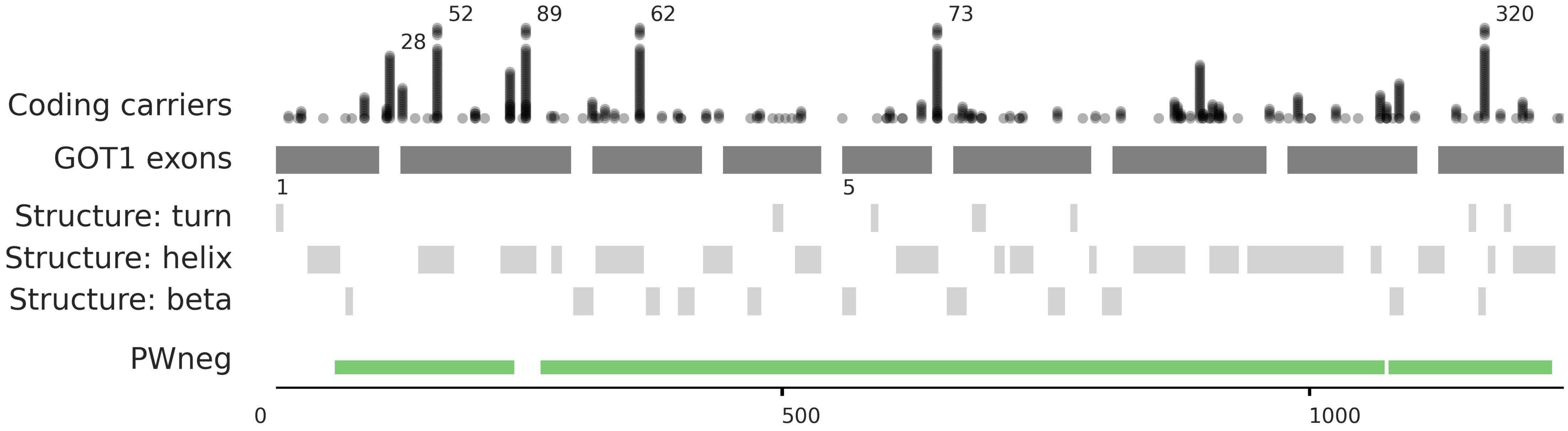


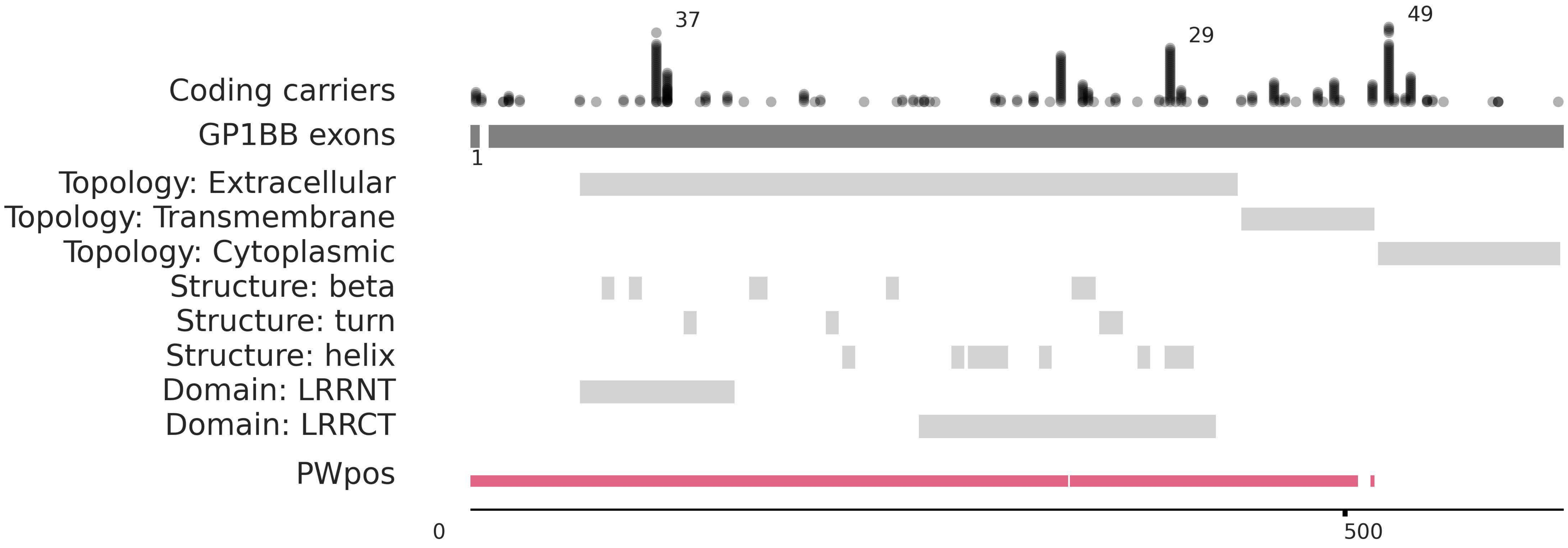
38
43
38
Coding carriers
CST3 exons
1
Structure: beta
Structure: helix
Structure: turn
Domain: Cystatin
PWneg
0


28
47
61
306
39
Coding carriers
FCGRT exons
2
5
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: turn
Structure: helix
Domain: MHC_I-like_Ag-...
Domain: Ig-like C1-type
PWneg
0
500
1000

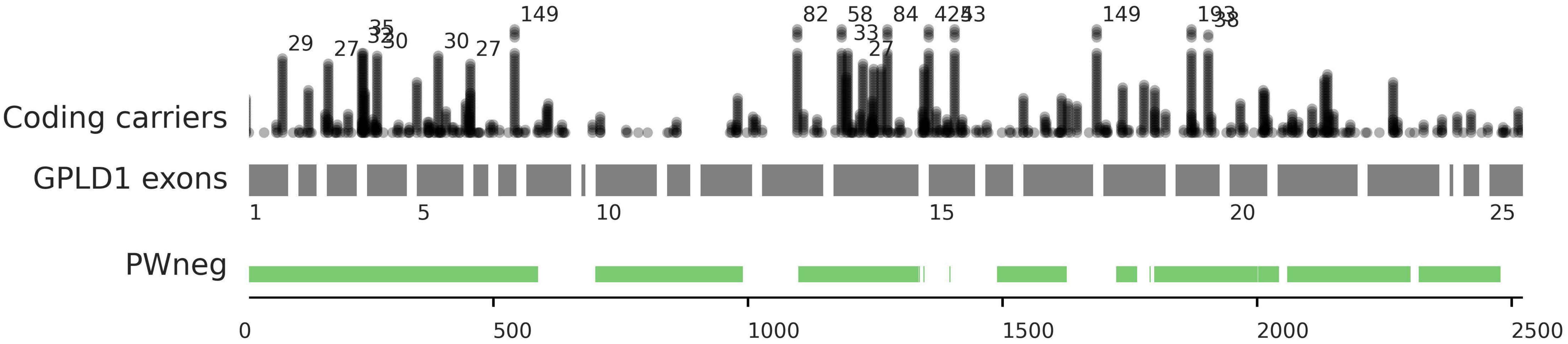

44
38
78
44
53
Coding carriers
GCK exons
1
5
10
Structure: helix
Structure: coiled-coil
Structure: turn
Structure: beta
Domain: Hexokinase
Domain: Hexokinase_1
Domain: Hexokinase_2
PWpos
PWneg
0
500
1000

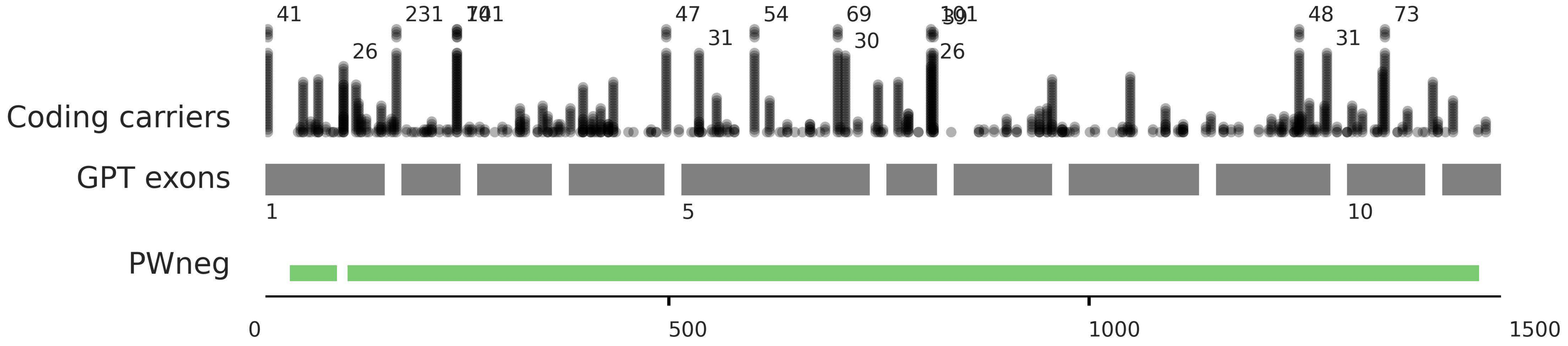

53
127
9
160
81
Coding carriers
GFI1B exons
2
5
Domain: zinc finger
Domain: C2H2-type
PWpos
0
500
1000

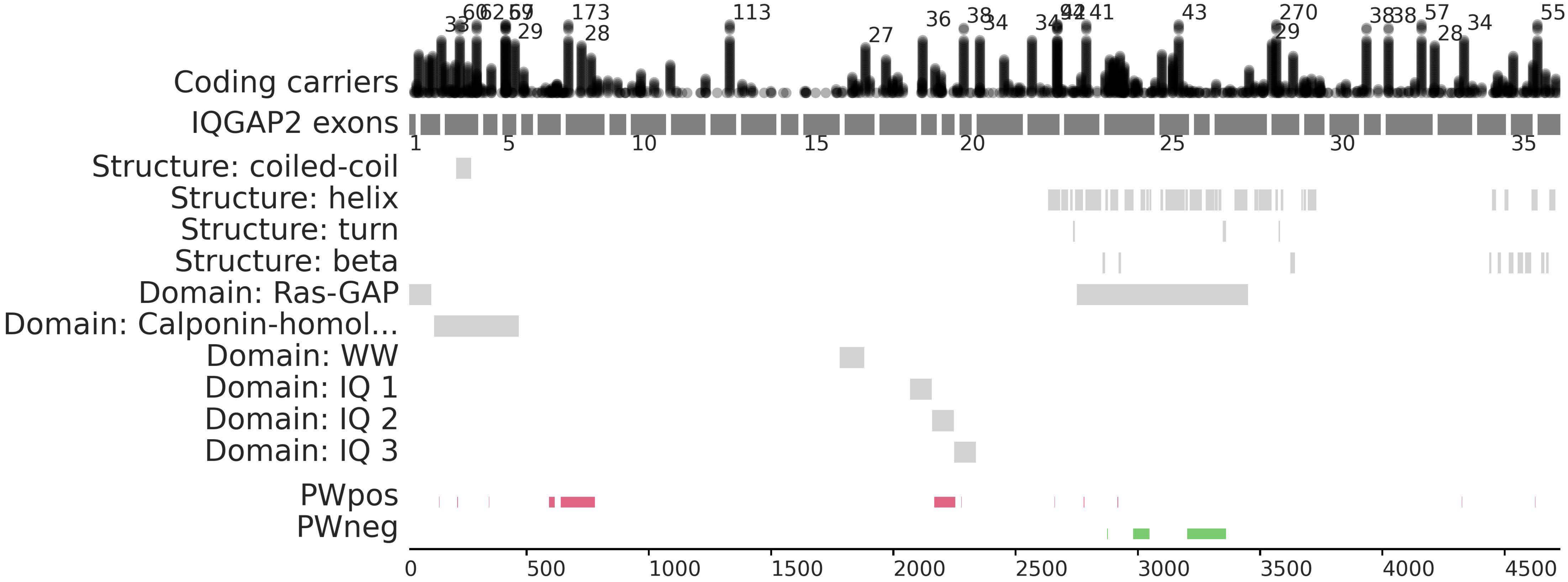

Coding carriers
GOT1 exons
Structure: turn
Structure: helix
Structure: beta
PWneg
28
52
89
62
73
320
1
5
0
500
1000

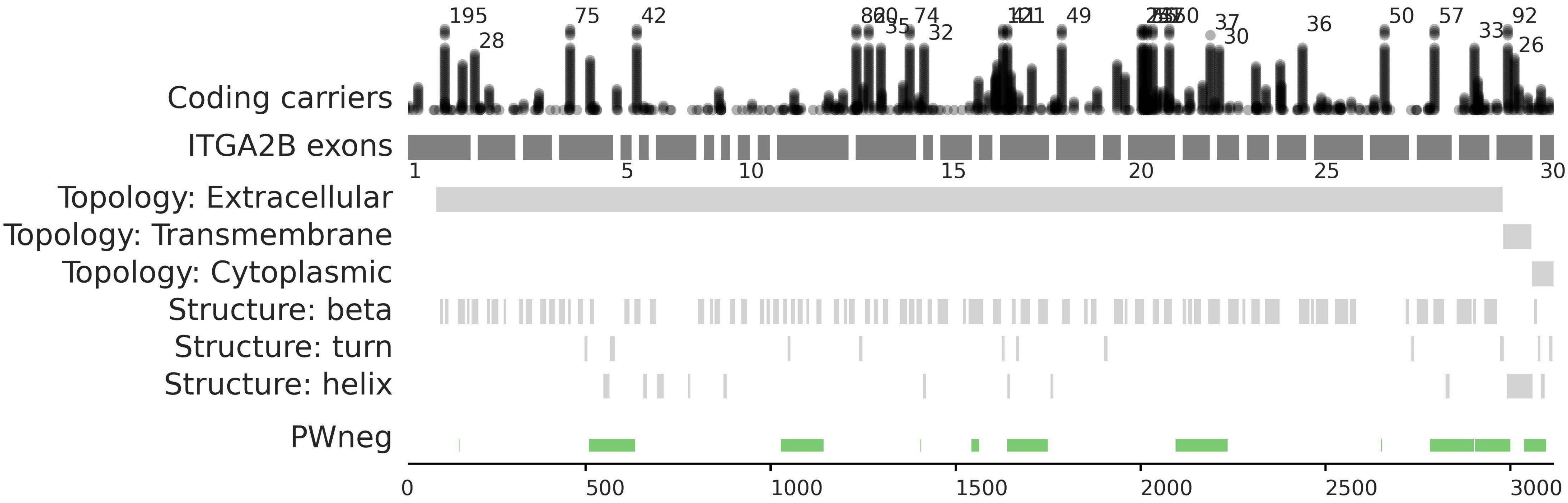


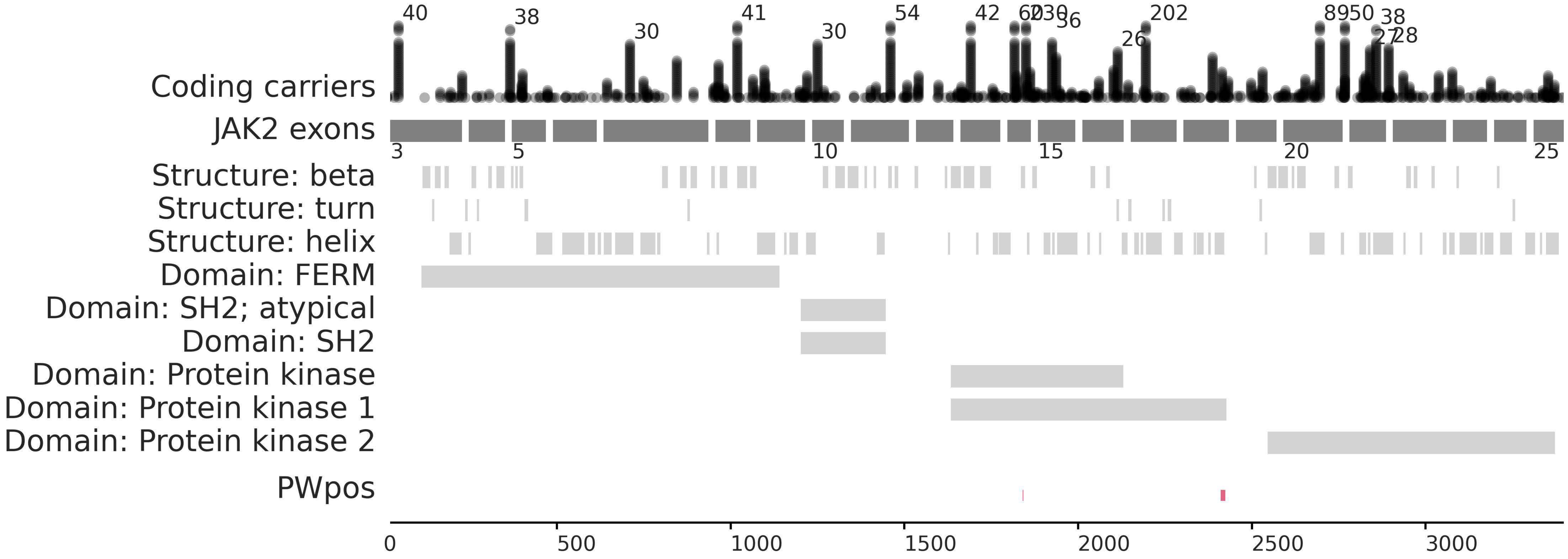
37
29
49
Coding carriers
GP1BB exons
1
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: turn
Structure: helix
Domain: LRRNT
Domain: LRRCT
PWpos
0
500

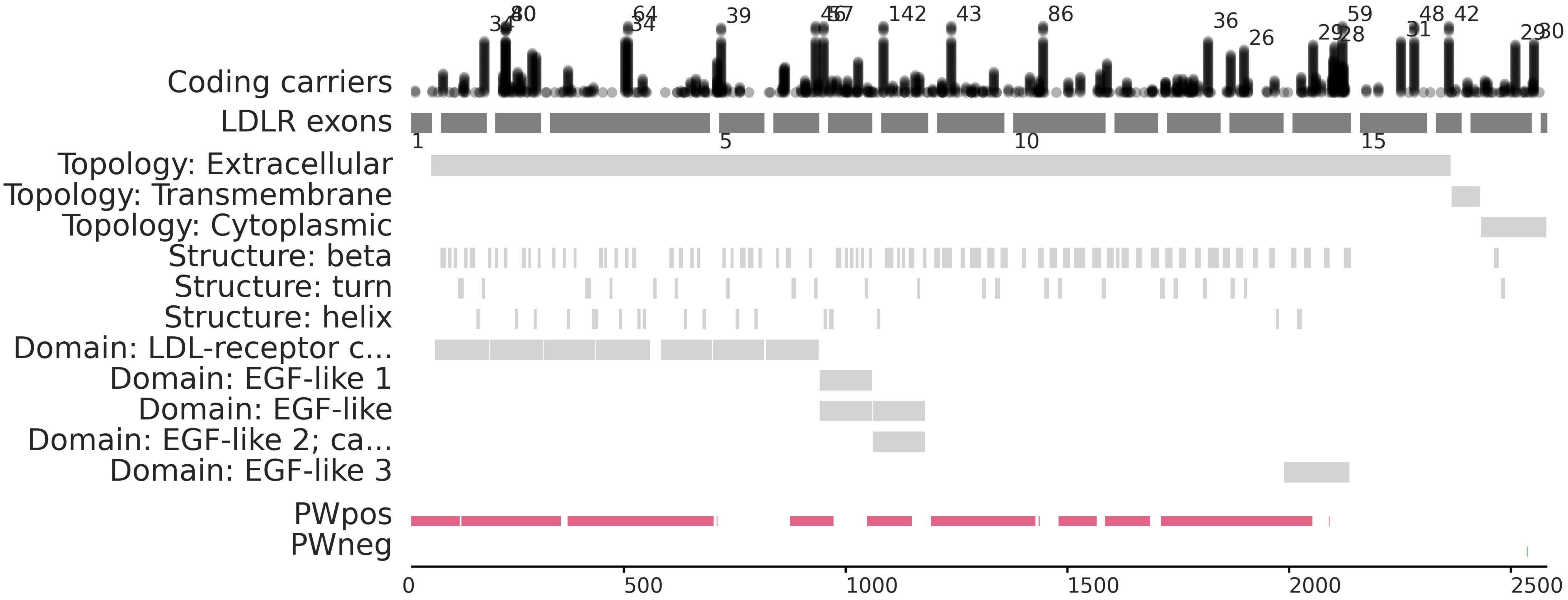

Coding carriers
GPLD1 exons
PWneg
29
27
35
32
30
30
27
149
82
58
33
27
84
425
43
149
193
38
1
5
10
15
20
25
0
500
1000
1500
2000
2500

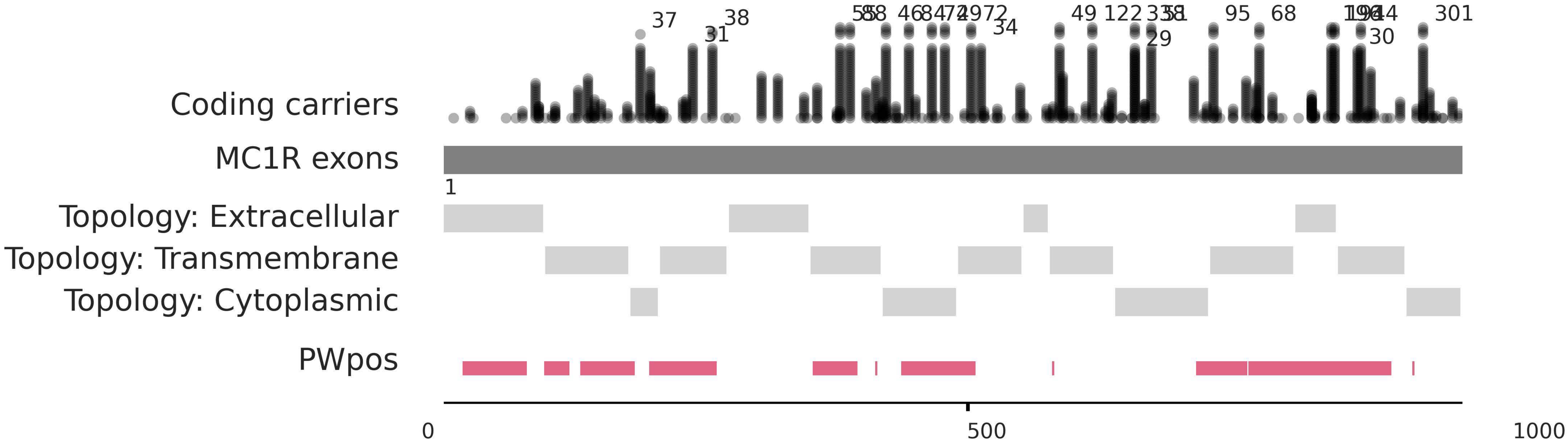

41
26
231
70
141
47
31
54
69
30
101
39
26
48
31
73
Coding carriers
GPT exons
1
5
10
PWneg
0
500
1000
1500

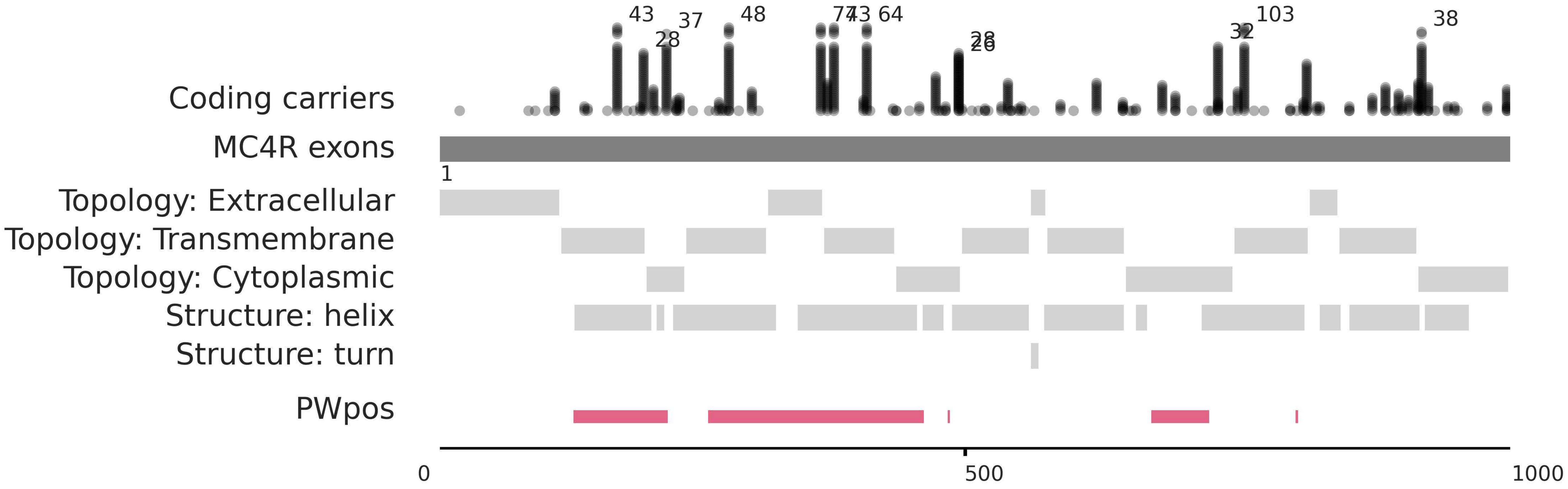

Coding carriers
IQGAP2 exons
Structure: coiled-coil
Structure: helix
Structure: turn
Structure: beta
Domain: Ras-GAP
Domain: Calponin-homol...
Domain: WW
Domain: IQ 1
Domain: IQ 2
Domain: IQ 3
PWpos
PWneg
33
60
62
59
57
29
173
28
113
27
36
38
34
34
91
42
41
43
270
29
38
38
57
28
34
55
1
5
10
15
20
25
30
35
0
500
1000
1500
2000
2500
3000
3500
4000
4500

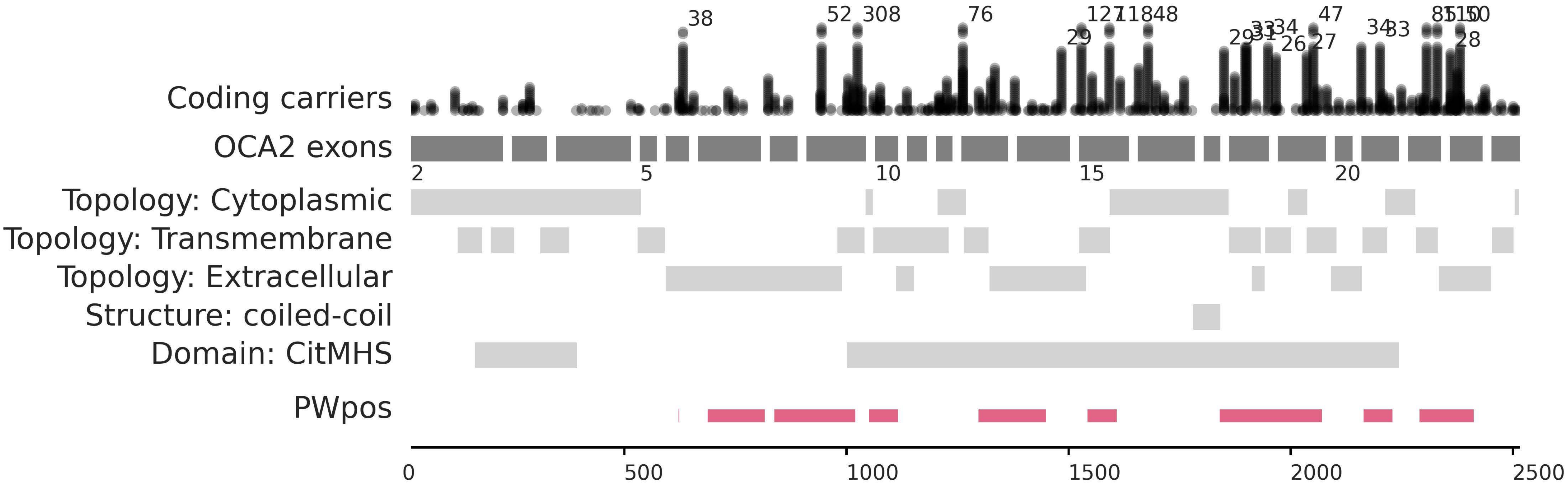

195
28
75
42
35
74
32
49
37
30
36
50
57
33
92
26
Coding carriers
ITGA2B exons
1
5
10
15
20
25
30
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: turn
Structure: helix
PWneg
0
500
1000
1500
2000
2500
3000

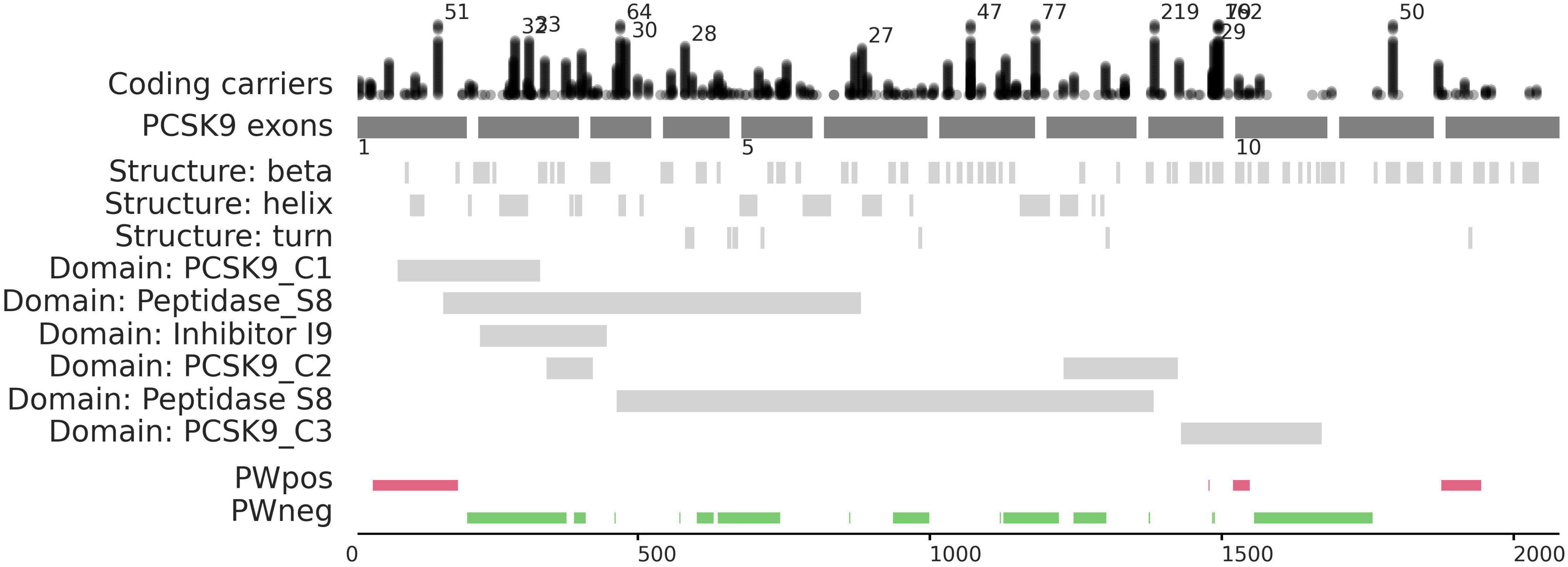

40
38
30
41
30
54
42
62
30
30
36
26
202
89
50
38
27
28
Coding carriers
JAK2 exons
3
5
10
15
20
25
Structure: beta
Structure: turn
Structure: helix
Domain: FERM
Domain: SH2; atypical
Domain: SH2
Domain: Protein kinase
Domain: Protein kinase 1
Domain: Protein kinase 2
PWpos
0
500
1000
1500
2000
2500
3000

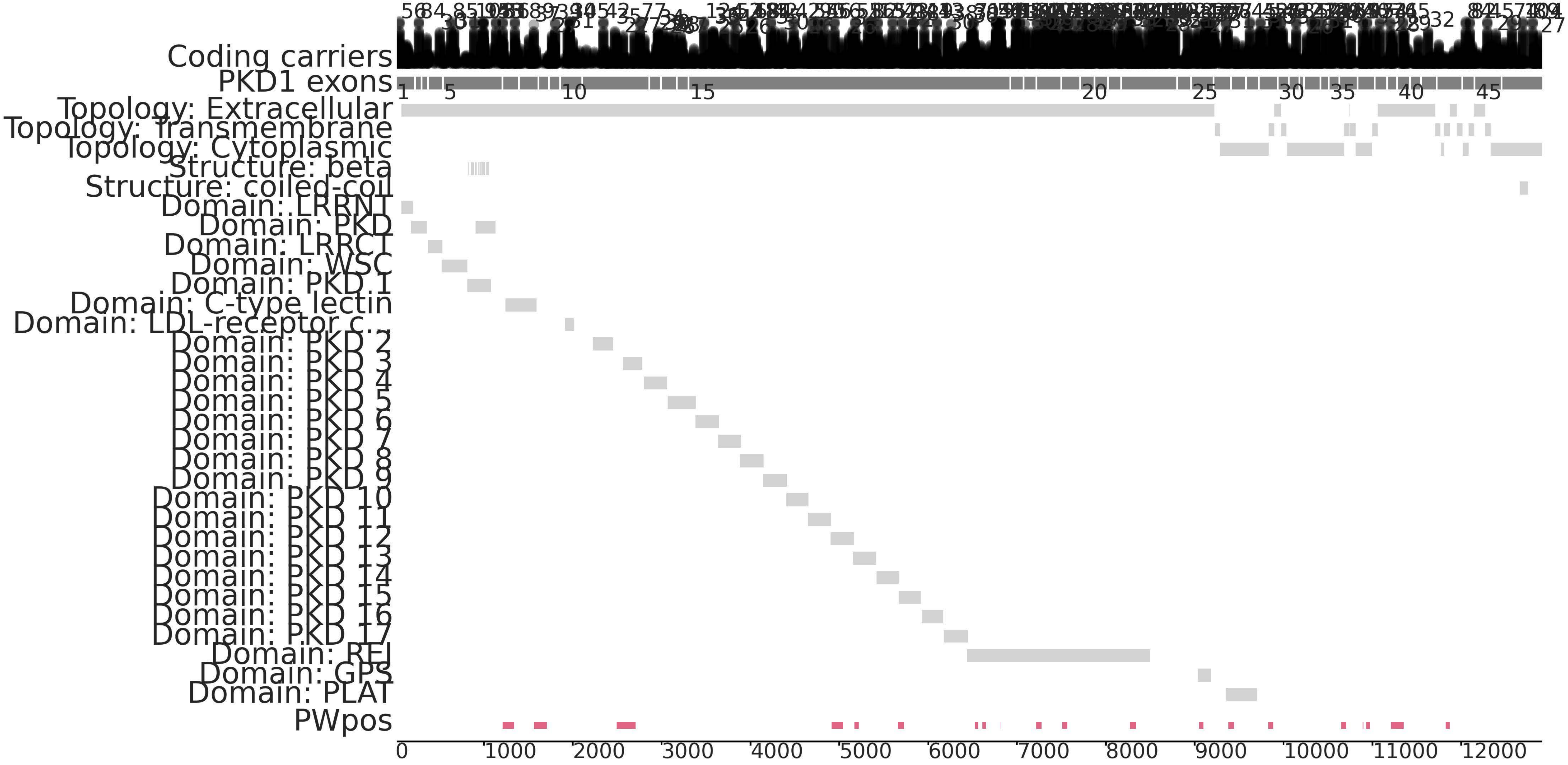

34
80
64
34
39
45
67
142
43
86
36
26
29
28
59
31
48
42
29
30
Coding carriers
LDLR exons
1
5
10
15
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: turn
Structure: helix
Domain: LDL-receptor c...
Domain: EGF-like 1
Domain: EGF-like
Domain: EGF-like 2; ca...
Domain: EGF-like 3
PWpos
PWneg
0
500
1000
1500
2000
2500

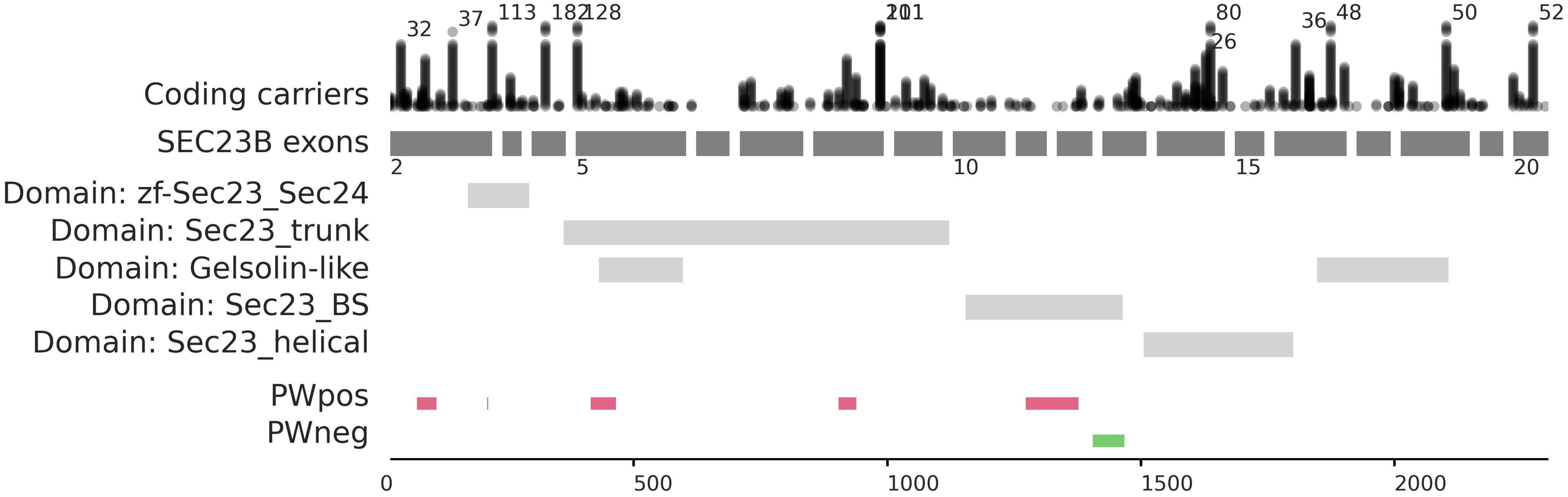

37
31
38
5858
46847472972
34
49
122
33581
29
95
68
1961044
30
301
Coding carriers
MC1R exons
1
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
PWpos
0
500
1000

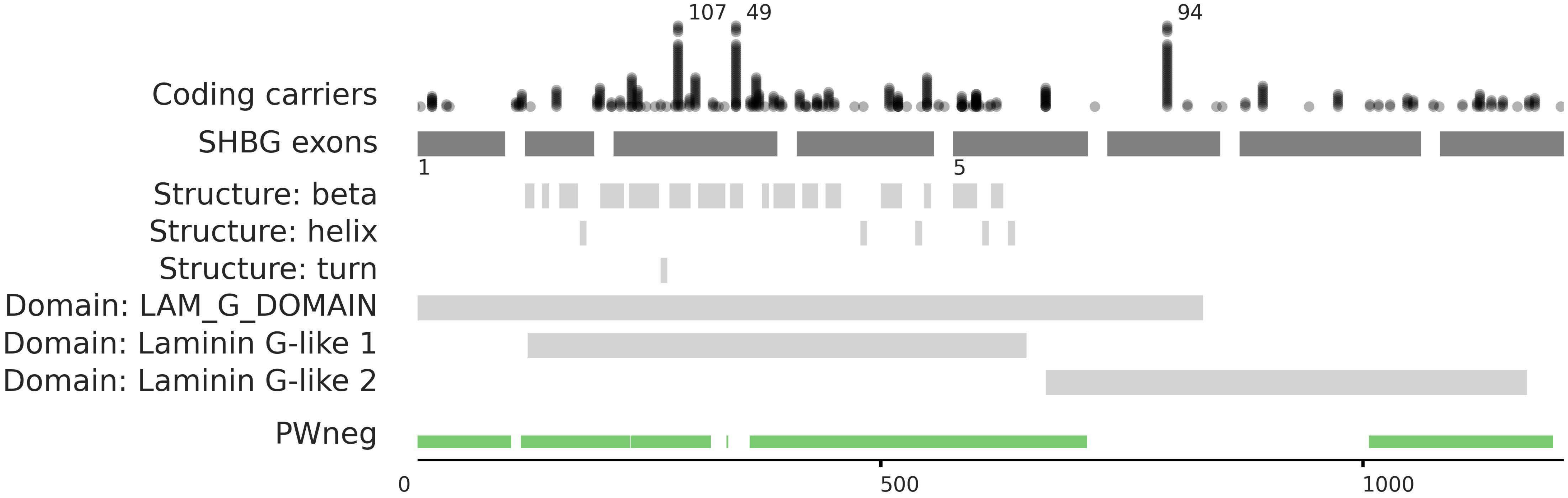

43
28
37
48
77
43
64
28
26
32
103
38
Coding carriers
MC4R exons
1
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: helix
Structure: turn
PWpos
0
500
1000

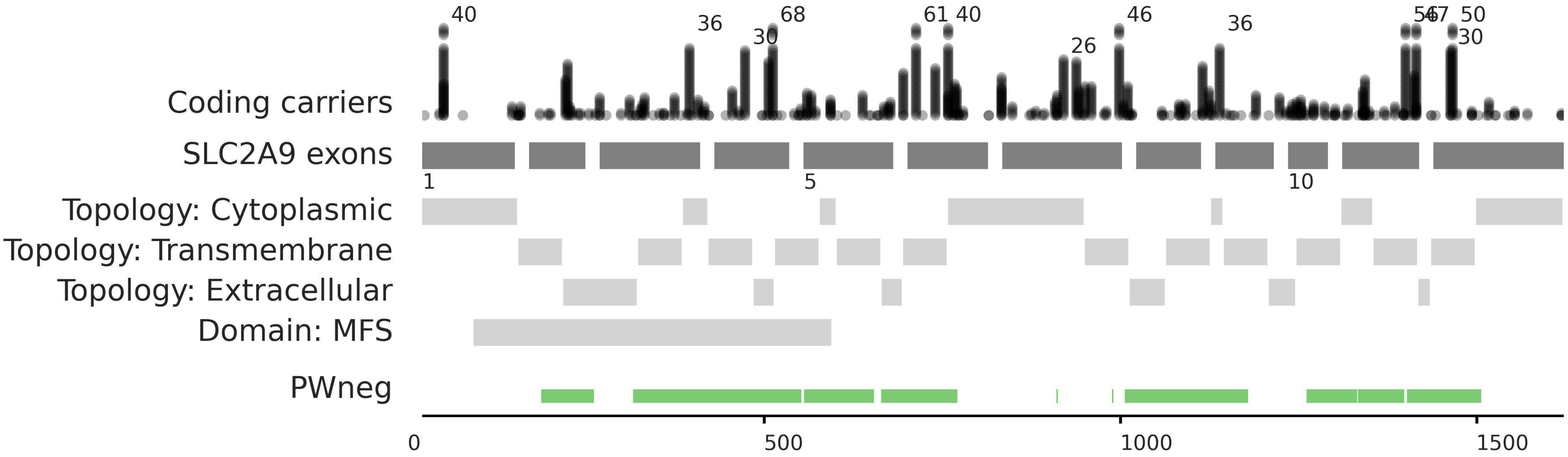

38
52
308
76
29
127
118
48
29
33
31
34
26
47
27
34
33
85
150
28
Coding carriers
OCA2 exons
2
5
10
15
20
Topology: Cytoplasmic
Topology: Transmembrane
Topology: Extracellular
Structure: coiled-coil
Domain: CitMHS
PWpos
0
500
1000
1500
2000
2500

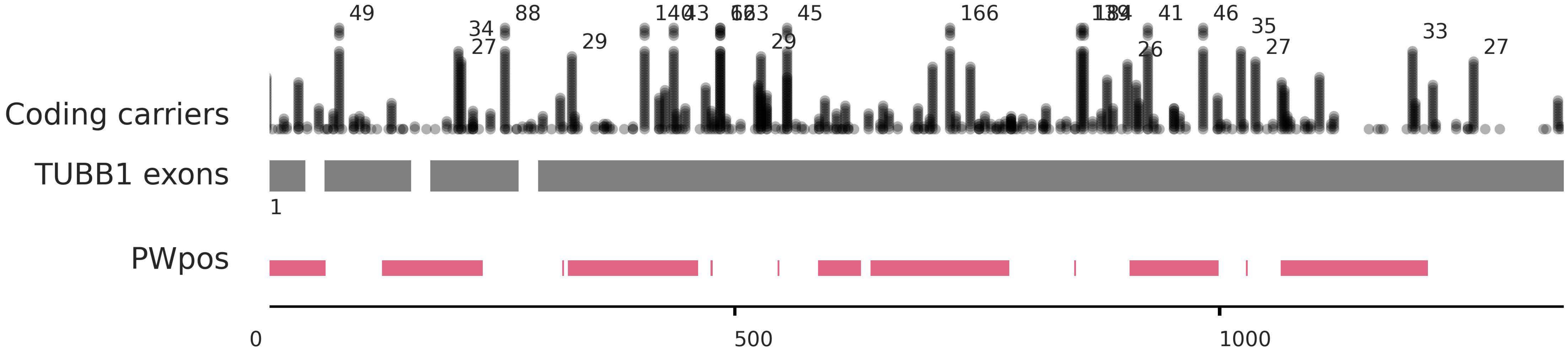

51
32
33
64
30
28
27
47
77
219
76
102
29
50
Coding carriers
PCSK9 exons
1
5
10
Structure: beta
Structure: helix
Structure: turn
Domain: PCSK9_C1
Domain: Peptidase_S8
Domain: Inhibitor I9
Domain: PCSK9_C2
Domain: Peptidase S8
Domain: PCSK9_C3
PWpos
PWneg
0
500
1000
1500
2000


Coding carriers
PKD1 exons
1
5
10
15
20
25
30
35
40
45
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: coiled-coil
Domain: LRRNT
Domain: PKD
Domain: LRRCT
Domain: WSC
Domain: PKD 1
Domain: C-type lectin
Domain: LDL-receptor c...
Domain: PKD 2
Domain: PKD 3
Domain: PKD 4
Domain: PKD 5
Domain: PKD 6
Domain: PKD 7
Domain: PKD 8
Domain: PKD 9
Domain: PKD 10
Domain: PKD 11
Domain: PKD 12
Domain: PKD 13
Domain: PKD 14
Domain: PKD 15
Domain: PKD 16
Domain: PKD 17
Domain: REJ
Domain: GPS
Domain: PLAT
PWpos
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000

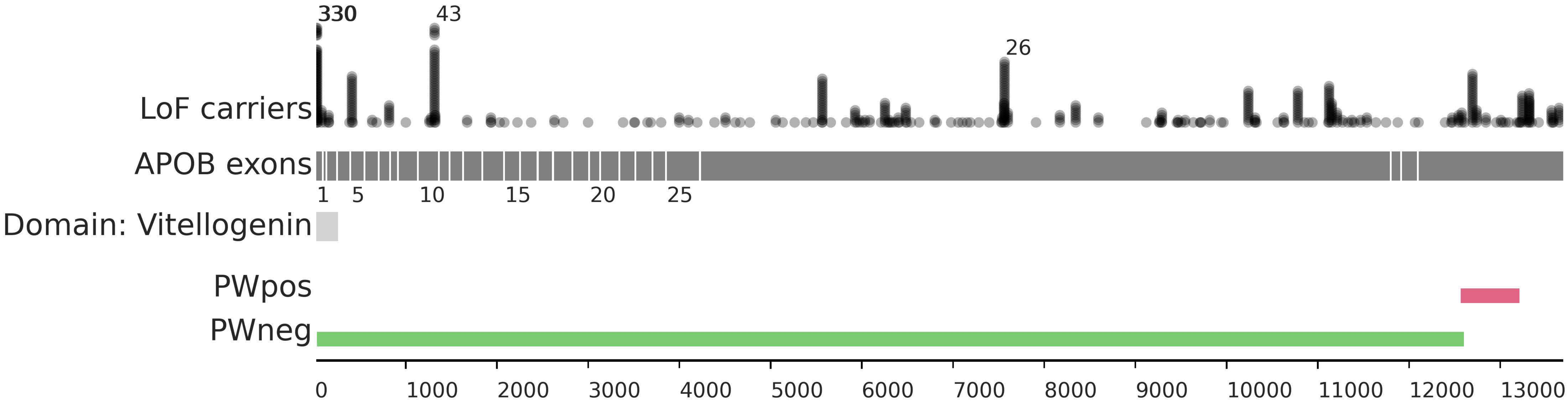

32
37
113
182
128
201
80
26
36
48
50
52
Coding carriers
SEC23B exons
2
5
10
15
20
Domain: zf-Sec23_Sec24
Domain: Sec23_trunk
Domain: Gelsolin-like
Domain: Sec23_BS
Domain: Sec23_helical
PWpos
PWneg
0
500
1000
1500
2000

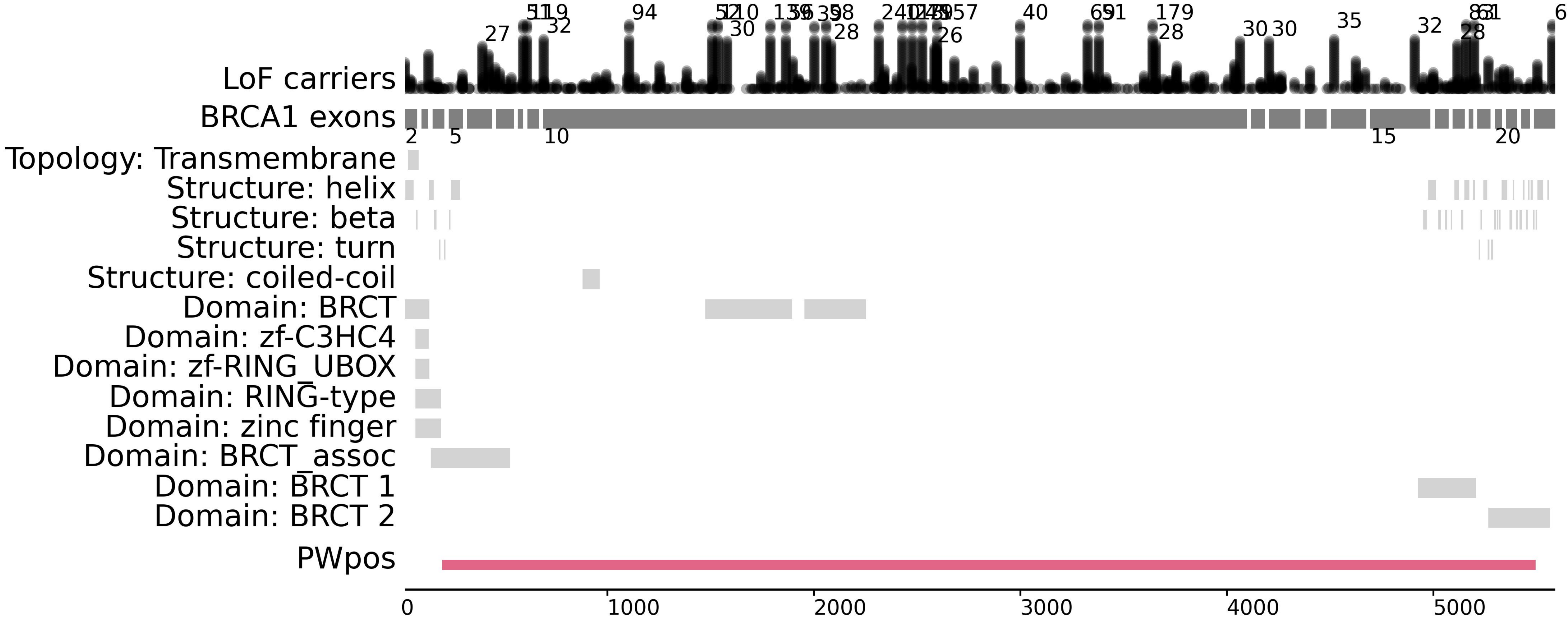

107
49
94
Coding carriers
SHBG exons
1
5
Structure: beta
Structure: helix
Structure: turn
Domain: LAM_G_DOMAIN
Domain: Laminin G-like 1
Domain: Laminin G-like 2
PWneg
0
500
1000


40
36
30
68
61
40
26
46
36
5467
50
30
Coding carriers
SLC2A9 exons
1
5
10
Topology: Cytoplasmic
Topology: Transmembrane
Topology: Extracellular
Domain: MFS
PWneg
0
500
1000
1500

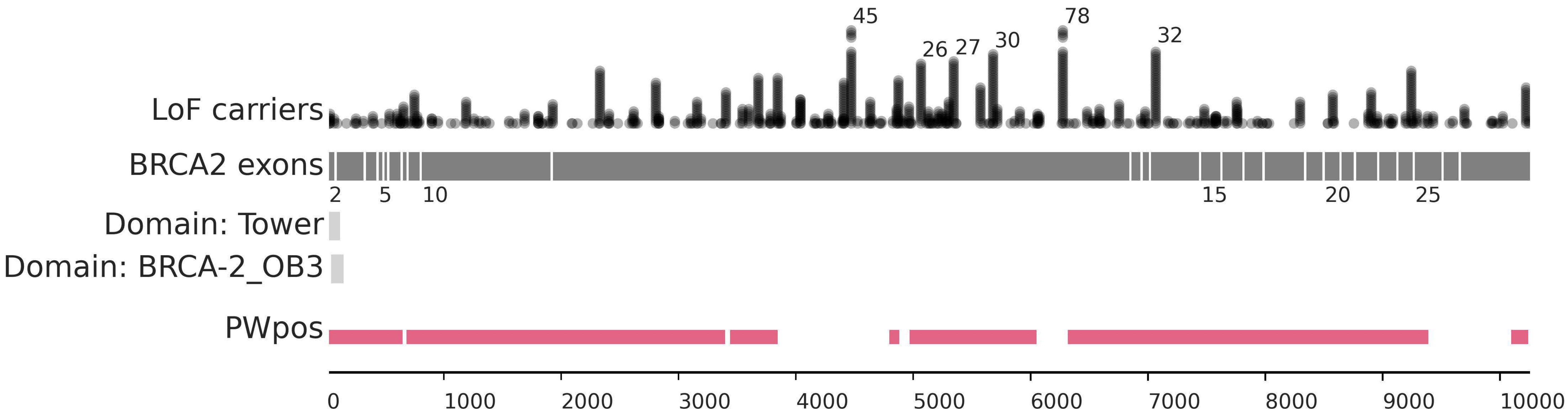


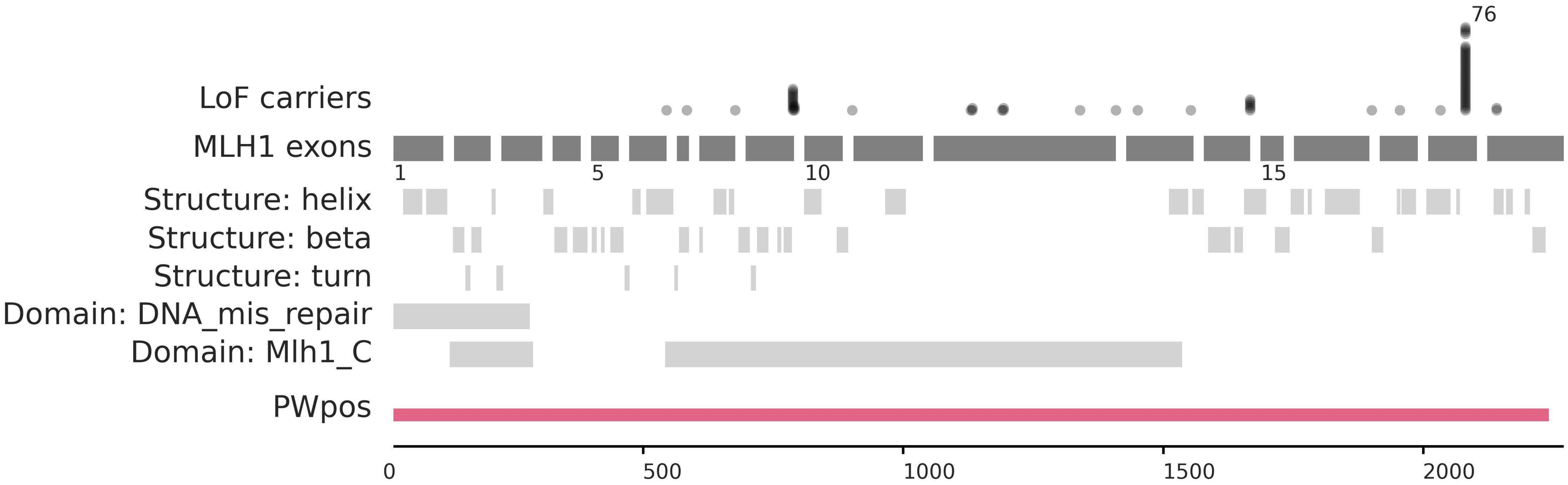
49
34
27
88
29
140
43
62
163
29
45
166
139
184
41
26
46
35
27
33
27
Coding carriers
TUBB1 exons
1
PWpos
0
500
1000

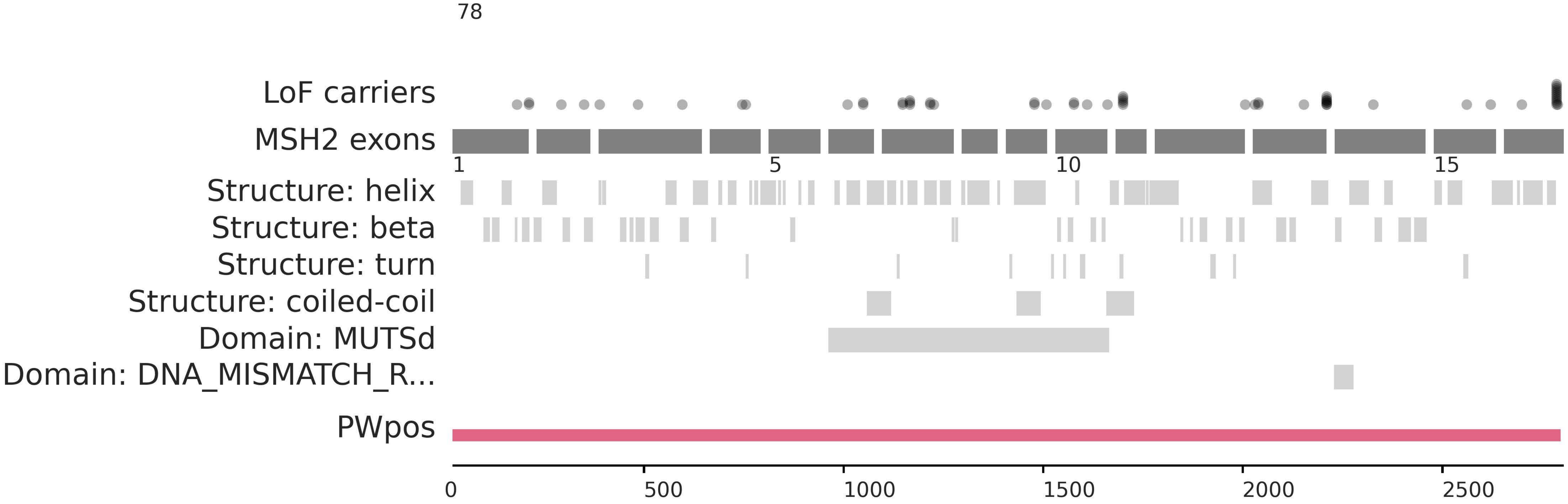

330
43
26
LoF carriers
APOB exons
1
5
10
15
20
25
Domain: Vitellogenin
PWpos
PWneg
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000

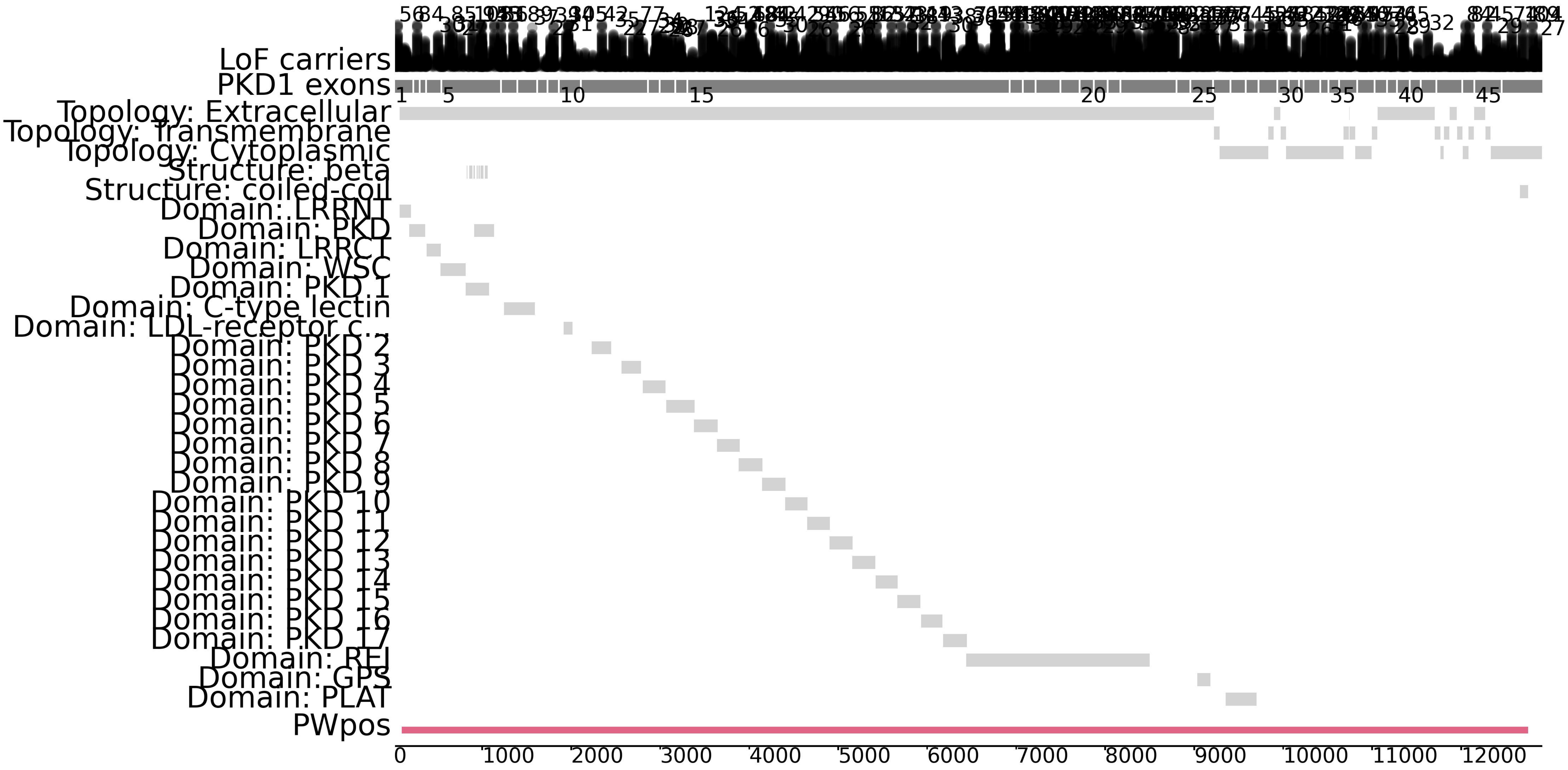


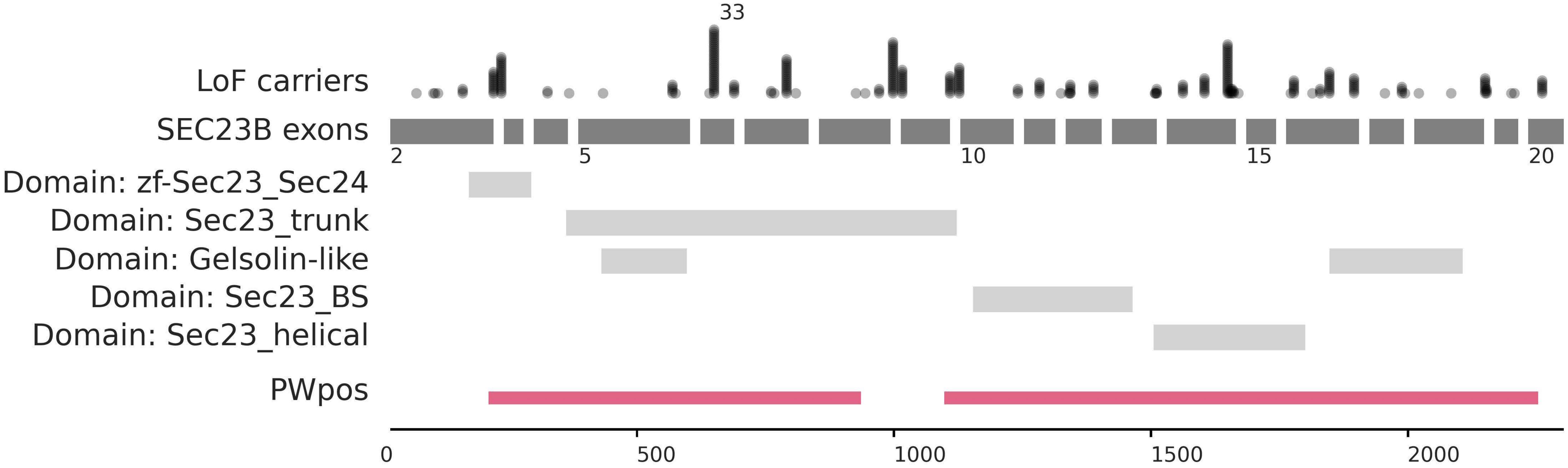
27
51
119
32
94
52
110
30
139
56
35
58
28
240
26
57
40
65
91
179
28
30
30
35
32
83
28
61
64
LoF carriers
BRCA1 exons
2
5
10
15
20
Topology: Transmembrane
Structure: helix
Structure: beta
Structure: turn
Structure: coiled-coil
Domain: BRCT
Domain: zf-C3HC4
Domain: zf-RING_UBOX
Domain: RING-type
Domain: zinc finger
Domain: BRCT_assoc
Domain: BRCT 1
Domain: BRCT 2
PWpos
0
1000
2000
3000
4000
5000

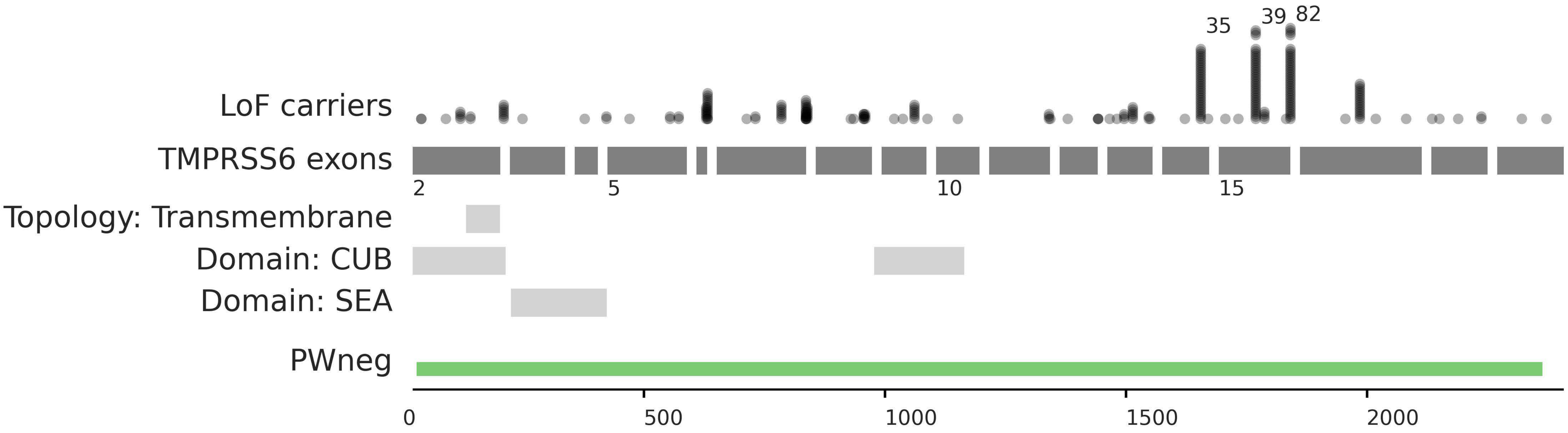

45
78
26
27
30
32
LoF carriers
BRCA2 exons
2
5
10
15
20
25
Domain: Tower
Domain: BRCA-2_OB3
PWpos
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000


76
LoF carriers
MLH1 exons
1
5
10
15
Structure: helix
Structure: beta
Structure: turn
Domain: DNA_mis_repair
Domain: Mlh1_C
PWpos
0
500
1000
1500
2000


78
LoF carriers
MSH2 exons
1
5
10
15
Structure: helix
Structure: beta
Structure: turn
Structure: coiled-coil
Domain: MUTSd
Domain: DNA_MISMATCH_R...
PWpos
0
500
1000
1500
2000
2500


LoF carriers
PKD1 exons
1
5
10
15
20
25
30
35
40
45
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: coiled-coil
Domain: LRRNT
Domain: PKD
Domain: LRRCT
Domain: WSC
Domain: PKD 1
Domain: C-type lectin
Domain: LDL-receptor c...
Domain: PKD 2
Domain: PKD 3
Domain: PKD 4
Domain: PKD 5
Domain: PKD 6
Domain: PKD 7
Domain: PKD 8
Domain: PKD 9
Domain: PKD 10
Domain: PKD 11
Domain: PKD 12
Domain: PKD 13
Domain: PKD 14
Domain: PKD 15
Domain: PKD 16
Domain: PKD 17
Domain: REJ
Domain: GPS
Domain: PLAT
PWpos
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000


33
LoF carriers
SEC23B exons
2
5
10
15
20
Domain: zf-Sec23_Sec24
Domain: Sec23_trunk
Domain: Gelsolin-like
Domain: Sec23_BS
Domain: Sec23_helical
PWpos
0
500
1000
1500
2000


35
39
82
LoF carriers
TMPRSS6 exons
2
5
10
15
Topology: Transmembrane
Domain: CUB
Domain: SEA
PWneg
0
500
1000
1500
2000

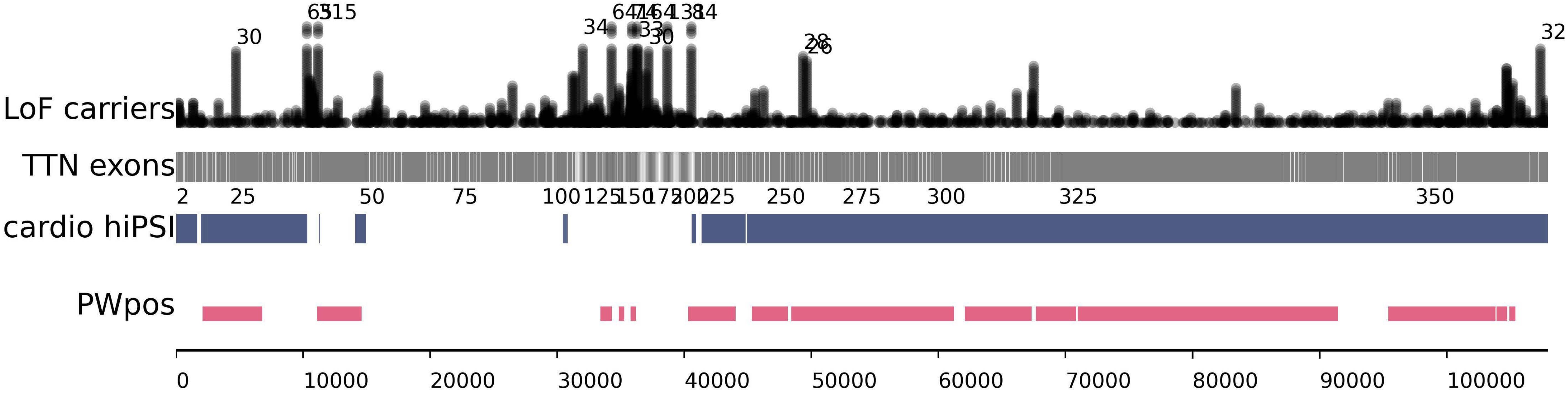

LoF carriers
TTN exons
cardio hiPSI
PWpos
30
65
315
34
64
71
46
41
38
14
33
30
28
26
32
2
25
50
75
100
125
150
175
200
225
250
275
300
325
350
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000

Figure S5. Power Window results for the 37 significant models. Tracks are drawn against each gene's canonical coding transcript. Coding position is shown at the bottom scale. Carriers: each dot represents an individual carrier and are stacked for each carrier at a given position. For brevity, carriers are trimmed to 35 and total number of carriers is indicated when total carriers>40. Exons: exons (dark grey to scale; introns not to scale). Exon number indicated below exon track. Topology, secondary structures and major domains are annotated according to UniProt. PWpos: merger of all significant windows with a positive direction of effect (beta>0.5 or OR>cutoff for that gene; pink). PWneg: merger of all significant windows with a negative direction of effect (beta<-0.5; green). Significant associations are determined as indicated in the main text.

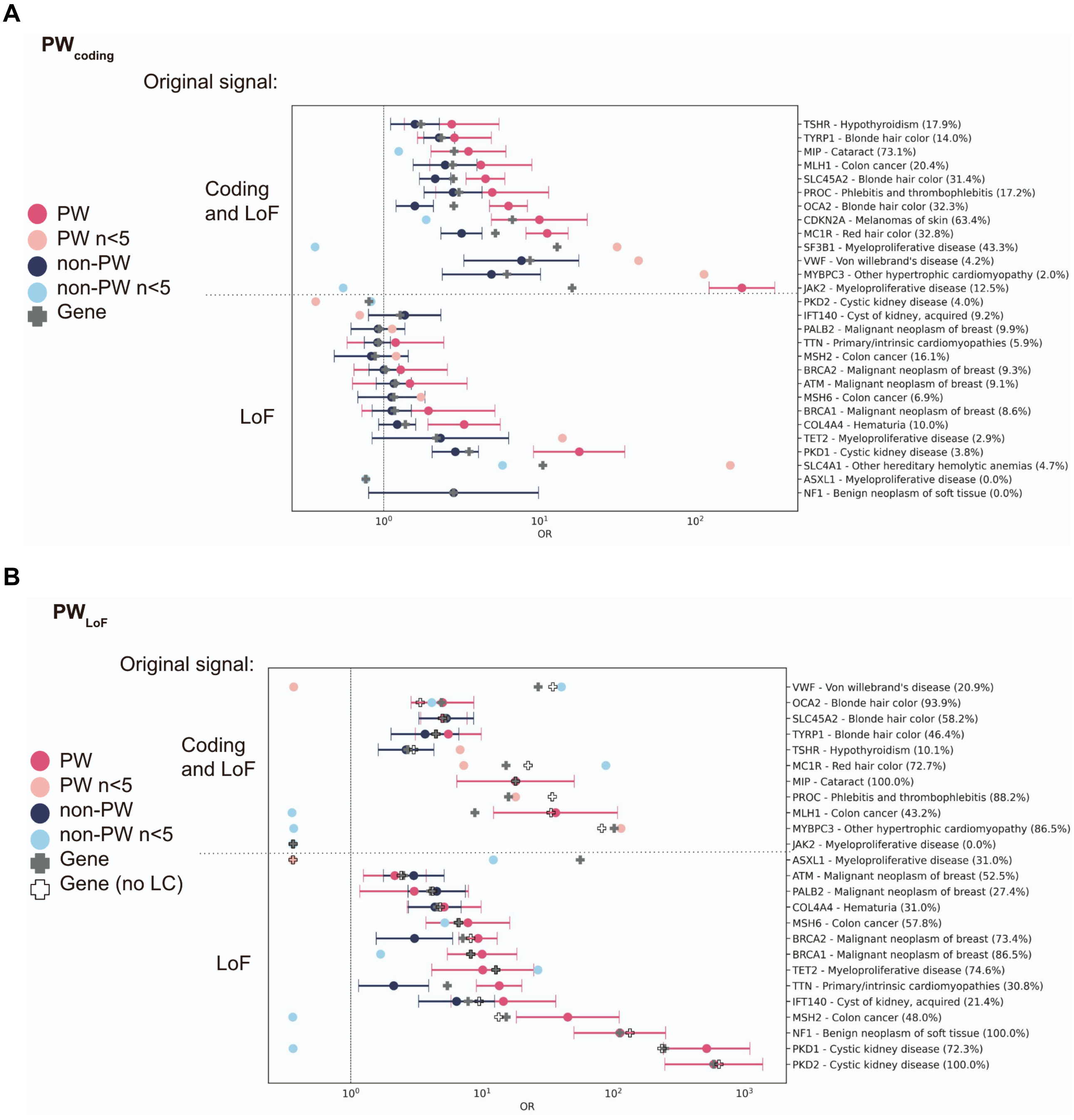


Figure S6. Performance of Power Window for all binary traits in the 117k UKB test set. A) Coding model B) LoF model. The odds ratios for PW models are shown in pink, and the excluded regions are shown in blue. The gene-phenotype label includes in parentheses an indication of what percent of rare variant carriers in the gene were included in the PW model. For brevity, 95% confidence intervals are only shown when there are at least 5 case carriers: datapoints from <5 carriers are not well supported and require additional data to confirm. The genes were grouped based on the original genome-wide association as follows: **LoF**: original whole-gene associations had an absolute beta at least 3x as high in the LoF model as the Coding model; **Coding and LoF**: original whole-gene associations had a beta <3x as high for LoF as for Coding (no gene-phenotype combinations had a whole-gene Coding model absolute beta that was at least 3x higher than the whole-gene LoF beta). Two whole-gene models are shown: gray filled cross, which includes all qualifying variants in the gene, and black empty cross, in which LOFTEE LC variants are excluded.

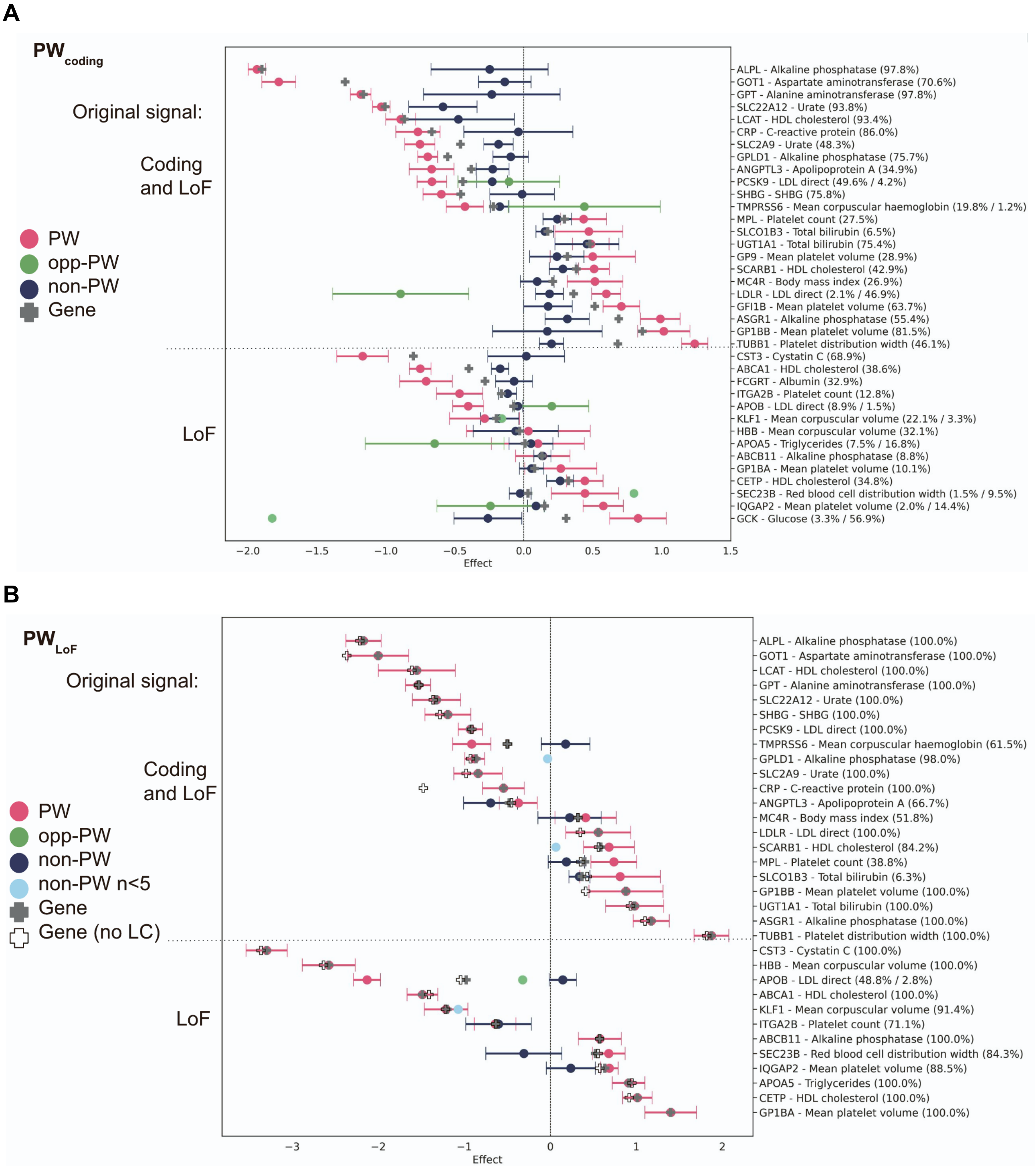


Figure S7. Performance of Power Window for all quantitative traits in the 117k UKB test set. A) Coding model B) LoF model. The effect sizes (normalized phenotypes) for PW models are shown in pink, and the excluded regions are shown in blue. The gene-phenotype label includes in parentheses an indication of what percent of rare variant carriers in the gene were included in the PW model (when there are 2 models for a gene, the percent shown is for negative model / positive model). When a model for the opposite direction of the main effect was built (opp-PW), it is shown in green. For brevity, 95% confidence intervals are only shown when there are at least 5 carriers: datapoints from <5 carriers are not well supported and require additional data to confirm. The genes were grouped based on the original genome-wide association as follows: **LoF**: original whole-gene associations had an absolute beta at least 3x

as high in the LoF model as the Coding model; **Coding and LoF**: original whole-gene associations had a beta <3x as high for LoF as for Coding (no gene-phenotype combinations had a whole-gene Coding model absolute beta that was at least 3x higher than the whole-gene LoF beta). Two whole-gene models are shown: gray filled cross, which includes all qualifying variants in the gene, and black empty cross, in which LOFTEE LC variants are excluded.

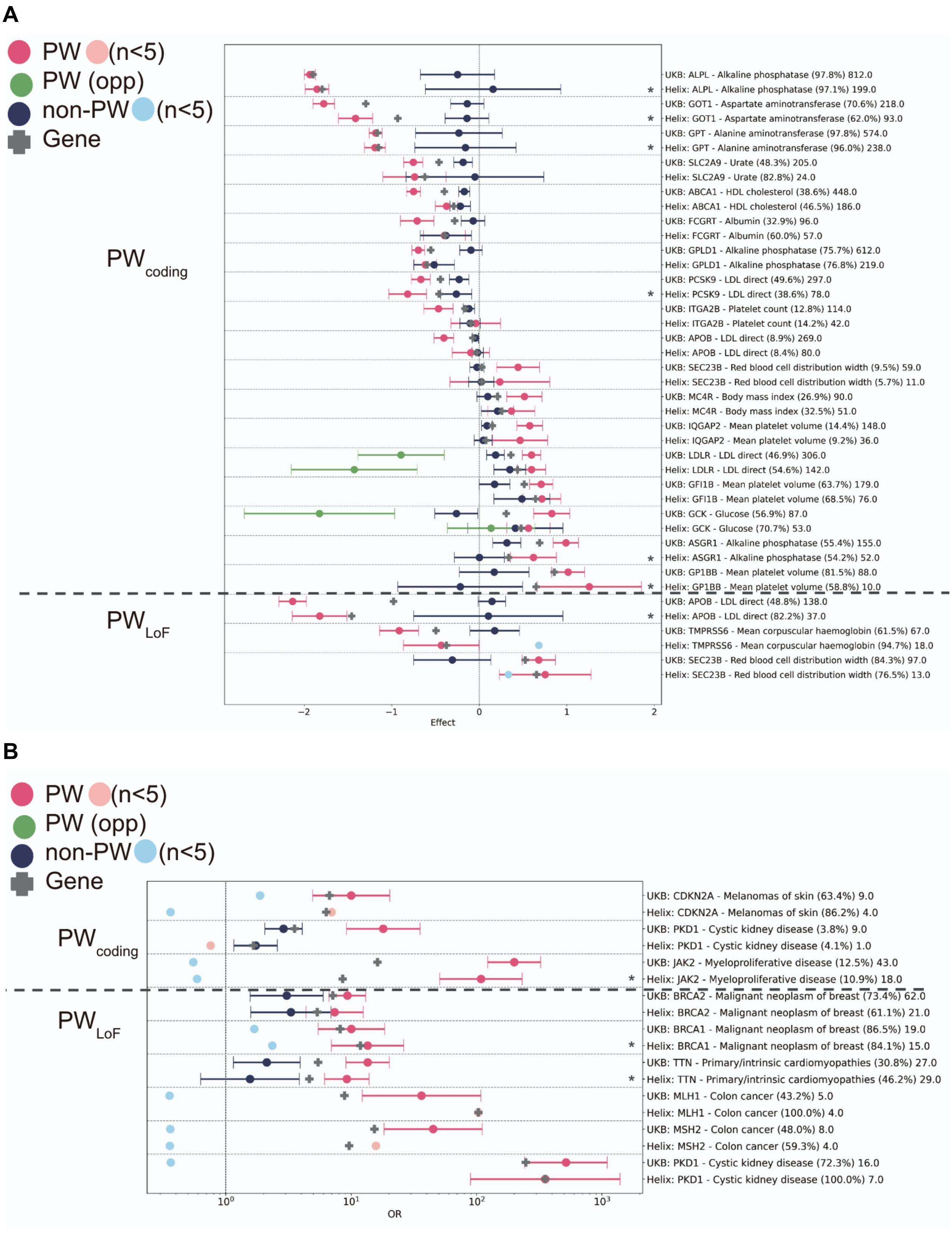

A
PW (n<5)
PW (opp)
non-PW (n<5)
Gene
PWcoding
PWLoF
UKB: ALPL - Alkaline phosphatase (97.8%) 812.0
Helix: ALPL - Alkaline phosphatase (97.1%) 199.0
UKB: GOT1 - Aspartate aminotransferase (70.6%) 218.0
Helix: GOT1 - Aspartate aminotransferase (62.0%) 93.0
UKB: GPT - Alanine aminotransferase (97.8%) 574.0
Helix: GPT - Alanine aminotransferase (96.0%) 238.0
UKB: SLC2A9 - Urate (48.3%) 205.0
Helix: SLC2A9 - Urate (82.8%) 24.0
UKB: ABCA1 - HDL cholesterol (38.6%) 448.0
Helix: ABCA1 - HDL cholesterol (46.5%) 186.0
UKB: FCGRT - Albumin (32.9%) 96.0
Helix: FCGRT - Albumin (60.0%) 57.0
UKB: GPLD1 - Alkaline phosphatase (75.7%) 612.0
Helix: GPLD1 - Alkaline phosphatase (76.8%) 219.0
UKB: PCSK9 - LDL direct (49.6%) 297.0
Helix: PCSK9 - LDL direct (38.6%) 78.0
UKB: ITGA2B - Platelet count (12.8%) 114.0
Helix: ITGA2B - Platelet count (14.2%) 42.0
UKB: APOB - LDL direct (8.9%) 269.0
Helix: APOB - LDL direct (8.4%) 80.0
UKB: SEC23B - Red blood cell distribution width (9.5%) 59.0
Helix: SEC23B - Red blood cell distribution width (5.7%) 11.0
UKB: MC4R - Body mass index (26.9%) 90.0
Helix: MC4R - Body mass index (32.5%) 51.0
UKB: IQGAP2 - Mean platelet volume (14.4%) 148.0
Helix: IQGAP2 - Mean platelet volume (9.2%) 36.0
UKB: LDLR - LDL direct (46.9%) 306.0
Helix: LDLR - LDL direct (54.6%) 142.0
UKB: GFI1B - Mean platelet volume (63.7%) 179.0
Helix: GFI1B - Mean platelet volume (68.5%) 76.0
UKB: GCK - Glucose (56.9%) 87.0
Helix: GCK - Glucose (70.7%) 53.0
UKB: ASGR1 - Alkaline phosphatase (55.4%) 155.0
Helix: ASGR1 - Alkaline phosphatase (54.2%) 52.0
UKB: GP1BB - Mean platelet volume (81.5%) 88.0
Helix: GP1BB - Mean platelet volume (58.8%) 10.0
UKB: APOB - LDL direct (48.8%) 138.0
Helix: APOB - LDL direct (82.2%) 37.0
UKB: TMPRSS6 - Mean corpuscular haemoglobin (61.5%) 67.0
Helix: TMPRSS6 - Mean corpuscular haemoglobin (94.7%) 18.0
UKB: SEC23B - Red blood cell distribution width (84.3%) 97.0
Helix: SEC23B - Red blood cell distribution width (76.5%) 13.0
*
−2
−1
0
1
2
Effect
B
PW (n<5)
PW (opp)
non-PW (n<5)
Gene
PWcoding
PWLoF
UKB: CDKN2A - Melanomas of skin (63.4%) 9.0
Helix: CDKN2A - Melanomas of skin (86.2%) 4.0
UKB: PKD1 - Cystic kidney disease (3.8%) 9.0
Helix: PKD1 - Cystic kidney disease (4.1%) 1.0
UKB: JAK2 - Myeloproliferative disease (12.5%) 43.0
Helix: JAK2 - Myeloproliferative disease (10.9%) 18.0
UKB: BRCA2 - Malignant neoplasm of breast (73.4%) 62.0
Helix: BRCA2 - Malignant neoplasm of breast (61.1%) 21.0
UKB: BRCA1 - Malignant neoplasm of breast (86.5%) 19.0
Helix: BRCA1 - Malignant neoplasm of breast (84.1%) 15.0
UKB: TTN - Primary/intrinsic cardiomyopathies (30.8%) 27.0
Helix: TTN - Primary/intrinsic cardiomyopathies (46.2%) 29.0
UKB: MLH1 - Colon cancer (43.2%) 5.0
Helix: MLH1 - Colon cancer (100.0%) 4.0
UKB: MSH2 - Colon cancer (48.0%) 8.0
Helix: MSH2 - Colon cancer (59.3%) 4.0
UKB: PKD1 - Cystic kidney disease (72.3%) 16.0
Helix: PKD1 - Cystic kidney disease (100.0%) 7.0
*
10^0
10^1
10^2
10^3
OR

Figure S8. Replication in the Helix cohorts. A) Quantitative traits. The effect sizes (normalized phenotypes) for PW models are shown in pink, and the excluded regions (non-PW) are shown in blue. When a model for the opposite direction of the main effect was built (PW-opp), it is shown in green. Significant models ($p<0.002$ with Bonferroni correction for multiple tests) are marked with an asterisk. B) Binary traits. The odds ratios for PW models are shown in pink, and the excluded regions (non-PW) are shown in blue. The gene-phenotype label includes in parentheses an indication of what percent of rare variant carriers in the gene were included in the PW model (when there are 2 models for a gene, the percent shown is for negative model / positive model). For brevity, 95% confidence intervals are only shown when there are at least 5 case carriers. Only models that were significant in the UKB117k test cohort are tested. The UKB data shown are from the 117k test cohort. The total number of PW variant carriers (or case carriers for binary traits) is shown at the end of each row. The results for *PKD1* coding and LoF as well as *MLH1* LoF were not included in the final counts as they did not have case carriers in both PW and non-PW regions for comparison.

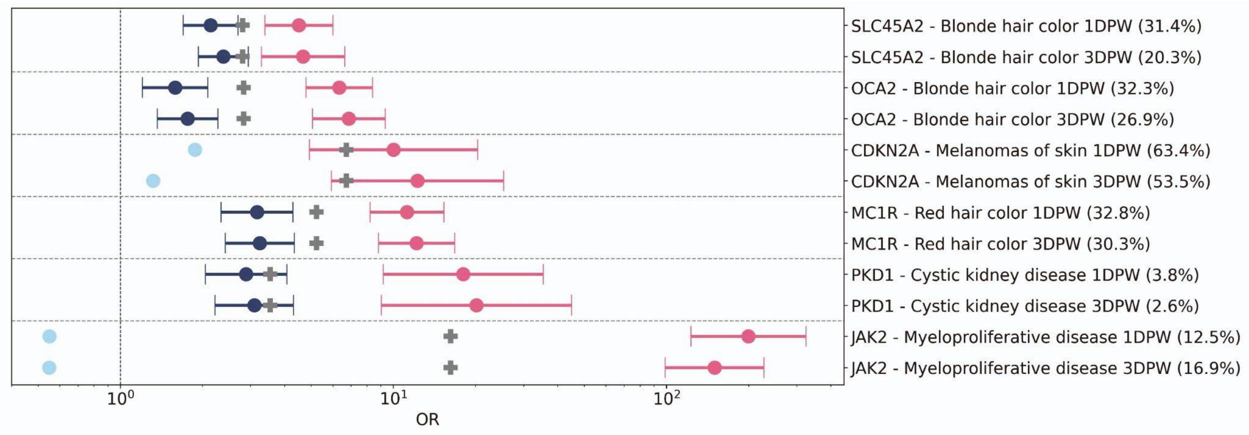
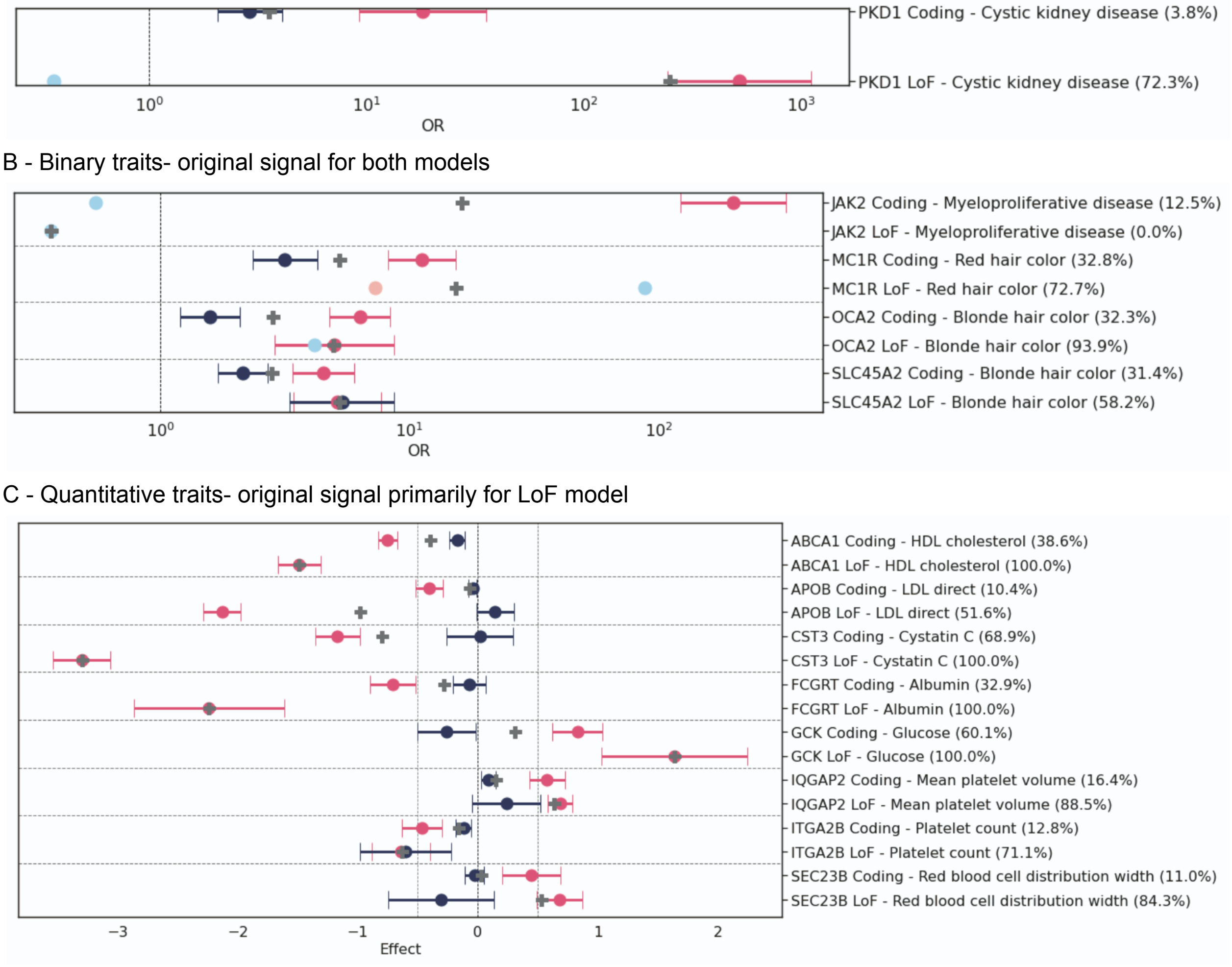


Figure S9. 3D Power Window for binary traits. As in Figure 5, PW models are shown in the test set of 117k individuals for significant PW_{coding} models. Each gene-phenotype pair is shown twice, for the 1D and 3DPW models, with 95% confidence intervals. The model results are very similar in terms of percent of carriers kept in the model and final effect sizes. The percentage to the right of the phenotype is the percent carriers retained in the model.


A - Binary traits- original signal primarily for LoF model

B - Binary traits- original signal for both models

C - Quantitative traits- original signal primarily for LoF model

D - Quantitative traits- original signal for both models

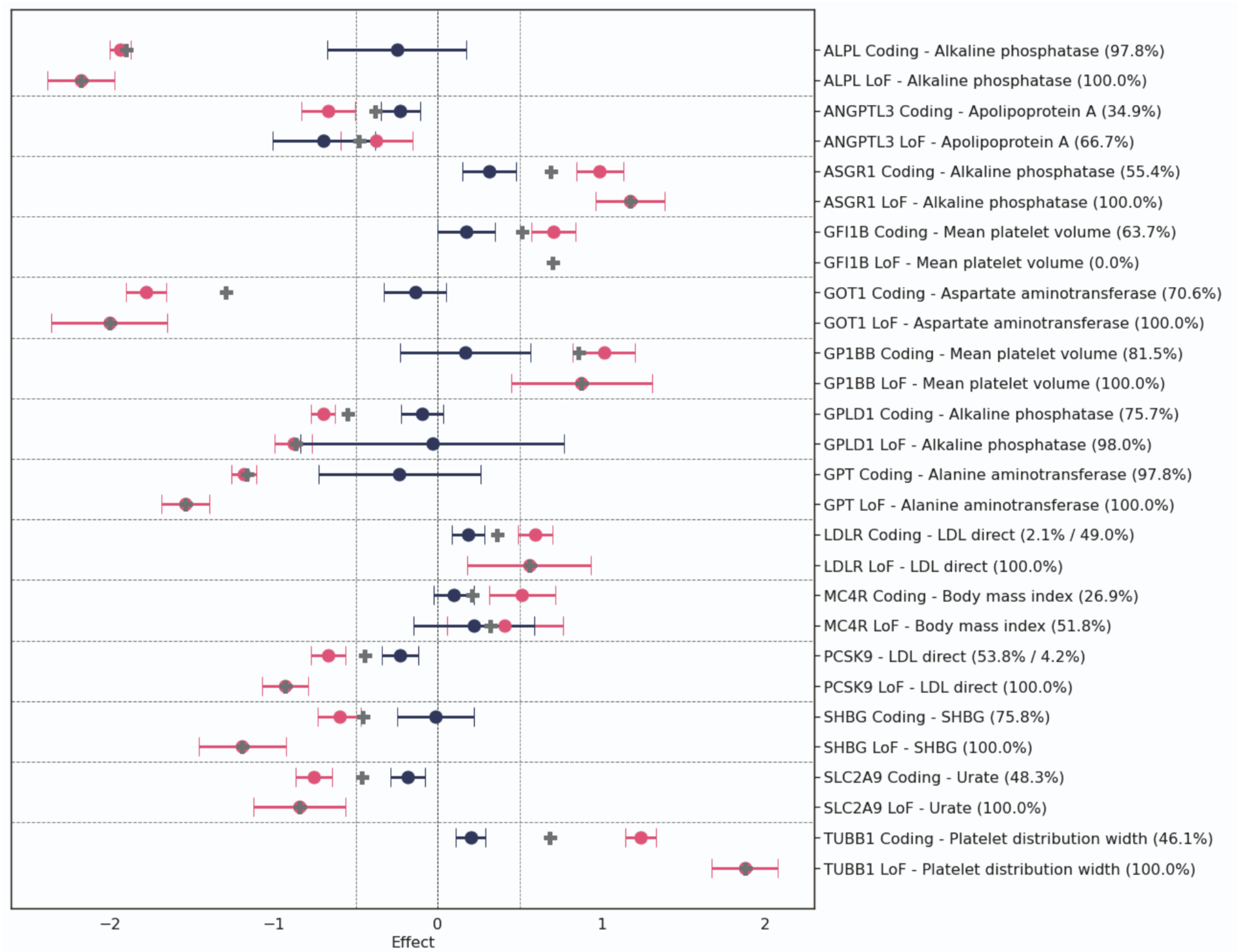


Figure S10. Performance of significant PW$_{\text{coding}}$ models compared to PW$_{\text{LoF}}$ models for the same phenotype. A) Binary traits where the original whole-gene associations were primarily for the LoF model (LoF beta >3x Coding beta); B) binary traits where the original whole-gene associations were for both Coding and Lof models (beta<3x as high in LoF as in Coding) C) Quantitative traits where the original whole-gene associations were primarily for the LoF model (LoF beta >3x Coding beta); D) quantitative traits where the original whole-gene associations were for both Coding and Lof models (beta<3x as high in LoF as in Coding). Data are shown for the 117k UKB test set, wth 95% confidence intervals. These PWcoding models fit the criteria for statistical significance and are also displayed in Figure 3. For the PWcoding models, the effect size of the PW$_{\text{coding}}$ model is similar to that of the PWLoF model for 1 of the 3 binary models (A) and 3 of 8 quantitative models (C) where the whole-gene association was primarily LoF; this was also true for all 4 binary models (B, excluding *JAK2* where there was no LoF signal) and 14 of 19 quantitative models (D) where the whole-gene association was for both coding and LoF. In these cases, the model was able to identify coding variants with what appear to be complete LoF effects. For the remaining models, the PW$_{\text{coding}}$ effect size is a substantial improvement over the whole-gene coding model, but is still not as extreme as those with LoF variants. These may reflect variants that reduce function without completely losing it.

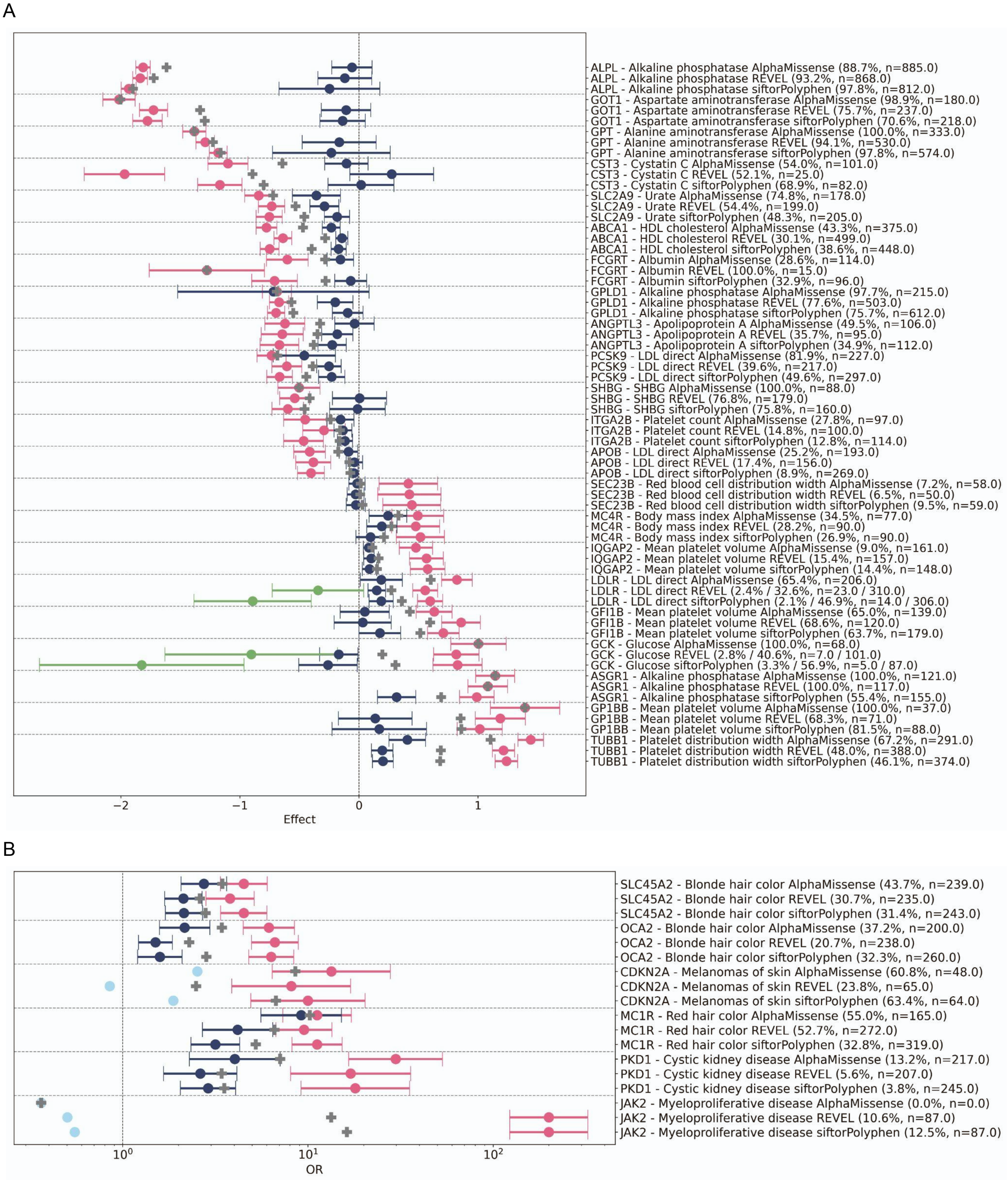


Figure S11. Comparison of significant PWcoding models built with REVEL, AlphaMissense, or sift/Polyphen. As in Figure 3, PW models are shown in the test set of 117k individuals with 95% confidence intervals for significant PW$_{coding}$ models for A) quantitative and B) binary traits. Each gene-phenotype pair is shown three times, for when REVEL was used as the bioinformatic predictor of what is damaging (cutoff 0.25), when AlphaMissense was, and when sift/Polyphen were (default model used in the paper, remove missense variants that were benign by both of these). The model results are very similar in terms of percent of carriers kept in the model and final effect sizes. The

percentage to the right of the phenotype is the percent carriers retained in the model; when there is both a negative and positive model, the percent carriers retained in the negative model is shown first.