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

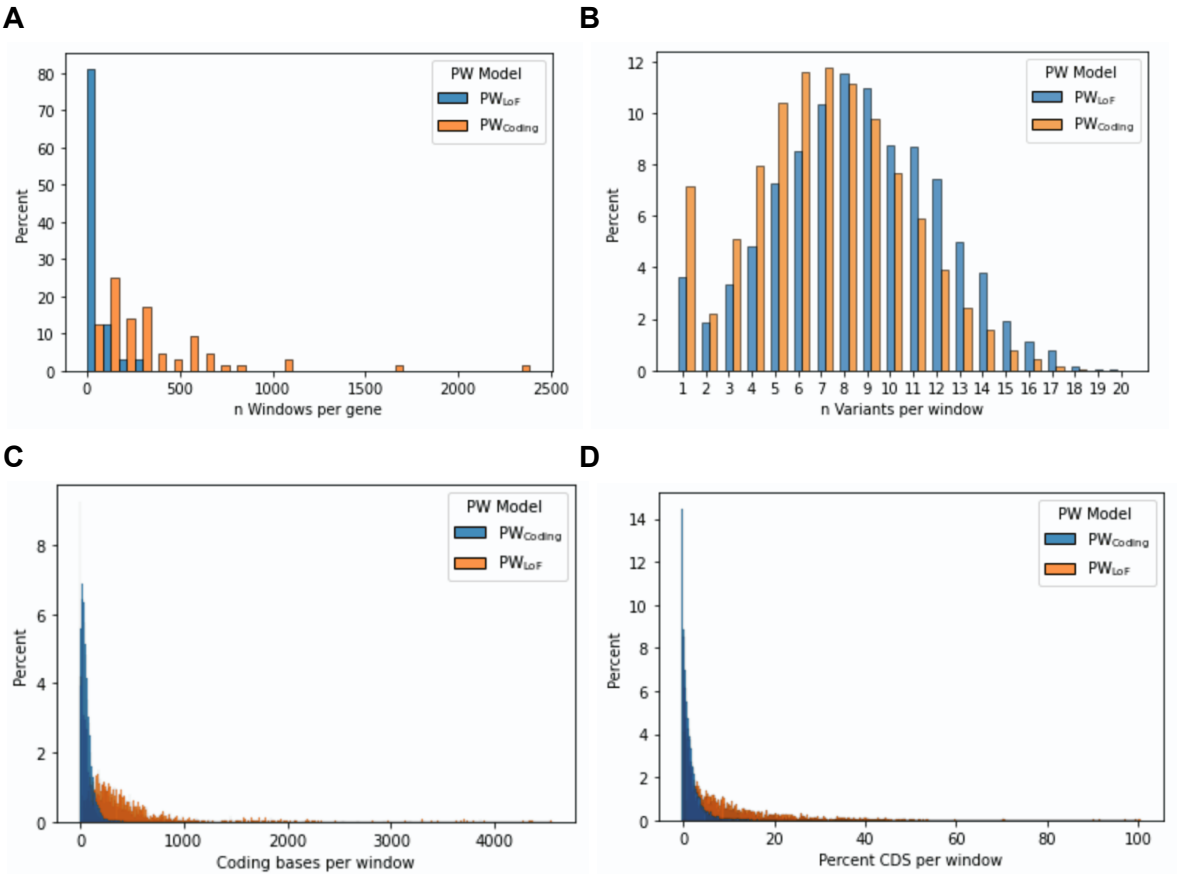
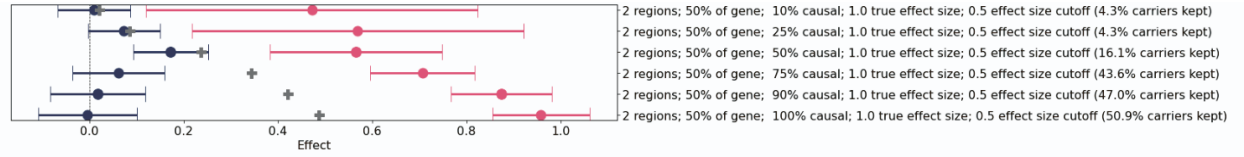
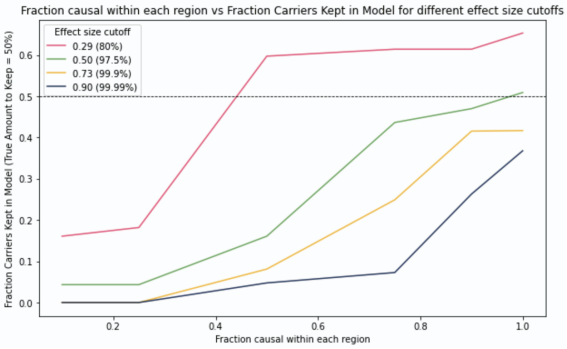


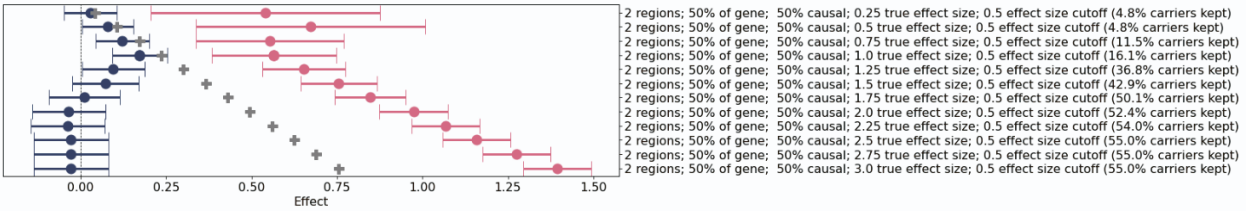
Figure S1. Stats for power windows. Because of the size of *TTN*, the stats for this gene are not displayed and instead are written separately in the legend. A) Histogram of number of windows analyzed per gene. The mean number of windows analyzed per gene was 357 / 67 for PW_{Coding} / PW_{LoF} (range 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**

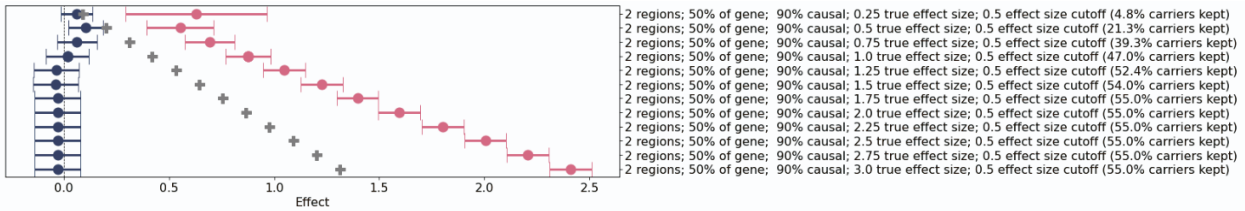




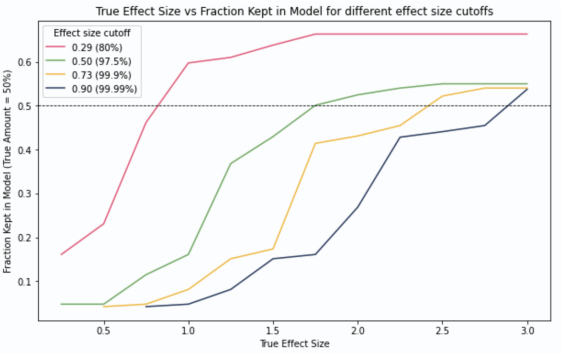
B **Different effect sizes: 50% causal variants within 2 regions**



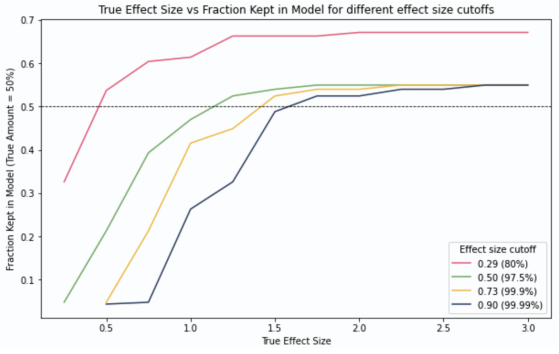
Different effect sizes: 90% causal variants within 2 regions



50% causal variants within 2 regions

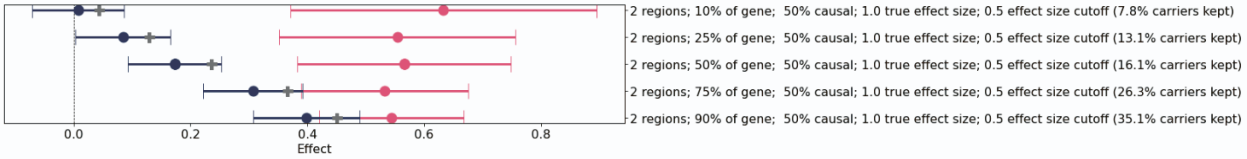


90% causal variants within 2 regions

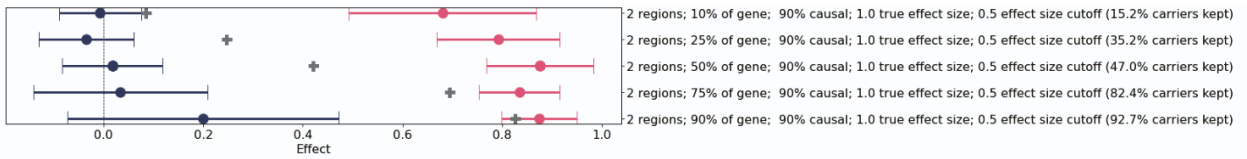


C

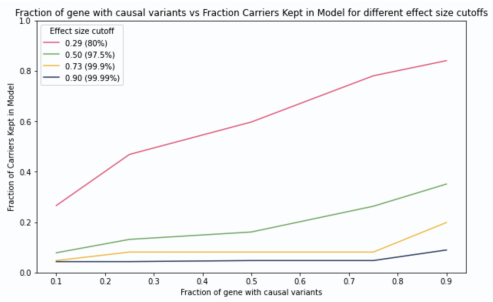
Different percents of gene with causal variants: 2 regions in which 50% are causal variant



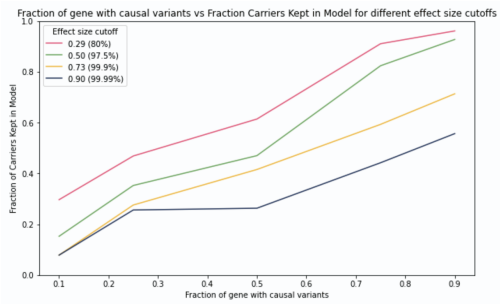
Different percents of gene with causal variants: 2 regions in which 90% are causal variants



50% causal variants within 2 regions

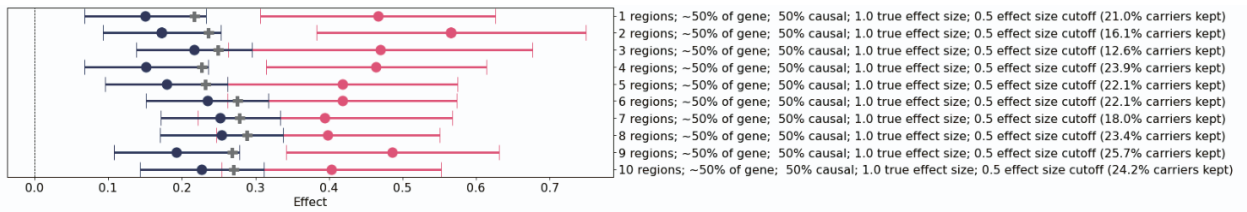


90% causal variants within 2 regions

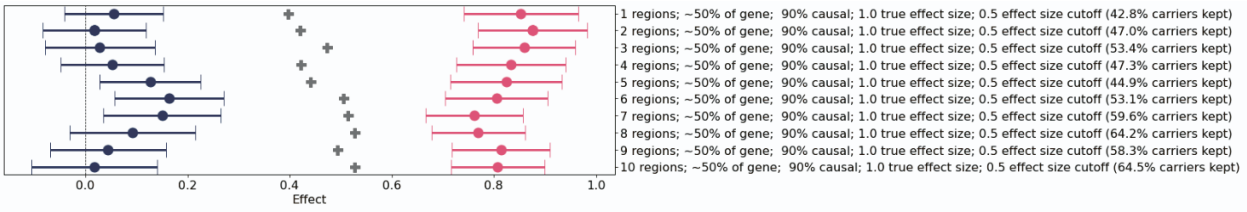


D

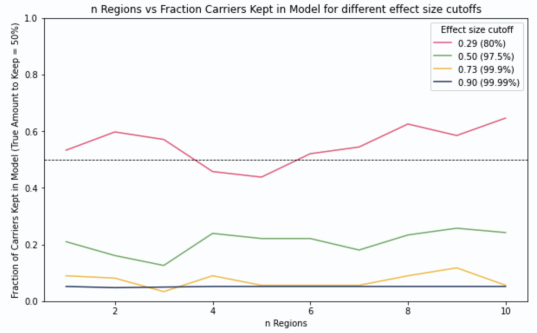
Different number of causal regions per gene: 50% causal variants in each region (effect size 1)



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

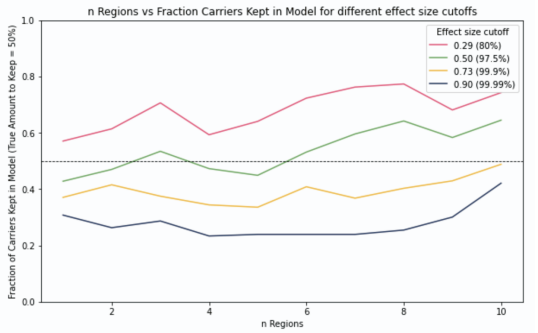


Figure S2. Simulations showing the effect of tuning different parameters for Power Window in the example gene *GCK* and phenotype glucose with a coding model. For the simulations, causal regions were defined within *GCK* and glucose levels for each carrier were reassigned according to the parameters 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.

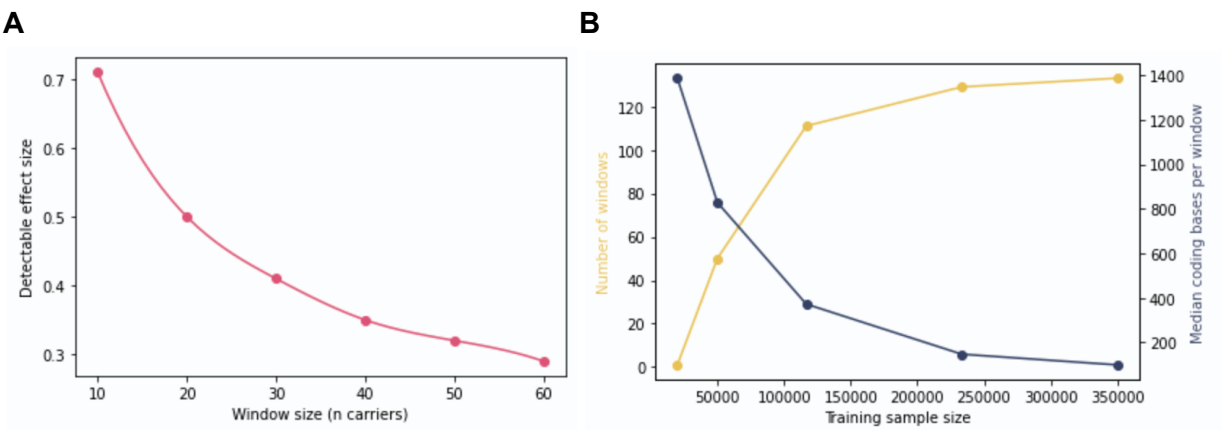


Figure S3. Properties of Power Window size. **A)** For a quantitative trait in a training set of 350k samples, the window size (number of carriers in the window) is compared to the effect size for which there is power to identify (with 97.5% confidence) that the beta is not 0. **B)** For the example 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.

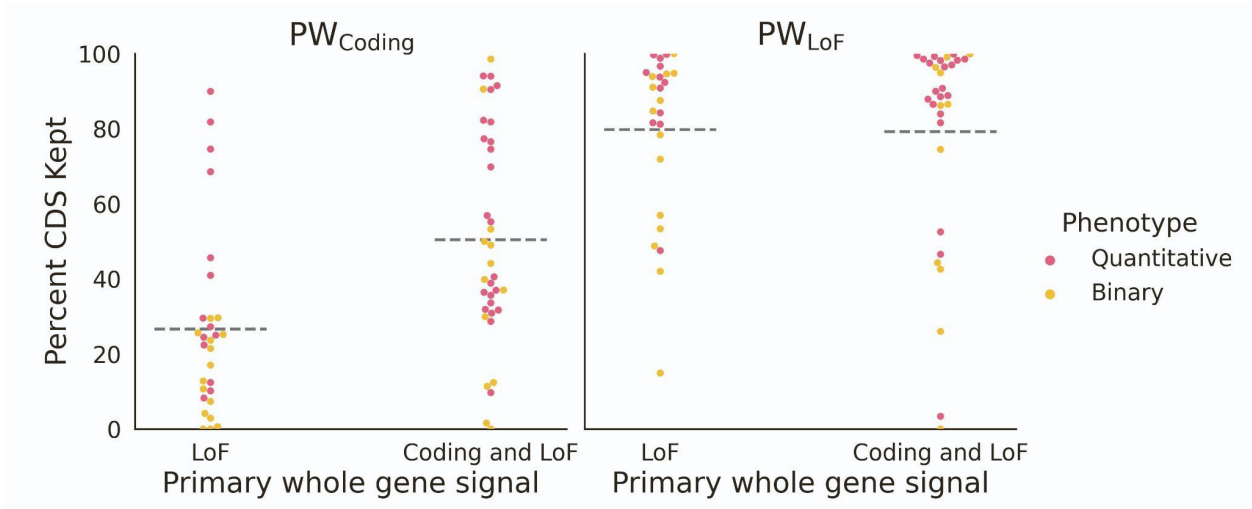
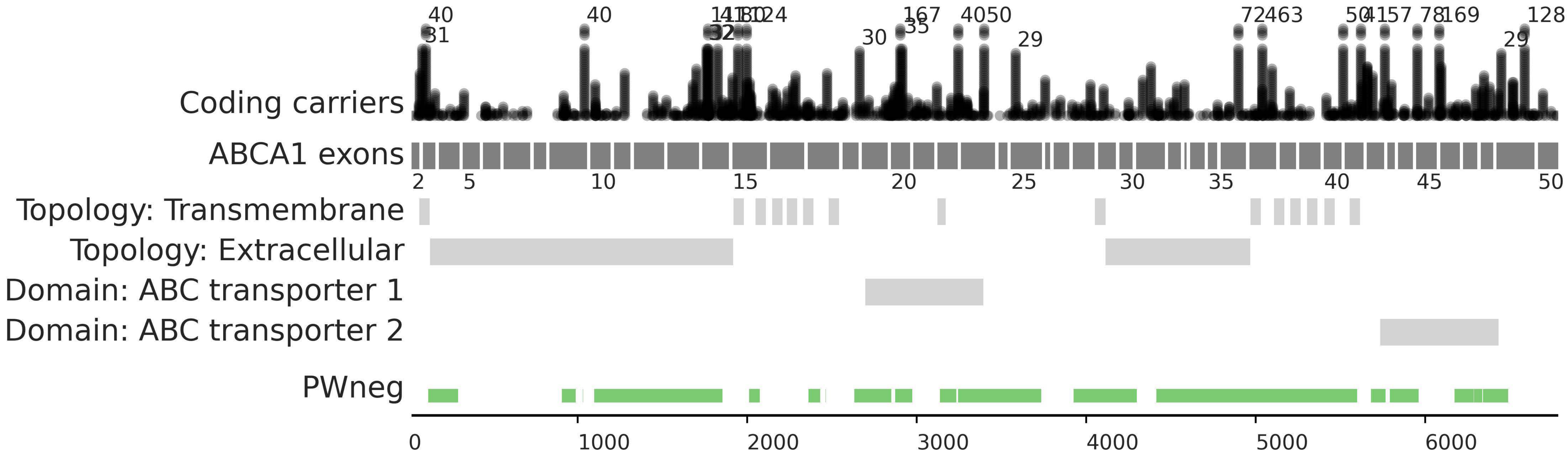
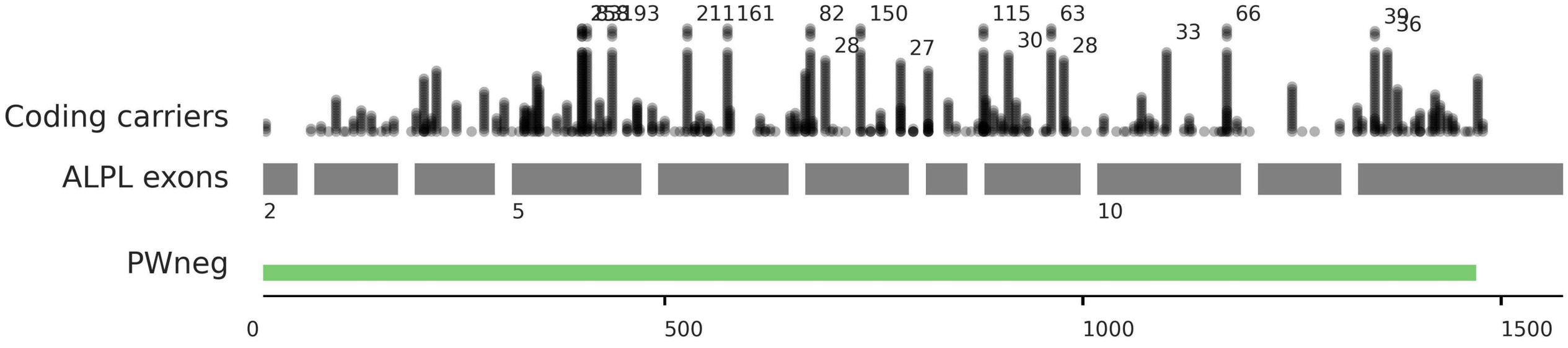
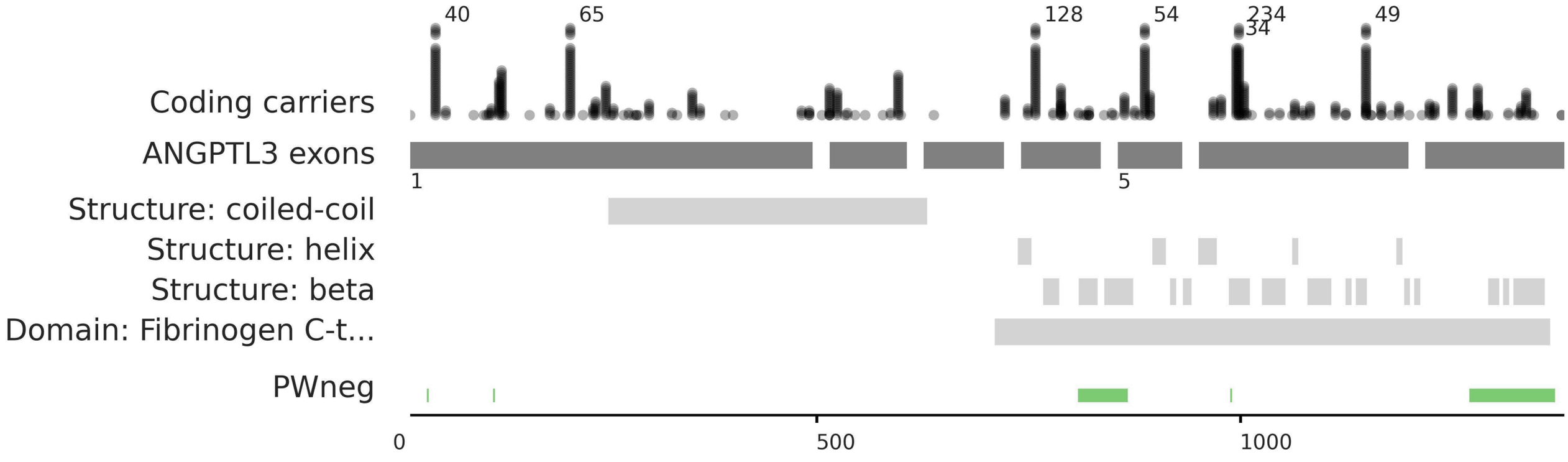


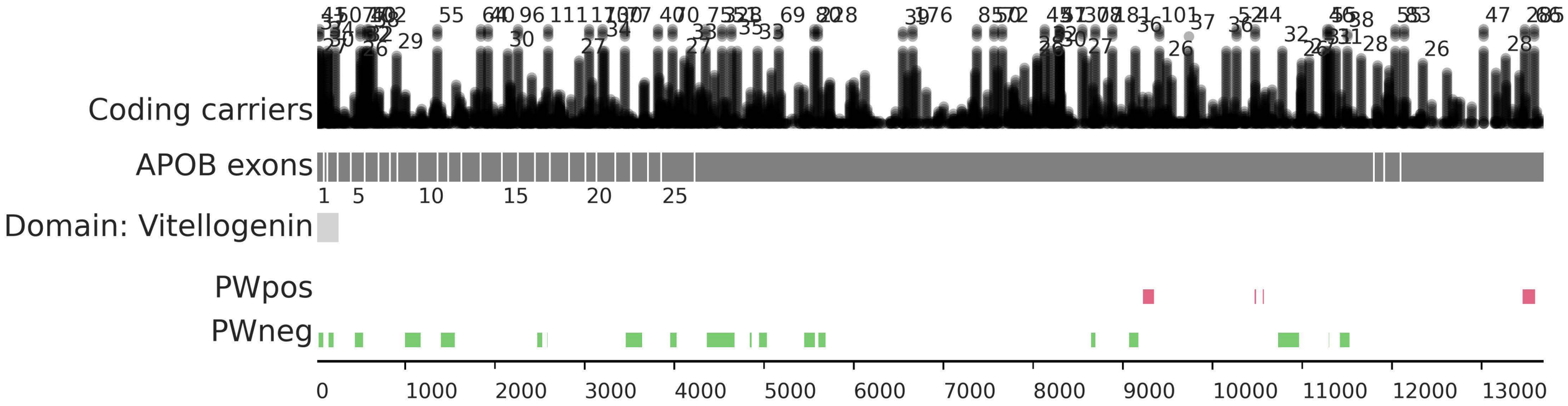
Figure S4. Percent CDS kept in Power Window models. For each gene (n=65), the percent of the CDS for the gene that was retained by the PW model was evaluated for (A) PW_{Coding} and (B) PW_{LoF}. Within each model, the phenotypes were grouped into quantitative (pink) or binary (yellow) traits, and the 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.

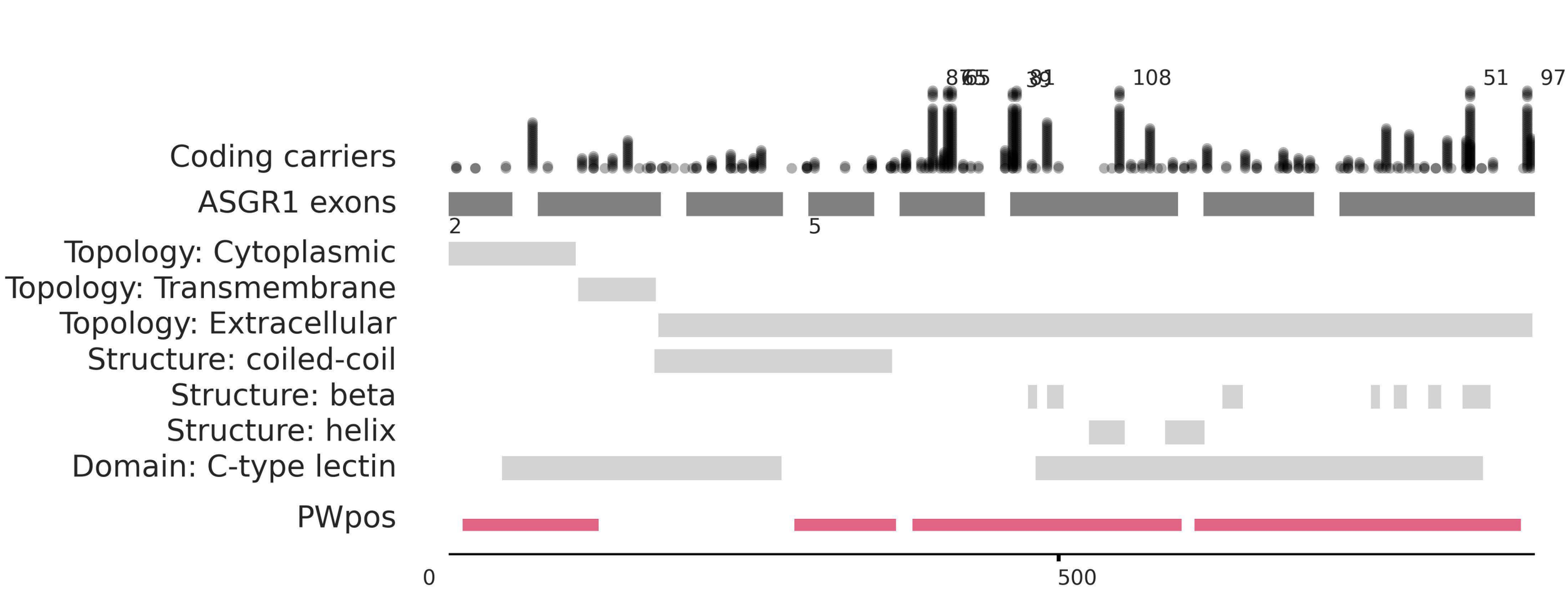


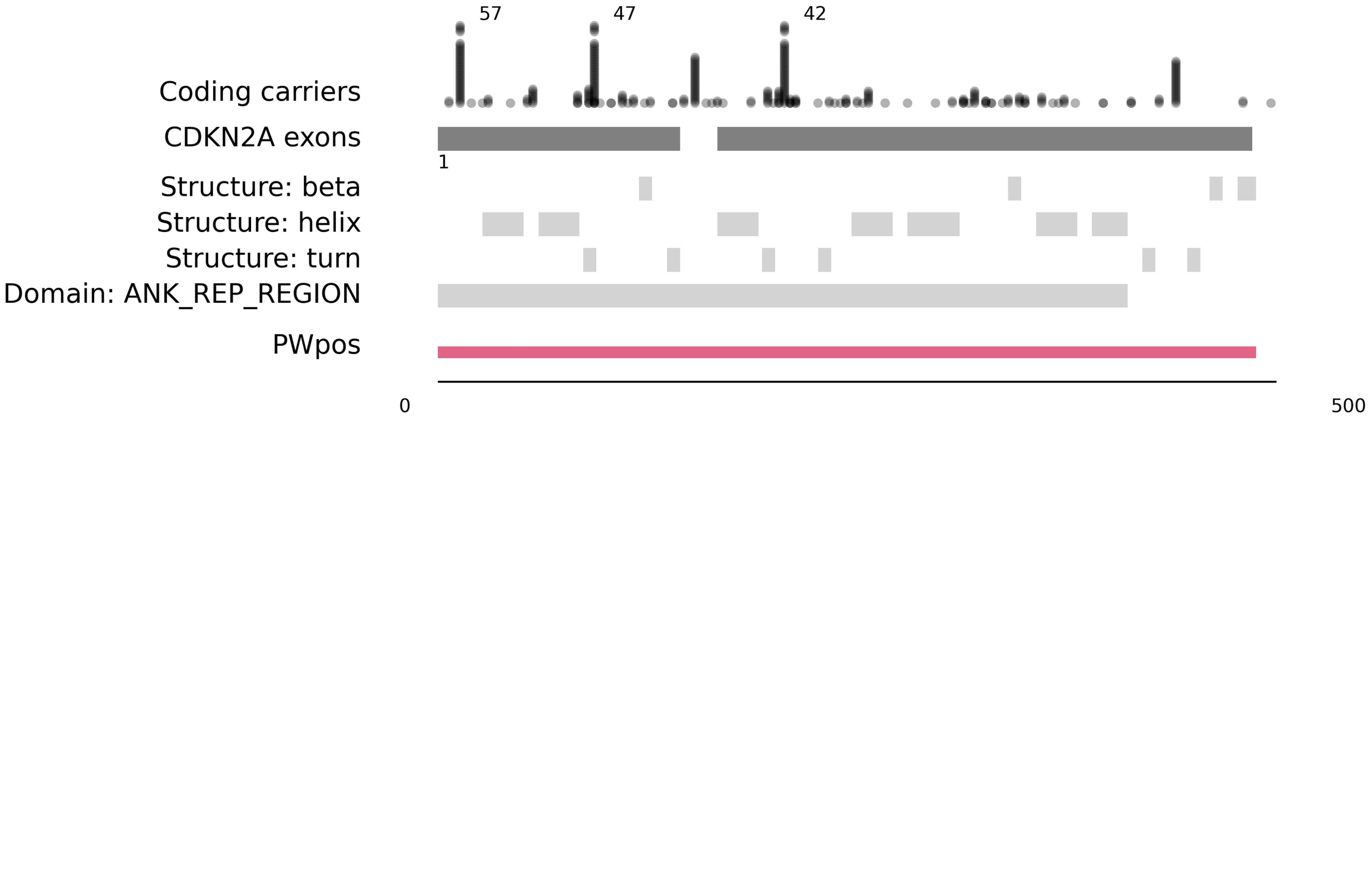
Coding carriers
ABCA1 exons
Topology: Transmembrane
Topology: Extracellular
Domain: ABC transporter 1
Domain: ABC transporter 2
PWneg
0
1000
2000
3000
4000
5000
6000

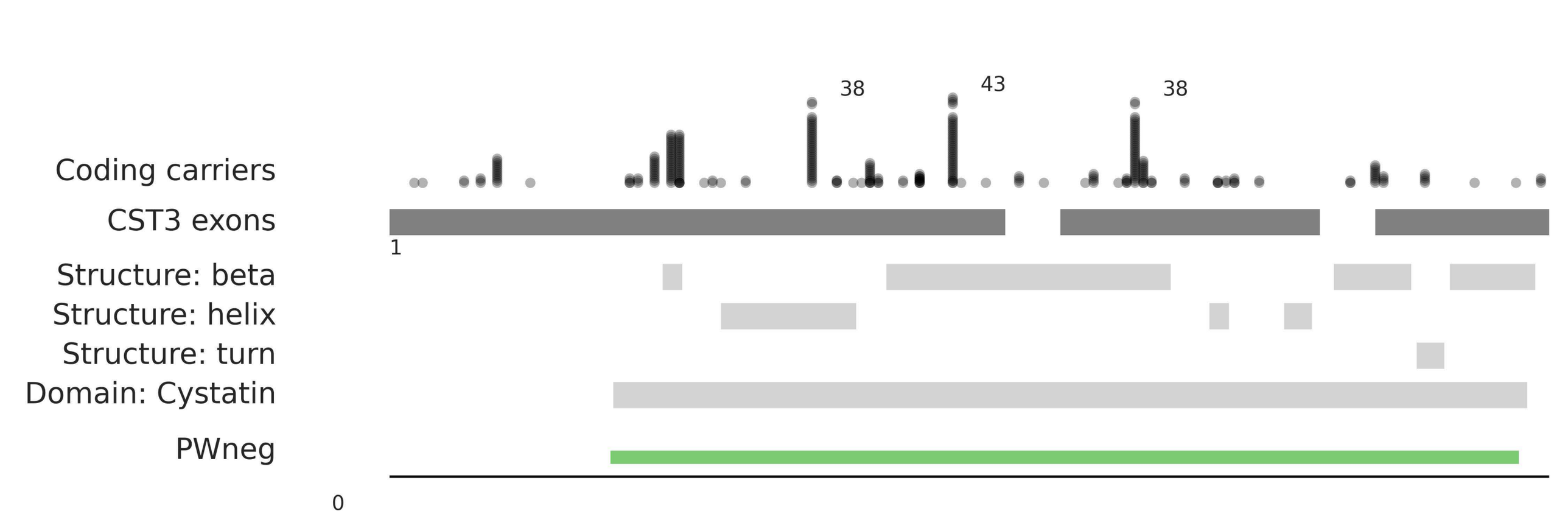

Coding carriers
ALPL exons
PWneg
2
5
10
0
500
1000
1500

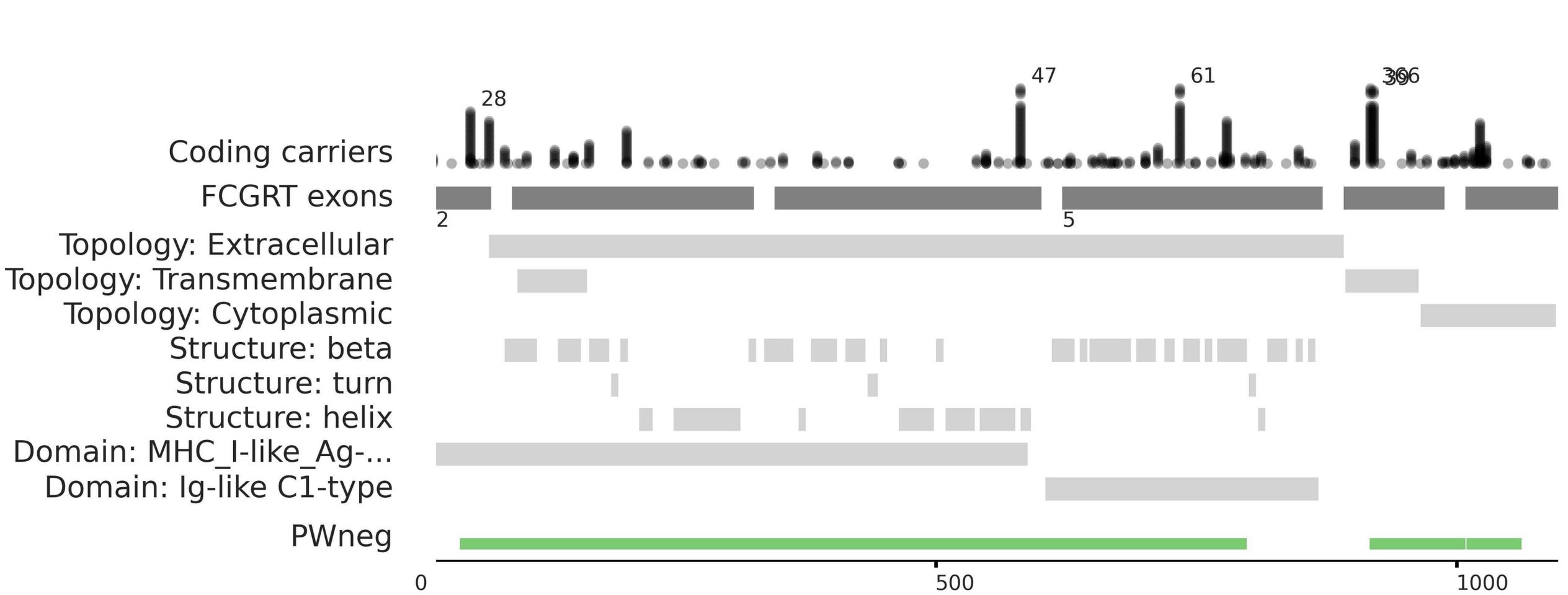

40
65
128
54
234
34
49
Coding carriers
ANGPTL3 exons
1
5
Structure: coiled-coil
Structure: helix
Structure: beta
Domain: Fibrinogen C-t...
PWneg
0
500
1000

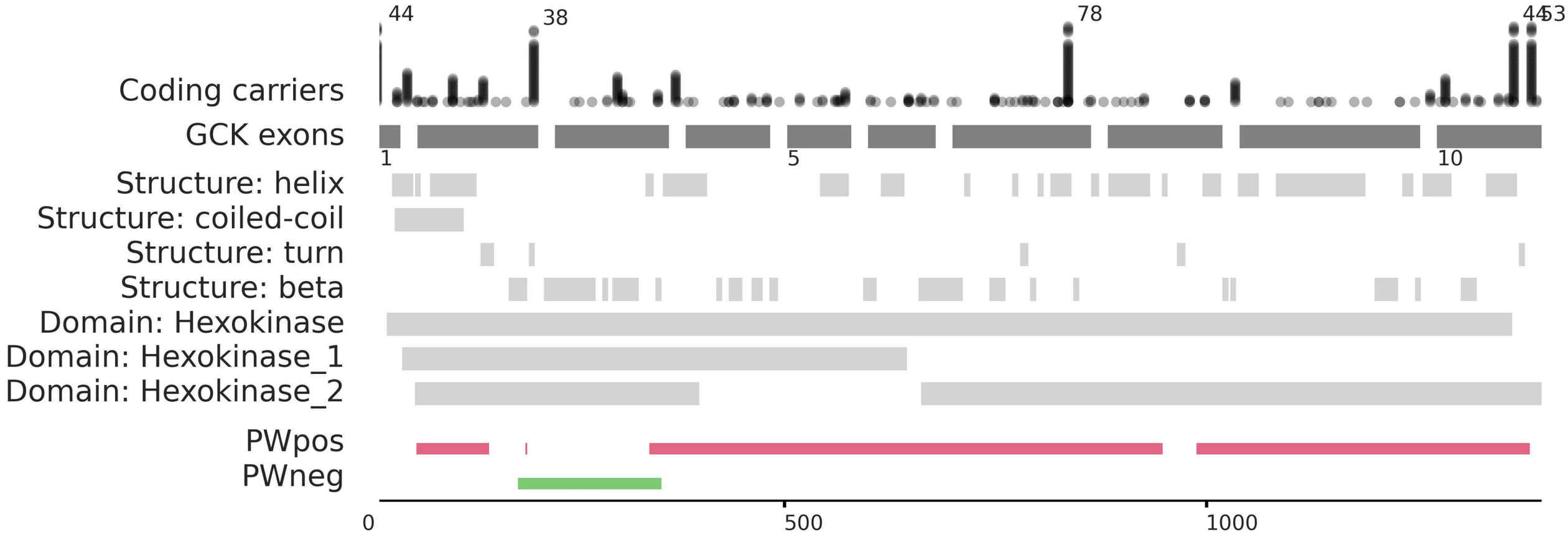

Coding carriers
APOB exons
1 5 10 15 20 25
Domain: Vitellogenin
PWpos
PWneg
0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000

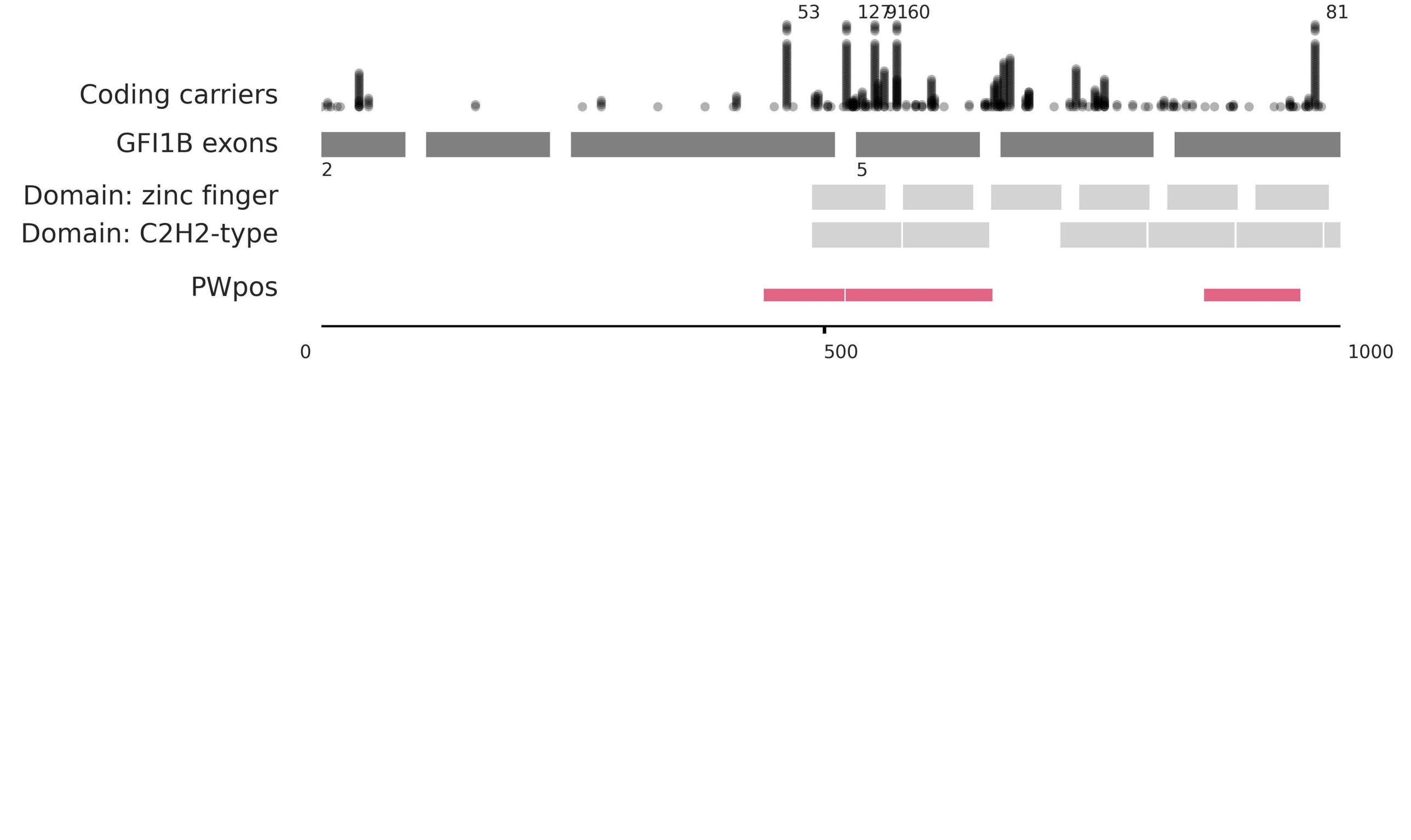

Coding carriers
ASGR1 exons
Topology: Cytoplasmic
Topology: Transmembrane
Topology: Extracellular
Structure: coiled-coil
Structure: beta
Structure: helix
Domain: C-type lectin
PWpos
8
7
6
5
3
9
1
108
51
97
2
5
0
500

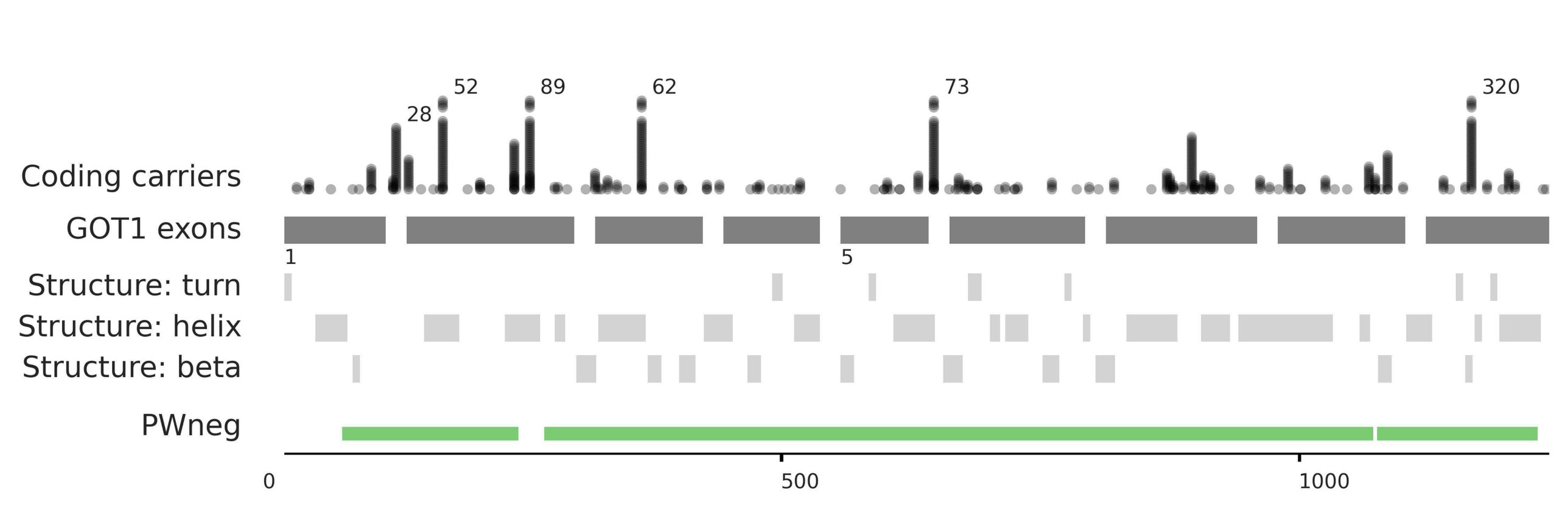

57
47
42
Coding carriers
CDKN2A exons
1
Structure: beta
Structure: helix
Structure: turn
Domain: ANK_REP_REGION
PWpos
0
500

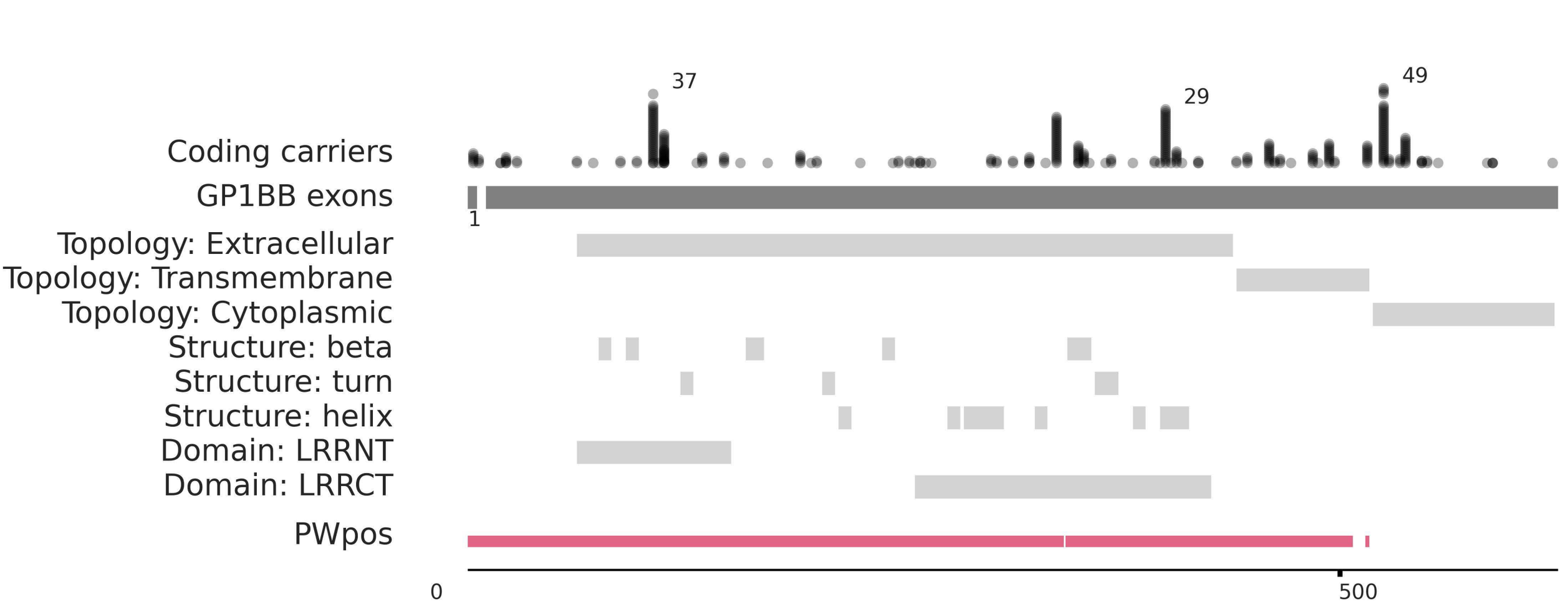

38
43
38
Coding carriers
CST3 exons
1
Structure: beta
Structure: helix
Structure: turn
Domain: Cystatin
PWneg
0

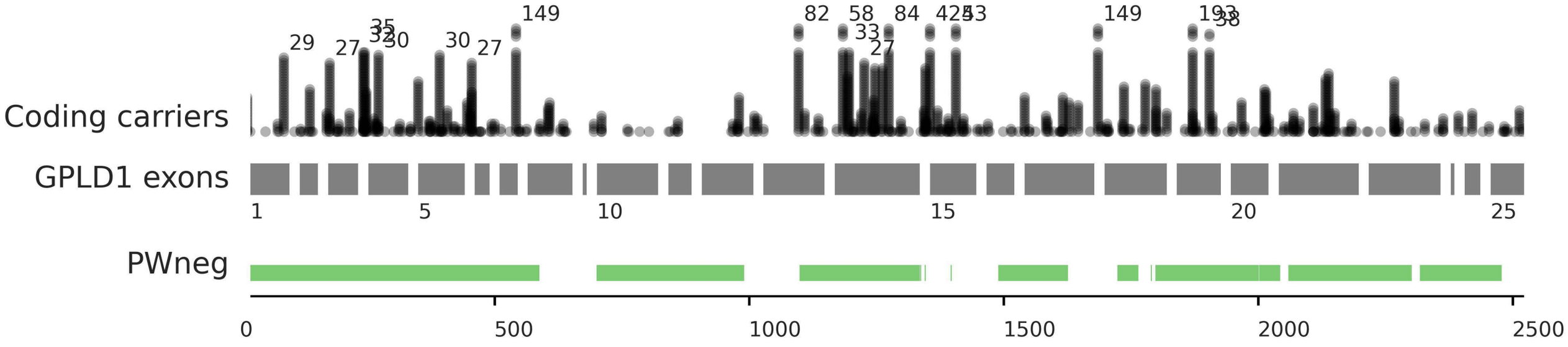

Coding carriers
FCGRT exons
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: turn
Structure: helix
Domain: MHC_I-like_Ag-...
Domain: Ig-like C1-type
PWneg
28
47
61
366
2
5
0
500
1000

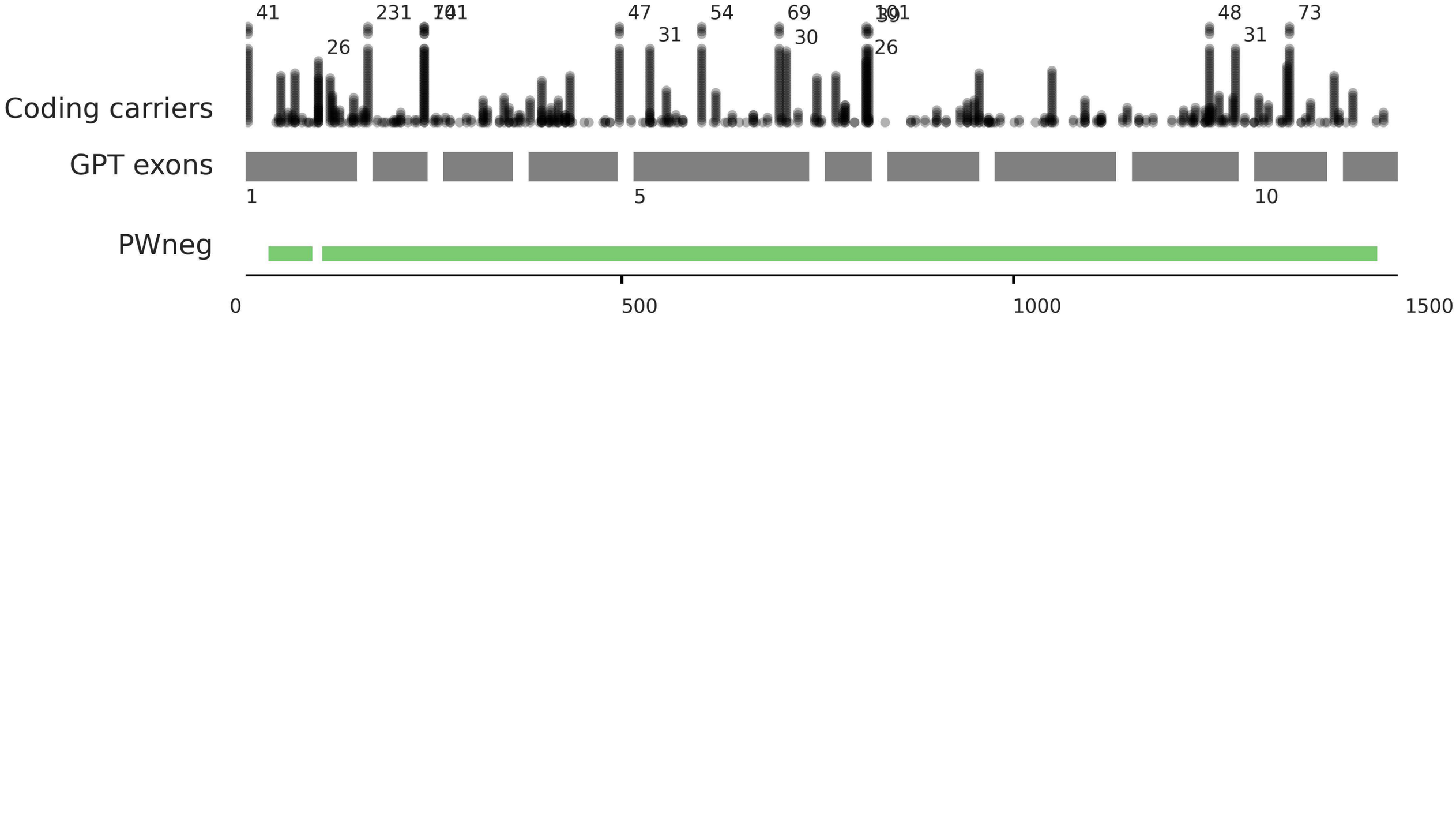

Coding carriers
GCK exons
Structure: helix
Structure: coiled-coil
Structure: turn
Structure: beta
Domain: Hexokinase
Domain: Hexokinase_1
Domain: Hexokinase_2
PWpos
PWneg
44
38
78
445
53
1
5
10
0
500
1000

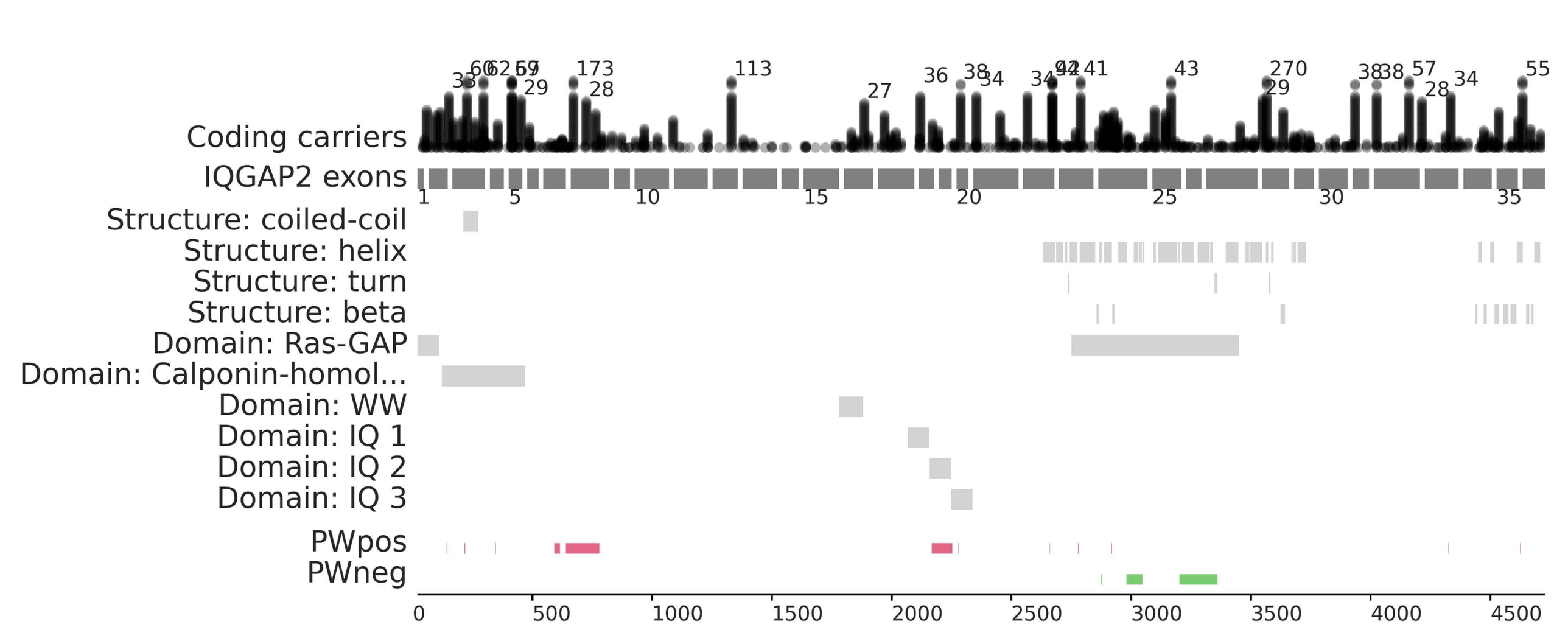

53
127
9160
81
Coding carriers
GFI1B exons
2
5
Domain: zinc finger
Domain: C2H2-type
PWpos
0
500
1000

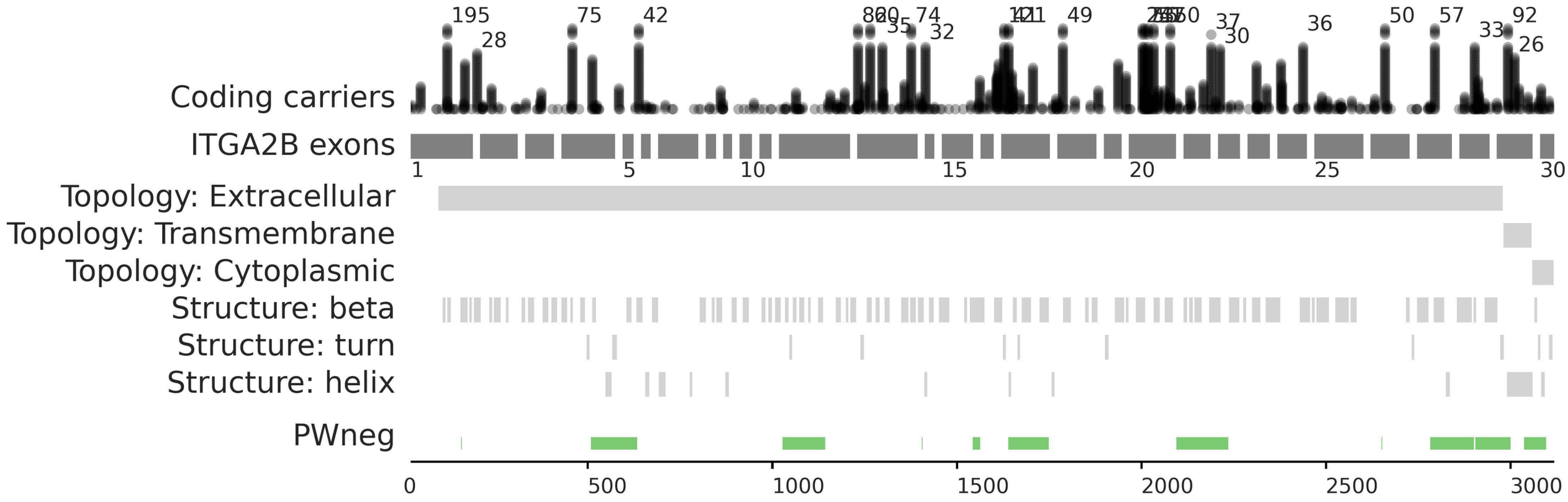

Coding carriers
GOT1 exons
Structure: turn
Structure: helix
Structure: beta
PWneg
28
52
89
62
73
320
1
5
0
500
1000

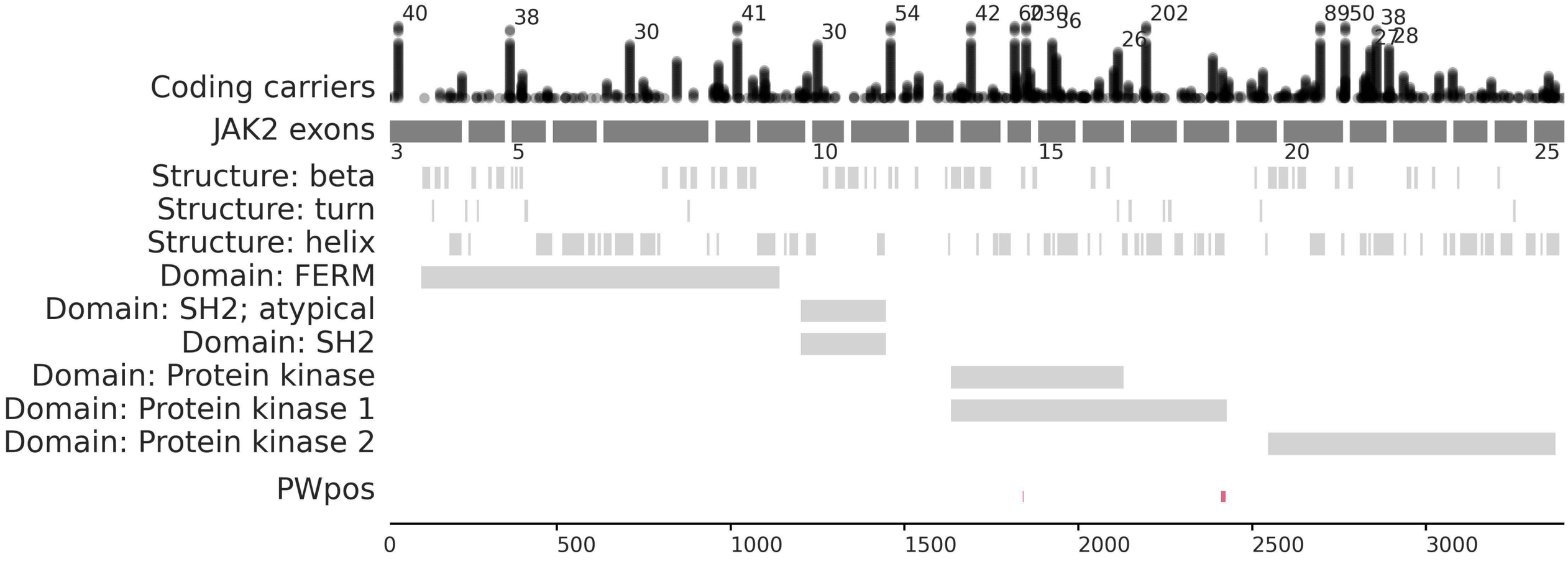

Coding carriers
GP1BB exons
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: turn
Structure: helix
Domain: LRRNT
Domain: LRRCT
PWpos
37
29
49
1
0
500

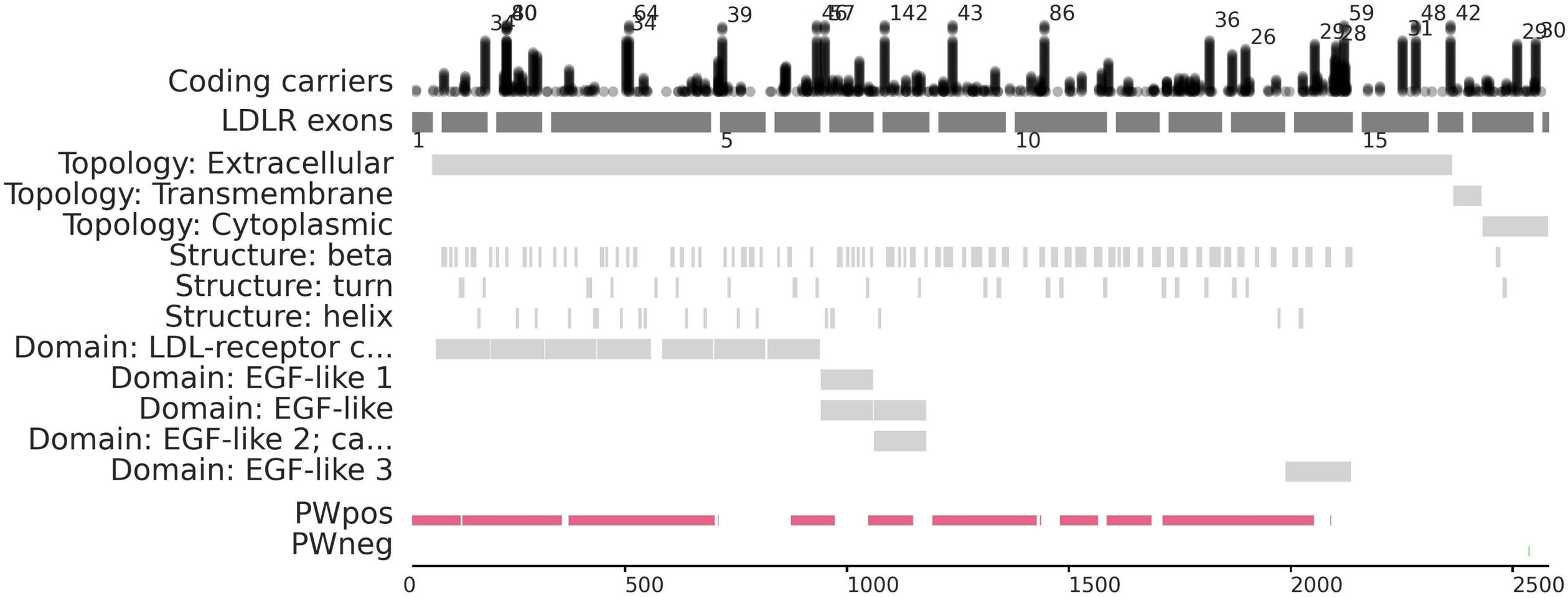

Coding carriers
GPLD1 exons
PWneg
29
27
32
35
30
30
27
149
82
58
33
84
27
42
43
149
193
38
1
5
10
15
20
25
0
500
1000
1500
2000
2500

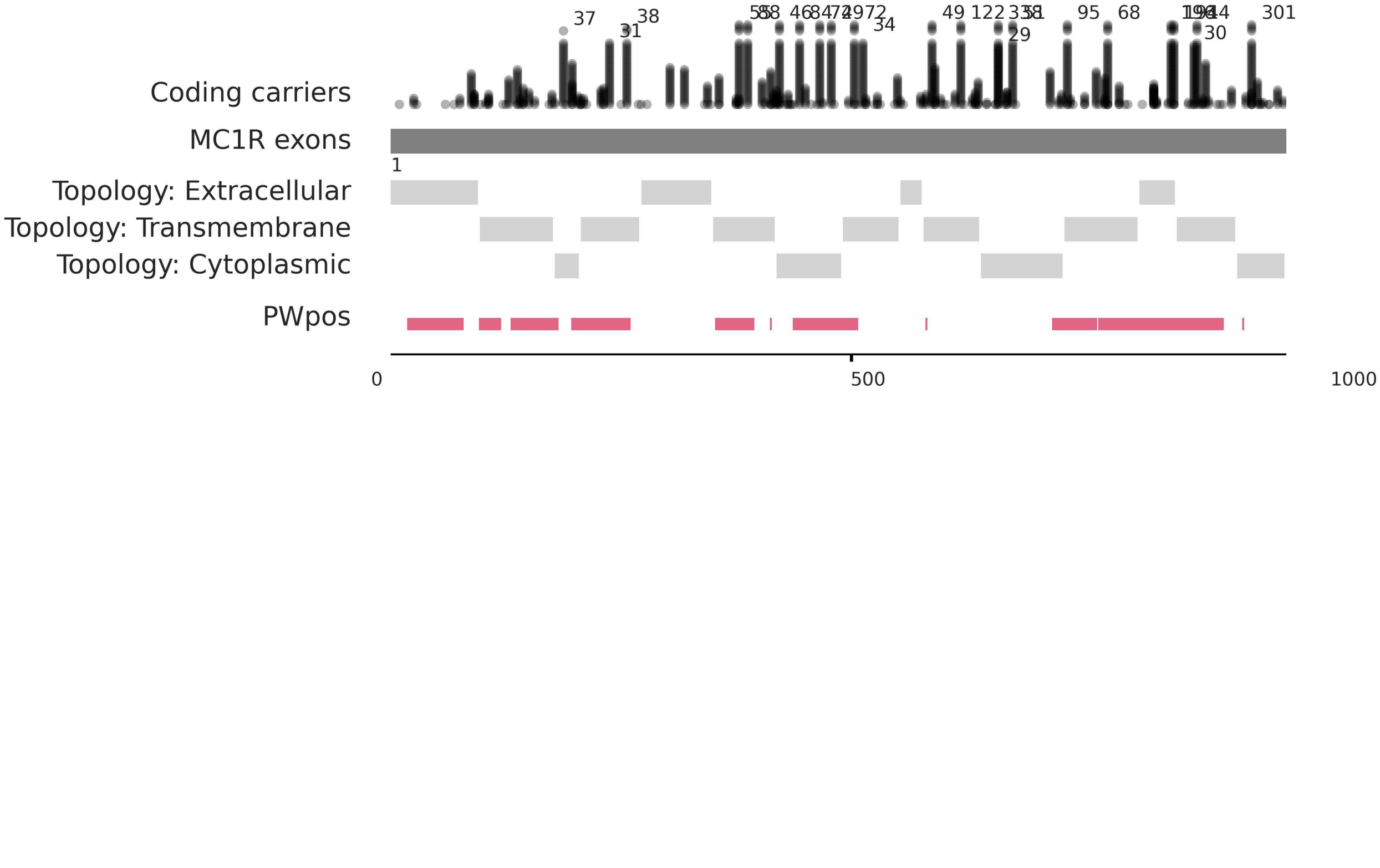

Coding carriers
GPT exons
PWneg
41
26
231
14
101
47
31
54
69
30
101
39
26
48
31
73
1
5
10
0
500
1000
1500

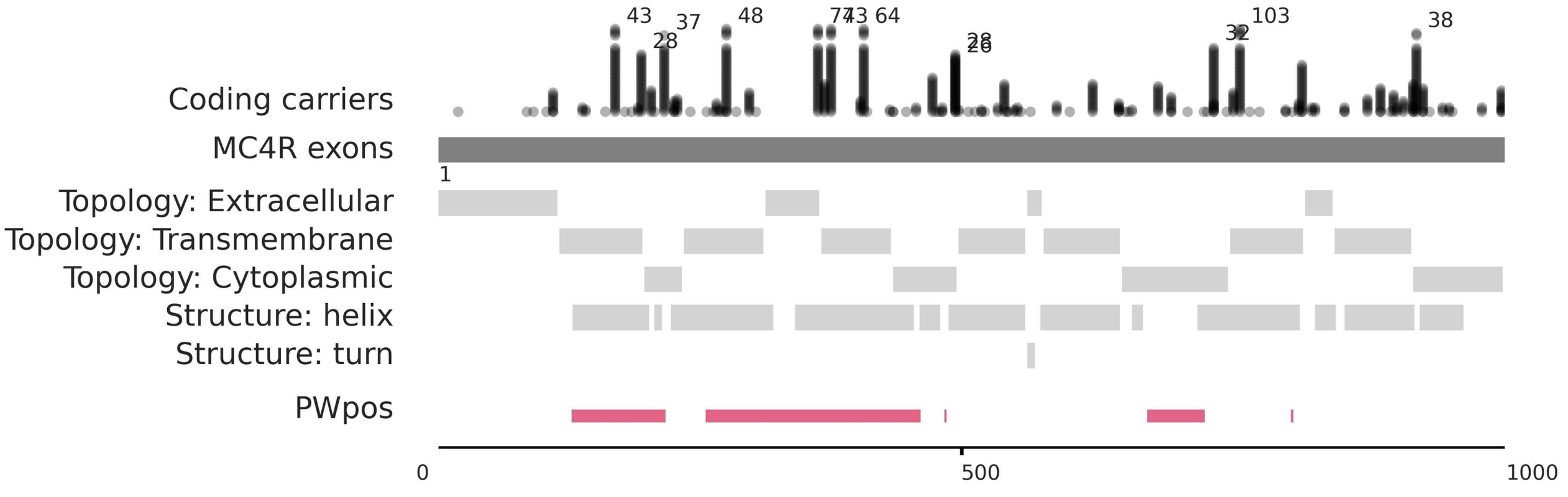

Coding carriers
IQGAP2 exons
Structure: coiled-coil
Structure: helix
Structure: turn
Structure: beta
Domain: Ras-GAP
Domain: Calponin-homol...
Domain: WW
Domain: IQ 1
Domain: IQ 2
Domain: IQ 3
PWpos
PWneg
1
5
10
15
20
25
30
35
0
500
1000
1500
2000
2500
3000
3500
4000
4500

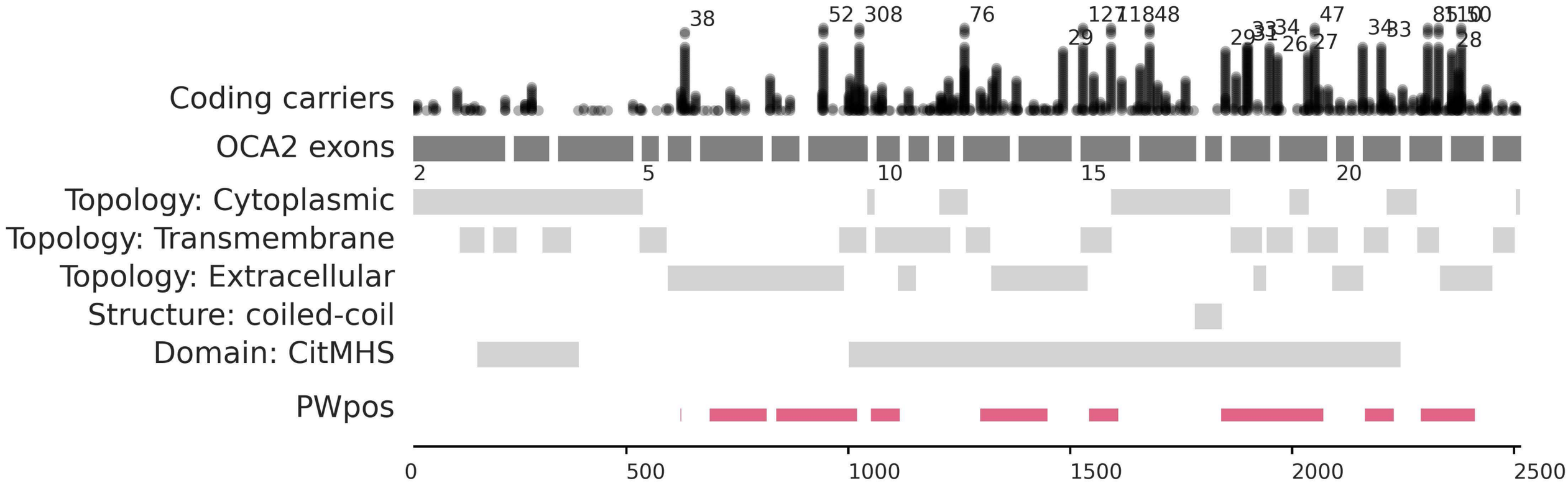

Coding carriers
ITGA2B exons
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: turn
Structure: helix
PWneg
195
28
75
42
74
32
49
37
30
36
50
57
33
92
26
1
5
10
15
20
25
30
0
500
1000
1500
2000
2500
3000

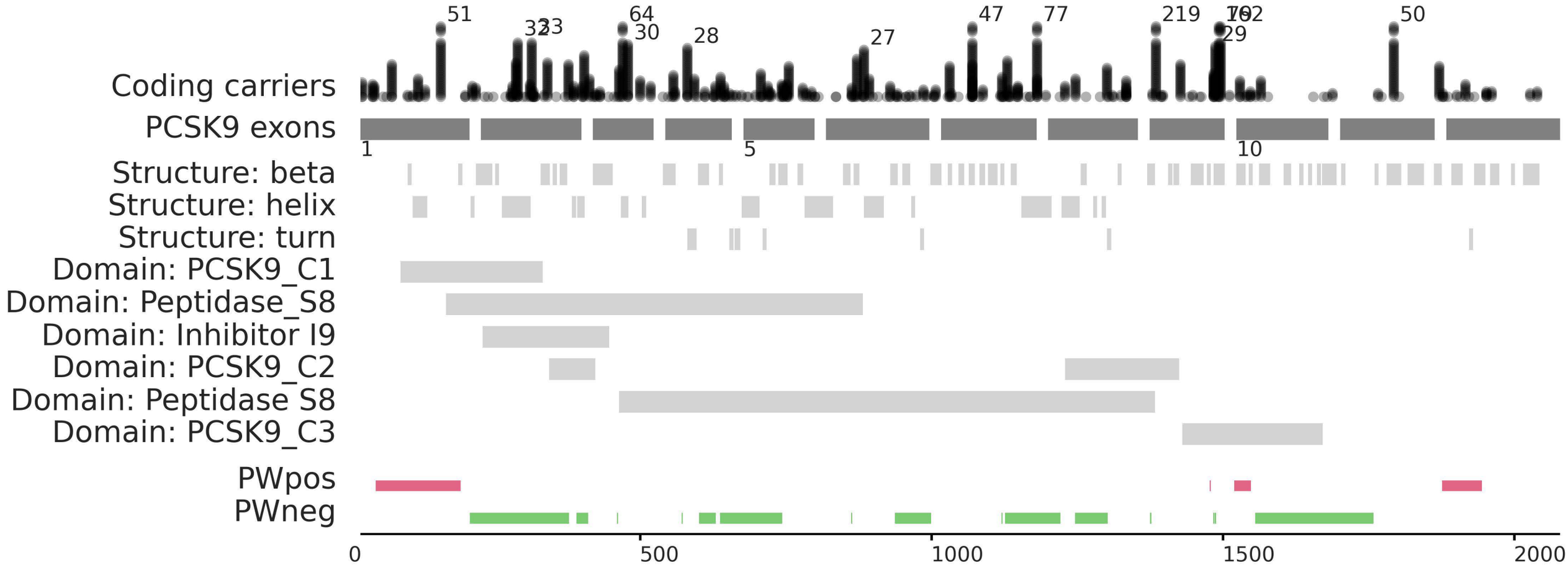

Coding carriers
JAK2 exons
Structure: beta
Structure: turn
Structure: helix
Domain: FERM
Domain: SH2; atypical
Domain: SH2
Domain: Protein kinase
Domain: Protein kinase 1
Domain: Protein kinase 2
PWpos
40
38
30
41
30
54
42
62
30
36
26
202
89
50
38
27
28
3
5
10
15
20
25
0
500
1000
1500
2000
2500
3000

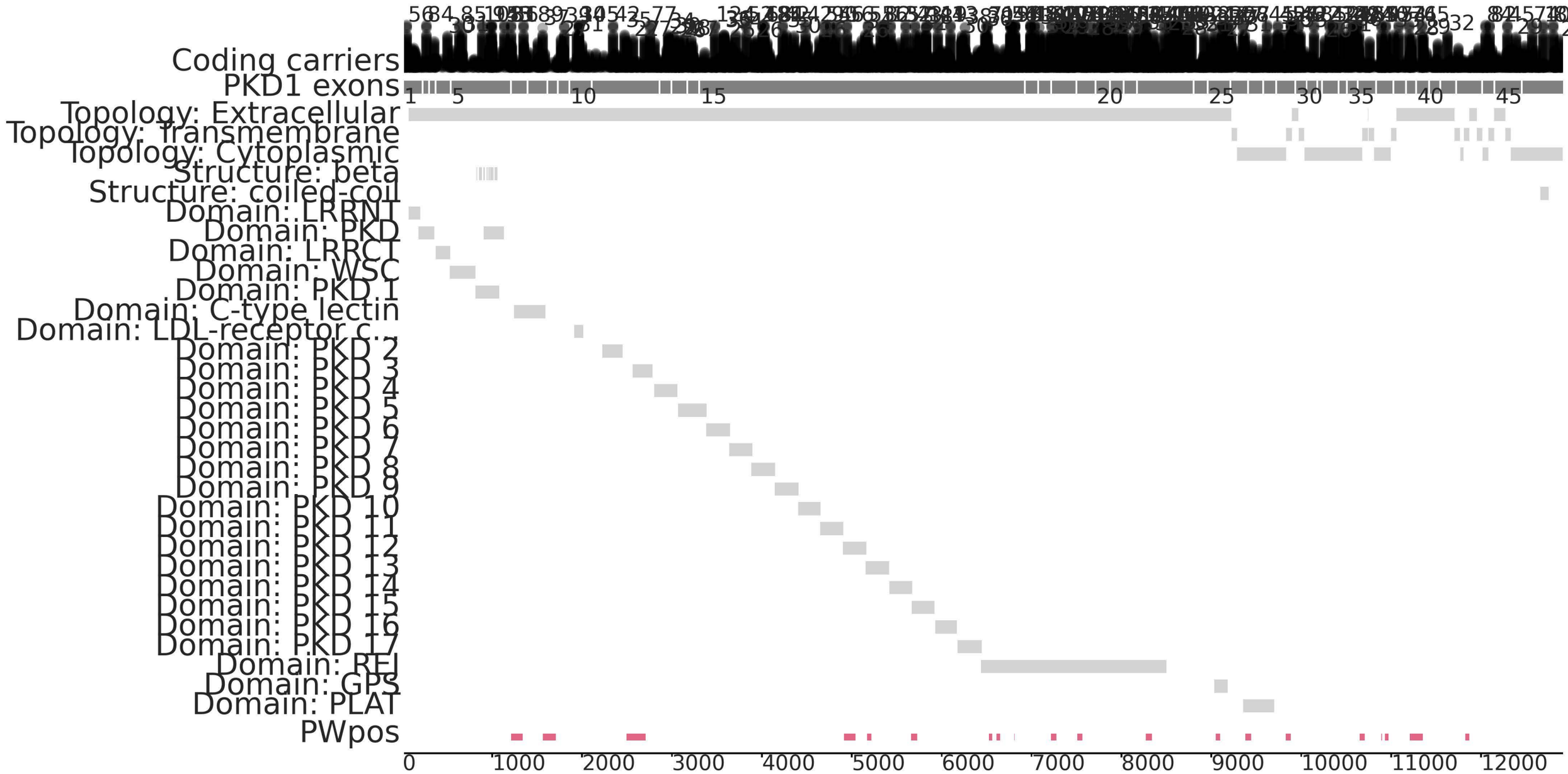

Coding carriers
LDLR exons
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: turn
Structure: helix
Domain: LDL-receptor c...
Domain: EGF-like 1
Domain: EGF-like
Domain: EGF-like 2; ca...
Domain: EGF-like 3
PWpos
PWneg
1
5
10
15
0
500
1000
1500
2000
2500

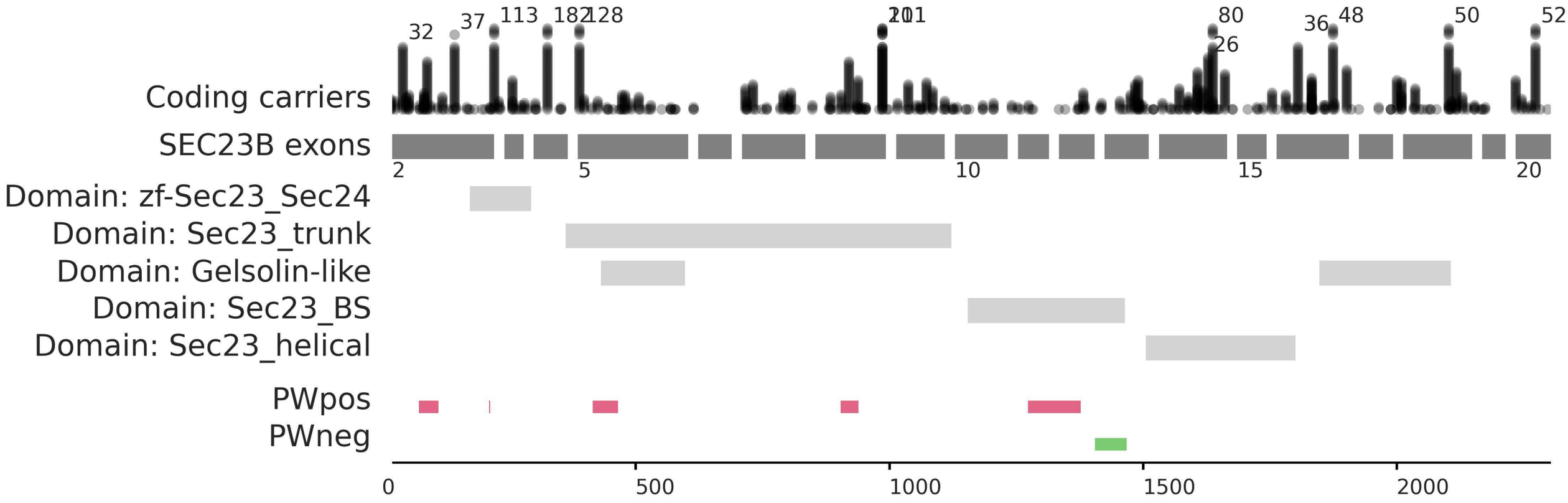

Coding carriers
MC1R exons
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
PWpos
1
0
500
1000

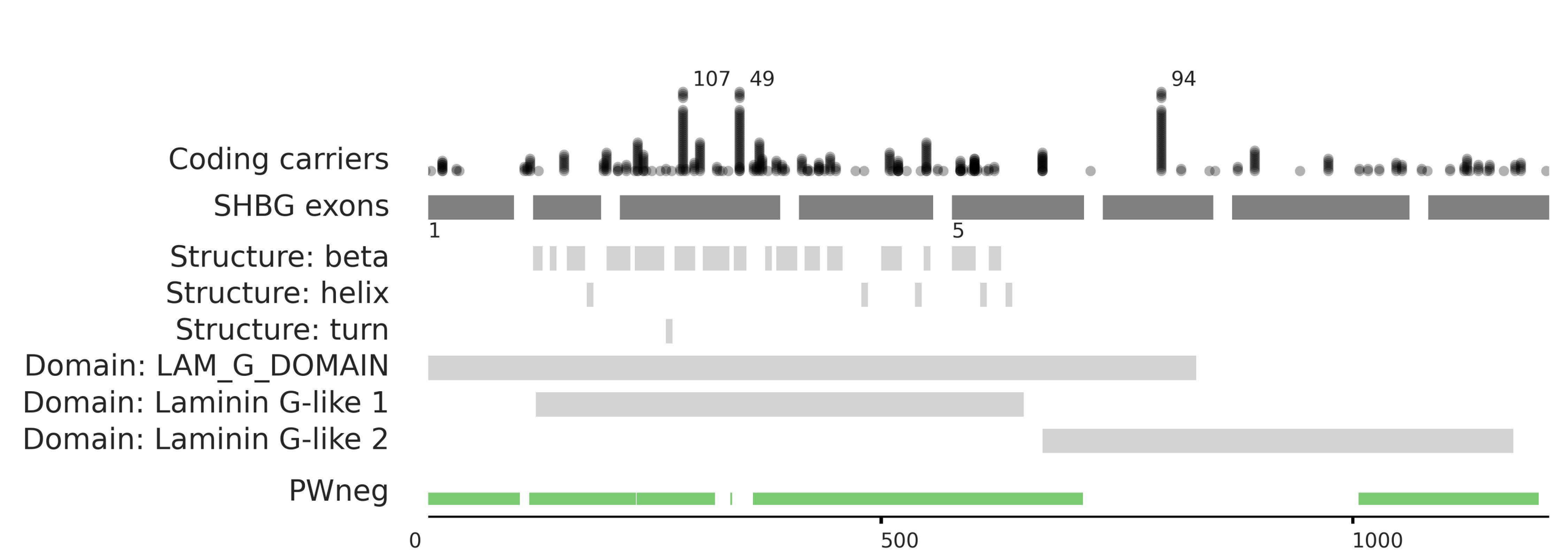

Coding carriers
MC4R exons
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: helix
Structure: turn
PWpos
43
37
28
48
74
43
64
38
103
32
1
0
500
1000

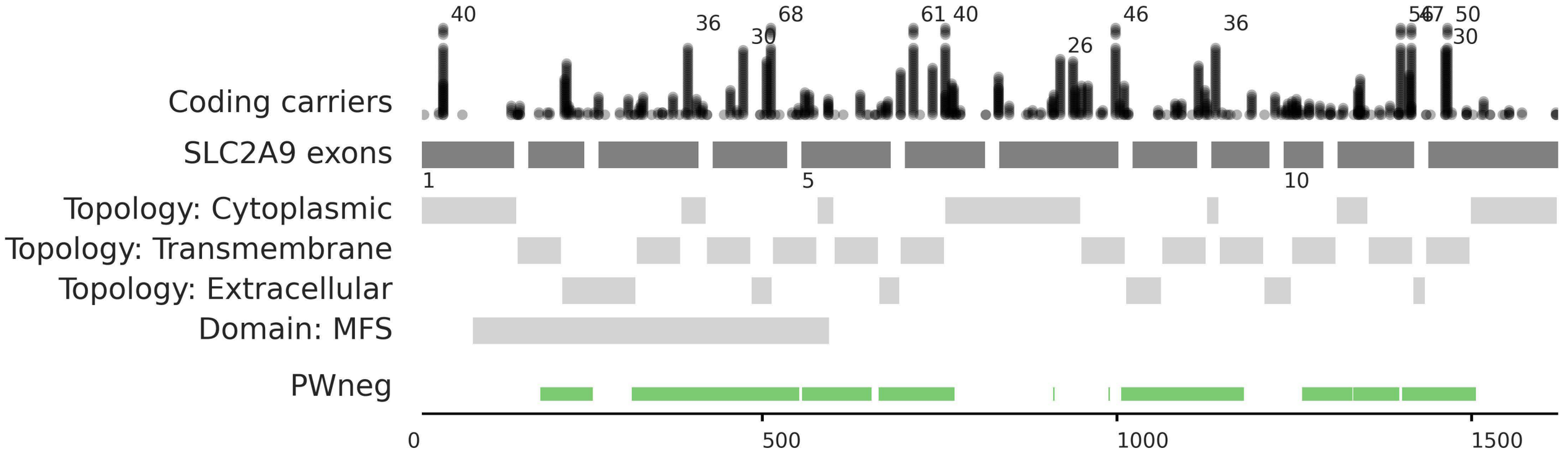

Coding carriers
OCA2 exons
Topology: Cytoplasmic
Topology: Transmembrane
Topology: Extracellular
Structure: coiled-coil
Domain: CitMHS
PWpos
38
52
308
76
127
118
48
29
47
34
33
85
515
50
28
2
5
10
15
20
0
500
1000
1500
2000
2500

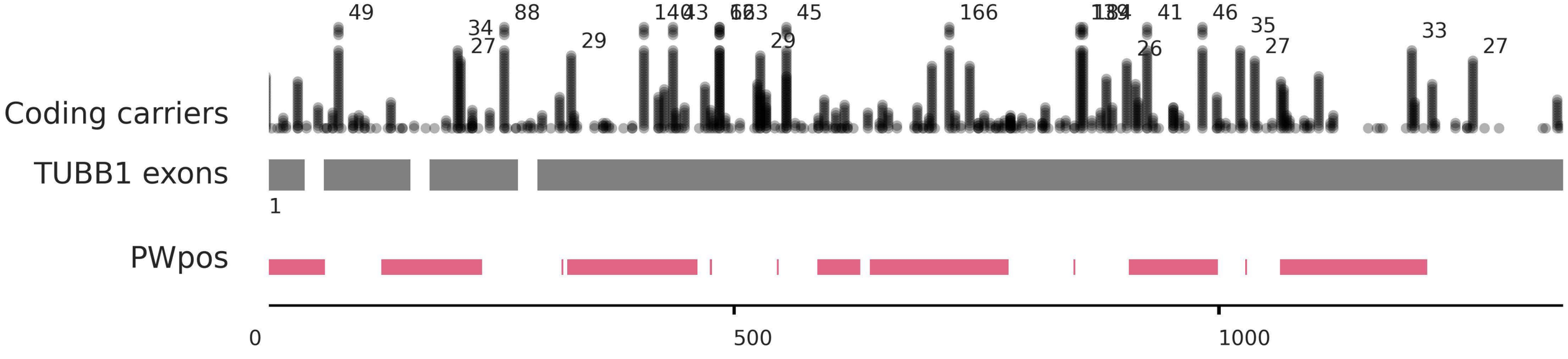

Coding carriers
PCSK9 exons
Structure: beta
Structure: helix
Structure: turn
Domain: PCSK9_C1
Domain: Peptidase_S8
Domain: Inhibitor I9
Domain: PCSK9_C2
Domain: Peptidase S8
Domain: PCSK9_C3
PWpos
PWneg
51
32
23
64
30
28
27
47
77
219
70
29
102
50
1
5
10
0
500
1000
1500
2000

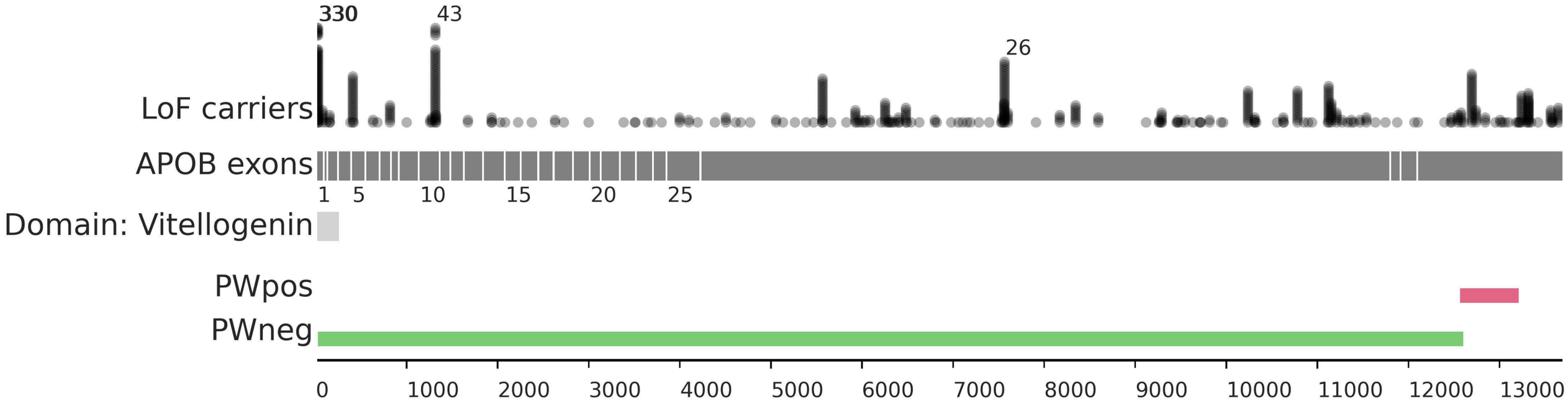

Coding carriers
PKD1 exons
Topology: Extracellular
Topology: Transmembrane
Topology: Cytoplasmic
Structure: beta
Structure: coiled-coil
Domain: LRRNT
Domain: PKD
Domain: LRRCT
Domain: WSC
Domain: PKD 1
Domain: C-type lectin
Domain: LDL-receptor c...
Domain: PKD 2
Domain: PKD 3
Domain: PKD 4
Domain: PKD 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
1
5
10
15
20
25
30
35
40
45
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000

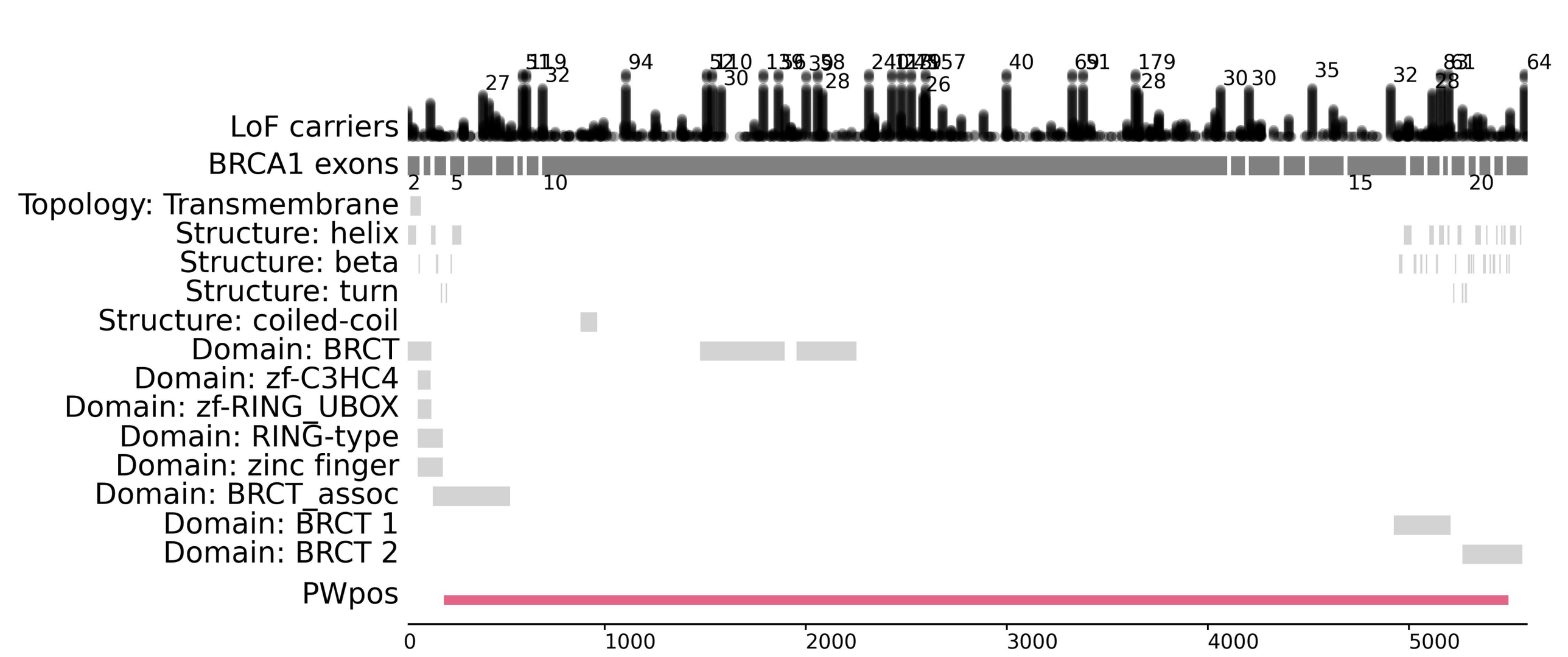

Coding carriers
SEC23B exons
Domain: zf-Sec23_Sec24
Domain: Sec23_trunk
Domain: Gelsolin-like
Domain: Sec23_BS
Domain: Sec23_helical
PWpos
PWneg
32
37
113
182
128
201
80
26
36
48
50
52
2
5
10
15
20
0
500
1000
1500
2000

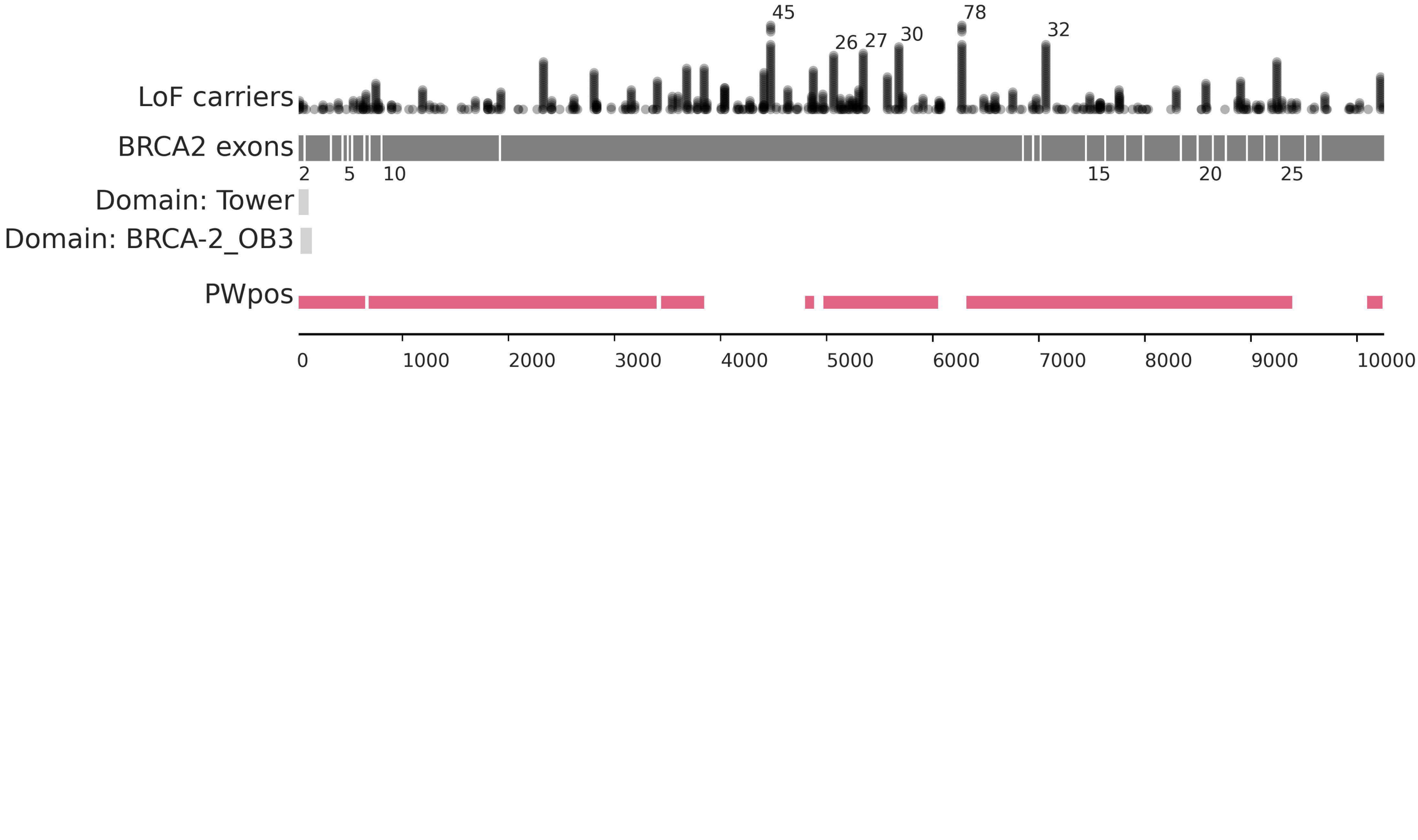

107
49
94
Coding carriers
SHBG exons
1
5
Structure: beta
Structure: helix
Structure: turn
Domain: LAM_G_DOMAIN
Domain: Laminin G-like 1
Domain: Laminin G-like 2
PWneg
0
500
1000

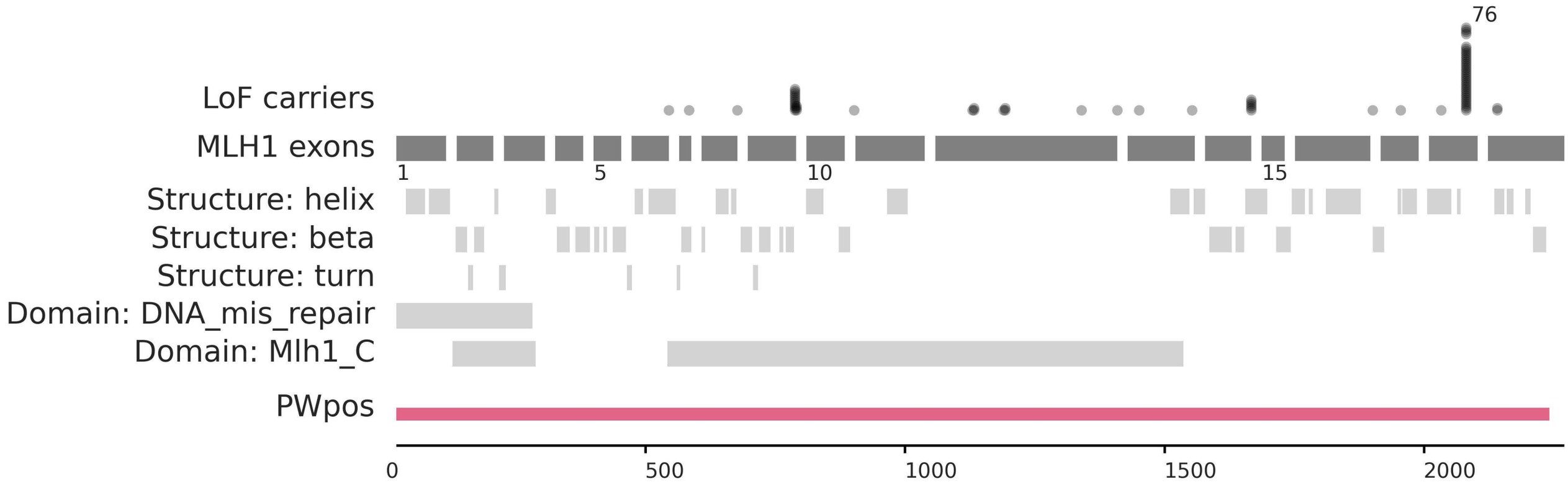

Coding carriers
SLC2A9 exons
Topology: Cytoplasmic
Topology: Transmembrane
Topology: Extracellular
Domain: MFS
PWneg
40
36
30
68
61
40
26
46
36
54
67
50
30
1
5
10
0
500
1000
1500

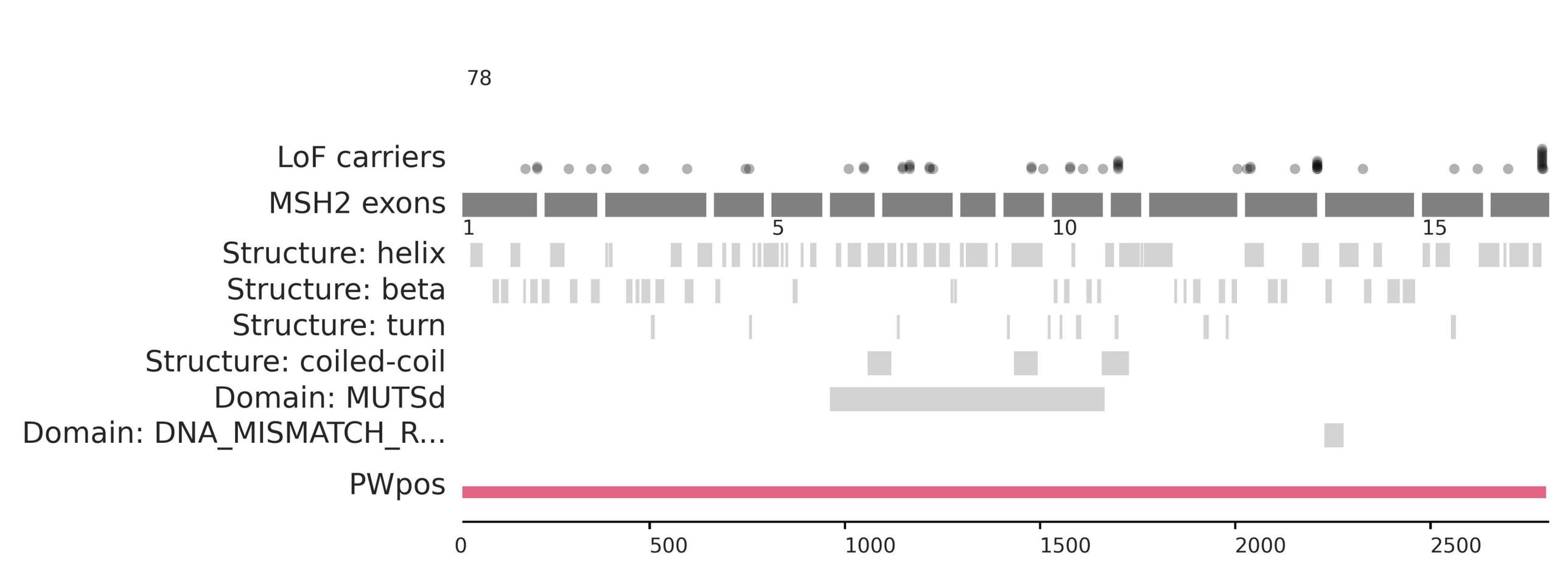

Coding carriers
TUBB1 exons
PWpos
1
0
500
1000
49
34
27
88
29
140
43
66
33
45
29
166
184
41
26
46
35
27
33
27

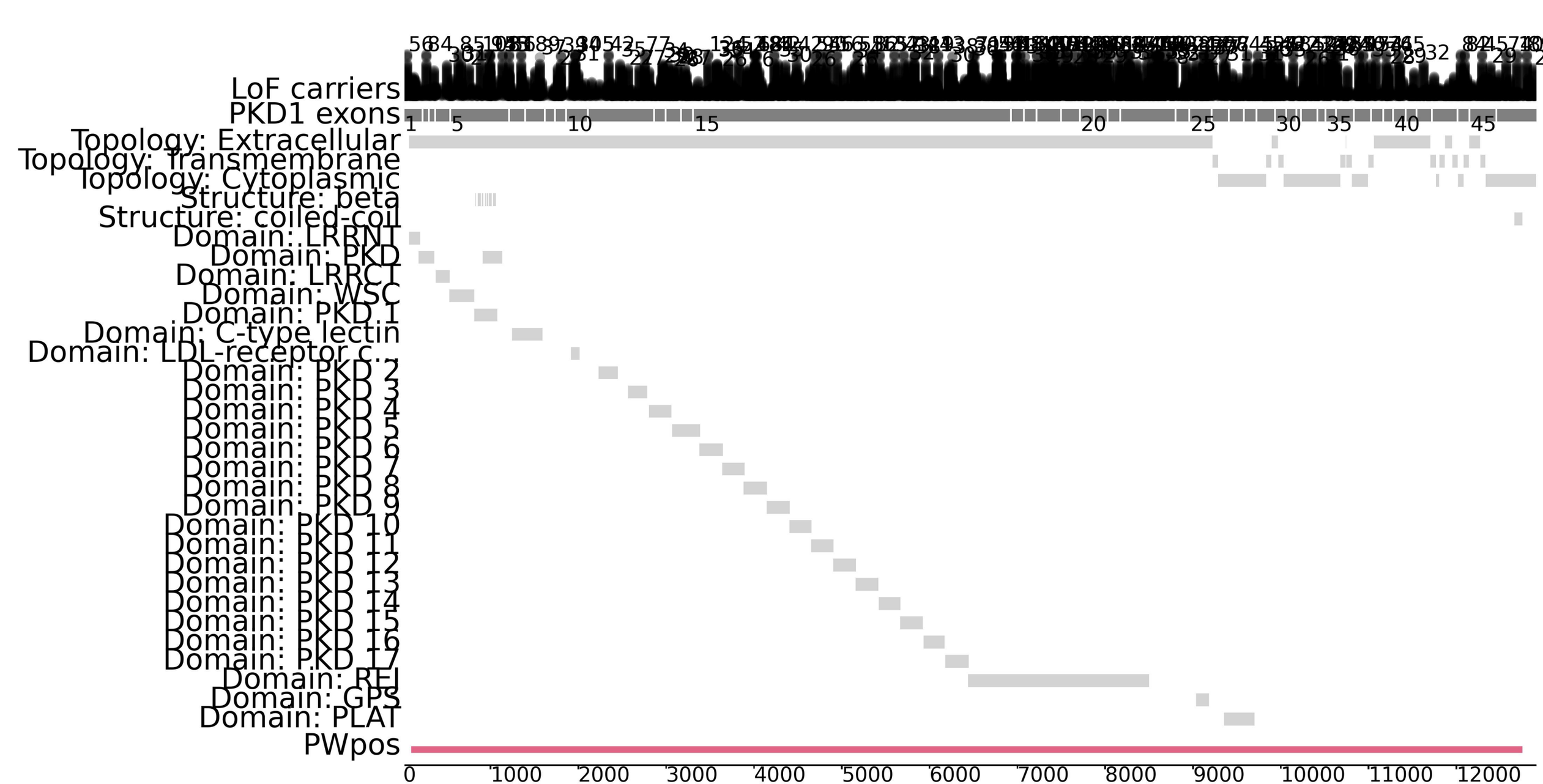

330
43
26
LoF carriers
APOB exons
1
5
10
15
20
25
Domain: Vitellogenin
PWpos
PWneg
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000

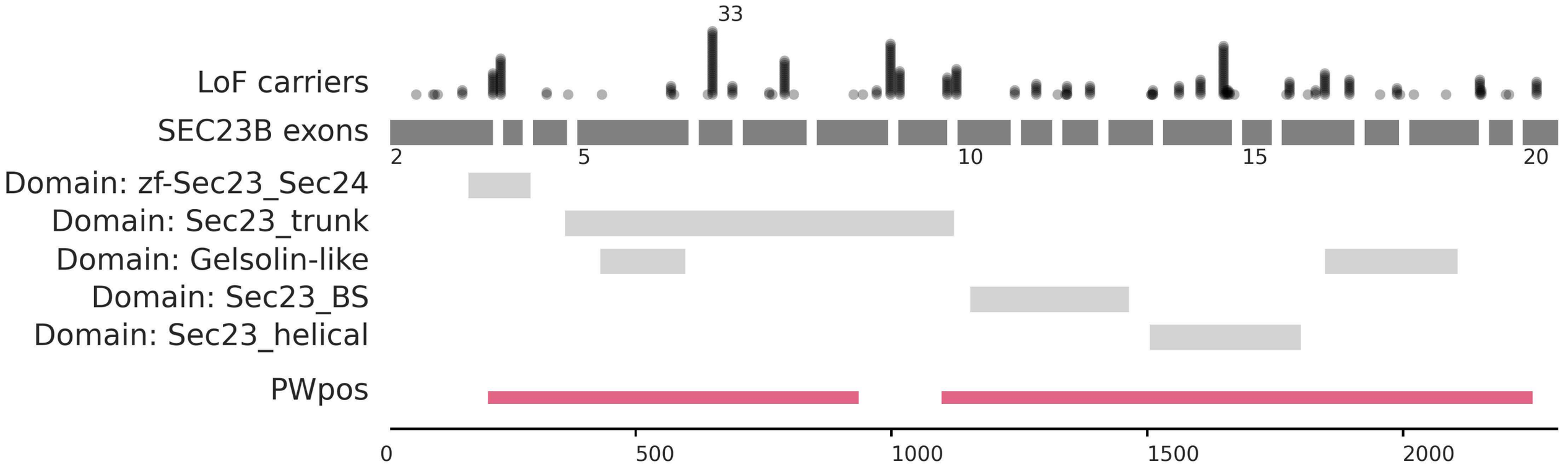

LoF carriers
BRCA1 exons
Topology: Transmembrane
Structure: helix
Structure: beta
Structure: turn
Structure: coiled-coil
Domain: BRCT
Domain: zf-C3HC4
Domain: zf-RING_UBOX
Domain: RING-type
Domain: zinc finger
Domain: BRCT_assoc
Domain: BRCT 1
Domain: BRCT 2
PWpos
0
1000
2000
3000
4000
5000

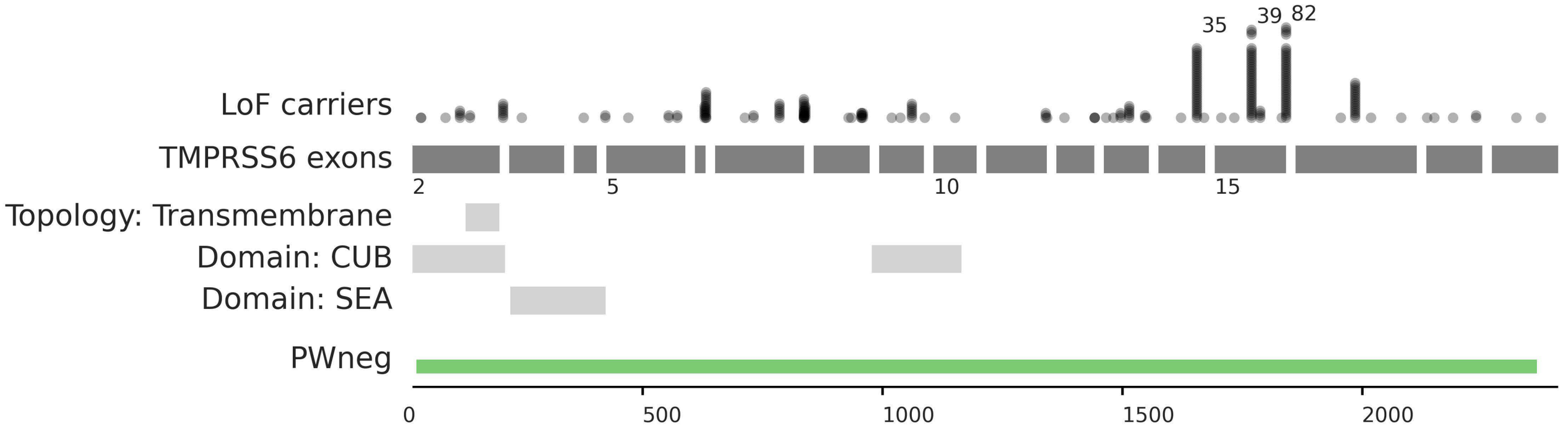

45
78
26
27
30
32
LoF carriers
BRCA2 exons
2
5
10
15
20
25
Domain: Tower
Domain: BRCA-2_OB3
PWpos
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000


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
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
1
5
10
15
20
25
30
35
40
45
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

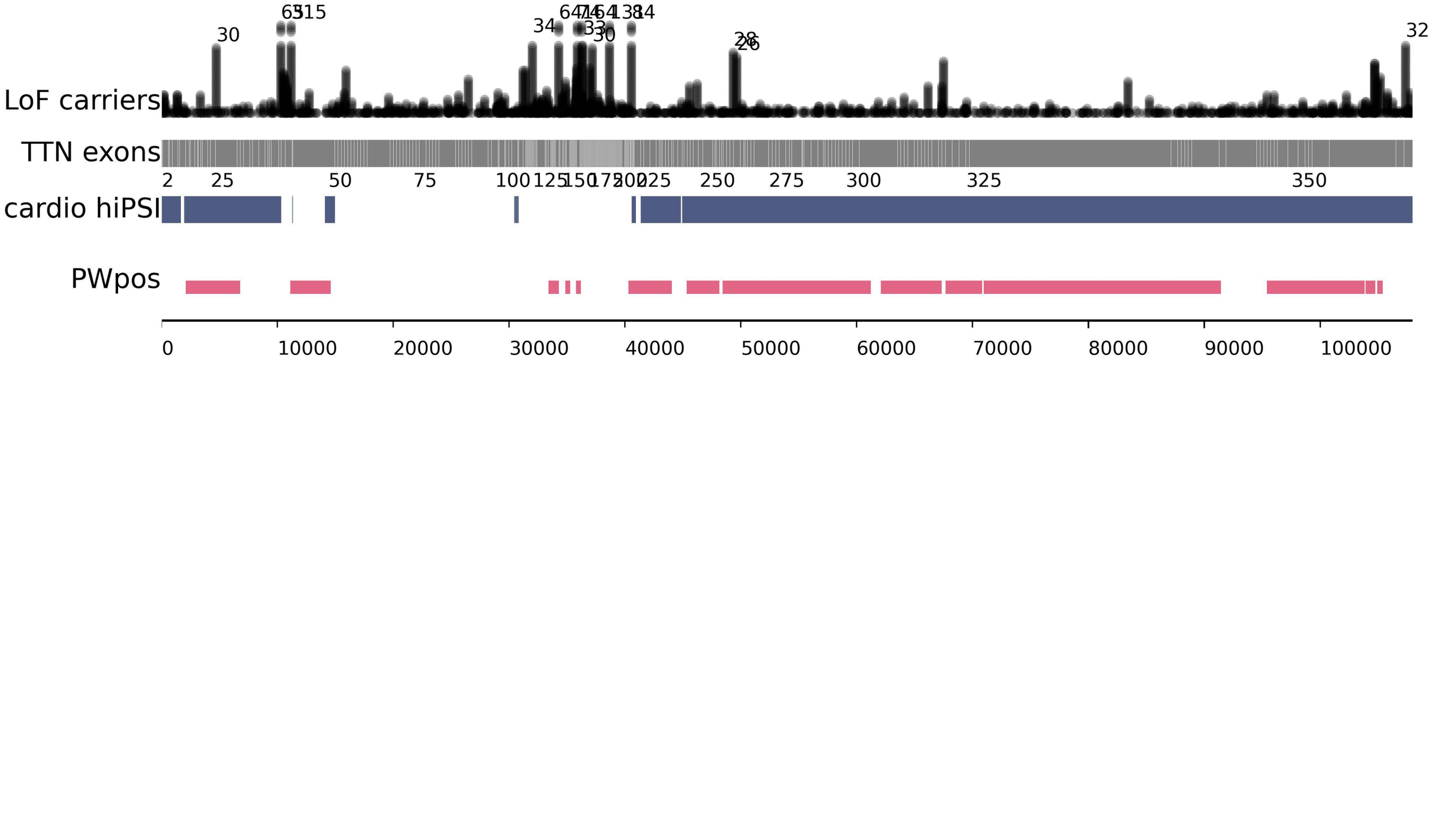
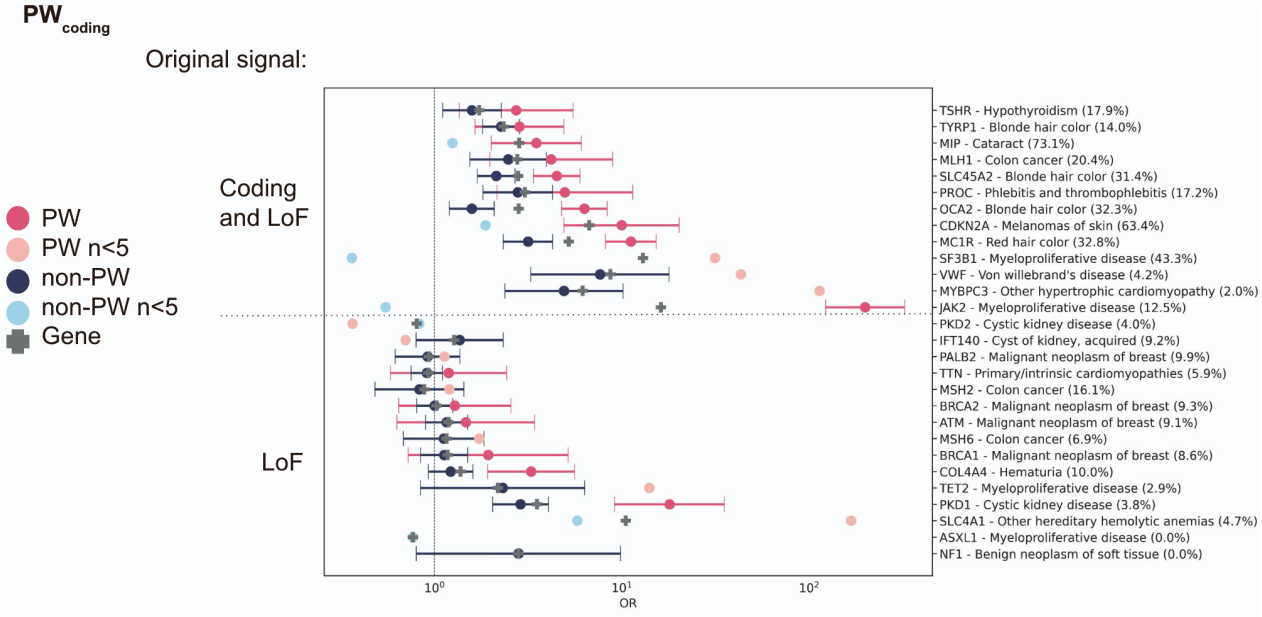

LoF carriers
TTN exons
cardio hiPSI
PWpos
30
65
15
34
64
7
14
64
13
8
14
33
30
28
26
32
2
25
50
75
100
125
150
175
200
225
250
275
300
325
350
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000

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.

A



B

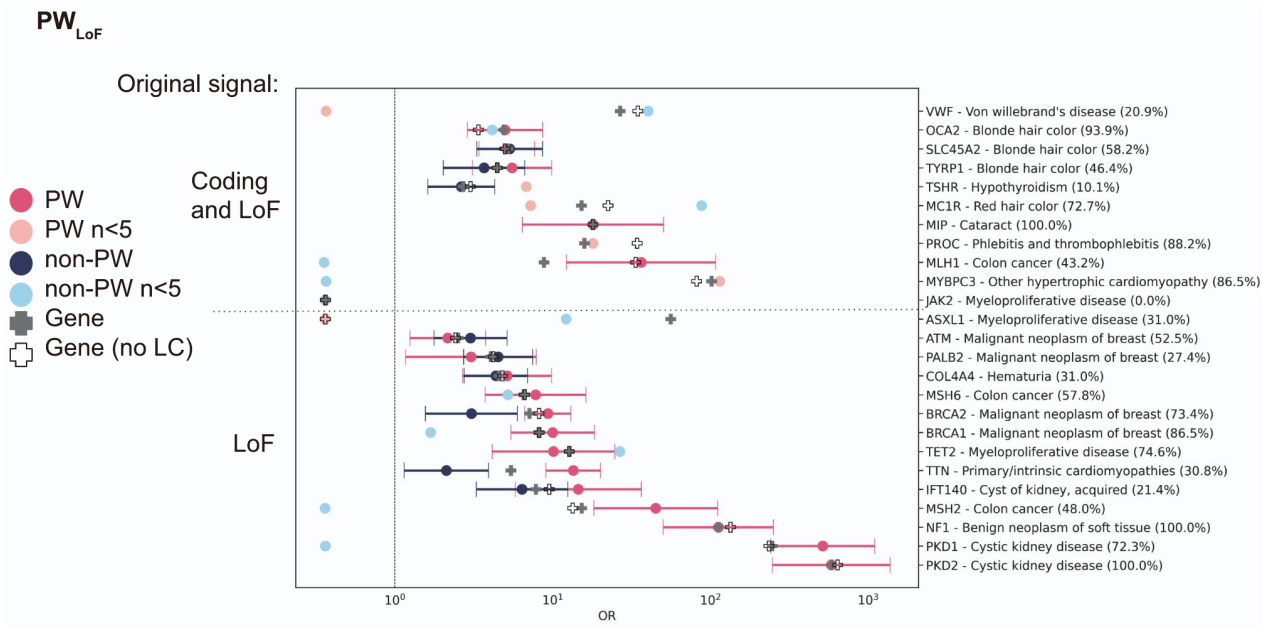


Figure S6. Performance of Power Window for all binary traits in the 117k UKB test set. A) Coding model B) LoF model. The odds ratios for PW models are shown in pink, and the excluded regions are shown in blue. The gene-phenotype label includes in parentheses an indication of what percent of rare variant 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.

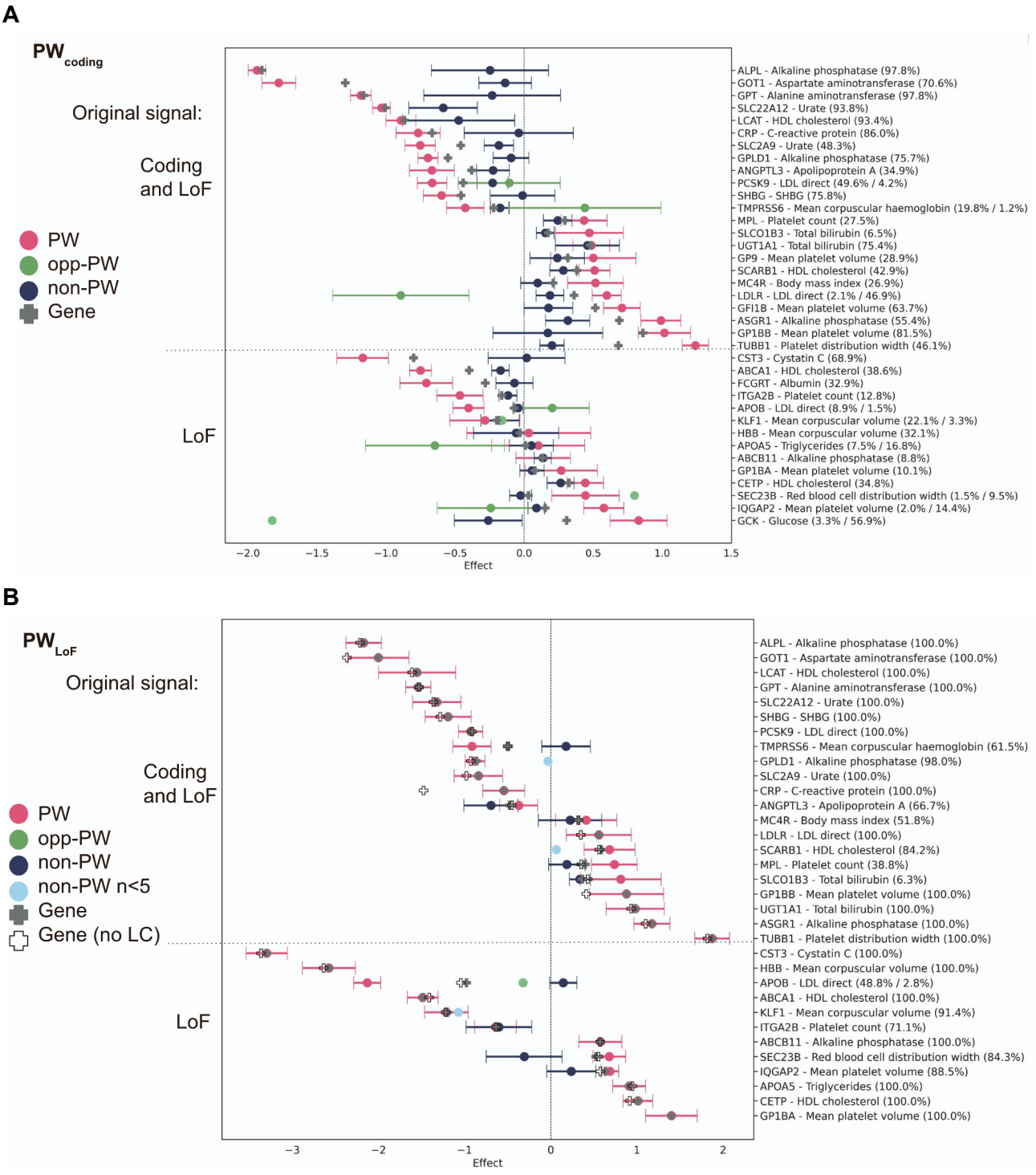
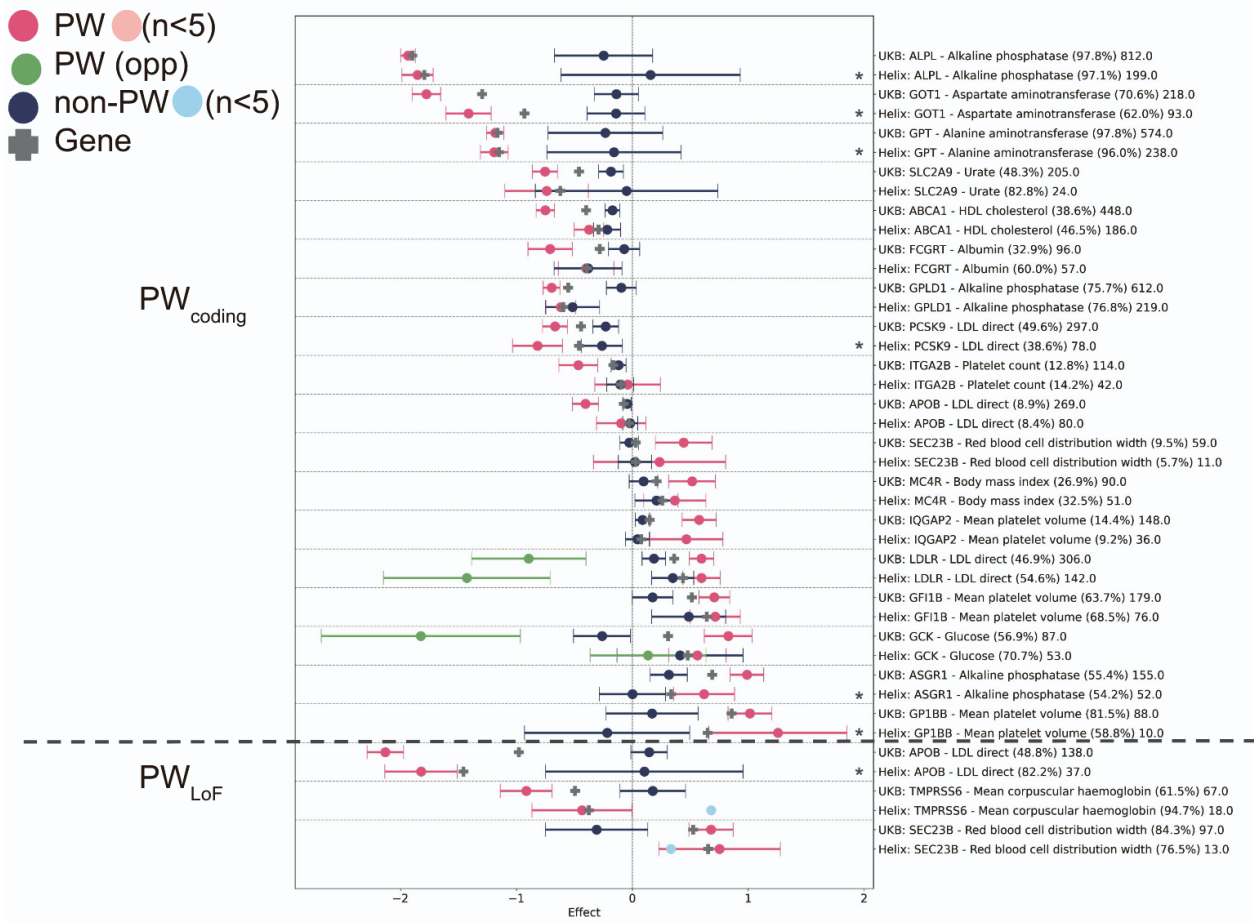


Figure S7. Performance of Power Window for all quantitative traits in the 117k UKB test set. A) Coding model B) LoF model. The effect sizes (normalized phenotypes) for PW models are shown in pink, and the excluded regions are shown in blue. The gene-phenotype label includes in parentheses an indication 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
PW coding
PW LoF
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

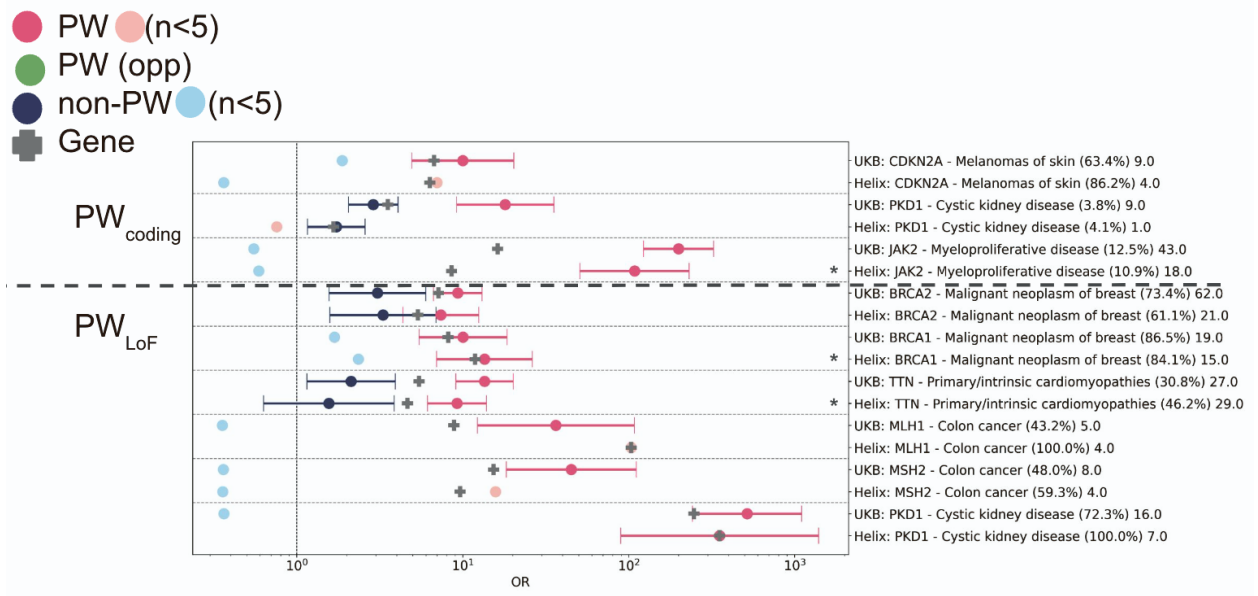

PW (n<5)
PW (opp)
non-PW (n<5)
Gene
PW coding
PW LoF
UKB: CDKN2A - Melanomas of skin (63.4%) 9.0
Helix: CDKN2A - Melanomas of skin (86.2%) 4.0
UKB: PKD1 - Cystic kidney disease (3.8%) 9.0
Helix: PKD1 - Cystic kidney disease (4.1%) 1.0
UKB: JAK2 - Myeloproliferative disease (12.5%) 43.0
Helix: JAK2 - 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.

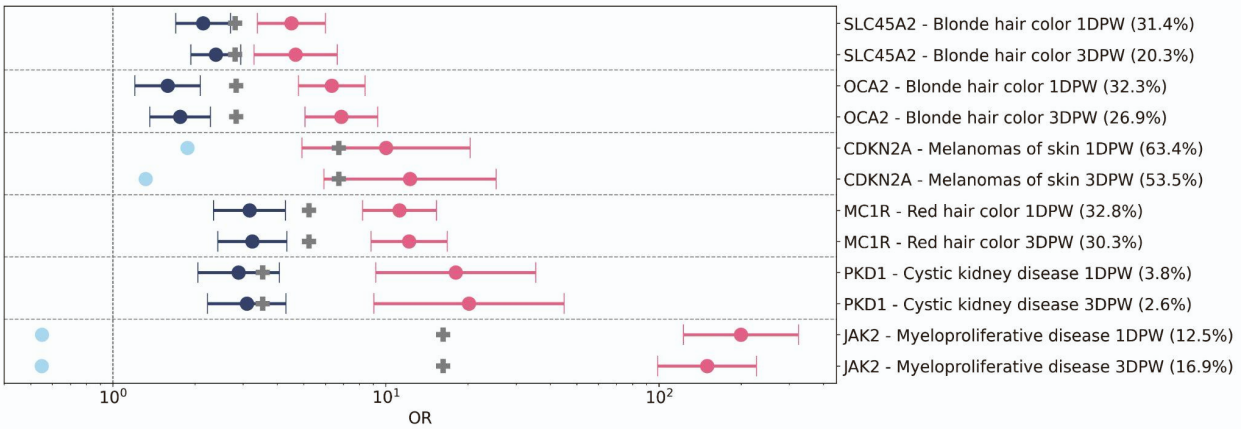
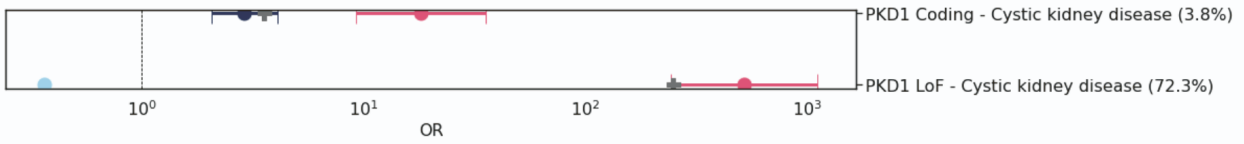
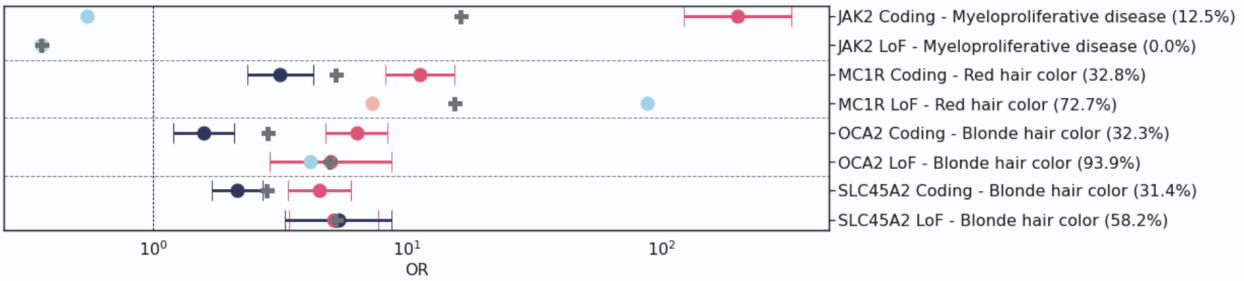


Figure S9. 3D Power Window for binary traits. As in Figure 5, PW models are shown in the test set of 117k individuals for significant PW_{coding} models. Each gene-phenotype pair is shown twice, for the 1D and 3DPW models, with 95% confidence intervals. The model results are very similar in terms of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

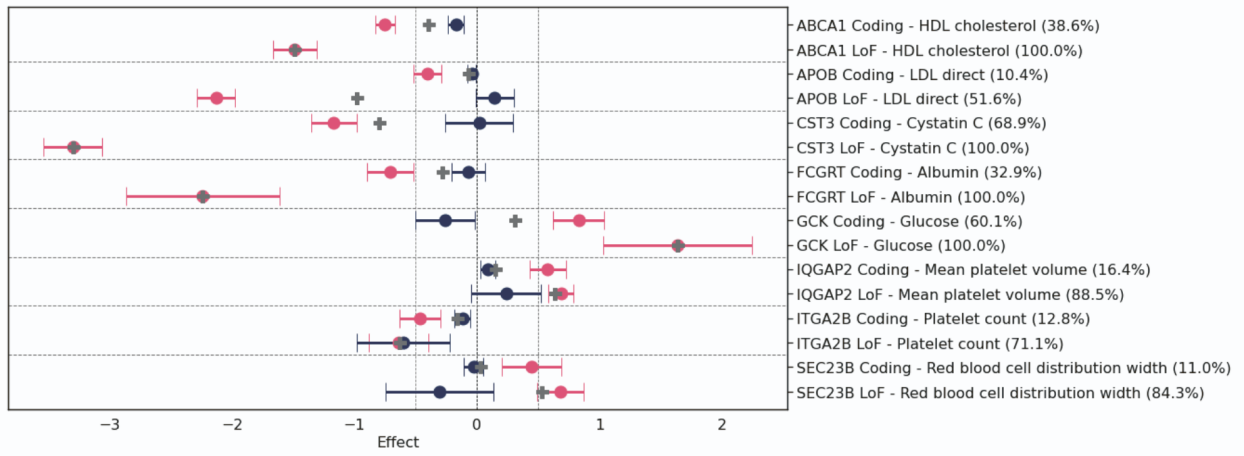
A - Binary traits- original signal primarily for LoF model



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

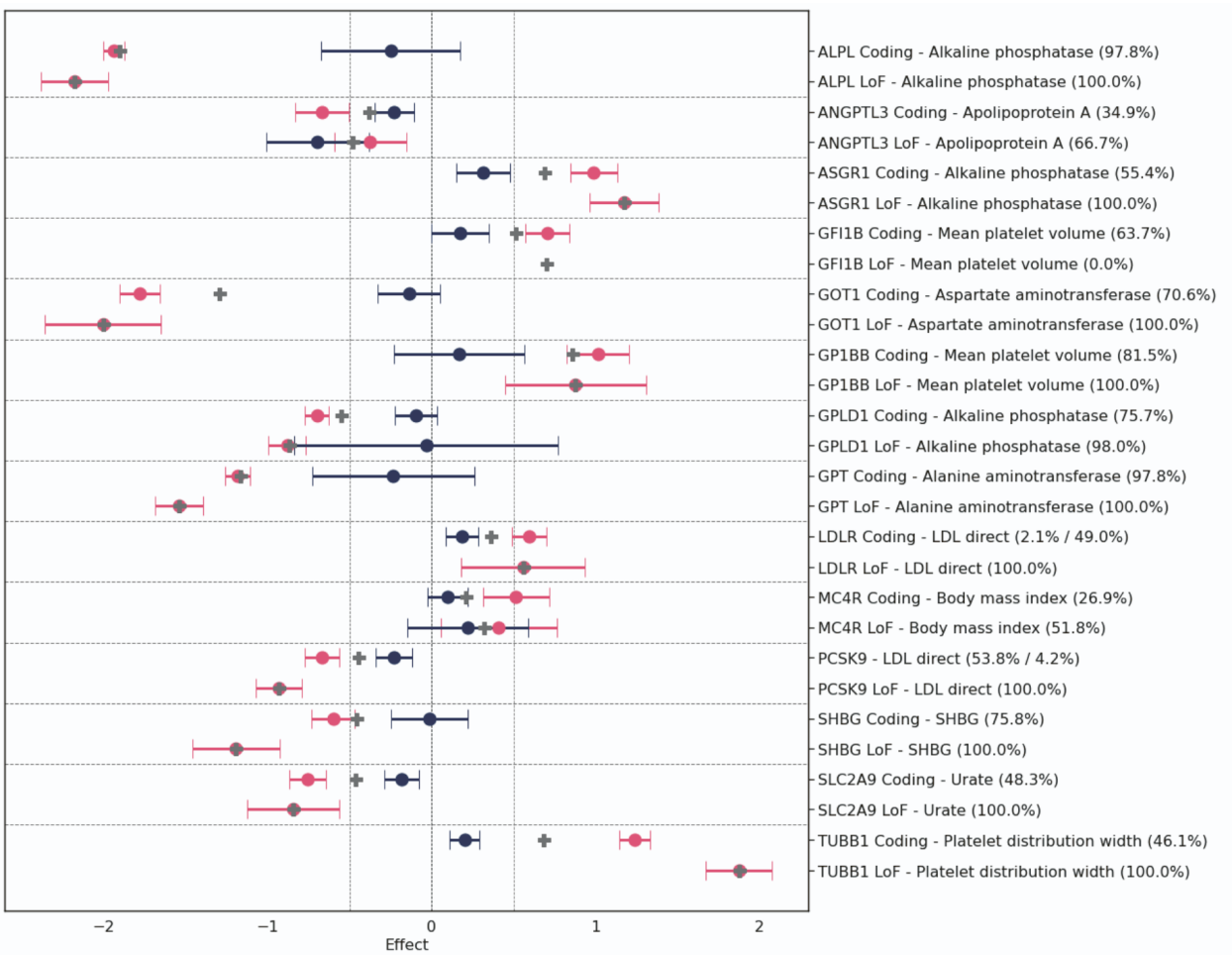
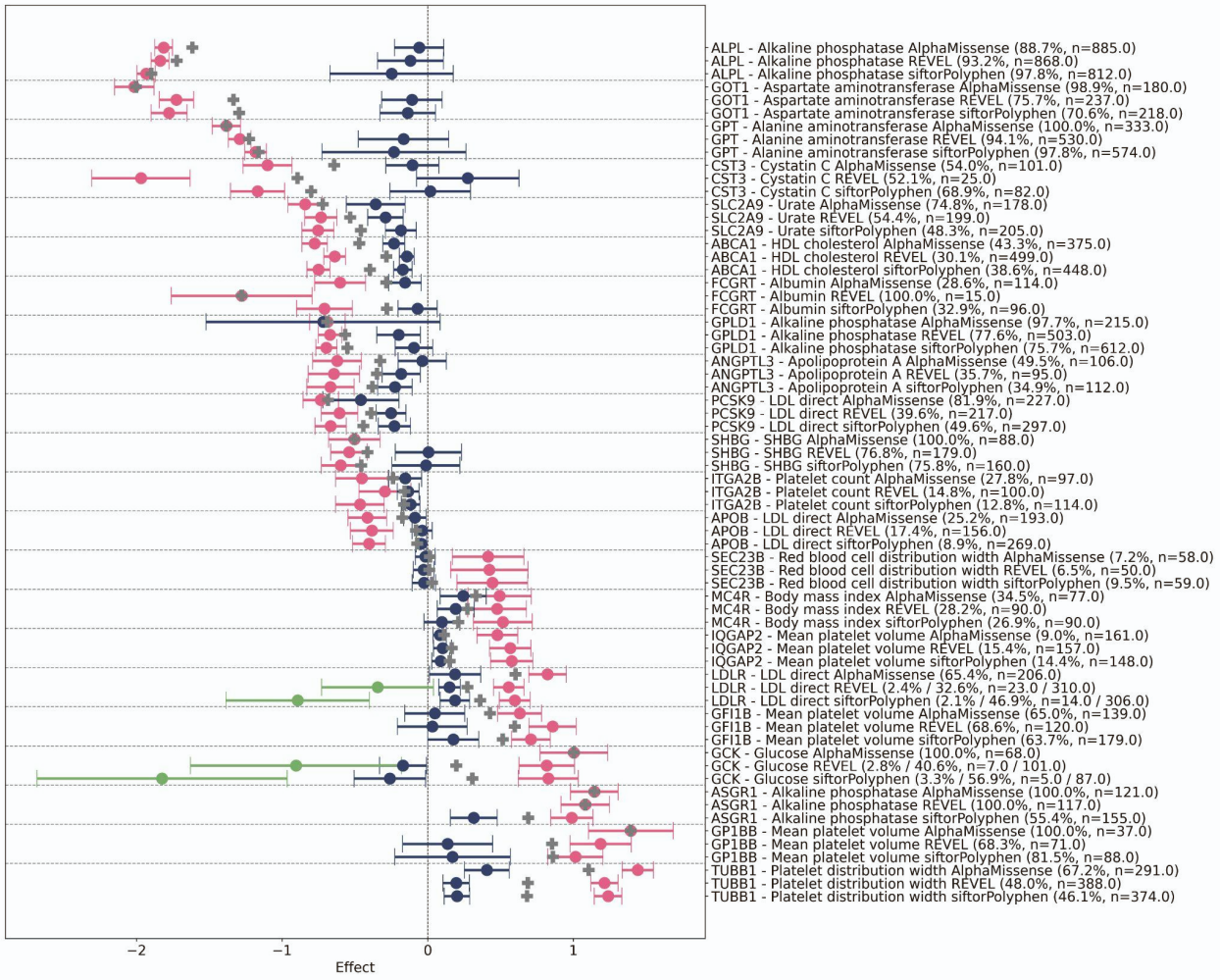


Figure S10. Performance of significant PW_{coding} models compared to PW_{LoF} models for the same phenotype. A) Binary traits where the original whole-gene associations were primarily for the LoF model (LoF beta >3x Coding beta); B) binary traits where the original whole-gene associations were for both 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_{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_{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.

A



B

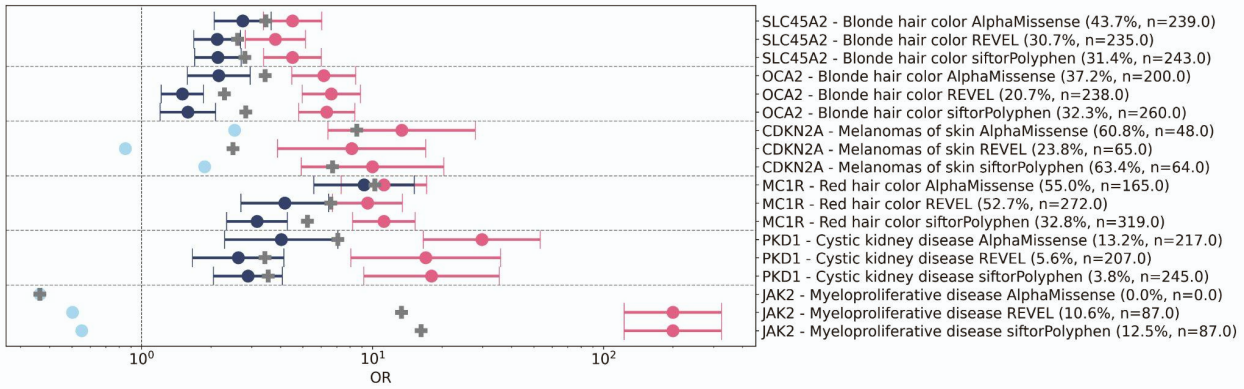


Figure S11. Comparison of significant PWcoding models built with REVEL, AlphaMissense, or sift/Polyphen. As in Figure 3, PW models are shown in the test set of 117k individuals with 95% confidence intervals for significant PW_{coding} models for A) quantitative and B) binary traits. Each gene-phenotype 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.