nature genetics

SUPPLEMENTARY INFORMATION

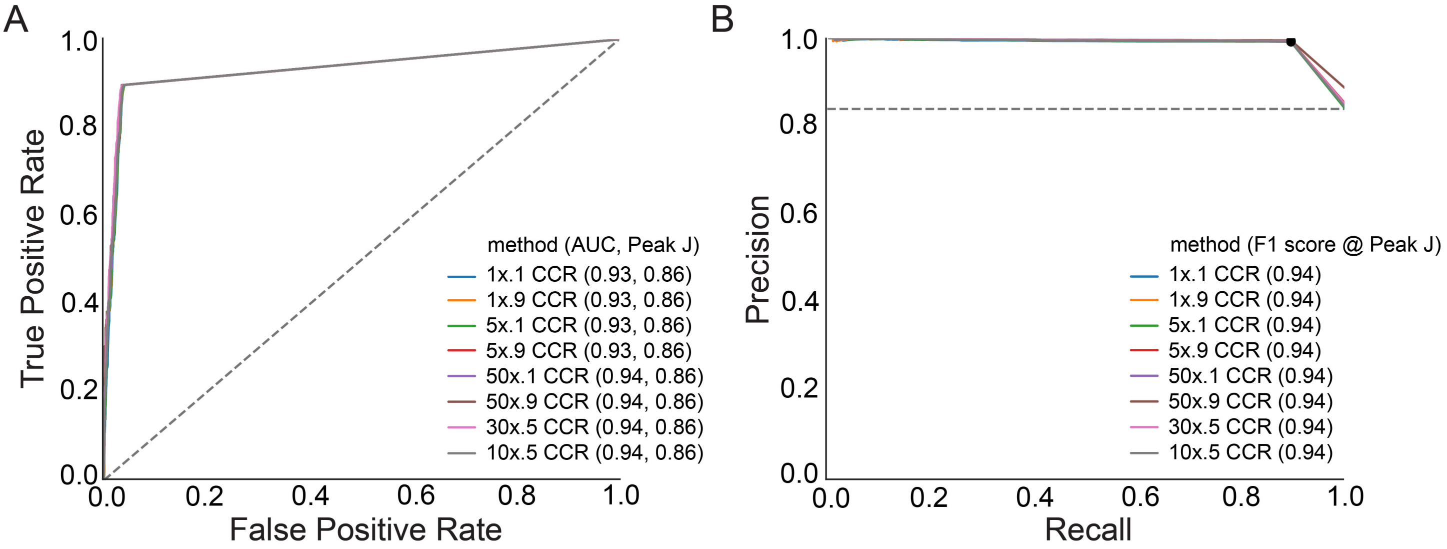
https://doi.org/10.1038/s41588-018-0294-6

In the format provided by the authors and unedited.

A map of constrained coding regions in the human genome

James M. Havrilla[1,2], Brent S. Pedersen[1,2], Ryan M. Layer[3,4] and Aaron R. Quinlan[1,2,5*]

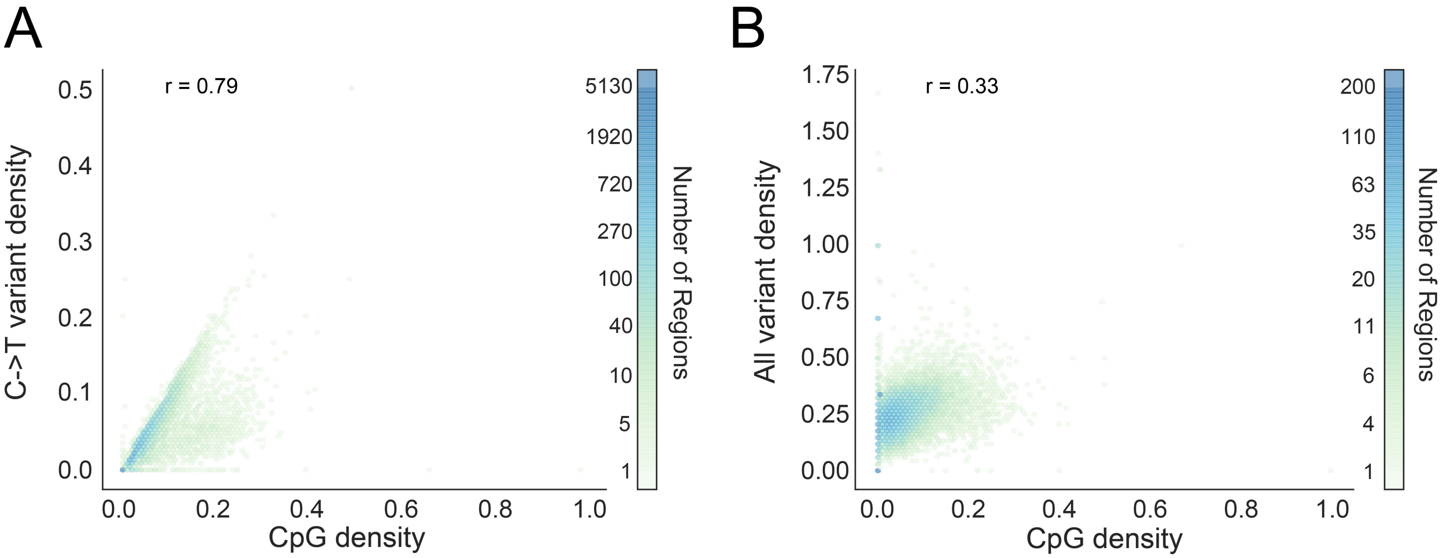
[1]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. [2]USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA. [3]BioFrontiers Institute, University of Colorado, Boulder, CO, USA. [4]Department of Computer Science, University of Colorado, Boulder, CO, USA. [5]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA. *e-mail: aaronquinlan@gmail.com



Supplementary Figure 1

Evaluation of CCR models by sequencing coverage threshold.

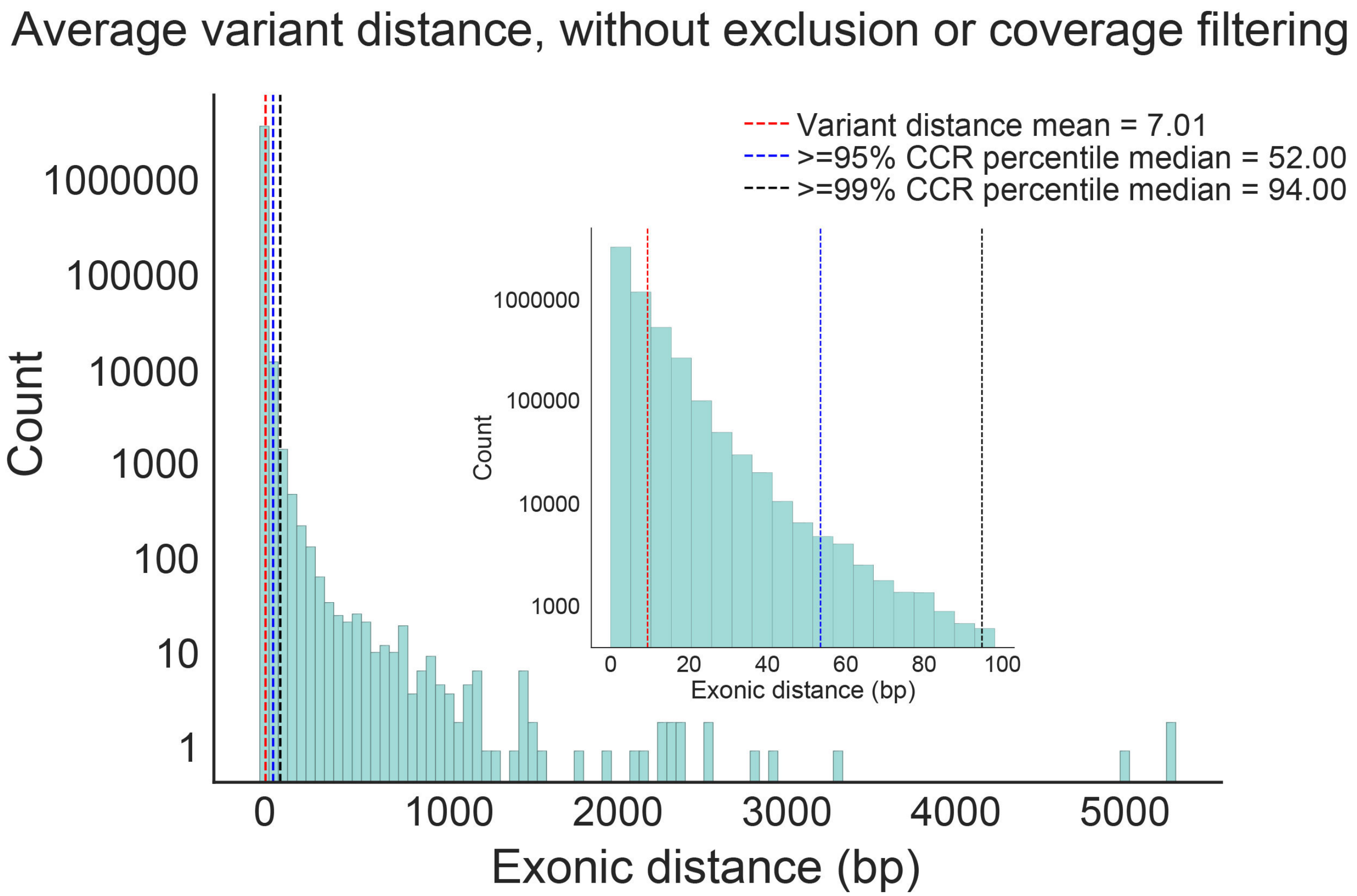
Evaluation of CCR models constructed using different coverage thresholds and different thresholds for the percentage of gnomAD individuals meeting the minimum coverage depth. For example, “10x.5 CCR” reflects a CCR model where every position in a CCR region was required to have 10× coverage in at least 50% of gnomAD individuals. **a**, ROC curve based on the ClinVar variant set. **b**, PR curve based on ClinVar. True positives are pathogenic variants and likely pathogenic variants from ClinVar. True negatives are variants labeled as benign from ClinVar. The performance of each model is clearly very similar, and the “10x.5 CCR” model imposed the most relaxed coverage requirement while exhibiting the highest performance. It was therefore chosen as the coverage threshold for the final model. 24,554 pathogenic variants from ClinVar were used, and 4,689 benign variants were used for the evaluation dataset.



Supplementary Figure 2

Correlation between exonic CpG density and genetic variation.

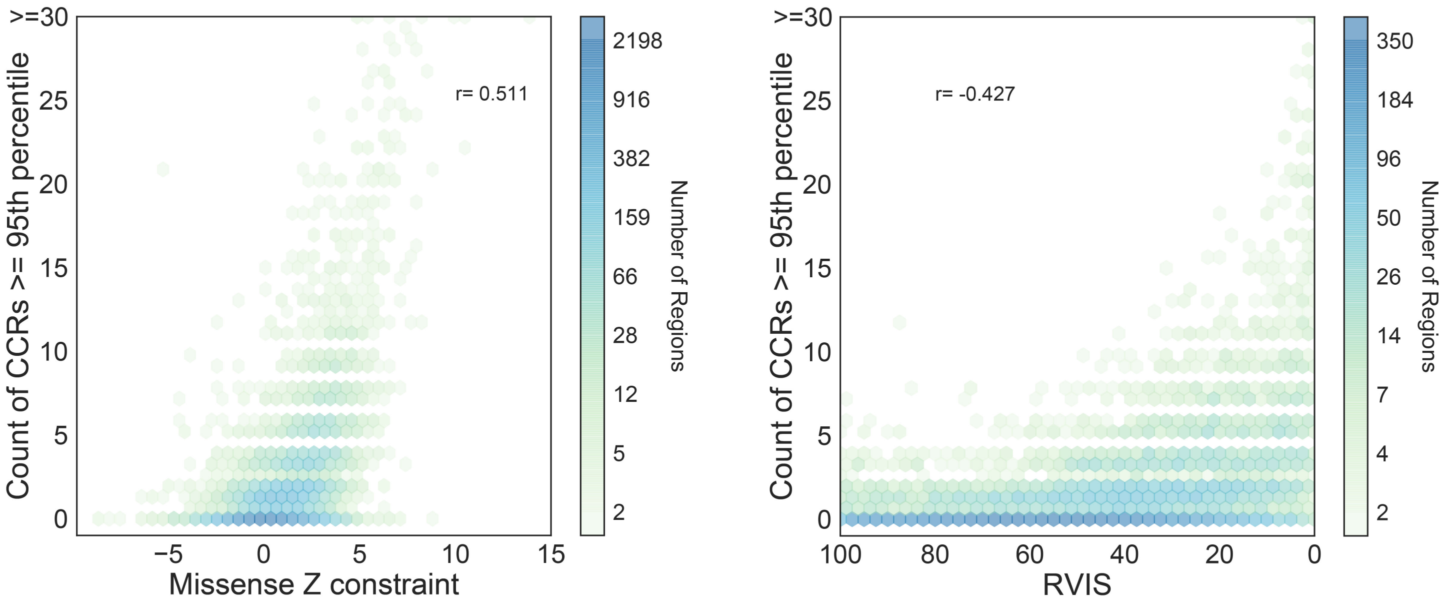
The sample size is the number of CCRs, which is 8,065,333 unique regions. Pearson's correlation was used. **a**, Exonic CpG density compared to the density of exonic C>T or G>A transitions. **b**, Exonic CpG density compared to the density of all exonic variant types.



Supplementary Figure 3

Average exonic distance for adjacent gnomAD variants.

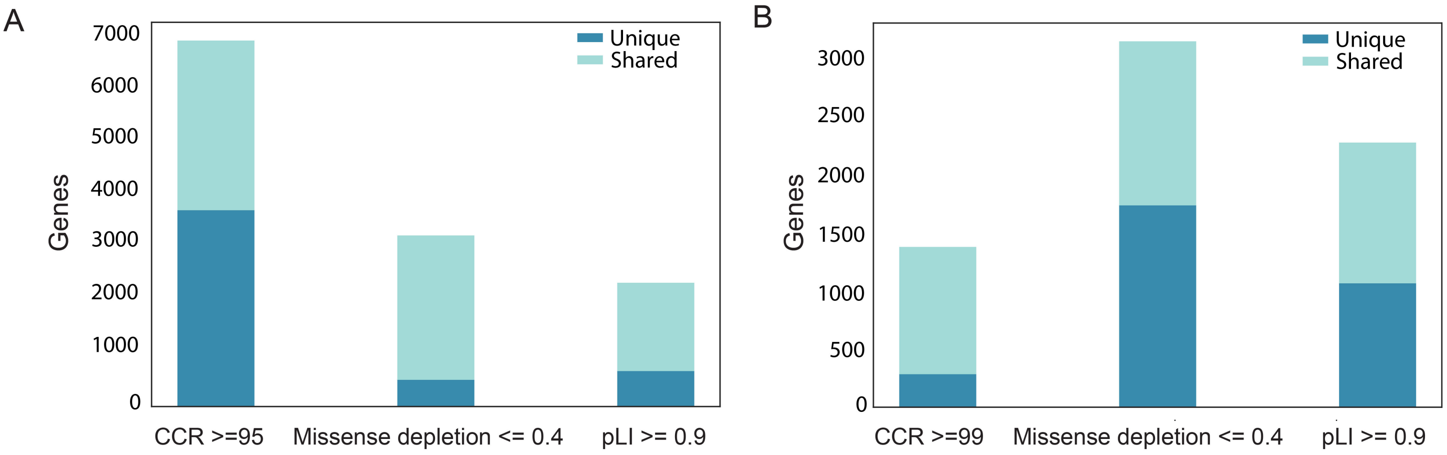
Distribution of the exonic distance between protein-changing (missense or LoF) variants in gnomAD without filtering regions by coverage, segmental duplications, or self-chains. The red dashed line is the average distance between protein-changing variants. The blue and black dashed lines represent the average length of CCRs in the 95th and 99th percentile, respectively.



Supplementary Figure 4

Correlation of constrained coding regions to other models of genic constraint.

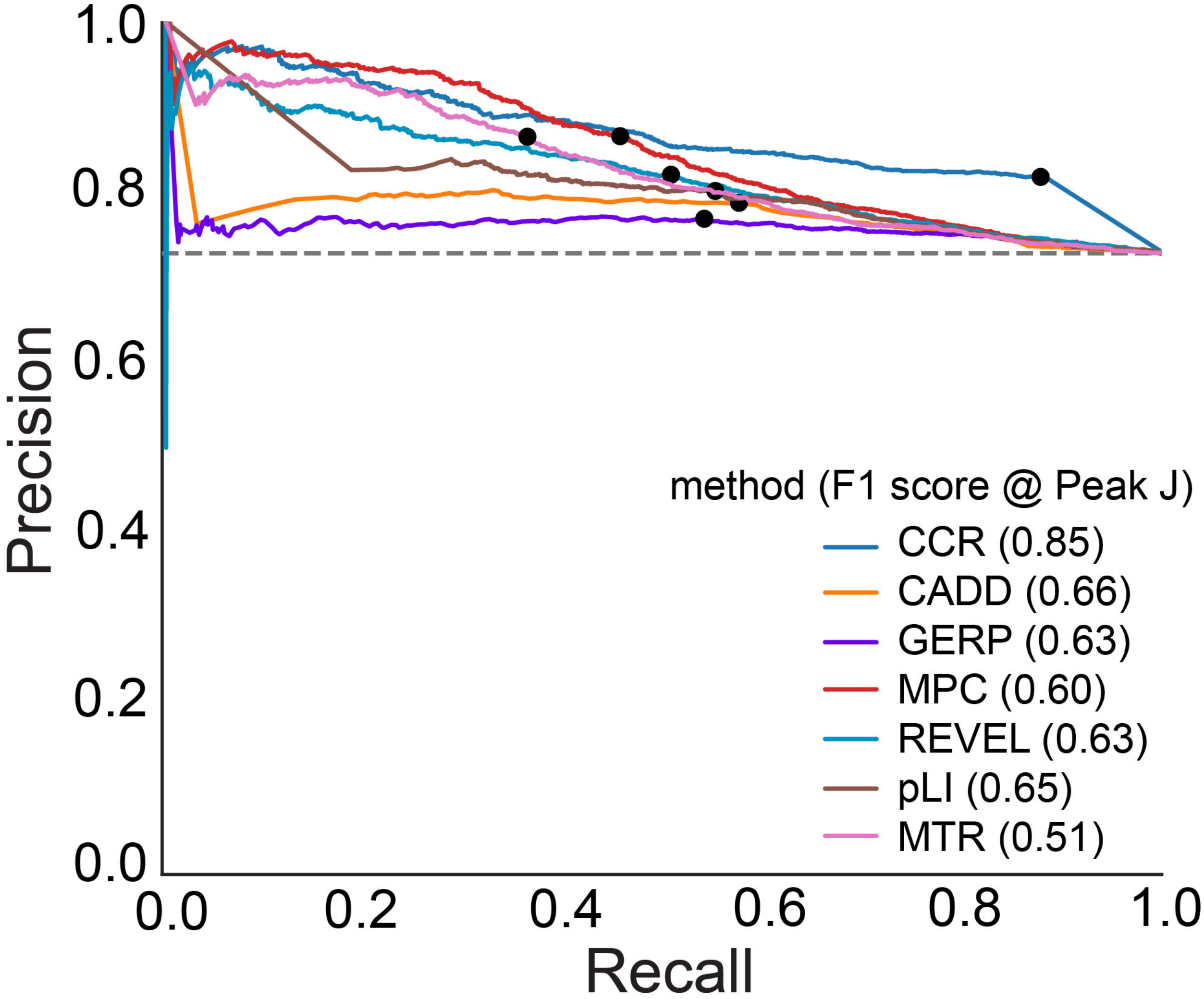
The sample size is the number of CCRs ≥95%, which is 21,650 unique regions, and the number of genes with a Missense Z constraint score or pLI score is 18,225 genes for both sets. **a**, The correlation between a gene's Missense Z metric (least to most constrained from left to right) and the number of CCRs in the 95th percentile or higher observed in the gene. **b**, The correlation between a gene's RVIS metric (least to most constrained from left to right) and the number of CCRs in the 95th percentile or higher observed in the gene.



Supplementary Figure 5

Total number of shared and unique genes across metrics for predetermined constraint metric cutoffs.

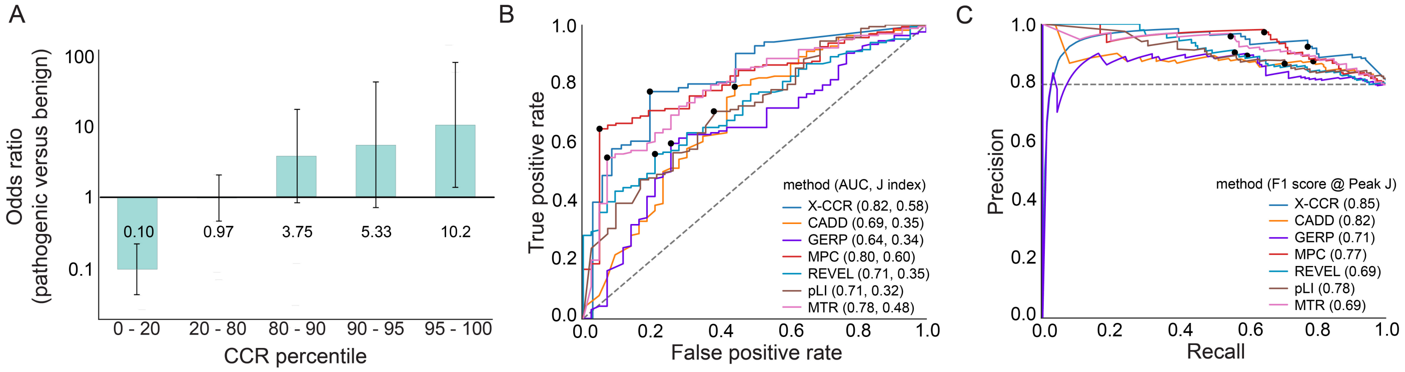
a,**b**, Comparison of genes covered by each metric's cutoff for constraint (CCR ≥ 95 (**a**) or 99 (**b**), pLI ≥ 0.9, and missense depletion ≤ 0.4). The dark blue bar indicates how many genes are unique to a particular metric's cutoff for constraint, and the light blue-green bar represents how many of the genes for that cutoff are shared with at least one of the other two metrics.



Supplementary Figure 6

Precision–recall (PR) curves for the developmental disorder de novo variant evaluation set.

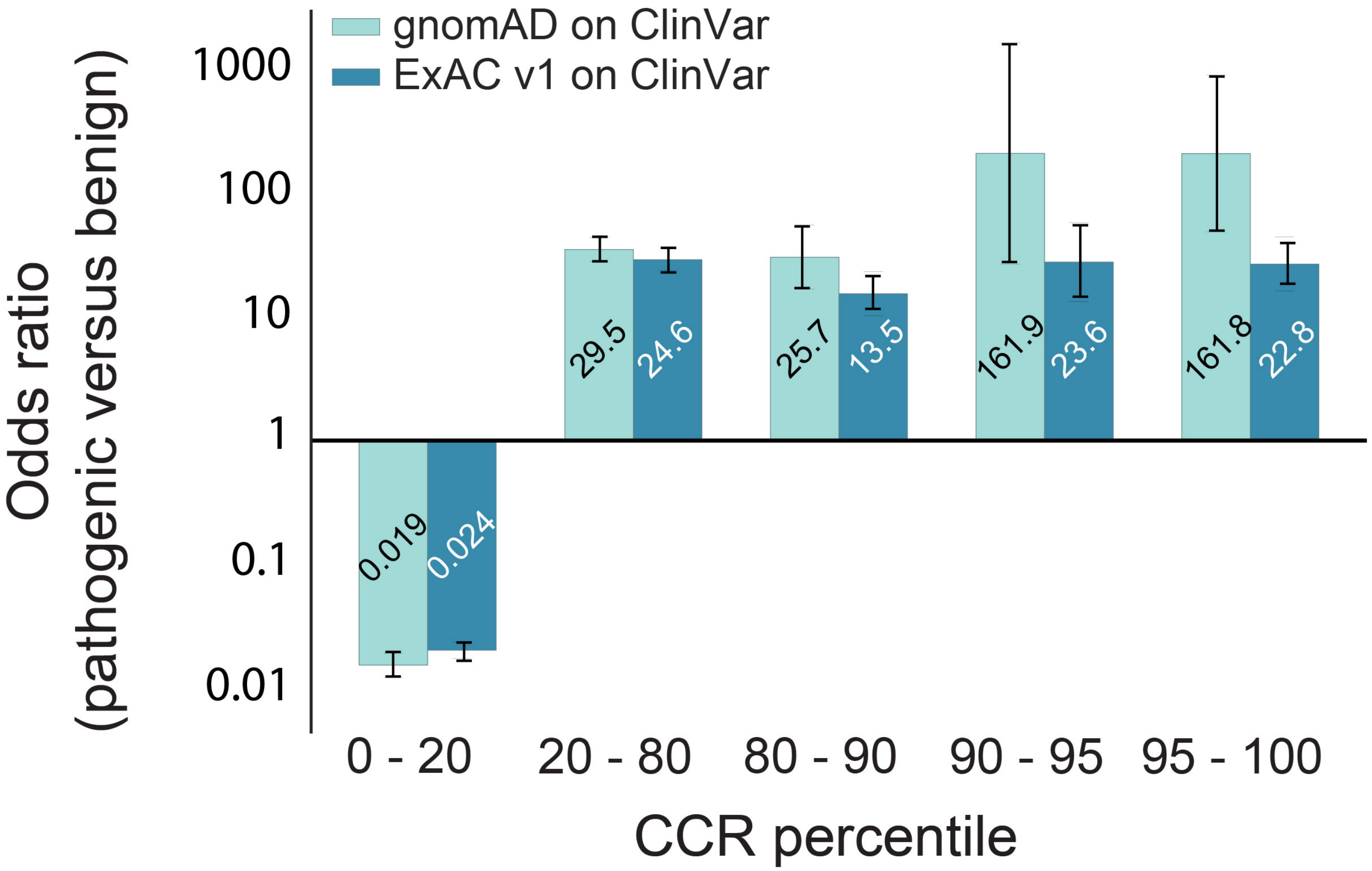
The true positives are 3,400 missense-only de novo variants from patients with developmental disorders. The true negatives are 1,269 missense de novo variants from the unaffected siblings of autism patients. The dots indicate the score cutoff with the maximal Youden J statistic for each tool. Values in parentheses indicate the F1 score, the weighted average of recall and precision, at the J-score cutoff.



Supplementary Figure 7

X-chromosome variant pathogenicity prediction comparison for CCR versus other metrics.

a, Enrichment of 166 pathogenic de novo mutations on the X chromosome in the most constrained X-CCRs and 43 benign mutations in the least constrained X-CCRs. The error bars represent 95% confidence intervals of 0.043–0.226 for the 0–20 bin, 0.46–2.07 for the 20–80 bin, 0.85–16.5 for the 80–90 bin, 0.69–41.1 for the 90–95 bin, and 1.35–77.2 for the 95–100 bin. **b**, ROC curve for the developmental disorder de novo variant evaluation set. The true positives are 166 missense-only de novo variants from patients with developmental disorders. The true negatives are 43 missense de novo variants from the unaffected siblings of autism patients. **c**, PR curve for X-CCR versus other metrics for the de novo set. The dots in **b** and **c** indicate the score cutoff with the maximal Youden J statistic for each tool. Values in parentheses indicate AUC and peak J score (respectively) for **b** and the F1 score, the weighted average of recall and precision, at the J-score cutoff for **c**.



Supplementary Figure 8

Odds ratio comparison between ExAC-based CCR and gnomAD-based CCR for the ClinVar variant set.

True positives are 24,554 pathogenic variants and likely pathogenic variants from ClinVar. True negatives are 4,689 variants labeled as benign from ClinVar. For ExAC v1, the 95% confidence intervals are 0.021–0.028 for the 0-20 bin, 20.5-29.6 for the 20–80 bin, 9.09–20.0 for the 80–90 bin, 11.8–47.4 for the 90–95 bin, and 14.1–36.8 for the 95–100 bin. For gnomAD, the 95% confidence intervals are 0.015–0.023 for the 0–20 bin, 23.9–36.6 for the 20–80 bin, 14.6–45.4 for the 80–90 bin, 22.8–1151.0 for the 90–95 bin, and 40.4–647.5 for the 95–100 bin.