The American Journal of Human Genetics, Volume *102*

Supplemental Data

Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures

R. Michael Sivley, Xiaoyi Dou, Jens Meiler, William S. Bush, and John A. Capra

Supplementary Figures

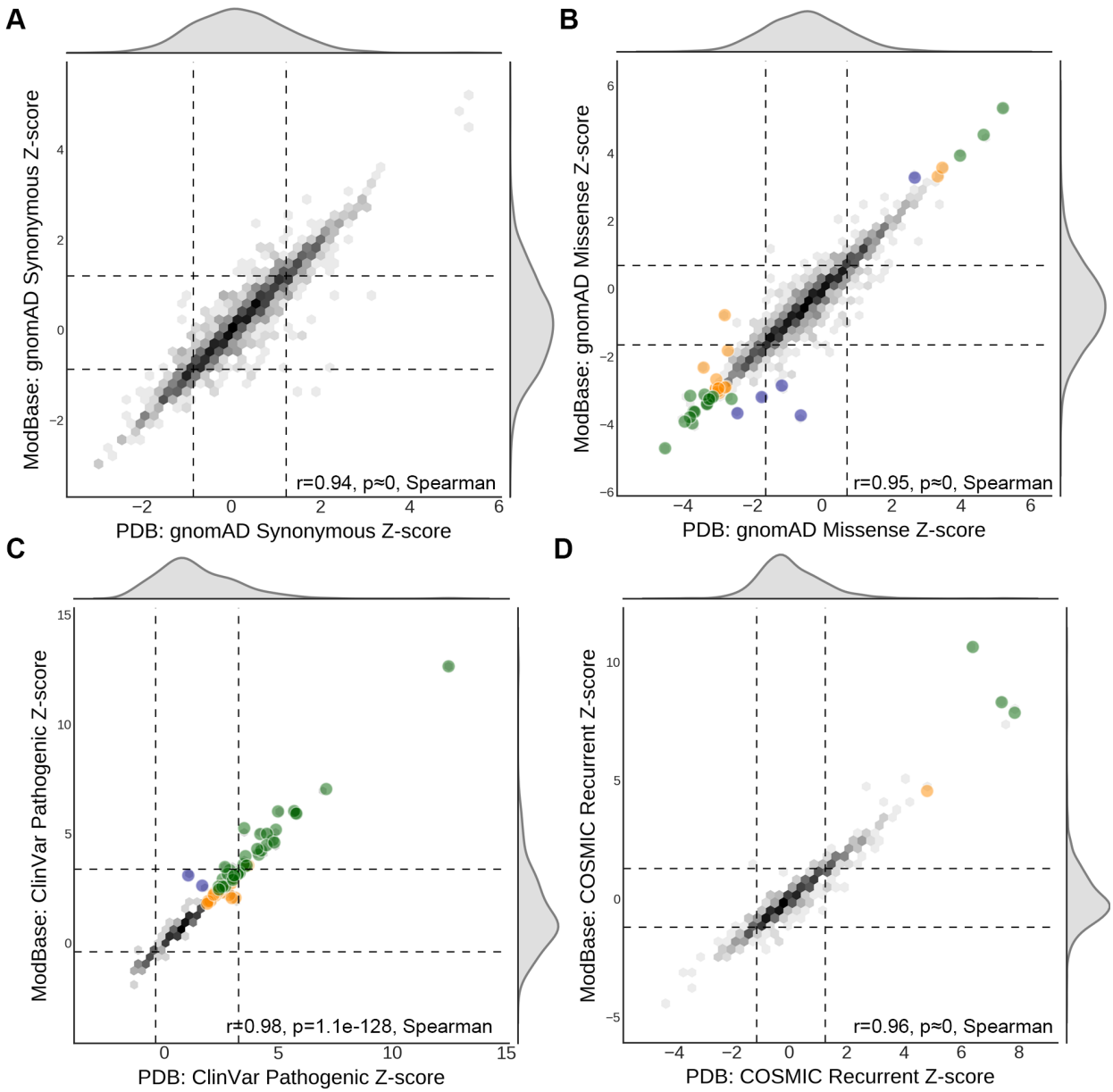

Figure S1: Spatial statistics derived from PDB structures and ModBase homology models are significantly correlated. PDB-derived spatial statistics (Ripley's K Z-score) are plotted against ModBase-derived spatial statistics on shared, sequence-matched proteins for each genetic dataset: (A) gnomAD synonymous, (B) gnomAD missense, (C) ClinVar pathogenic, and (D) COSMIC recurrent. The distribution over all pairs is shown as a density plot, with black indicating higher density. Proteins significant in the PDB analysis are shown in yellow, significant in the ModBase analysis shown in blue, and significant in both in green. We required >95% sequence overlap for each pair of PDB and ModBase structural models, and excluded any pair where the PDB structure was used as the initial template for the ModBase model.

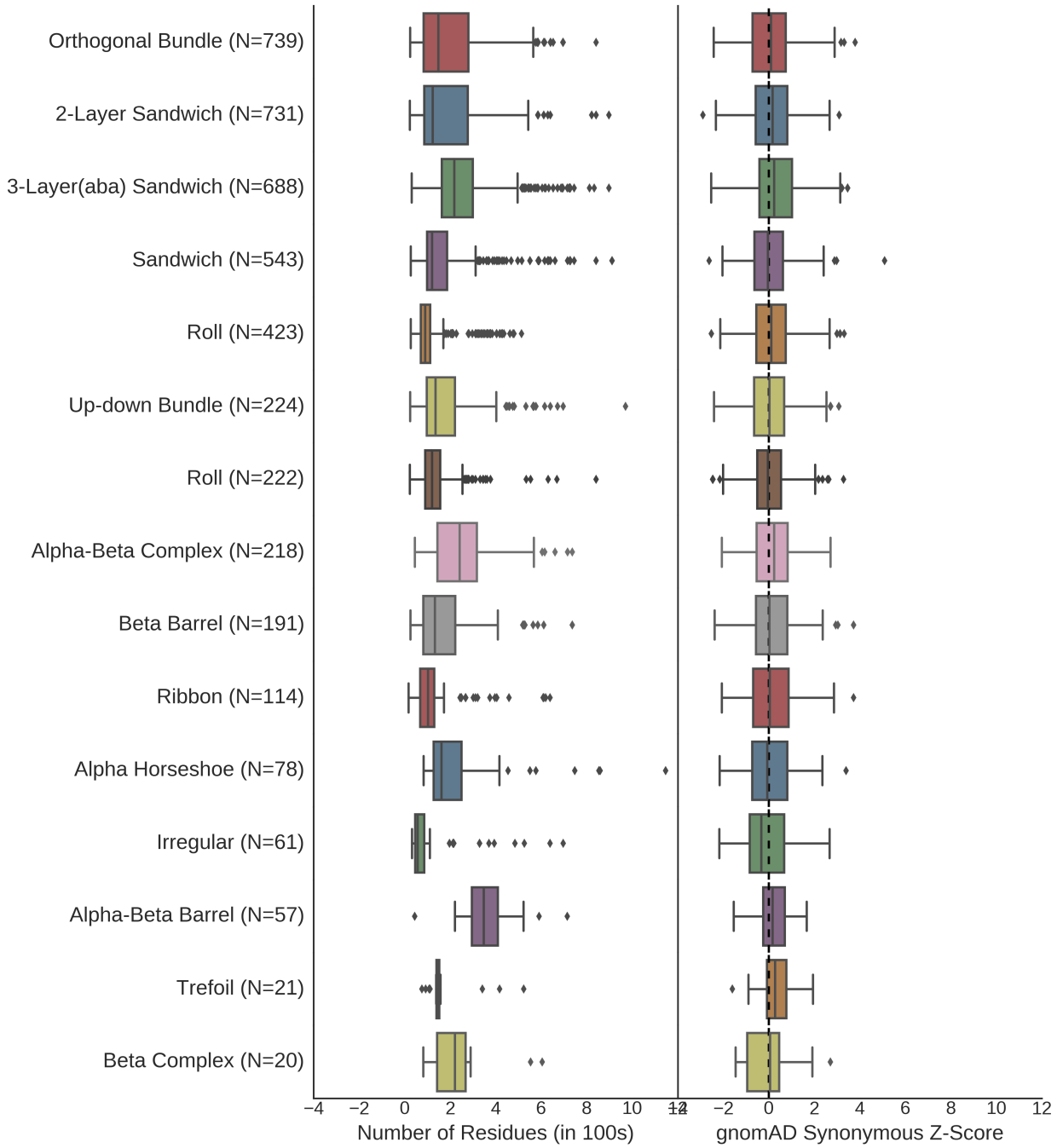

Figure S2: Synonymous variants display similar unconstrained spatial patterns across proteins in different structural domains. Structural domains are defined by CATH (Class Architecture Topology Homology). The number of experimentally derived proteins analyzed from each class is provided to the right of the class label. Domains with fewer than 20 analyzed protein structures were excluded. The distribution of the length of the proteins in each class is summarized on the left, and the distribution of Ripley's K Z-scores for gnomAD synonymous variants is summarized on the right. The stability of Z-scores across distinct structural domains and sizes confirms that our permutation procedure accurately corrects for the background distribution of amino acids in each structure.

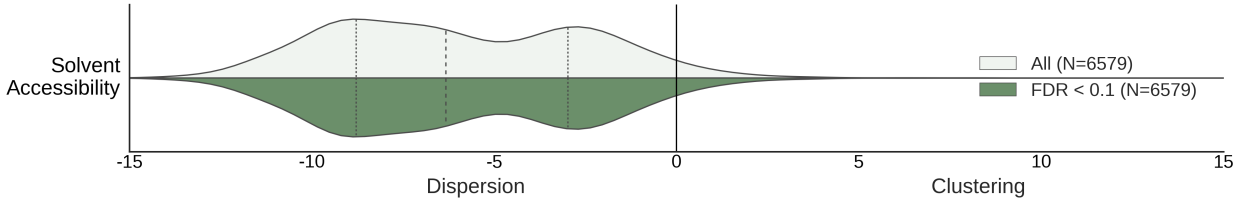

Figure S3: Tolerance of missense variation in solvent accessible residues produces significant spatial dispersion. Distribution of protein Z-scores for the PDB structures from the weighted Ripley's K analysis of relative solvent accessibility (RSA). In 96% of proteins, observed spatial distributions of RSA values are significantly more dispersed than expected by chance; this indicates that, as expected, surface-exposed residues are identified as spatially dispersed. This is a useful benchmark for interpreting the significant spatial dispersion observed among missense variants from gnomAD. It suggests that neutral missense variants may preferentially affect amino acids at the protein surface (Figure S4), consistent with previously observed patterns of 1000 Genomes missense variants[1].

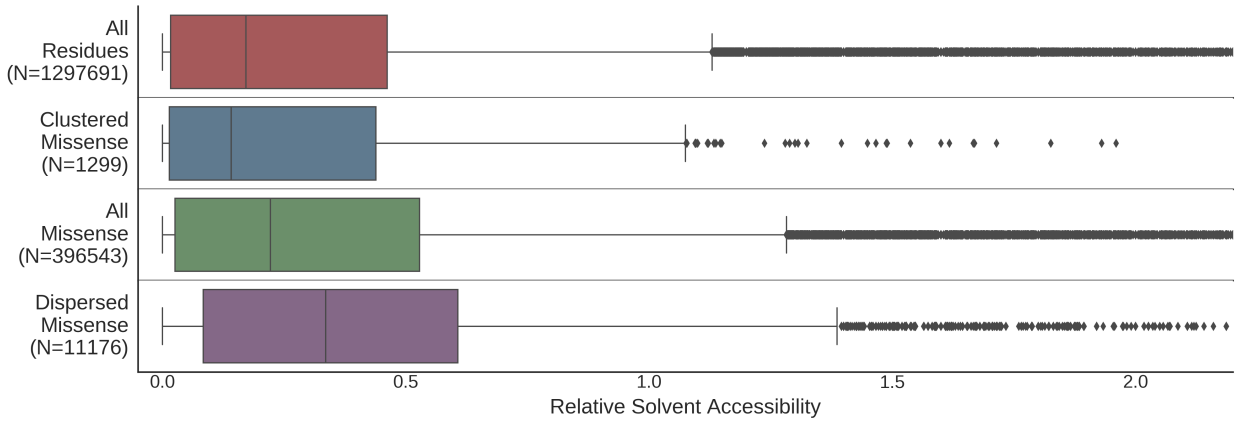

Figure S4: Significantly dispersed missense variants are also significantly more solvent accessible. Residues at which missense variants are observed are significantly more solvent accessible than residues overall (Median $RSA_{missense}$=0.22, Median RSA_{all}=0.17, $p \approx 0$, Mann-Whitney U test). Furthermore, dispersed missense variants are significantly more solvent accessible than all missense (Median $RSA_{dispersed}$=0.34, $p = 1.6x10^{-71}$). This is consistent with constraint against missense mutations in the core of these proteins. In contrast, significantly clustered missense variants have similar solvent accessibility patterns to all residues (Median $RSA_{clustered}$=0.14, $p = 0.19$), suggesting that missense variant clusters commonly occur throughout the protein. Solvent accessibility was calculated with DSSP[2] and normalized by the maximum solvent accessible surface area of each amino acid in an Ala-X-Ala tripeptide.

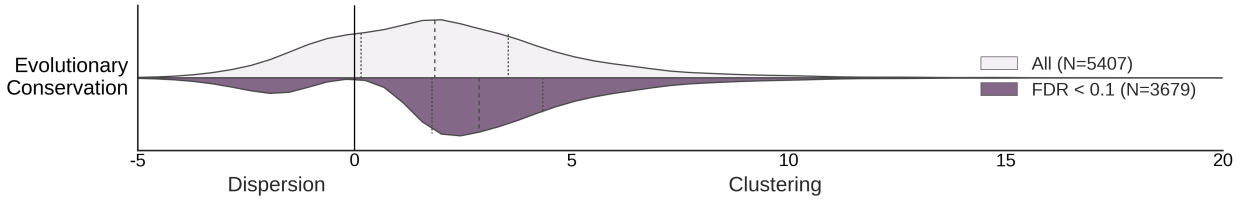

Figure S5: Evolutionarily conserved residues are significantly clustered in protein structures. Evolutionary conservation is a predictor of functionally important residues, and it has been shown to cluster within protein structure at functionally important sites within a limited set of proteins[3–6]. To evaluate this effect comprehensively, we quantified the evolutionary conservation of all amino acids in our PDB dataset using Jensen-Shannon divergence[5] across multiple sequence alignments from HSSP[2] and performed a weighted, spatial analysis of the conservation scores. We identified significant clustering of evolutionary conservation in 3,193 of 5,407 proteins (59%, FDR<0.1) and significant dispersion in 486 proteins (9%). (Figure S5). These results suggest strong spatial constraint on protein function and suggest that functionally important residues are commonly clustered within protein structure.

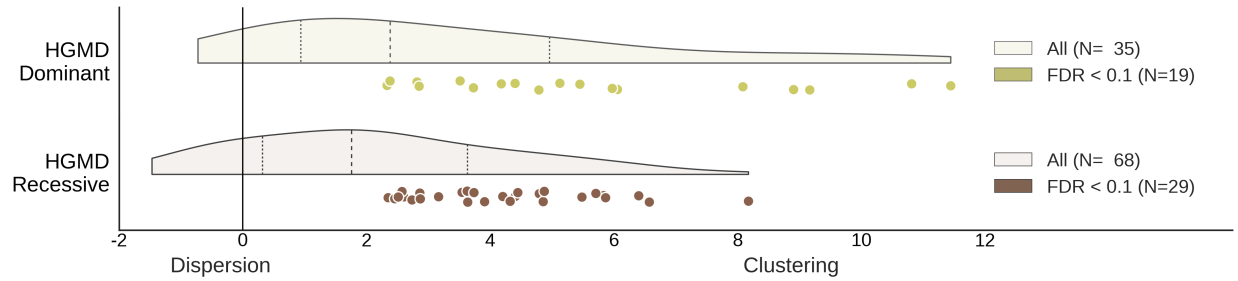

Figure S6: Autosomal dominant and recessive missense variants from the Human Gene Mutation Database (HGMD) are both spatially clustered in protein structure, but dominant variants form smaller clusters. Within proteins with significantly clustered variation, dominant variants (N_{AD}=19) formed significantly smaller clusters (median peak significance distance: 8Å) than recessive variants (N_{AR}=29; median peak significance: 14Å; p = 0.0005, Mann–Whitney U test). These findings support previous conclusions that both gain- and loss-of-function variants are more clustered than neutral variants. The smaller clusters formed by dominant variants additionally support the hypothesis that gain-of-function mutations are localized to specific sites with functional potential, while loss-of-function mutations more generally disrupt regions of functional importance.

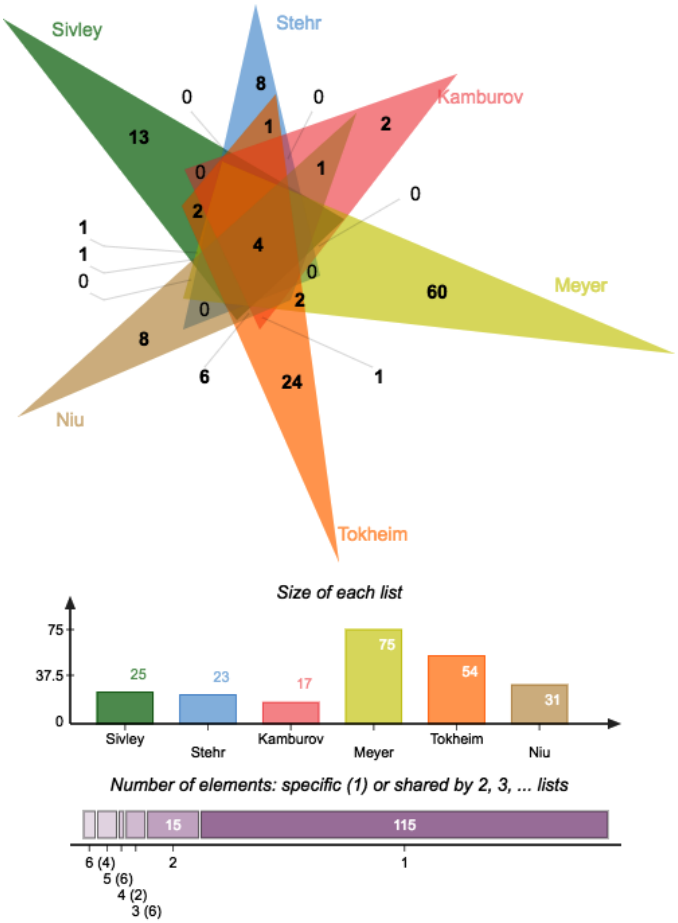

Figure S7: Comparison of our findings with previous studies of somatic mutation clustering. The Venn diagram gives the overlap in genes found to harbor significantly clustered somatic missense variation between related studies. AR, CBL, CCDC160, COMP, CREBBP, DDX3X, ITLN2, MROH2B, PCDHAC1, SEZ6, SIRPA, SMO, and TET2 were uniquely identified by our analysis of COSMIC recurrent somatic missense variation. All studies identified significant clustering in BRAF, FBXW7, EGFR, and PIK3CA. See the Discussion for a description of differences in the goals, methodologies, and datasets in each analysis.

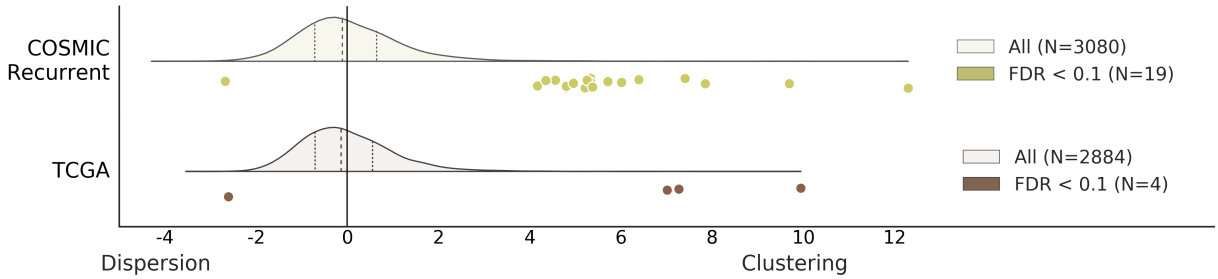

Figure S8: COSMIC recurrent somatic mutations and somatic mutations from TCGA display similar overall spatial trends. General conclusions about the spatial distribution of somatic mutations are consistent between COSMIC and TCGA. There is no statistically significant difference between the distributions of spatial constraint (Ripley's K Z-scores) on somatic variants from COSMIC and TCGA ($p = 0.185$ Mann-Whitney U). In general, analysis of COSMIC identified more proteins with significant constraint, likely due to the larger number of mutations in COSMIC. Nonetheless, analysis of TCGA variants identified clusters in two known cancer proteins that were not detected in COSMIC.

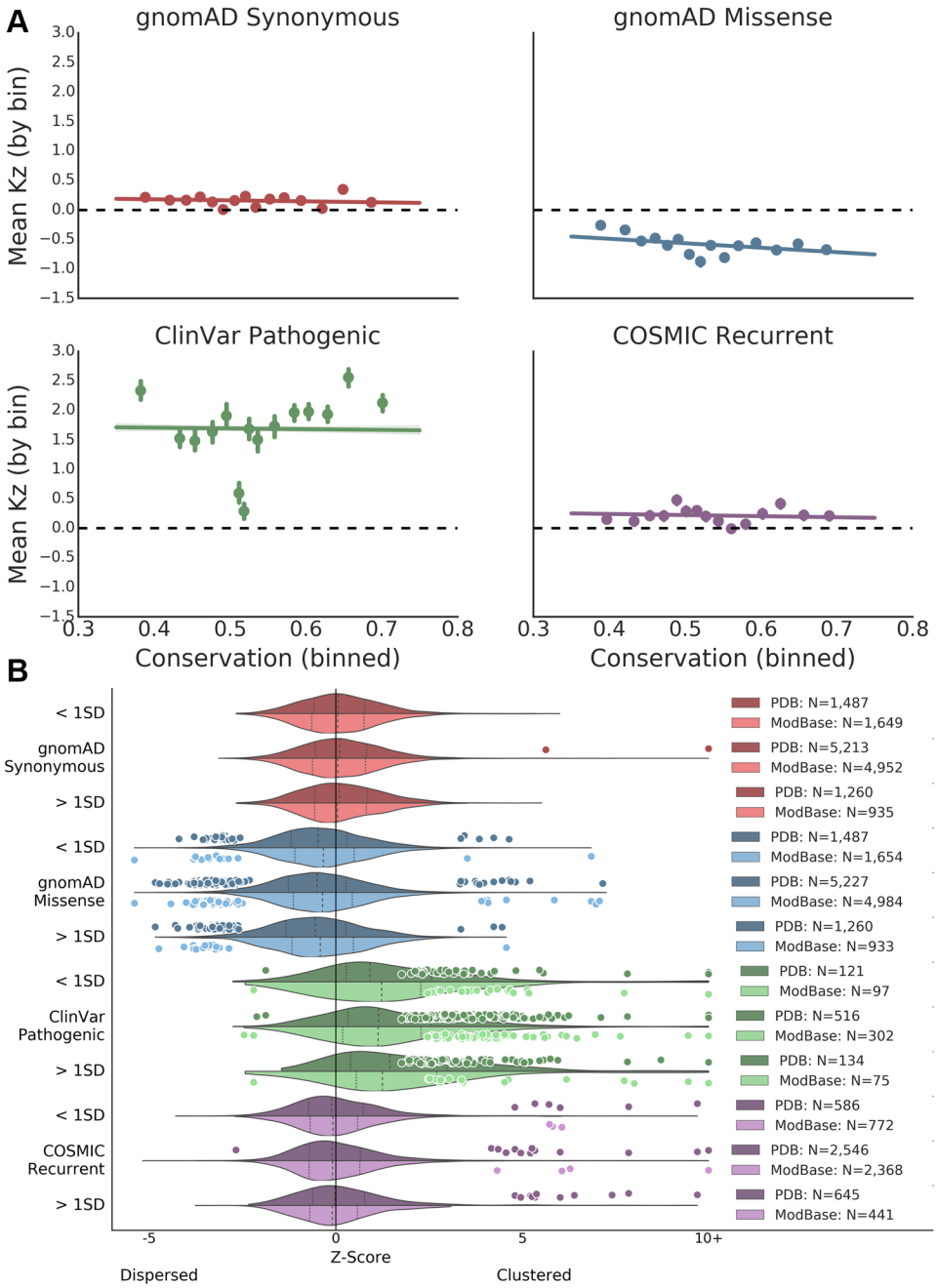

Figure S9: Quantifying the impact of protein evolutionary conservation on spatial statistics. To evaluate the impact of protein evolutionary conservation between species on the observed patterns of variant spatial constraint, we evaluated the correlation between evolutionary conservation and the K Z-score (Kz) for each class of variant and compared K Z-score distributions over proteins stratified by evolutionary conservation. Residue-level evolutionary conservation scores were calculated using Jensen-Shannon divergence[5], and protein conservation was defined as the mean residue-level conservation score. (A) Evolutionary conservation (binned into equally sized groups) plotted against the K Z-score (Kz, mean and 95% confidence intervals plotted for each bin). Evolutionary conservation explained very little of the overall variance in spatial distributions (R^2 between 0.0001 and 0.004). However, due to the large sample size, the modest associations with synonymous (R^2=0.0003; p=0.0001), missense (R^2=0.004; p=3.17e-42), and recurrent somatic dispersion (R^2=0.0002; p=0.0138) were statistically significant. (B) Our conclusions about the spatial distributions of all variant sets held when analyzing proteins at the extremes of the conservation score distribution (+/– 1 standard deviation), and no significant differences were observed between the stratified sets (p between 0.06 and 0.98, Mann-Whitney U test).

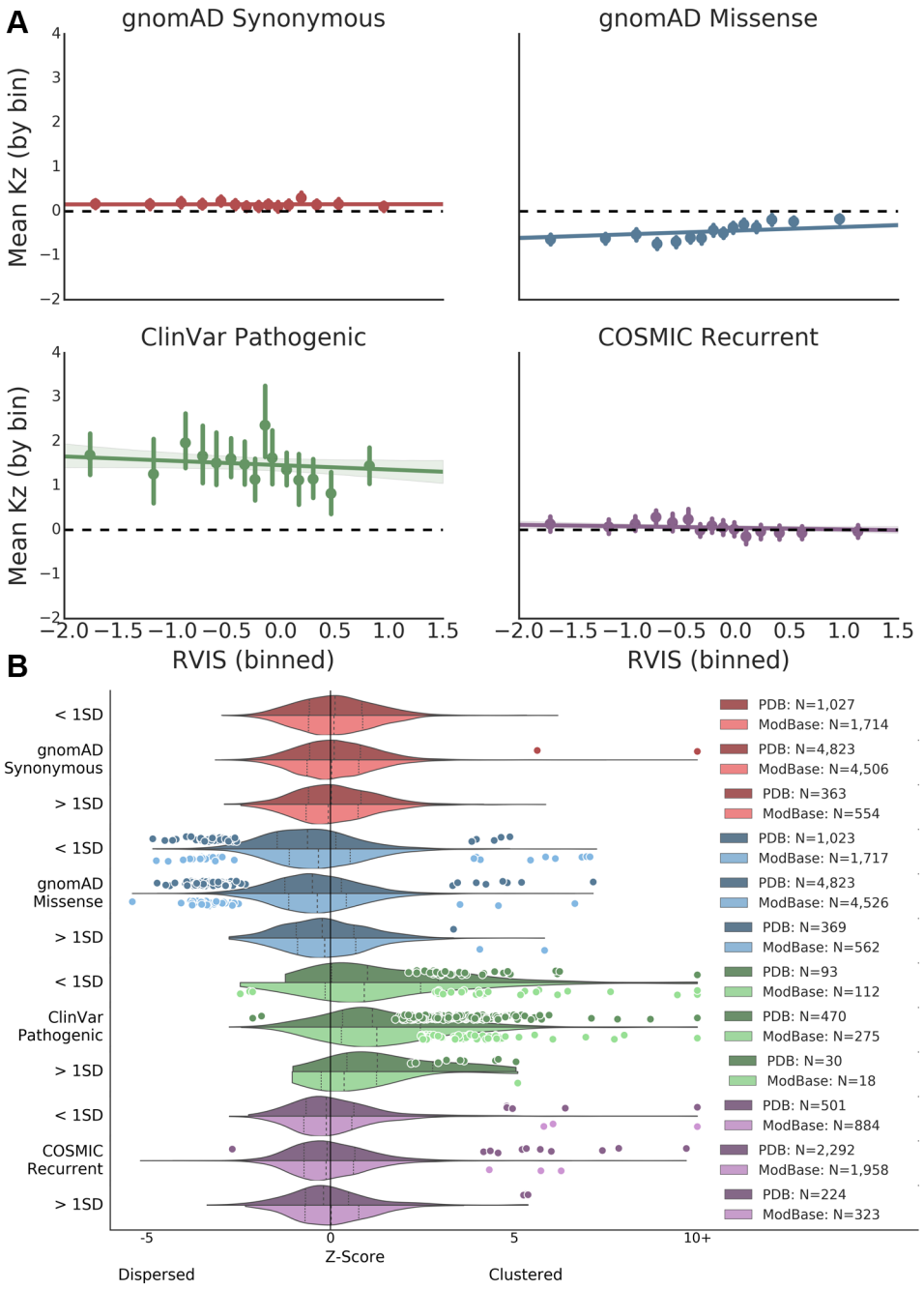

Figure S10: Quantifying the impact of genic intolerance to variation on spatial statistics. To evaluate the impact of genic intolerance to variation on the observed patterns of variant spatial constraint, we evaluated the correlation between Residual Variance Intolerance Score (RVIS)[7] and the K Z-score (Kz) for each class of variant and compared K Z-score distributions over proteins stratified by RVIS. Genic intolerance to variation (RVIS) was mapped to each protein using UniProt cross-references[8]. Proteins with high RVIS have more common functional variation, and those with negative scores are more intolerant to functional variation. (A) RVIS (binned into equally sized groups) plotted against the K Z-score (Kz, mean and 95% confidence intervals plotted for each bin). RVIS explained very little of the overall variance in spatial distributions (R^2 between 0 and 0.009). However, due to the large sample size, the modest RVIS associations with missense clustering (R^2=0.009; p=2.57e-14) and with recurrent somatic dispersion (R^2=0.001; p=0.0418) were statistically significant. (B) gnomAD missense variants in proteins with high tolerance to variation (RVIS > 1 standard deviation from the mean) are significantly less dispersed than those in proteins with lower RVIS (p=2.25e–10 PDB, p=5.00e–05 ModBase, Mann-Whitney U test). Nonetheless, the overall spatial trends hold when proteins are stratified by RVIS.

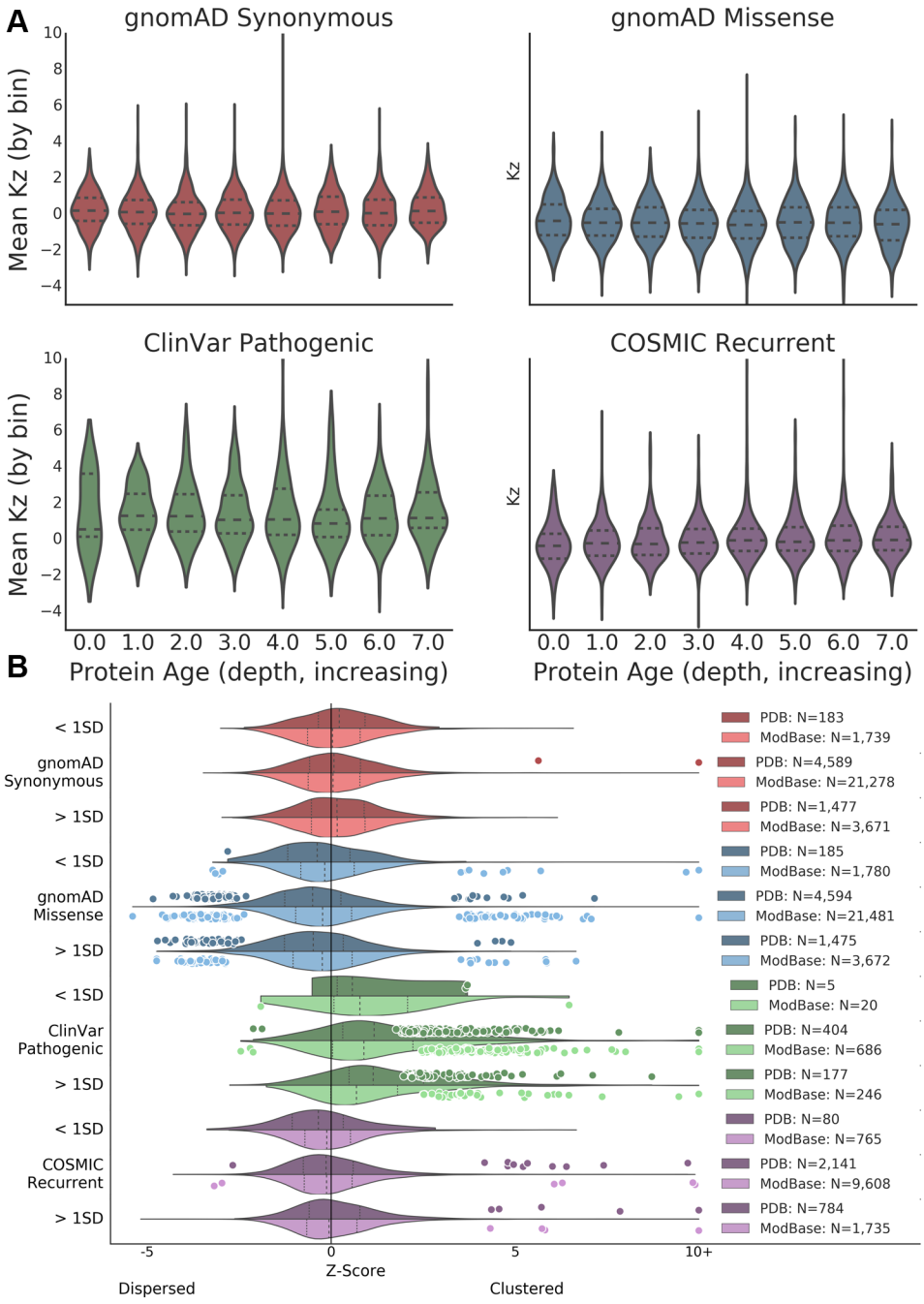

Figure S11: Quantifying the impact of protein age on spatial statistics. To evaluate the impact of protein evolutionary age on the observed patterns of variant spatial constraint, we evaluated the correlation between protein age and the K Z-score (Kz) for each class of variant and compared K Z-score distributions over proteins stratified by age. Protein age was quantified by ProteinHistorian using the PPODv4_PTHR7-OrthoMCL_wagner1.0 dataset[9]. (A) Protein ages (binned into equally-sized groups) plotted against the K Z-score (Kz, mean plotted for each bin). Protein age explained very little of the overall variance in spatial distributions (R^2 between 0.0001 and 0.0058). However, due to the large sample size, protein age is significantly associated with missense dispersion (R^2=0.0008; p=0.0282) and with recurrent somatic clustering (R^2=0.0058; p=2.72e-05). (B) The spatial distributions of all variant sets were qualitatively similar when analyzing proteins at the extremes of the protein age distribution (+/– 1 standard deviation).

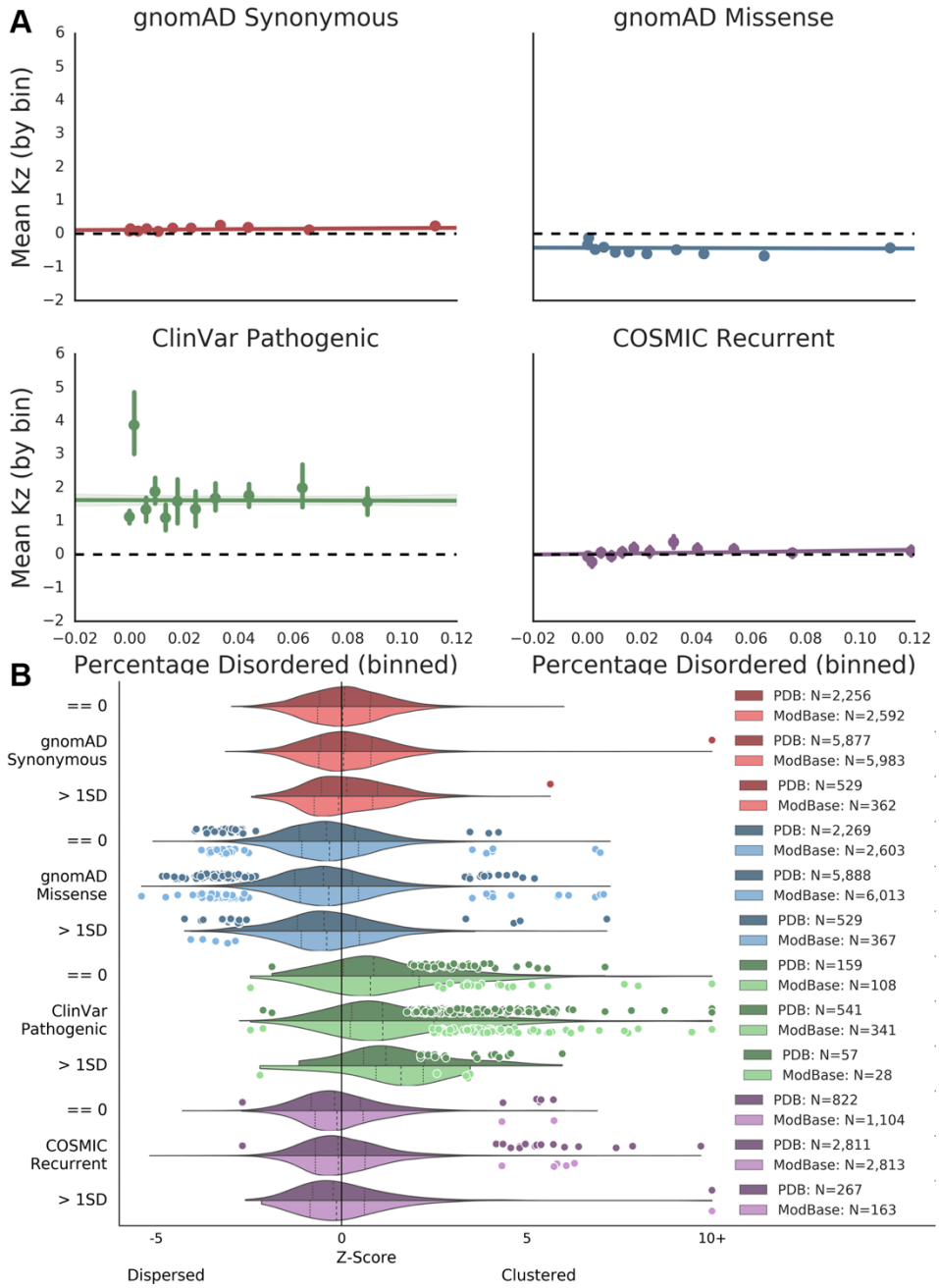

Figure S12: Quantifying the impact of protein disorder on spatial statistics. To evaluate the impact of protein disorder on the observed patterns of variant spatial constraint, we evaluated the correlation between disorder and the K Z-score (Kz) for each class of variant and compared K Z-score distributions over proteins stratified by amount of disorder. The proportion of disorder per protein is calculated from annotations in MOBIdb[10], and defined as the proportion of the total protein sequence annotated as disordered. (A) Protein disorder (binned into equally-sized groups) plotted against the K Z-score (Kz, mean and 95% confidence intervals plotted for each bin). Protein disorder explained very little of the overall variance in spatial distributions (R^2 between 0 and 0.0023). However, due to the large sample size, protein disorder is significantly associated with synonymous (R^2=0.0008; p=0.0025) and recurrent somatic clustering (R^2=0.002; p=0.001). (B) Our conclusions about the spatial distributions of all variant sets held when analyzing proteins at the extremes of the disorder distribution (> 1 standard deviation above the mean and no disorder). However, modest but significant, differences in the spatial distributions of germline missense, pathogenic, and recurrent somatic variants were detected when stratifying each group by proportion of disordered sequence (p between 0.001 to 0.03).

Supplementary Tables

		Spearman Correlation		Significant Proteins				
	N	**rho**	**p-value**	**PDB**	**ModBase**	**Both**	**Precision**	**Recall**
gnomAD synonymous	1826	0.94	0	0	0	0	-	-
gnomAD missense	1824	0.95	0	36	23	18	0.78	0.50
ClinVar pathogenic	177	0.98	1.06E-128	59	40	38	0.95	0.64
COSMIC recurrent	961	0.96	0	4	3	3	1.00	0.75

Table S1: ModBase homology models accurately identify spatial patterns observed in experimentally derived structures. Quantifications of 3D spatial constraint (Ripley's K Z-score) calculated from experimentally derived structures (PDB) and homology models (Modbase) of the same protein are significantly correlated (also see Figure S1). Precision and recall were calculated by evaluating the agreement of significance as determined by analysis of ModBase-derived models with results on the corresponding experimentally derived (PDB) structures. The moderate recall of structures with significant spatial constraint suggests that analyses of homology models are less powered to detect significant spatial patterns. The high precision, especially for pathogenic variants, indicates that significant spatial patterns detected in homology models are also found in solved structures. We required >95% sequence overlap for each pair of PDB and ModBase structural models, and excluded any pair where the PDB structure was used as the initial template for the ModBase model.

References

1. de Beer, T. a P., Laskowski, R. a, Parks, S.L., Sipos, B., Goldman, N., and Thornton, J.M. (2013). Amino Acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. PLoS Comput. Biol. *9*, e1003382.
2. Touw, W.G., Baakman, C., Black, J., Te Beek, T.A.H., Krieger, E., Joosten, R.P., and Vriend, G. (2015). A series of PDB-related databanks for everyday needs. Nucleic Acids Res. *43*, D364–D368.
3. Schueler-furman, O., and Baker, D. (2003). Conserved Residue Clustering and Protein Structure Prediction. *235*, 225–235.
4. Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., and Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J. Mol. Biol. *316*, 139–154.
5. Capra, J.A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. Bioinformatics *23*, 1875–1882.
6. Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M., and Funkhouser, T.A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput. Biol. *5*,.
7. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet. *9*,.
8. The UniProt Consortium (2014). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204-212.
9. Capra, J.A., Williams, A.G., and Pollard, K.S. (2012). Proteinhistorian: Tools for the comparative analysis of eukaryote protein origin. PLoS Comput. Biol. *8*,.
10. Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Mičetić, I., Camilloni, C., Davey, N., Dosztányi, Z., Mészáros, B., Monzon, A.M., et al. (2017). MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. Nucleic Acids Res. 1–6.