AlphaCluster: Coevolutionary driven residue-residue interaction models enable quantifiable clustering analysis of de novo variants to enhance predictions of pathogenicity

Joseph Obiajulu
Columbia University https://orcid.org/0000-0002-9240-789X
Ranger Kuang
Columbia University https://orcid.org/0000-0001-5936-0586
Guoije Zhong
Columbia University
Jake Hagen
Columbia University
Chang Shu
Columbia University
Wendy Chung
Columbia University
Yufeng Shen (ys2411@cumc.columbia.edu)
Columbia University https://orcid.org/0000-0002-1299-5979

Article

Keywords: missense variants, protein structure, statistical genetics, pathogenicity

Posted Date: August 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1910518/v2

AlphaCluster: Coevolutionary driven residue-residue interaction models enable quantifiable clustering analysis of *de novo* variants to enhance predictions of pathogenicity

Joseph Obiajulu[2,1], Ranger Kuang[2], Guoije Zhong[2], Jake Hagen[2,1], Chang Shu[2,1], Wendy K. Chung[1, #], Yufeng Shen[2,2, #]

1. Department of Pediatrics, Columbia University, New York, NY, USA
2. Department of Systems Biology, Columbia University, New York, NY USA
3. Department of Biomedical Informatics, Columbia University, New York, NY, USA

Corresponding authors: W.K.C (wkc15@columbia.edu) and Y.S. (ys2411@cumc.columbia.edu)

Abstract

Missense variants have highly variable effects and effect size, which often makes it challenging to distinguish pathogenic and non-pathogenic variants and subsequently implicate new genes for disease association in studies of *de novo* and inherited rare variants. Importantly, missense variants can be the sole molecular mechanism for some genetic disorders, and so statistical approaches tailored for the analysis of missense variants are critical. Analysis of the clustering of missense variants is a promising approach which leverages the fact that missense variants in protein domains often have similar effects on function. Here we describe a new clustering analysis approach, AlphaCluster, a statistical method which quantifiably analyzes the spatial clustering of *de novo* variants by mapping missense residues onto the protein tertiary structure. We show that our approach can quantify the evidence supporting pathogenic missense variants and increase the power to detect clustering when compared to available genomic clustering tools. Using AlphaCluster, we identified genes newly implicated in autism spectrum disorder and neurodevelopmental disorders (NDD). We also apply AlphaCluster to protein complexes and detect an association between the gamma aminobutyric acid receptor complex (GABA-A $\alpha 1 \beta 2 \gamma 2$ receptor).

Introduction

De novo genetic variants are a significant contributing factor to early onset human diseases and conditions that impact reproductive fitness, such as neurodevelopmental disorders (NDD)[1,2], autism[3–9] and congenital anomalies[10–13]. *De novo* variants which result in complete loss-of-function (LoF) of the protein have traditionally been the main focus of *de novo* analyses. LoF variants often result in nonsense mediated decay and lead to haploinsufficiency, which can be severely biologically damaging and have consistent effect, making LoF variants both impactful and amendable to statistical analysis by aggregating LoF variants across most of a gene. On the other hand, missense variants, much more abundant than LoF variants, are variable in effect

and effect size. It is difficult to differentiate between benign and deleterious missense variants, so gene-wide aggregation of missense variants frequently exhibits low signal-to-noise ratios for missense variants, leading to challenges implicating such genes with a large proportion of missense variants as disease associated.

Nevertheless, missense variants are the main contributors to the disease mechanism or mode of action for certain genes. For example, a recent study of *PTEN* identified multiple molecular mechanisms underlying protein dysfunction, including several missense variants which appear to be dominant negative, resulting in less overall protein function than a monoallelic LoF variant[14]. Current methods to identify genes for which missense variants contribute to risk have focused on detecting and analyzing the enrichment of missense variants predicted to be damaging by computational algorithms, a means to increase the signal-to-noise ratio. For example, the TADA statistical method treats damaging missense variants (Dmis) as a particular class among others in a mixture model. However, the downside of this approach is that missense variants are essentially treated as less damaging LoF variants and nothing more, which misses an opportunity to leverage the unique aspects of location of missense variants as another important data element. Recurrent and/or clustered missense variants can help elucidate the genetic causes of conditions for which association from LoF variants has not been shown but will require more than a simple re-application of LoF driven statistical tools. Indeed, there are fundamental differences between LoF and some missense variants, especially their relative effect sizes, necessitate fundamentally different approaches in analysis.

One missense-specific approach is to exploit the locations of missense variants across genes in a gene family, and to search for significant clustering of missense variants within regions/domains[15]. Clustering of pathogenic missense variants is expected to often result in similar protein function. For example, it has recently been shown that damaging missense variants in *LONP1* which contribute to congenital diaphragmatic hernia and CODAS syndrome are located in distinct regions with different genetic modes of inheritance (dominant and recessive, respectively)[13]. Thus, clustering of pathogenic variants is not only non-random but may be phenotype specific, and thus can be used to establish phenotype and disease association.

Analysis of clustering of *de novo* variants to establish disease association is a nascent approach [1,16,17]. Here, we further expand the clustering analysis approach by quantitatively analyzing the clustering of missense variants by the three-dimensional locations of relevant amino acids within the folded protein and the functional relatedness between residues. Additionally, using predicted models for protein multimers, we also examine clustering of missense variants within an entire protein complex, the most relevant biological unit. We name the tool for these analyses "AlphaCluster."

This new approach is enabled by recent major advance in accuracy of protein folding prediction, such as AlphaFold[18,19] and RoseTTAFold[20,21], and the increase in publicly available genomic

data. While these predictions are not ground truth, they are highly accurate as demonstrated in CASP14 and provide meaningful information about the structure of proteins. The increasing availability of genomic data from large cohorts such as SPARK (Simons Foundation Powering Autism Research for Knowledge) provide sufficient numbers of individuals with specific conditions to allow for robust assessment of variant clustering.

Results

AlphaCluster Overview: Leveraging predicted tertiary structures for missense clustering analysis

AlphaCluster is a novel clustering analysis tool which enables statistically rigorous measurement of the degree of clustering of missense variants within the tertiary structure of a protein. The tertiary structure is user specified. The tool comes pre-loaded with tertiary structures from the AlphaFold Protein Structure Database developed by DeepMind and EMBL-EBI which contains protein folding predictions for 992,316 structures from the human proteome, to examine variant clustering in three-dimensional space. It draws inspiration from *denovonear*, which performs clustering analysis strictly based on genomic positions of variants, calibrated to background mutation rates. In addition to performing clustering analysis based on three-dimensional positioning in tertiary structures, AlphaCluster incorporates scores to predict alteration in function (such as gMVP[22] or CADD[23]) to put greater or lesser weight on variants predicted to be more or less damaging.

The main intuition behind the tool is that significant spatial clustering of missense variants of a tertiary structure, similar to the primary structure or genomic precursor, can be detected through a frequentist simulation approach. The general "closeness" of all variants is captured in a distance metric, and the observed distance is compared for extremity against a background distribution of distances observed from simulation under the null hypothesis of no spatial clustering. Ideally, this distance metric primarily captures the Euclidean distance between affected residues, as well as inherent properties of the variants which suggest potential pathogenicity. The distance metric can then be used to detect clustering of pathogenic missense variants.

The algorithm of AlphaCluster works as follows: The N variants of interest are fetched for a specified gene of interest, from a user defined list of *de novo* variants. For example, this may be a set of *de novo* variants from an autism cohort, NDD cohort, or some other condition. The critical information which must be included is the chromosome, genomic position, reference allele, alternative allele, and gene effect (i.e. LoF, missense or synonymous) of each variant. Optionally, if the user specifies, variants which fall below a certain score (such as CADD = 25 or gMVP rank score = 0.7) can be excluded from further analysis. By default, AlphaCluster chooses gMVP rank score = 0.7 as a floor threshold. Next, the tertiary structure is parsed for its residue sequence and the Cartesian coordinates of each residue. The appropriate transcript

which maps to the residue sequence is chosen, and if none so aligns, the tool halts because there is no function to map between genomic variants and tertiary structure. The Euclidean distance between all pairs of residues $\{R_i\}_{i=1}^{N}$ in which an observed variants maps, is calculated, which are then used to calculate a generalized mean of degree p, which is by default the geometric mean (p = -1):

$$d_{R_iR_j} = \sqrt{\left(x_{R_i} - x_{R_j}\right)^2 + \left(y_{R_i} - y_{R_j}\right)^2 + \left(z_{R_i} - z_{R_j}\right)^2}$$

$$\text{generalized mean}(p, \left\{d_{R_iR_j}\right\}_{1 \le i < j \le N}) = \left(\prod_{1 \le i < j \le N} d_{R_iR_j}^{\frac{1}{p}} \right)^p$$

In the special case where there are duplicate residues with variants, we additively increase each observed distance $d_{R_iR_j}$ by 3.5 Å, the approximate average length of one amino acid, and subsequently subtract 3.5 Å from the final geometric mean. This is a conservative approach, which essentially treats duplicate variants as neighboring variants to shift the zero distance to a small non-zero value. Additionally, the Euclidean distances between all pairs can be scaled based on a damaging-ness score (such as CADD {Kircher 2014} and gMVP {Zhang 2021}),

$$\tilde{d}_{ij} = \frac{d_{ij}}{s_i + s_j}$$

and these scaled distances used in the mean calculation if the user so wishes. By default, AlphaCluster scales the distances with gMVP rank scores. With the geometric mean metric for the clustering of the observed variants, or the observed geometric mean for shorthand, calculated, a null distribution of geometric means of N *de novo* variants is formed to deduce a p-value. Namely, a simulation is run to generate samples of N *de novo* variants under the null hypothesis (namely, variants occur in conformance to the background mutation rate). The geometric mean or scaled geometric mean is calculated for each sample, and a distribution for the geometric means or scaled geometric means under the null hypothesis is thus formed. We run our simulation with 1E9 iterations by default, but this value can be user specified. Finally, the p-value of the observed geometric mean or scaled geometric mean is computed from the simulated distribution. The workflow for the entire method is schematically depicted in Error! Reference source not found..

Increase in evidence for disease association using AlphaCluster compared to conventional burden analysis and 1D sequence-based clustering methods

Previous missense clustering methods have generated p-values from exclusively examining the genomic coordinates of variants. We explored the increase in evidence of pathogenicity which our new three-dimensional methods provide over the previous 1D approach. We selected proteins which were known to demonstrate *de novo* missense enrichment in autism and NDD

cohorts, to be used as true positives in our power analysis. For autism, we selected the genes *DNMT3A*[24], *CHD8*[4], *PTEN*[25] and *KDM5B*[5], and for NDD we selected *MAP3K7*[1], *TFE3*[1], *GRIN2A*[1,26], and *DEAF1*[1,27]. See **Error! Reference source not found.** and **Error! Reference source not found.** for a detailed overview of the previous evidence of clustering of missense variants in these eight proteins from Kaplanis et al[1] and Zhou et al[9].

Across various clustering tests (1D clustering, 3D clustering and 3D clustering with gMVP score scaling and a threshold of rank score 0.7), we calculated the mean p-values for a given random subsample of the total missense variants of the known gene over 100 trials, obtaining missense variants from random cohort samples of fixed sizes (ranging from 100, 500, 5,000, 10,000, 15,000, 20,000, and the full cohort of 21,020 for the autism cohort, and 100, 500, 5,000, 10,000, 15,000, 20,000, 25,000 and the full cohort of 31,783 for the NDD cohort). Subsampling the cohort enabled us to also run burden analyses in the form of the Poisson enrichment test, which is important, because our full AlphaCluster test combines clustering p-values (from 3D clustering with gMVP score scaling and a threshold of rank score 0.7) with these Poisson test p-values to detect likely missense disease mechanisms. Correspondingly, we performed analysis with Fisher combined p-values of the Poisson enrichment test and these different missense clustering tests, as well as compared against the p-values of the Poisson enrichment test as a baseline (Error! Reference source not found.**a** for autism and **Error! Reference source not found.a** for NDD), as a way to benchmark AlphaCluster. AlphaCluster showed a marked decrease of average p-values compared to tests which used 1D clustering or simple 3D clustering (without use of predictive damaging scores) in *DNMT3A, CHD8, PTEN* and *KDM5B* for the autism cohort, and *MAP3K7, TFE3, GRIN2A* and *DEAF1* for the NDD cohort.

Additionally, we estimated statistical power of risk gene discovery by de novo missense variants only. We combined the evidence from the clustering and enrichment tests at various significance thresholds using Fisher's method. We observed an increase in power over the 1D and simple 3D clustering analyses (Error! Reference source not found.**b** for autism and **Error! Reference source not found.b** for NDD) for all four true positive cases in the autism and NDD cohort.

AlphaCluster reveals several new candidate genes for NDD and autism

We reran the 1D clustering analysis which was performed in Kaplanis et al. using the *de novo* missense variants from the NDD cohort and enrichment p-values from the previous analysis. We reproduced 186 positive results compared to the original 188. The discrepancies are negligible (*MMGT1* at p-value = 3.70E-6, and *NR4A2* at p-value = 2.52E-6).

We turned to the entire set of 204 genes which reached genome-wide significance from 1D clustering analysis Fisher combined with DeNovoWEST enrichment analysis from Kaplanis et al[1] or through AlphaCluster. A substantial proportion of these genes are likely to have altered function mode of action[2], such as gain of function or dominant negative effects. Of the genes which reached genome-wide significance through either of these two methods, AlphaCluster showed more evidence of pathogenicity in 194 of the total 251 genes (**Error! Reference source not found.a**). Additionally, for completeness, we show the counts of genes which reached genome-wide significance through the 1D approach employed by Kaplanis et al., AlphaCluster, and a customized version of AlphaCluster which used CADD annotation scores (in place of gMVP rank scores) with scale scoring and a threshold of a CADD score of 25 (**Error! Reference source not found.b**).

When AlphaCluster was applied (3D clustering analysis which is further enhanced with gMVP annotation scores as described), we identified 50 genes which were not formerly identified at the genome-wide level from the 1D clustering analysis. Of these, 34 only reached significance by AlphaCluster and were not identified from Kaplanis et al., either from the missense driven analysis nor the LoF and missense driven enrichment test (**Supplemental Data 1**), whereas 16 reached genome-wide significance when LoF variants were considered**.** *YWHAG, PPPC3A,* and *DHX30* are three such genes. **Figure 4** highlights how our 3D clustering captures variant clustering which 1D analysis cannot. Finally, eleven of these 34 genes (*BMPR2*, *SLC18A3, KBTBD7, MAST3, PSMC3, KIAA0100, ZBTB39, CAMK4, TMEM63B, KAT8* and *ATF2*) are novel candidate genes in the sense that they were not identified by Kaplanis et al., nor are they listed with an associated phenotype in the Development Disorder Genotype - Phenotype Database (DDG2P), though other supporting studies may exist. These eleven genes are presented in **Error! Reference source not found.**. We analyzed these genes for evidence of NDD association, as well as molecular function and known protein interactions (see **Error! Reference source not found.**).

This same analysis was performed for autism, with a similar result. We reproduced the analysis of Zhou et al., in which eleven genes reached genome wide significance from missense enrichment p-values Fisher combined with 1D clustering analysis p-values. In our reproduced analysis, all of these genes reached genome wide significance, except the near-miss of *MYT1L* (p-value = 2.73E-06). An additional eight candidate genes which did not reach genome-wide significance from the missense enrichment combined with 1D clustering reached genome-wide significance through AlphaCluster (*GRIN2B, ADNP, CHD2, TAOK1, CLCN4, GABBR2, TBL1XR1* and

SATB2). Of these 8 genes, *GABBR2* and *SATB2* are novel risk genes in the sense that they did not reach genome-wide significance in Zhou et al or Satterstrom et al, nor had they conclusively been shown to be associated with autism. These eight genes are presented in **Error! Reference source not found.**, and existing supporting evidence for *GABBR2* and *SATB2* are summarized in Error! Reference source not found.**.**

Several complexes show significant clustering of missense variants in NDD

Thus far, we have demonstrated the ability and power of AlphaCluster, and 3D clustering analysis more generally, to provide evidence of pathogenic clustering in protein models and to detect potential missense disease mechanisms. Here, we show that these approaches can be extended to quantify the clustering of missense variants within a protein complex.

We applied the same clustering analysis as in the protein singleton case, aggregating *de novo* missense variants from the component proteins of a complex, running simulation aided analysis, using gMVP rank scores for scaling and a rank score of 0.7 as a lower threshold. An example of clustering analysis over a multimeric complex is presented for the GABA-A $\alpha1\beta2\gamma2$ pentamer (Error! Reference source not found.**a** and **b**), which is a GABA-A $\alpha1\beta2\gamma2$ receptor relevant to autism and NDD. It is composed of two GABA-A receptor $\alpha1\beta2\gamma2$ 1, two GABA-A receptor $\beta2$ and one GABA-A receptor $\gamma2$ protein subunits, encoded by *GABRA1, GABRB2*, and *GABRG2* respectively. It should be noted that while there was a pre-existing human model of GABA-A $\alpha1\beta2\gamma2$ (PDB ID: 6D6T[28]), we also generated multimeric predictions of GABA-A from scratch using AlphaFold's multimeric capabilities. We found a high level of congruence between the models created through both models. Our analysis of the clustering of the 14 *de novo* missense variants from our autism cohort on the five composite proteins (3 in *GABRA1*, 3 in *GABRB2* and 2 in *GABRG2*) yielded a 3D protein clustering p-value of 0.065, whereas the same analysis run separately with the 65 *de novo* missense variants from our NDD cohort (13 in *GABRA1*, 18 in *GABRB2* and 3 in *GABRG2*) yielding a 3D protein clustering p-value of 4.7e-4. The variants from NDD and autism lie in similar regions, namely the α-helices of the transmembrane domain, and so the clustering results are likely to be more robust for large autism cohorts as the number of observed DNVs in these proteins increases with larger sample sizes. Importantly, in NDD, the protein subunits of GABA-A show no observed LoF variants but only missense variants. Whereas *GABRA1* and *GABR2* reached genome wide significance both from AlphaCluster and 1D clustering analyses, *GABRG2* did not reach genome wide significance when considered as a singleton through AlphaCluster or previous 1D clustering analyses.

Discussion

We present AlphaCluster, a new method aimed to facilitate elucidation of the role of missense variants in genetic diseases. AlphaCluster allows users to quantify and statistically assess the clustering of missense variants for a given protein model, thus identifying proteins which show higher than expected clustering of variants that may alter protein function in a similar manner.

As demonstrated, this can be used to identify risk genes for disease and to identify pathogenic variants in established disease associated genes.

Additionally, our approach can study protein complexes and aggregate *de novo* variants across the protein complex and detect interactions between regions of different proteins within the complex. Previous work has shown that missense variants disrupting protein-protein interaction interface is enriched in autism[29] and other conditions. AlphaCluster is an advance to perform statistical test which more closely approximates the biologically functional unit. We anticipate that with the greater accuracy and availability of protein-protein interaction models, our method will have even greater impact.

Through use of AlphaCluster, we have identified several new candidate genes associated with NDD and autism. It is important to note that these candidate genes result entirely from our new method, and not any increase in sample size (as the trio cohorts we used were those from two previous studies). These candidate genes reached genome-wide significance, though functional assessment of the missense variants is still required to understand the molecular mechanism. It is also important to note that AlphaCluster can be applied to any disease cohort with *de novo* variant calls and can serve as an additional tool in the standard collection of WES/WGS statistical tests, alongside TADA and DeNovoWEST.

The core method of AlphaCluster presents opportunities for further expansion. Noticeably, the current method is applicable only to *de novo* variants, since the background mutation rate of *de novo* variants is well established. AlphaCluster could be extended to inherited variants, given a careful choice of inherited background variant frequency. Additionally, AlphaCluster uses static protein structures to provide locations for residues impacted by missense variants, whereas most proteins are truly dynamic in nature. The use of static models does limit the ability to detect clustering which may be more apparent in a different protein configuration than what one static model present. As the field of protein folding predictions produces more dynamic modeling of proteins, such as the dynamic modelling of an entire nuclear pore complex[30], AlphaCluster should be expanded to test for clustering on dynamic models.

In general, our results suggest new opportunities for the dual application of predicted protein models and large genomic cohort data. Looking ahead, we anticipate continued advancement on both fronts, with increasing genomic data availability and more precise protein and protein complex models. The power and applicability of AlphaCluster should increase with those advances.

Methods

Software implementation

AlphaCluster is a comprehensive expansion of denovonear[1]. It is a python script which wraps a core C++ library which performs the simulation calculations (the computational heavy lifting), and interfaces with this library through a cython intermediate layer. We create additional python scripts for the processing of PDB files for the Cartesian locations all residues.

Protein representation of preloaded models

AlphaCluster uses the canonical UniProtKB sequence to create the models of human proteome. Thus, for the preloaded PDB models are of the canonical UniProtKB sequence, although it is well known that proteins often exhibit multiple isoforms. These alternative isoforms can be explored given a protein model of this alternative isoform.

Mapping genomic variants to residue positions

Much care was taken to ensure a proper mapping of genomic variants to the impacted residue on the given protein. Our approach was to be conservative. We translated the canonical transcript of a given gene to its corresponding amino acid sequence and checked if this sequence was in perfect agreement with the protein sequence of the selected protein model. If it was not, but there was a subsequence of both which was in perfect alignment, we reduced the scope of our analysis to this subsequence and variants within. If there was still no alignment, another transcript is attempted. If all transcripts show no alignment, AlphaCluster returns a misalignment error.

Virtually all the relevant genes (the 6060 and 2468 genes for the NDD and autism cohort with at least two missense variants, respectively) were amendable to AlphaCluster, displaying perfect mapping between the canonical transcript and the canonical UniProtKB sequence.

Handling of repeat missense variants in clustering test

It is often the case that true risk genes and proteins present the identical missense variants, or the identical residue impacted by the missense. This introduces the difficulty of how to measure the distance between repeatedly impacted residues. Given our choice of metric, if there was not special handling, the geometric mean in the case of repeat residues would be zero. To correct for this, in the case of repeat residues, we increment each distance by 3.5, because the average length of a residue is 3.5 angstrom. Then, we proceed with calculating the mean, but and decrement this value by 3.5 after calculation:

$$d'_{R_iR_j} = d_{R_iR_j} + 3.5 = \sqrt{\left(x_{R_i} - x_{R_j}\right)^2 + \left(y_{R_i} - y_{R_j}\right)^2 + \left(z_{R_i} - z_{R_j}\right)^2} + 3.5$$

$$\text{generalized mean}\left(p, \left\{d_{R_iR_j}\right\}_{1 \le i < j \le N}\right) = \left(\prod_{1 \le i < j \le N} {d'_{R_iR_j}}^{\frac{1}{p}}\right)^p - 3.5$$

We note that this is a very conservative handling of this case, which approximates a repeatedly impacted residue with the case of two neighboring residues being impacted, whereas, in reality, the former is much more of a significant phenomenon.

Protein complex mode of AlphaCluster

The protein complex model of AlphaCluster runs in essentially the same manner as its singleton counterpart. It is important to note, however, that in tests where the complex has proteins which appear two or more times in the complex (such as a homodimer, or a trimer with a repeated protein) the identical residue is selected for each copy of the protein in the complex in our simulation. This prevents the case of having the observed missense variants be symmetrically coordinated in a way that increases the observed amount of clustering, whereas the simulated missense variants would not have this symmetry if each individual protein had uniquely simulated impacted residues.

Best use of damaging scores in AlphaCluster

The traditional use of predicted damaging scores, such as CADD, for the systematic analysis of missense variants is a thresholding approach, in which some threshold is used to categorize Dmis and below damaging missense (Bmis), which are thought to be noise, and to exclude the Bmis variants from further analysis. In AlphaCluster, we propose to not treat all missense variants the same, as do current 1D clustering approaches, but to scale the distance between two variants by the inverse sum of their damaging scores. This in effect puts more weight on two nearby Dmis variants more so than two nearby Bmis variants. We call this approach score-scaling. As already seen in the previous section, score-scaling shows significant decreases in mean p-value and increases in power over the 3D clustering approach unaided by damaging scores (in which the distance between two variants is the true Euclidean distance).

We tested if score-scaling was also more powerful than the traditional score threshold approach. We ran a power analysis, like those of the previous section, except where for the traditional score threshold approach, the enrichment test was the Poisson test where background mutation rate is that rate for the given classification of Dmis used for the clustering analysis. We determine that the scale thresholding provides a more powerful statistical test than does thresholding the missense variants (**Error! Reference source not found.**).

Poisson test used for enrichment

We elected to simply use the Poisson test to arrive at a significance for enrichment of missense variants.

Fisher combination of Poisson test p-value and clustering test p-value

The p-values for both the Poisson enrichment test and the clustering analysis are presumed to be independent under the null hypothesis. Indeed, if a given gene is not a risk gene, then neither is it expected to have any significant enrichment for missense variants, nor significant spatial clustering of missense variants. This assumption of independence under the null enables us to arrive at our final p-value, using Fisher's combined probability test for independent tests:

$$\chi^2_4 \sim -2\left(\log(p_{\text{enrichment}}) + \log(p_{\text{clustering}})\right)$$

The resulting p-value is the final p-value returned by AlphaCluster.

Cohorts and de novo variants

Several autism cohorts were used as a source of *de novo* variants from affected probands (SPARK[9,31], SSC[32] and ASC[33]), whereas the cohort from Kaplanis et al. was used to source NDD *de novo variants*. Cohort information for autism an NDD is presented in **Error! Reference source not found.**.

For NDD, we use the *de novo* variants of "31,058 parent–offspring trios of individuals with developmental disorders".

All *de novo* variants used in this study are from previously released cohorts. Possible duplicate proband inclusion was screen by identical variants, sex, and self-reported race and no duplicates were identified. Variant information for autism an NDD is presented in **Error! Reference source not found.**.

Resources

AlphaCluster: https://github.com/ShenLab/AlphaCluster
Denovonear: https://github.com/jeremymcrae/denovonear
AlphaFold Database: https://alphafold.ebi.ac.uk
gMVP: https://github.com/ShenLab/gMVP
CADD: https://cadd.gs.washington.edu
ChimeraX: https://www.cgl.ucsf.edu/chimerax
dbNSFP: https://sites.google.com/site/jpopgen/dbNSFP
UniProt: https://www.uniprot.org
lollipops: https://github.com/joiningdata/lollipops
Development Disorder Genotype - Phenotype Database (DDG2P): https://www.deciphergenomics.org/ddd/ddgenes
SFARI Gene: https://gene.sfari.org

Data availability

All of the ASD *de novo* variants used in this paper are presented in the supplementary table. For those of NDD, we direct the reader to Kaplanis 2020, where the *de novo* variants used were those reported.

Code availability

AlphaCluster is available on GitHub, along with code necessary to reproduce the results presented here. The repository is intended to be user friendly, and easily applicable to other WES/WGS cohorts. The *de novo* variants for NDD and autism come pre-loaded, along with those from CHD, CDH, epilepsy, and schizophrenia trio cohorts.

Acknowledgements

We thank members of Chung and Shen labs at Columbia University for helpful discussions. This work was supported by NIH grants R01GM120609 and Simons Foundation Autism Research Initiative (SIMONS606450).

References

1. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).

2. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).

3. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).

4. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).

5. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).

6. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).

7. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **180**, 568-584.e23 (2020).

8. Feliciano, P. *et al.* Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *NPJ Genomic Med.* **4**, 19 (2019).

9. Zhou, X. *et al. Integrating* de novo *and inherited variants in over 42,607 autism cases identifies mutations in new moderate risk genes.* http://medrxiv.org/lookup/doi/10.1101/2021.10.08.21264256 (2021) doi:10.1101/2021.10.08.21264256.

10. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262–1266 (2015).

11. Jin, S. C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).

12. Qi, H. *et al.* De novo variants in congenital diaphragmatic hernia identify MYRF as a new syndrome and reveal genetic overlaps with other developmental disorders. *PLOS Genet.* **14**, e1007822 (2018).

13. Qiao, L. *et al.* Likely damaging de novo variants in congenital diaphragmatic hernia patients are associated with worse clinical outcomes. *Genet. Med.* **22**, 2020–2028 (2020).

14. Post, K. L. *et al.* Multi-model functionalization of disease-associated PTEN missense mutations identifies multiple molecular mechanisms underlying protein dysfunction. *Nat. Commun.* **11**, 2073 (2020).

15. Liang, S., Mort, M., Stenson, P. D., Cooper, D. N. & Yu, H. *PIVOTAL: Prioritizing variants of uncertain significance with spatial genomic patterns in the 3D proteome*. http://biorxiv.org/lookup/doi/10.1101/2020.06.04.135103 (2020) doi:10.1101/2020.06.04.135103.

16. Lelieveld, S. H. *et al.* Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. *Am. J. Hum. Genet.* **101**, 478–484 (2017).

17. Goldmann, J. M. *et al.* Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50**, 487–492 (2018).

18. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

19. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

20. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

21. Humphreys, I. R. *et al.* Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).

22. Zhang, H., Xu, M. S., Chung, W. K. & Shen, Y. *Predicting functional effect of missense variants using graph attention neural networks.* http://biorxiv.org/lookup/doi/10.1101/2021.04.22.441037 (2021) doi:10.1101/2021.04.22.441037.

23. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021).

24. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).

25. Goffin, A., Hoefsloot, L. H., Bosgoed, E., Swillen, A. & Fryns, J.-P. PTEN mutation in a family with Cowden syndrome and autism. *Am. J. Med. Genet.* **105**, 521–524 (2001).

26. Endele, S. *et al.* Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat. Genet.* **42**, 1021–1026 (2010).

27. Vulto-van Silfhout, A. T. *et al.* Mutations Affecting the SAND Domain of DEAF1 Cause Intellectual Disability with Severe Speech Impairment and Behavioral Problems. *Am. J. Hum. Genet.* **94**, 649–661 (2014).

28. Zhu, S. *et al.* Structure of a human synaptic GABAA receptor. *Nature* **559**, 67–72 (2018).

29. Chen, S. *et al.* De novo missense variants disrupting protein–protein interactions affect risk for autism through gene co-expression and protein networks in neuronal cell types. *Mol. Autism* **11**, 76 (2020).

30. Mosalaganti, S. *et al. Artificial intelligence reveals nuclear pore complexity.* http://biorxiv.org/lookup/doi/10.1101/2021.10.26.465776 (2021) doi:10.1101/2021.10.26.465776.

31. Feliciano, P. *et al.* SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron* **97**, 488–493 (2018).

32. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. *Neuron* **68**, 192–195 (2010).

33. Buxbaum, J. D. *et al.* The Autism Sequencing Consortium: Large-Scale, High-Throughput Sequencing in Autism Spectrum Disorders. *Neuron* **76**, 1052–1056 (2012).

Acknowledgements

We are grateful to the SPARK Consortium for their diligent work in producing the largest autism WES trio cohort to date. We thank the DDD Consortium and GeneDx for curating the largest neurodevelopmental disease WES trio cohort. This work was supported by NIH grants R01GM120609 and U01HL153009, and Simons Foundation Autism Research Initiative (SIMONS606450).

Author information

Authors and Affiliations

Department of Systems Biology, Columbia University, New York, NY USA
Yufeng Shen, Joseph Obiajulu

Department of Pediatrics, Columbia University, New York, NY, USA
Wendy K. Chung, Joseph Obiajulu

Department of Biomedical Informatics, Columbia University, New York, NY, USA
Yufeng Shen

Contributions

J.O. and Y.S. conceived the project and designed the study. J.O. created the AlphaCluster codebase, performed the experiments, and analyzed the data. G.Z. assisted in the GO term gene enrichment analysis. J.O. and R.K. created the multimer protein predictions models. A.S. and J.H. assisted in autism WES data processing. Y.S. and W.K.C. oversaw the study. J.O. wrote the manuscript with input from Y.S. and W.K.C.

Extended data

Supplementary Figures and Tables (Supplementary_Figures_and_Tables.pdf)
Supplementary Data (Supplementary_Data_1.xlsx)

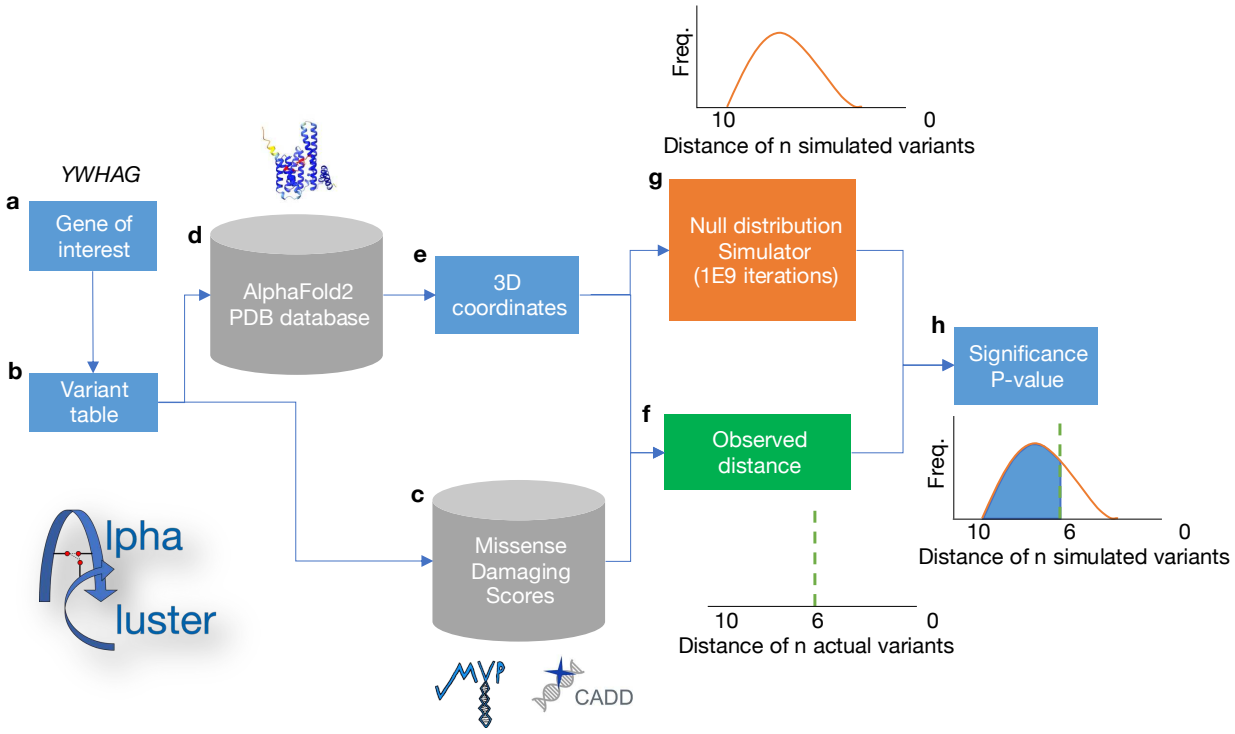

Figure 1: Schematic of AlphaCluster infrastructure. (a) A gene of interested is selected (b) the n variants of this gene of interest are fetched from the user specified variant table (with possible scale thresholding, which only selects some category of Dmis variants) (c) if specified, the missense damaging scores for these variants and all potential variants are fetched for later use (d) user specified protein or protein multimeric complex three-dimensional model is loaded (e) the 3D coordinates of the central carbon atom of each amino acid is retrieved (f) the observed geometric mean of the pairwise distances between each variant of the gene of interest is calculated; if desired, the pairwise distances can be inversely scaled by the sum of the damaging scores of the pair or variants which are below a given threshold can be excluded (g) for 1E9 iterations (or an otherwise user specified iteration count), n random variants are selected from all the possible variants in the gene of interest, with respect to the underlying background mutation rate. The geometric mean of the pairwise distances (by default with score scaling) is calculated and these geometric means are used to form a null distribution of geometric means (h) the null distribution is used to designate a p-value for the observed geometric mean of the n actual variants observed in the gene of interest.

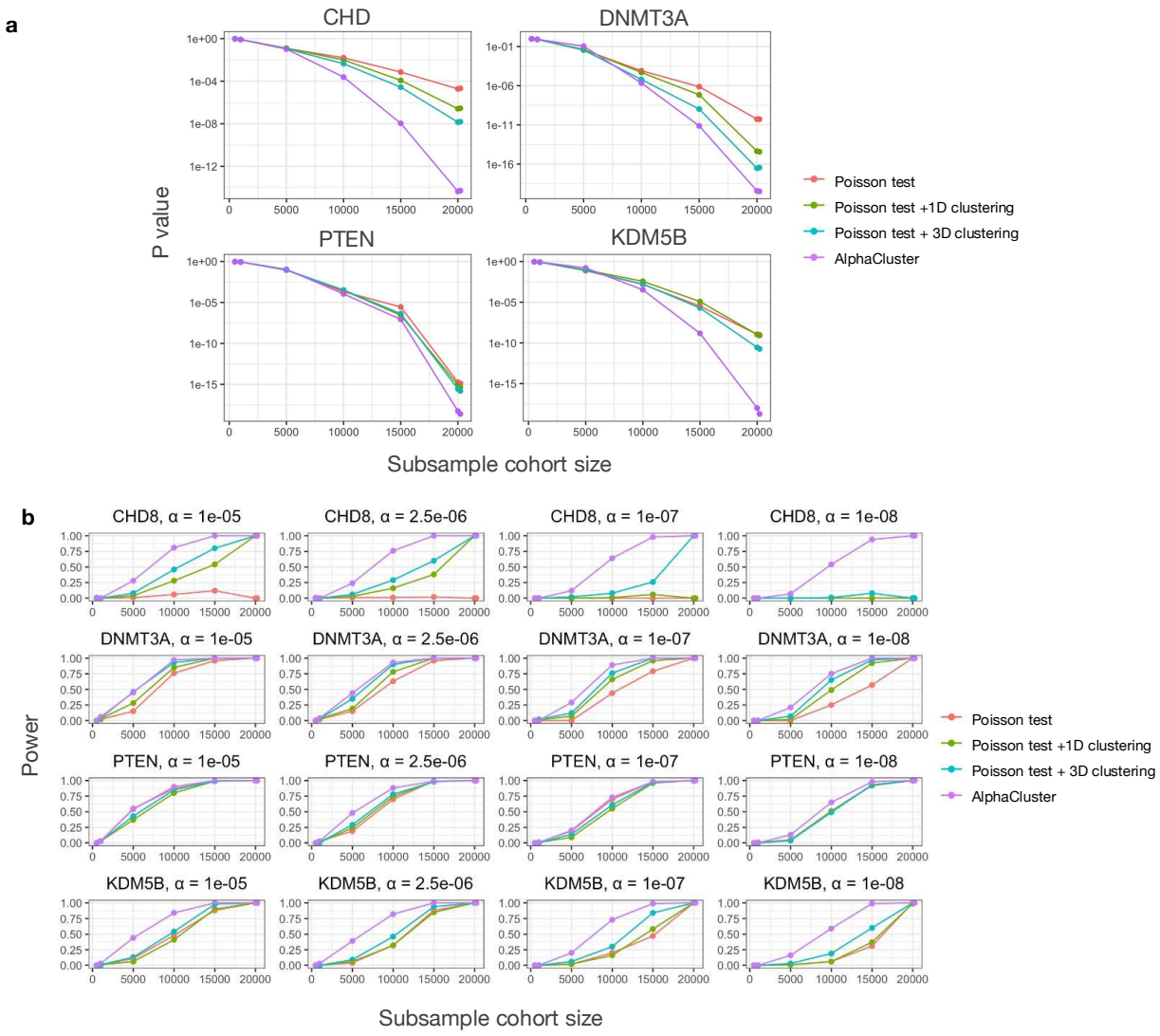

Figure 2: (a) Mean p-value and (b) power of 100 runs of various clustering tests performed over the missense variants of *CHD8*, *DNMT3A*, *PTEN*, and *KDM5B* from autism cohort across random cohort subsample sizes (100, 500, 1,000, 5,000, 10,000, 15,000, 20,000, and the full cohort of 21,020). Power was calculated at significance threshold 2.5E-6. The tests are 1D genomic clustering Fisher combined with Poisson test 3D protein clustering Fisher combined with Poisson, and our AlphaCluster test; additionally, these are compared to the baseline Poisson test.

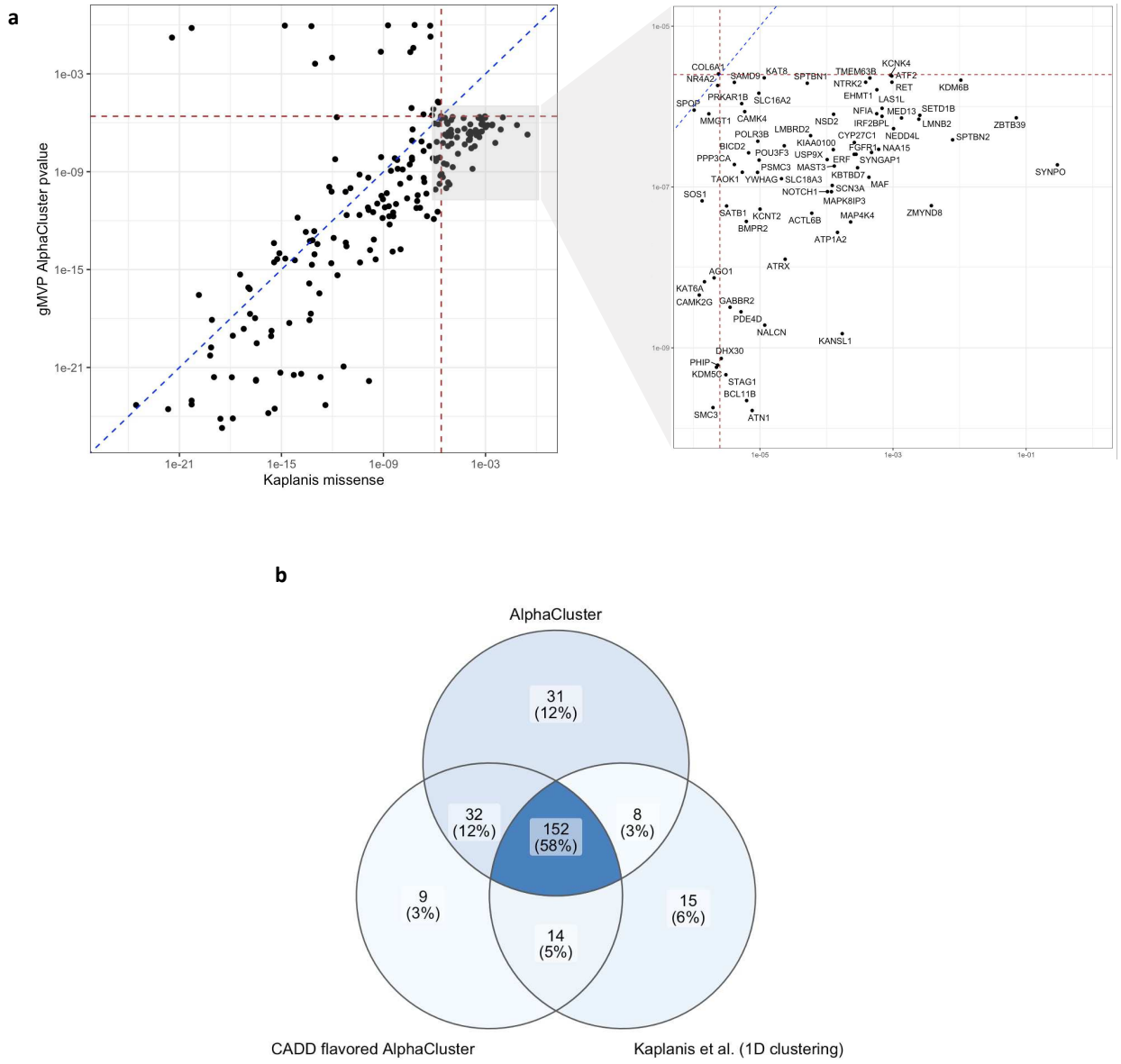

Figure 3: (a) A comparison of p-values from the 1D clustering combined with DeNovoWEST enrichment test versus AlphaCluster. Of the genes which reached genome-wide significance through either of these two methods, AlphaCluster showed more evidence of pathogenicity in 194 of the total 251 genes (b) a Venn diagram displaying comparative analysis of genes reaching genome-wide significance through AlphaCluster, AlphaCluster with CADD annotation scores instead of gMVP scores (CADD flavored AlphaCluster), and 1D clustering combined with DeNovoWEST missense enrichment test from Kaplanis et al.

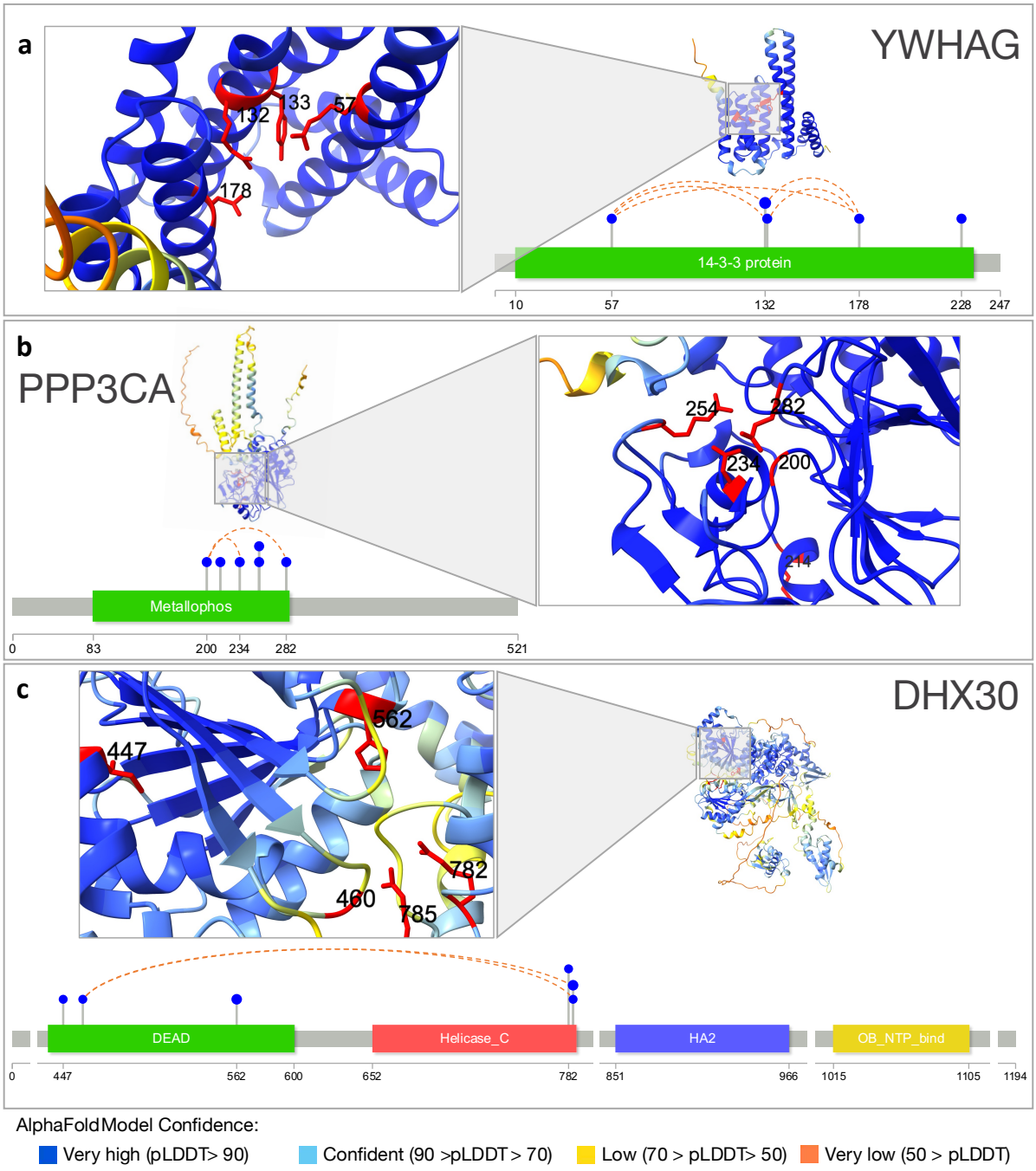

Figure 4: AlphaCluster lead to increased evidence of missense variant clustering due to better capturing of the true Euclidean distance between missense variants not properly represented from the genomic mapping in (a) *YWHAG* (b) *PPP3CA* and (c) *DHX30*. Views show affected amino acids (in red) which are closer than genomic distance would suggest and dotted red lines on the genomic map highlight which distances between missense variants (in blue) are much

closer in Euclidean space than genomic space. Open-source package lollipops[27] was used in creation of the lollipop graphs.

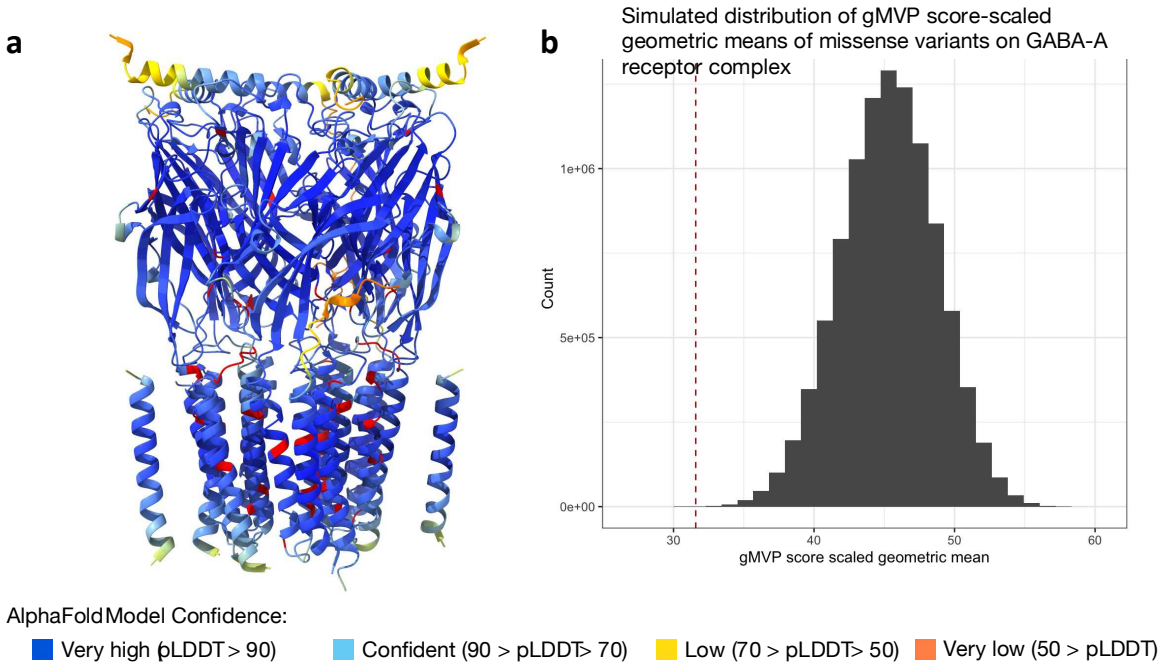

Figure 5: (a) Protein model of pentamer GABA-A $\alpha 1\beta 2\gamma 2$ subunits, with location of *de novo* variants from NDD (residues colored red). GABA-A alpha-1, GABA-A beta-2 and GABA-A gamma-2 have 13, 18, and 7 missense variants, respectively. (b) Histogram of geometric mean of distances between simulated variants choices, with choices calibrated with background mutation frequencies. The red line is the observed 3D geometric mean (uncalibrated by missense scores) of missense variants in NDD, corresponding to a p-value = 4.7E-4.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFiguresandTables.pdf
- SupplementaryData1.xlsx