Proteome-Wide Assessment of Clustering of Missense Variants in Neurodevelopmental Disorders Versus Cancer

Jeffrey K. Ng[1,#], Yilin Chen[2,#], Titilope M. Akinwe[1,3,#], Hillary B. Heins[1], Elvisa Mehinovic[1], Yoonhoo Chang[1,4], Zachary L. Payne[1,3], Juana G. Manuel[1], Rachel Karchin[2,5,6,7]*, and Tychele N. Turner[1,8]*

1. Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA.
2. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA.
3. Molecular Genetics & Genomics Graduate Program, Washington University School of Medicine, St. Louis, MO 63110, USA.
4. Human & Statistical Genetics Graduate Program, Washington University School of Medicine, St. Louis, MO 63110, USA.
5. Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA.
6. The Sidney Kimmel Comprehensive Cancer Center, School of Medicine, Johns Hopkins University, Baltimore, MD, USA.
7. Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, USA.
8. Intellectual and Developmental Disabilities Research Center, Washington University School of Medicine, St. Louis, MO, USA.

co-first authors

*Correspondence to: tychele@wustl.edu & karchin@jhu.edu

Tychele N. Turner, Ph.D.

Washington University School of Medicine

Department of Genetics

4523 Clayton Avenue

Campus Box 8232

St. Louis, MO 63110

Rachel Karchin, Ph.D.

Johns Hopkins University

Departments of Biomedical Engineering and Computer Science

Institute for Computational Medicine

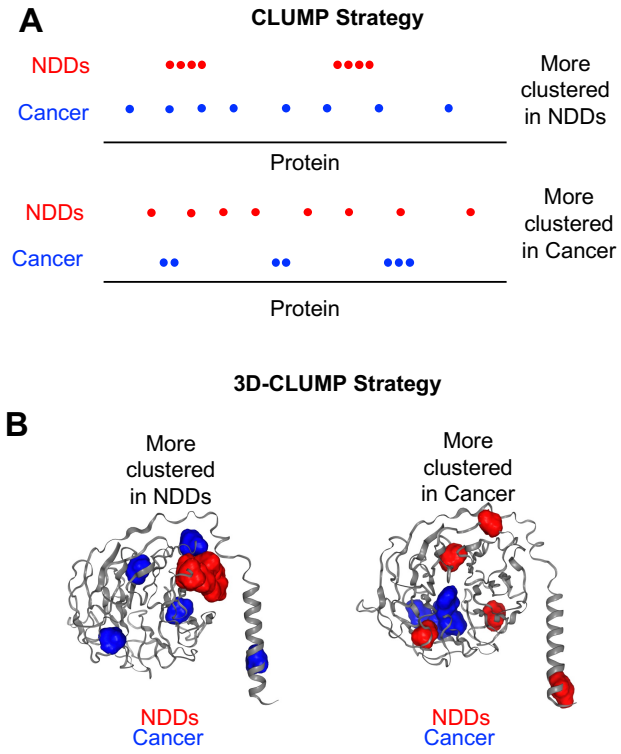
Baltimore, MD 21205

ABSTRACT

Missense de novo variants (DNVs) and missense somatic variants contribute to neurodevelopmental disorders (NDDs) and cancer, respectively. Proteins with statistical enrichment based on analyses of these variants exhibit convergence in the differing NDD and cancer phenotypes. Herein, the question of why some of the same proteins are identified in both phenotypes is examined through investigation of clustering of missense variation at the protein level. Our hypothesis is that missense variation is present in different protein locations in the two phenotypes leading to the distinct phenotypic outcomes. We tested this hypothesis in 1D protein space using our software CLUMP. Furthermore, we newly developed 3D-CLUMP that uses 3D protein structures to spatially test clustering of missense variation for proteome-wide significance. We examined missense DNVs in 39,883 parent-child sequenced trios with NDDs and missense somatic variants from 10,543 sequenced tumors covering five TCGA cancer types and two COSMIC pan-cancer aggregates of tissue types. There were 57 proteins with proteome-wide significant missense variation clustering in NDDs when compared to cancers and 79 proteins with proteome-wide significant missense clustering in cancers compared to NDDs. While our main objective was to identify differences in patterns of missense variation, we also identified a novel NDD protein BLTP2. Overall, our study is innovative, provides new insights into differential missense variation in NDDs and cancer at the protein-level, and contributes necessary information toward building a framework for thinking about prognostic and therapeutic aspects of these proteins.

INTRODUCTION

Neurodevelopmental disorders (NDDs) affect ~1% of the population and include autism, developmental delay, epilepsy, and intellectual disability. There have been several studies that have identified genetic contributions to NDDs including common variants (1), rare inherited variants (2, 3), and *de novo* variants (DNVs) (4-43). In particular, the study of DNVs through enrichment at the gene-level (44) has identified >300 genes involved in NDDs (36, 43). The initial discoveries were genes with high-impact likely gene-disruptive (LGD) DNVs (e.g., stop-gain, splice-site acceptor, splice-site donor, frameshift) (22, 45). The estimation of gene discovery from DNVs suggests a plateau at ~30,000 parent-child sequenced trios for LGD DNVs and for bioinformatically predicted severe missense DNVs a plateau at ~10,000 parent-child sequenced trios. However, where the plateau will be for gene discovery for the broader class of missense DNVs contributing to NDDs has been difficult to determine as they are harder to interpret than LGD DNVs (36). This highlights the importance for the development of computational and statistical tools to study this class of variation. Identification of relevant missense DNVs is critical for prognostics, functional testing, and future therapeutic strategies.



Our AlphaFold prediction for NP_002065.1 is used as an example protein in this figure.

Figure 1: Schematic of Examples of the CLUMP and 3D-CLUMP Methods. A) Two proteins are shown: one where there is more clustering in NDDs (Top) and one where there is more clustering in cancer (Bottom). B) Our AlphaFold prediction for NP_002065.1 (used as an example only) is shown in this image where variants are placed to exemplify more clustering in NDDs (Left) and more clustering in cancer (Right).

It is generally accepted that recurrent missense variants at a single amino acid residue position or in close proximity in a protein is a signature of strong positive selection (46). This hypothesis has motivated the search for clusters of driver mutations in cancers, both in one-dimensional gene or protein sequence and three-dimensional protein structures (47-50). While clusters of driver mutations have primarily been associated with oncogenes (51), we have previously shown that these clusters occur in both oncogenes and tumor suppressor genes (50). Furthermore, we developed the software CLUMP to compare clustering of germline variants in autosomal dominant vs. recessive Mendelian diseases and we identified more clustering of missense variants in autosomal dominant than autosomal recessive diseases (52). Application of CLUMP to NDDs has identified proteins with clustering of missense DNVs (53), and other groups have utilized other strategies to assess clustering of missense DNVs in NDDs at the gene or protein level (43, 54, 55).

In this study, we 1) apply the original CLUMP method to perform case-control testing to examine clustering in 1D protein space (**Figure 1A**), 2) and develop a new method called 3D-CLUMP to perform case-control testing to examine clustering in three-dimensional (3D) protein structure (**Figure 1B**). Both strategies are advantageous because they do not use *a priori* information about domains or known regions of functional importance within a protein.

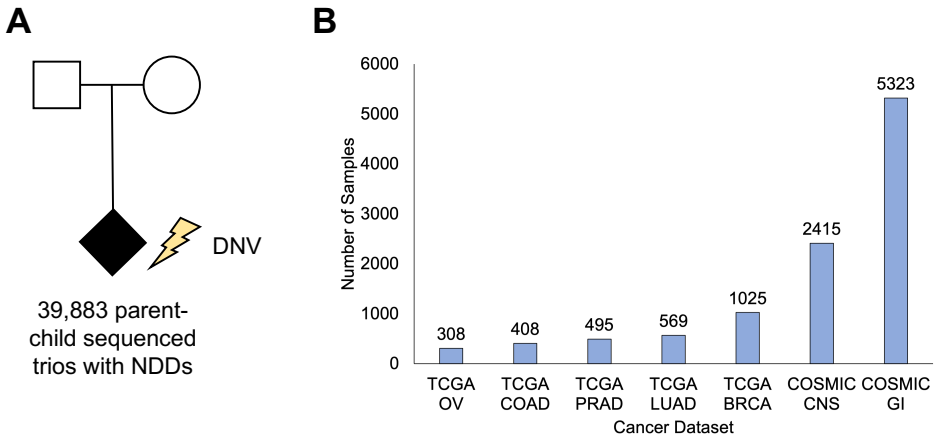


Figure 2: Variant Data Types in This Study. A) NDD data consisted of 39,883 parent-child sequenced trios (the lightning bolt is used to exemplify DNVs which, by definition, are only found in children). B) Cancer data consisted of 10,543 individuals from the TCGA and COSMIC databases.

As syndromic and genome-wide significant NDD genes have been identified, there has been an interesting observation that some of these genes are also implicated in cancers. Initial examples included *PTEN* in Cowden syndrome and *NF1* in Neurofibromatosis I (56, 57). More recent examples include *CHD8* in autism and in gastrointestinal cancers (45). However, for the majority of genes it remains unclear whether the specific variants identified in NDDs will later lead to cancer or not. In 2016, a forum paper by Crawley et al. (57) described an overlap of genes involved in both autism and cancer. They noted the centering of these genes on molecular pathways involved in gene regulation (e.g., chromatin, transcription, signaling). This was further examined in a review paper by Nussinov et al. in 2022 (58). Understanding why some of the same genes are involved in both NDDs and cancer is essential for prognostics and future therapeutics. In this study, we test the hypothesis that some genes involved in both NDDs, and cancer exhibit differential missense DNV clustering in the two phenotypes. To test this hypothesis, we aggregated NDD DNVs from the literature (37, 43) and cancer somatic variants from The Cancer Genome Atlas (TCGA) and Catalog Of Somatic Mutations in Cancer (COSMIC) databases and tested them using our existing (i.e., CLUMP (52) and our newly developed (3D-CLUMP) clustering tools.

In this study, our objectives are to identify 1) proteins which exhibit missense clustering in NDDs and not cancer and 2) proteins that exhibit missense clustering in cancer and not NDDs. This work involves the development of a new computational tool of broader use to the research community for assessing clustering of missense variation in 3D protein structures and provides novel insights into variation involved in cancer and NDDs with potential prognostic and therapeutic implications.

MATERIALS AND METHODS

Annotation of Variant Data

We aggregated DNV data from 39,883 parent-child sequenced NDD trios (37, 43) (**Figure 2A**). The DNVs were annotated with VEP (59) to RefSeq protein isoforms. Somatic variants (**Figure 2B**) were downloaded from TCGA at https://gdc.cancer.gov/about-data/publications/mc3-2017 for the following cancer types: Breast Cancer (BRCA), Colon Adenocarcinoma (COAD), Lung Adenocarcinoma (LUAD), PRAD (Prostate Adenocarcinoma), and Ovarian Cancer (OV). Somatic variants were also downloaded from the COSMIC database at https://cancer.sanger.ac.uk/cosmic/download in which multiple tissue types were aggregated in the categories of CNS (Central Nervous System) or GI (Gastrointestinal Track). CNS tissue types: Aqueduct of Sylvius, Basal ganglia, Brain Brainstem, Caudate nucleus, Cerebral cortex, Cerebral hemisphere, Cerebrum Chiasm, Choroid plexus, Cingulate gyrus , Cingulum, Corpus callosum, Diencephalon, Extra-central nervous system, Filum Foramen of Monro, Fourth ventricle, Frontal lobe, Frontobasal, Frontoparietal, Frontotemporal, Hypothalamus, Infratentorial, Intraventricular, Lateral ventricle, Lateral ventricle trigone, Left hemisphere, Medulla, Medullo cerebellar, Meninges, Midbrain, Middle cerebellar peduncle, Middle frontal gyrus, NS, Occipital lobe, Optic nerve, Optic pathway, Paracentral Parietal lobe, Parietooccipital, Parietotemporal, Periventricular, Pineal gland, Posterior fossa, Precentral gyrus, Sella turcica, Sellar suprasellar, Septum pellucidum, Spinal cord, Subcorticoparamedial, Superior frontal gyrus,

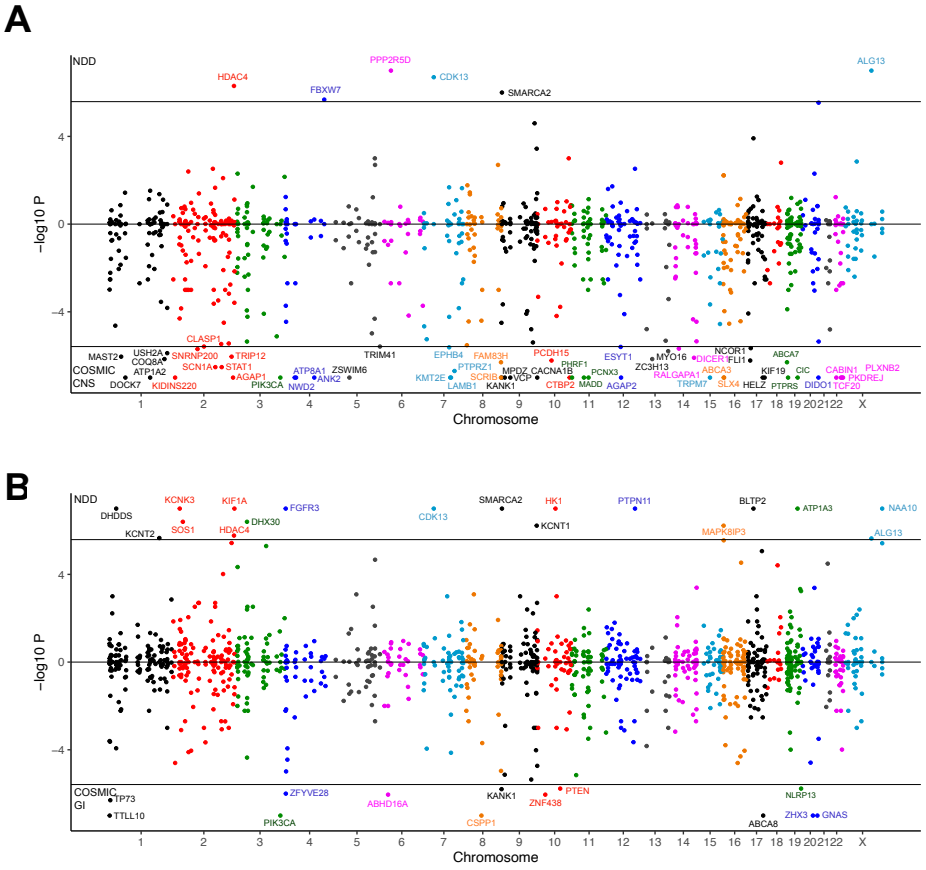


Figure 3: Chicago Plots for the 3D-CLUMP Results in the COSMIC Datasets. A) Chicago plot for 3D-CLUMP results in the NDD versus COSMIC CNS analyses. B) Chicago plot for 3D-CLUMP results in the NDD versus COSMIC GI analyses. For both A and B proteins that exhibit significant clustering in NDDs are shown on the top above the significance line and proteins that exhibit significant clustering in cancer are shown on the bottom below the significance line. Proteins are placed based on the genomic coordinates of the genes that encode them and all significant proteins are labeled on the plots.

Supratentorial, Tectum, Tegmentum, Temporal lobe, Temporobasal, Temporooccipital, Temporoparietal, Thalamus, Third ventricle, Trigone. GI tissue types: Large_intestine, Small_intestine, Gastrointestinal track undetermined site, Esophagus, and Stomach. The somatic variants were also annotated with VEP to RefSeq protein isoforms.

Application of Case-Control CLUMP

To assess clustering in 1D protein space, CLUMP was utilized for comparison of NDDs and cancer. CLUMP was run to compare missense variation identified in individuals with NDDs to missense variants identified in each cancer type, respectively. For this analysis, CLUMP was run in the case-control implementation using a *-m* value of 6 to denote >5 missense variants required in each dataset and a *-z* value of 10,000,000 to signify a permutation of 10 million for significance testing. This permutation level allows for proteome-wide significance testing in the dataset.

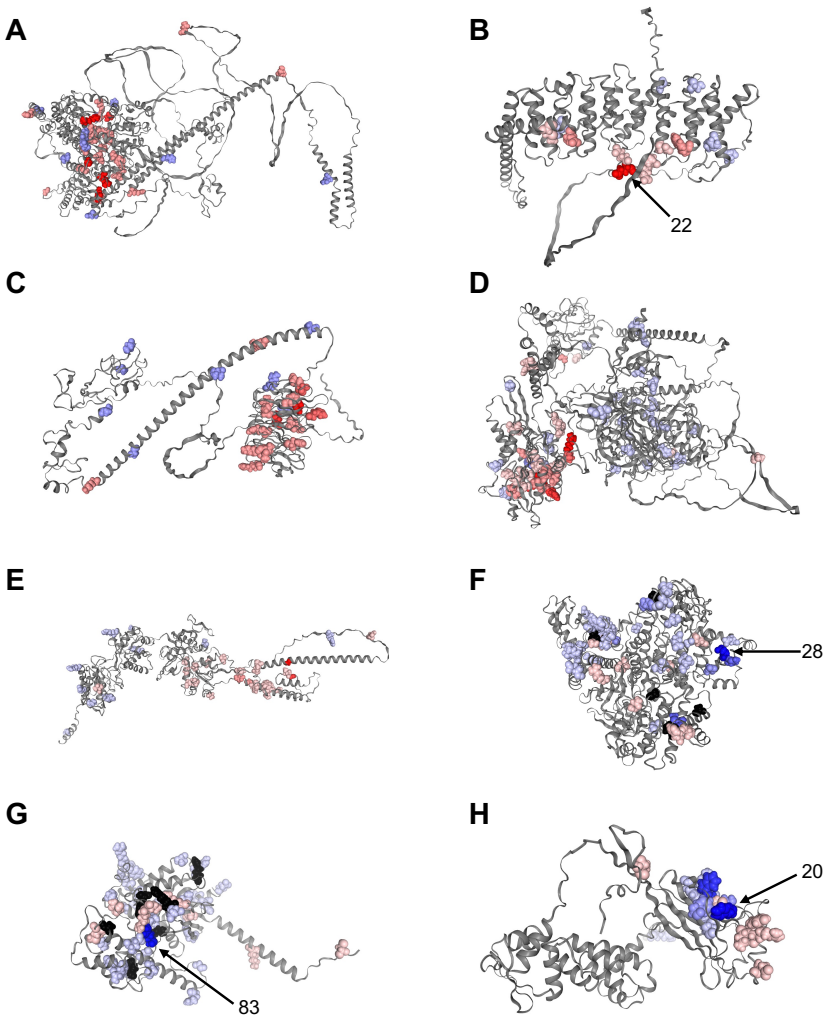


Figure 4: Examples of Proteins with Proteome-Wide Significant Clustering of Missense Variants. Subfigures A to E are significant in NDDs and subfigures F to H are significant in cancer. Red are variants seen in individuals with NDDs. Numbers are shown next to some residues to indicate the number of individuals. Blue are seen in individuals with cancer. Black are seen in both. The intensity of the color is scaled by the number of individuals with missense variants at the residue. A) NDD versus BRCA SMARCA2 (NP_620614), B) NDD versus CNS PPP2R5D (NP_851307), C) NDD versus COAD TRAF7 (NP_115647), D) NDD versus LUAD KIF1A (NP_004312), E) NDD versus COAD GRIN1 (NP_067544), F) NDD versus CNS PIK3CA (NP_006209), G) NDD versus GI GNAS (NP_000507), H) NDD versus PRAD SPOP (NP_003554).

AlphaFold Structure Prediction for Proteins with Missense Variation

We downloaded the RefSeq protein fasta file (https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_protein.faa.gz) from NCBI on May 18, 2021. For each protein, we extracted the protein sequence in fasta format from the RefSeq file. To generate a 3D protein structure for the protein, the protein fasta file was used as input to the AlphaFold (60) version 2.2.0 program. The following reference databases were used: *Reduced BFD*, *MGnify (2018_12)*, *Uniclust30 (2018_08)*, *UniRef90 (2022_01)*, and *PDB70 (2020-04-01)*. The AlphaFold program generates an MSA using the *reduced_dbs* setting, model generation produces 5 models with 3 recycles each and AMBER relaxation. The models are ranked by pLDDT, and the best scoring structure is chosen as the final

structure. Post-AlphaFold structure generation, each structure was deposited under accession "matur-clump" in ModelArchive. Proteins requiring >700 GB of memory were unable to be run through AlphaFold on our server.

The 3D-CLUMP Method

Development of 3D-CLUMP as a strategy to assess clustering of variation in 3D protein structures required modification of the method we originally developed for 1D clustering called CLUMP (52). As in the CLUMP method, 3D-CLUMP uses a partitioning around medoids (PAM) strategy specifically using the *pamk* function from the R package fpc (https://cran.r-project.org/web/packages/fpc/index.html) to identify clusters in the data. The advantage of integrating *pamk* in 3D-CLUMP is that it iteratively identifies the optimal number of clusters *k* in the data to find the optimal *k**. This is helpful as the users of the program may not know the optimal *k* for their dataset. The *pamk* function identifies both the number of clusters and estimates the 'medoid' that best represents each cluster. Importantly, in 3D-CLUMP the medoid is represented in 3D-space consisting of x,y,z coordinates on the 3D protein structure. The final 3D-CLUMP score (*Sp*) for a protein *p* is calculated as follows:

$$S_p = \sum_{i=1}^{k*} \sum_{j=1}^{n_i} \frac{ln\left(min_i \sqrt{\sum_j (m_i - X_{ij})^2} + 1\right)}{n_i},$$

where *k** is the optimal *k* for the protein, n_i is the number of variants in cluster *i*, m_i is the position of the medoid in the *i* cluster, and X_{ij} is the position of variant *j* in the *i* cluster. If all variants cluster at the exact same location as the medoid, this will yield optimal clustering and will result in a score of 0. In 3D-CLUMP, lower scores show maximal clustering of the variants.

To calculate a p-value for the test, we generate a null distribution of $\Delta S_p^{\emptyset}$ values, which is $S_p^{Controls} - S_p^{Cases}$. The $\Delta S_p^{\emptyset}$ values are calculated 10,000,000 times to enable testing for proteome-wide (i.e., genome-wide correction for 19,008 genes) significance. We use a Bonferroni corrected p-value of $\frac{0.05}{19{,}008} = 2.63 \times 10^{-6}$ as the threshold for proteome-wide significance.

Application of Case-Control 3D-CLUMP

To assess clustering in 3D protein space, 3D-CLUMP was utilized for comparison of NDDs and cancer. 3D-CLUMP was run to compare missense variation identified in individuals with NDDs to missense variants identified in each cancer type, respectively. For this analysis, 3D-CLUMP was run in the case-control implementation using a *-m* value of 6 to denote >5 missense variants required in each dataset and a *-z* value of 10,000,000 to signify a permutation of 10 million for significance testing. This permutation level allows for proteome-wide significance testing in the dataset.

Plotting of Variant Data

Missense variants were visualized on 3D structures using the R package *NGLVieweR* (https://cran.r-project.org/web/packages/NGLVieweR/vignettes/NGLVieweR.html). Missense variants identified in individuals with NDDs were colored red in the plots and missense variants

identified in individuals with cancer were colored blue in the plots. The intensity of the color corresponds to the number of individuals with DNVs at the specific amino acid residue.

RESULTS

Proteins Exhibiting Clustering in 1D Protein Space

For each of the seven cancer types, testing for significance of clustering in 1D space was done with CLUMP. For the comparison to BRCA, we identified two proteins reaching significance in NDDs (ADCY5, SETD2) and one protein reaching significance in cancer (PIK3CA). For the comparison to COAD, we identified four proteins reaching significance in NDDs (ALG13, CREBBP, PACS1, UPF1) and no protein reaching significance in cancer. For the comparison to LUAD, we identified 9 proteins reaching significance in NDDs (ADCY5, ALG13, CREBBP, EBF3, FOXP1, ITPR1, PPP2R1A, PPP2R5D, TRIO) and no protein reaching significance in cancer. For the comparison to PRAD, we identified one protein reaching significance in NDDs (DYNC1H1) and no protein reaching significance in cancer. For the comparison to OV, we did not identify any proteins

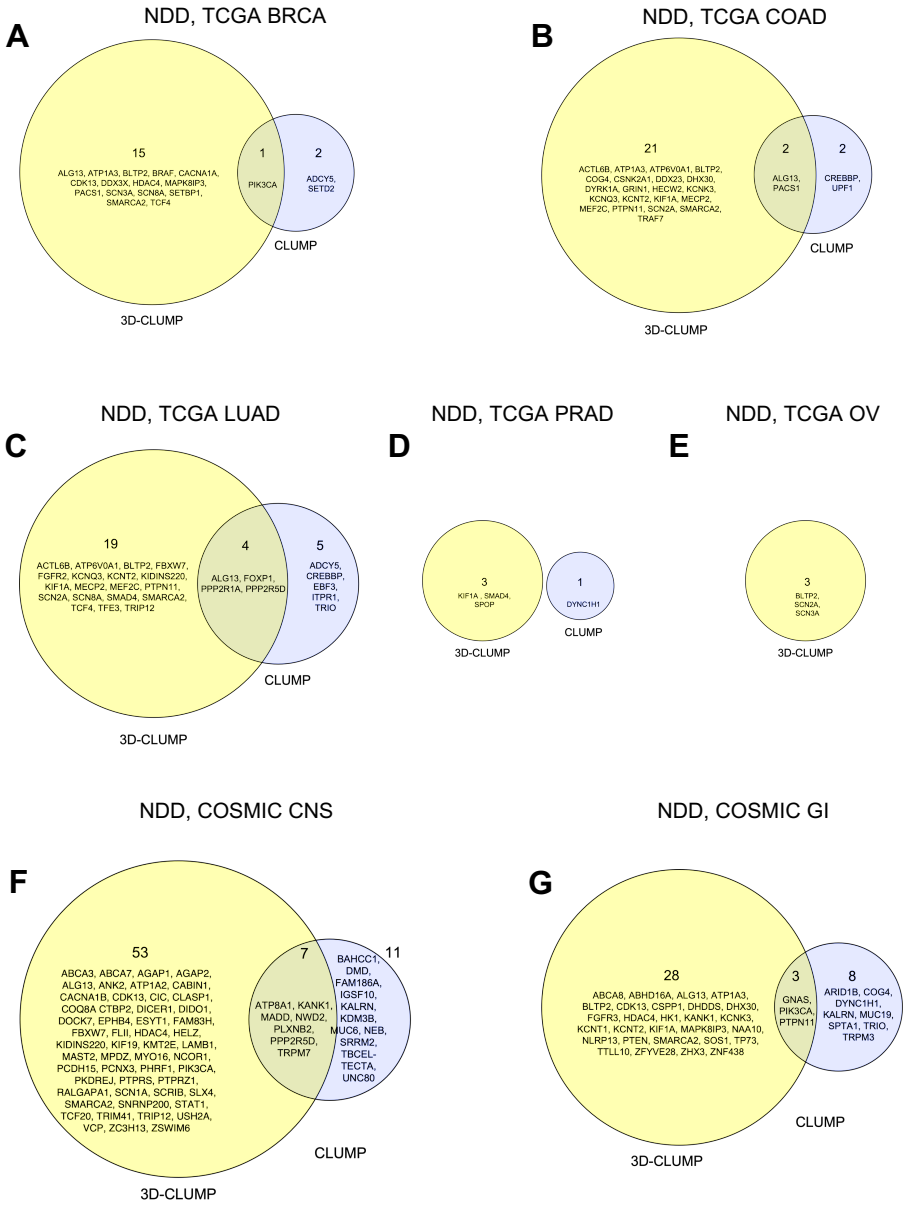


Figure 5: Discovery is Greater with 3D Structures. Shown are proteins exhibiting proteome-wide significance for clustering in either NDDs or the specified cancer type using the 3D-CLUMP and/or CLUMP methods.

reaching significance in either NDDs or cancer. For the comparison to COSMIC CNS, we identified one protein reaching significance in NDDs (PPP2R5D) and 17 protein reaching significance in cancer (ATP8A1, BAHCC1, DMD, FAM186A, IGSF10, KALRN, KANK1, KDM3B, MADD, MUC6, NEB, NWD2, PLXNB2, SRRM2, TBCEL-TECTA, TRPM7, UNC80). For the comparison to COSMIC GI, we identified six proteins reaching significance in NDDs (ARID1B, COG4, DYNC1H1, PTPN11, TRIO, TRPM3) and five protein reaching significance in cancer (GNAS, KALRN, MUC19, PIK3CA, SPTA1).

Metrics of AlphaFold Structures

Summary metrics of the AlphaFold structures generated for the dataset include a pLDDT of 69.79 ± 21.56 (mean ± standard deviation). Across the dataset, there was a negative correlation between the mean pLDDT value and the standard deviation pLDDT values ($r = -0.33$, $p < 2.2 \times 10^{-16}$), a negative correlation between the protein length and the mean pLDDT values ($r = -0.31$, $p < 2.2 \times 10^{-16}$), and a positive correlation between the protein length and the standard deviation pLDDT values ($r = 0.21$, $p < 2.2 \times 10^{-16}$). While the correlations were significant they were not strong. Observation of the data revealed there was an apparent inverse U-shape to the data when comparing the standard deviation pLDDT to the mean pLDDT. To test this we fit the standard deviation and mean to a model using the following formula in R:

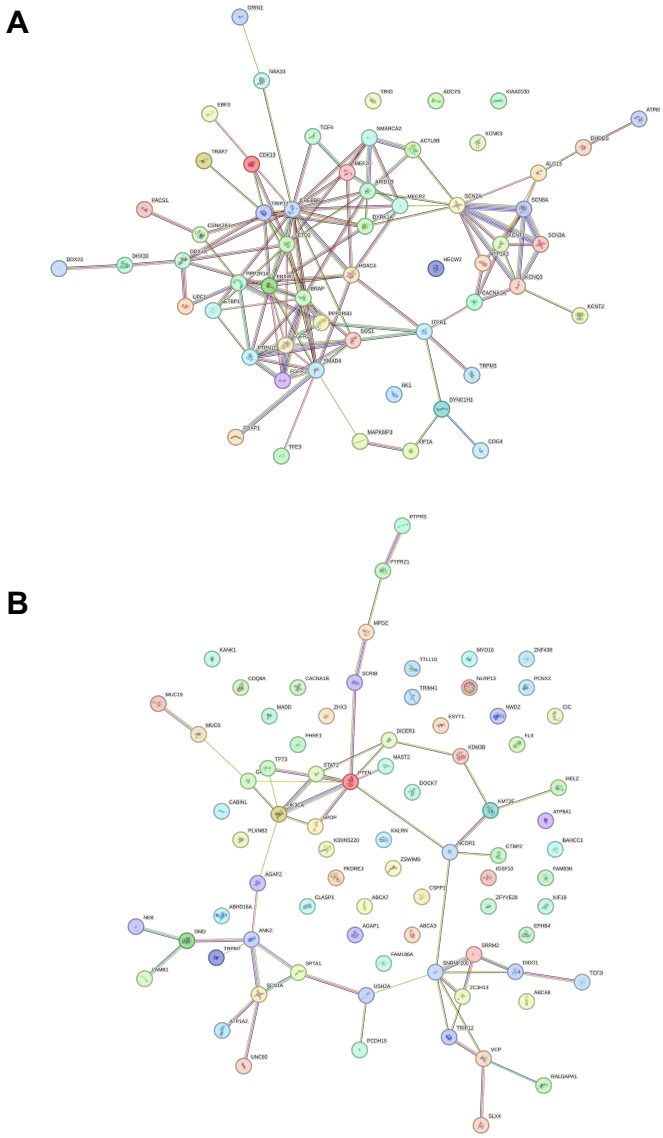


Figure 6: Protein-Protein Interaction Enrichment of Proteins with Significant Missense Clustering. A) PPI network of proteins with proteome-wide significant clustering in NDDs (number of nodes = 57, number of edges = 123, expected number of edges = 41, $p < 1 \times 10^{-16}$). B) PPI network of proteins with proteome-wide significant clustering in cancer (number of nodes = 78, number of edges = 51, expected number of edges = 33, $p = 2.5 \times 10^{-3}$).

$$model = lm(sdLDDT \sim meanLDDT + I(meanLDDT^2)$$

This model fit the data well with an adjusted $r^2 = 0.61$ and a $p < 2.2 \times 10^{-16}$. This result shows that structures with good predictions overall (i.e., high mean pLDDT scores) tend to have lower standard deviation for their pLDDT scores and structures with bad predictions overall (i.e., low mean pLDDT scores) tend to have lower standard deviation for their pLDDT scores. In short, really good structures are good overall and really bad structures are bad overall. There are then several structures with decent structures but lots of variability in how consistent they are in pLDDT overall.

Proteins Exhibiting Clustering in 3D Protein Space

For each of the seven cancer types, testing for significance of clustering in 3D space was done with 3D-CLUMP (**Figure 3**, **Figure 4**). For the comparison to BRCA, we identified 15 proteins reaching significance in NDDs (ALG13, ATP1A3, BLTP2, BRAF, CACNA1A, CDK13, DDX3X, HDAC4, MAPK8IP3, PACS1, SCN3A, SCN8A, SETBP1, SMARCA2, TCF4) and one protein reaching significance in cancer (PIK3CA). For the comparison to COAD, we identified 23 proteins reaching significance in NDDs (ACTL6B, ALG13, ATP1A3, ATP6V0A1, BLTP2, COG4, CSNK2A1, DDX23, DHX30, DYRK1A, GRIN1, HECW2, KCNK3, KCNQ3, KCNT2, KIF1A, MECP2, MEF2C, PACS1, PTPN11, SCN2A, SMARCA2, TRAF7) and no protein reaching significance in cancer. For the comparison to LUAD, we identified 22 proteins reaching significance in NDDs (ACTL6B, ALG13, ATP6V0A1, BLTP2, FBXW7, FGFR2, FOXP1, KCNQ3, KCNT2, KIF1A, MECP2, MEF2C, PPP2R1A, PPP2R5D, PTPN11, SCN2A, SCN8A, SMAD4, SMARCA2, TCF4, TFE3, TRIP12) and one protein reaching significance in cancer (KIDINS220). For the comparison to PRAD, we identified two proteins reaching significance in NDDs (KIF1A, SMAD4) and one protein reaching significance in cancer (SPOP). For the comparison to OV, we identified three proteins reaching significance in NDDs (BLTP2, SCN2A, SCN3A) and no protein reaching significance in cancer. For the comparison to COSMIC CNS, we identified six proteins reaching significance in NDDs (ALG13, CDK13, FBXW7, HDAC4, PPP2R5D, SMARCA2) and 54 proteins reaching significance in cancer (ABCA3, ABCA7, AGAP1, AGAP2, ANK2, ATP1A2, ATP8A1, CABIN1, CACNA1B, CIC, CLASP1, COQ8A, CTBP2, DICER1, DIDO1, DOCK7, EPHB4, ESYT1, FAM83H, FLII, HELZ, KANK1, KIDINS220, KIF19, KMT2E, LAMB1, MADD, MAST2, MPDZ, MYO16, NCOR1, NWD2, PCDH15, PCNX3, PHRF1, PIK3CA, PKDREJ, PLXNB2, PTPRS, PTPRZ1, RALGAPA1, SCN1A, SCRIB, SLX4, SNRNP200, STAT1, TCF20, TRIM41, TRIP12, TRPM7, USH2A, VCP, ZC3H13, ZSWIM6). For the comparison to COSMIC GI, we identified 18 proteins reaching significance in NDDs (ALG13, ATP1A3, BLTP2, CDK13, DHDDS, DHX30, FGFR3, HDAC4, HK1, KCNK3, KCNT1, KCNT2, KIF1A, MAPK8IP3, NAA10, PTPN11, SMARCA2, SOS1) and 13 protein reaching significance in cancer (ABCA8, ABHD16A, CSPP1, GNAS, KANK1, NLRP13, PIK3CA, PTEN, TP73, TTLL10, ZFYVE28, ZHX3, ZNF438).

Greater Statistical Discovery Through Assessment of Clustering in 3D-Structures

Using the 3D protein structure test (3D-CLUMP), greater significant protein discovery was observed in comparison to the 1D test. In every comparison of NDDs and the shown cancer type there were more proteins identified with proteome-wide significance in the 3D-CLUMP results (**Figure 5**). The number of proteins identified uniquely as proteome-wide significant in the 3D test was between 3 times to 10.5 times more than the number uniquely identified in the 1D test. In some instances, the same protein was identified in both the 1D and 3D tests (e.g., PIK3CA in

TCGA BRCA, ALG13 in TCGA COAD, PACS1 in TCGA COAD). However, most of the time the proteins discovered were unique to the specific test showing a benefit to using 3D structures in assessment of clustering of missense variation.

Protein-Protein Interaction Enrichment of Proteins with Significant Missense Clustering

For the proteins enriched for clustering in NDDs and cancer we tested for enrichment of protein-protein interactions (PPIs) in STRING-DB (**Figure 6**). For the proteins with enrichment in NDDs, we identified significant PPIs (number of nodes = 57, number of edges = 123, expected number of edges = 41, $p < 1 \times 10^{-16}$) with enrichment of proteins involved in the BAF complex and proteins that function as channels. In the proteins identified in the cancer analysis, we also so an enrichment of PPIs (number of nodes = 78, number of edges = 51, expected number of edges = 33, $p = 2.5 \times 10^{-3}$) and enrichment of proteins involved in ATP-related activities.

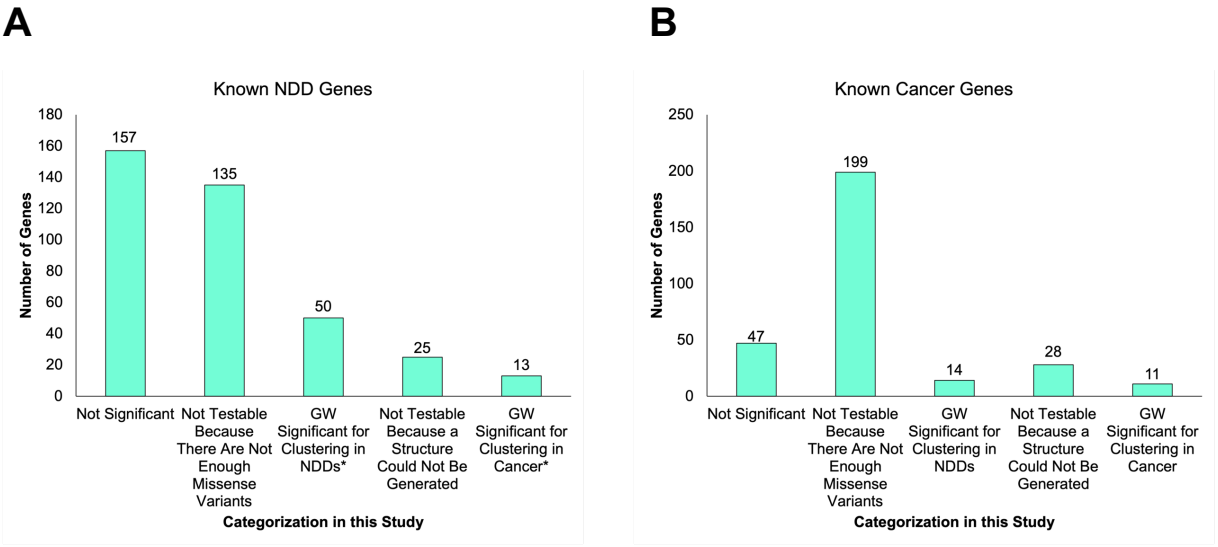


Figure 7: Comparison to Known NDD and Known Cancer Genes. A) Shown is the distribution of protein results in our study and comparison to 379 known NDD genes. *TRIP12 is significant in NDDs in comparison to LUAD and is significant in COSMIC CNS comparison to NDDs. B) Shown is the distribution of protein results in our study and comparison to 299 known Cancer genes

Comparison to Known NDD Genes

There are 379 NDD genes known to be genome-wide significant for enrichment of DNVs (36, 43). We examined what the status of these genes were in this study (**Figure 7**). There were 50 genes that were known to be significant in NDDs and exhibited significant missense DNV clustering in NDDs in this study. There were 13 genes that were known to be significant in NDDs and exhibited significant missense DNV clustering in cancer in this study. One gene (*TRIP12*) was found in both of these results. TRIP12 was significant in NDDs in the comparison to LUAD and TRIP12 is significant in COSMIC CNS in comparison to NDDs. This results in 62 genes exhibiting clustering in the known NDD set. The remaining 319 proteins that were not found significant by CLUMP or 3D-CLUMP either had a small number of independent missense variants (<6) for CLUMP/3D-CLUMP testing, or were too large to build an AlphaFold structure with our compute memory

limitation of 700 GB. There were 135 proteins with <6 independent missense variants and 25 for which we could not generate AlphaFold structures. There were 157 proteins that were not significant in any of our tests and likely did not harbor any clustered missense variation.

Based on the analyses above, for the 379 known NDD proteins there were 62 (16.4%) with clustering based on CLUMP/3D-CLUMP analyses, and 160 (42.2%) were not testable either because they did not have enough missense variants to perform the analyses or a protein structure could not be made, and 157 (41.4%) were not significant for clustering by any test.

BLTP2 as a Novel Proteome-Wide Significant Protein in NDDs

There were seven proteins (ACTL6B, BLTP2, DHX30, KCNT2, MAPK8IP3, SCN3A, SOS1) with significant clustering in NDDs that were not previously genome-wide significant for DNVs in NDDs. Six (ACTL6B, DHX30, KCNT2, MAPK8IP3, SCN3A, SOS1) of these proteins have been implicated in rare forms of NDDs. ACTL6B is involved in in "developmental and epileptic encephalopathy 76" (OMIM #618468) and "intellectual developmental disorder with severe speech and ambulation defects" (OMIM #618470). DHX30 is involved in "neurodevelopmental disorder with variable motor and speech impairment" (OMIM #617804). KCNT2 is involved in "developmental and epileptic encephalopathy 57" (OMIM #617771). MAPK8IP3 is involved in "neurodevelopmental disorder with or without variable brain abnormalities" (OMIM #618443). SCN3A is involved in "developmental and epileptic encephalopathy 62" (OMIM #617938) and "epilepsy, familial focal, with variable foci 4" OMIM #617935). SOS1 is involved in "Noonan syndrome 4" (OMIM #610733).

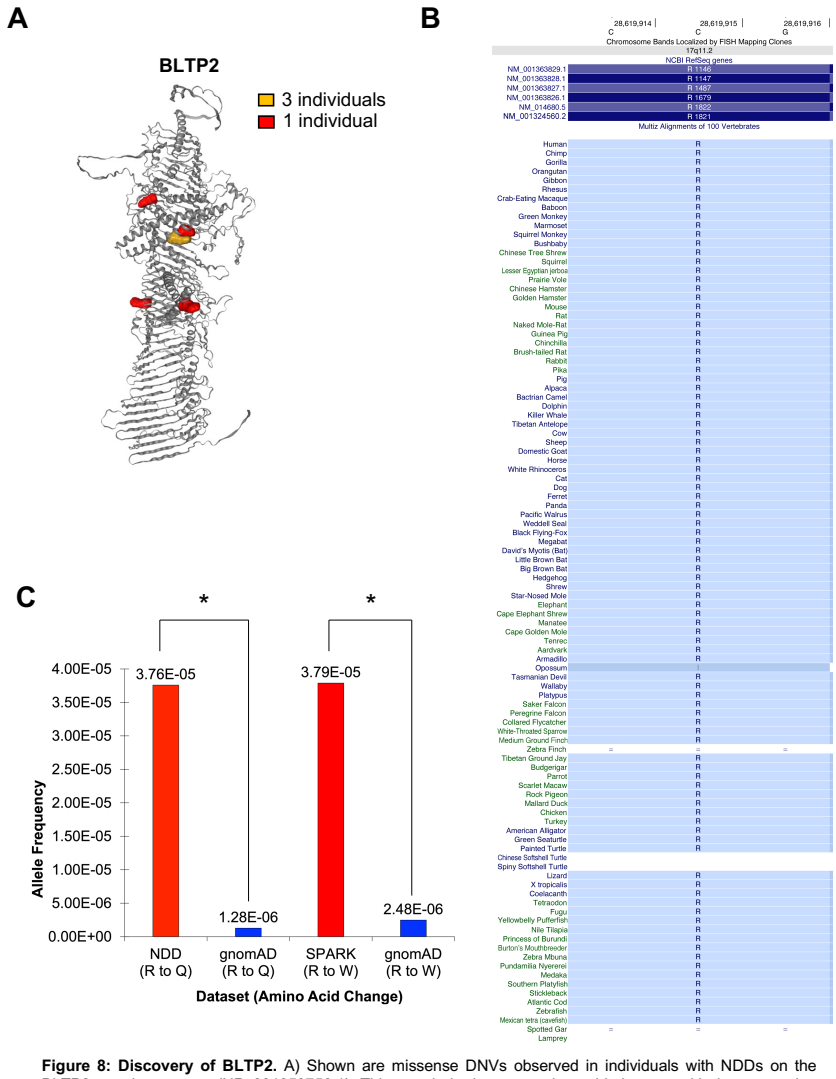


Figure 8: Discovery of BLTP2. A) Shown are missense DNVs observed in individuals with NDDs on the BLTP2 protein structure (NP_001350756.1). This protein had seven amino acid changes with three at amino acid position 1487 and one each at amino acid positions 605, 705, 1253, and 1483. B) The Arginine at position 1487 is highly conserved across several species. C) The Arginine to Glutamine missense variant is significantly enriched in NDDs (Fisher's Exact Test $p = 2.96 \times 10^{-3}$, OR = 29.4). Another missense variant (Arginine to Tryptophan) was identified at this amino acid position in an independent NDD cohort (SPARK) and is also enriched in NDDs (Fisher's Exact Test $p = 8.30 \times 10^{-4}$, OR = 15.2)

One protein (BLTP2) has not been implicated in NDDs before (**Figure 8**). BLTP2 was previously known as KIAA0100. We identified significant clustering of missense variants in NDDs in this protein using our 3D-CLUMP tool. It was significant, with a 3D-CLUMP Score in NDDs (S_p^{Cases}) of 0.28, in the comparison to BRCA ($p = 1 \times 10^{-7}$), COAD ($p = 2 \times 10^{-7}$), LUAD ($p = 5 \times 10^{-7}$), OV ($p = 1.2 \times 10^{-6}$), and COSMIC GI ($p < 1 \times 10^{-7}$). In particular, this protein had seven amino acid changes on isoform NP_001350756 with three at amino acid position 1487 and one each at amino acid positions 605, 705, 1253, and 1483. Running a conserved domain prediction on the protein, we found that the variants at positions 1483 and 1487 are within the Apt1 domain of this protein that is predicted to be involved in localization of the protein to the Golgi body. Since there were three individuals with a missense variant at amino acid position 1487, we also checked the gnomAD database for this variant (17-28619915-C-T). This allele is seen in 3 individuals in our study (3 alleles / 79766 total alleles, allele frequency = 3.76×10^{-5}) and is seen in 1 individual assessed in gnomAD (1 allele / 780820 total alleles in genome+exome samples, allele frequency = 1.28×10^{-6}). This allele is enriched in the NDD cohort (Fisher's Exact Test $p = 2.96 \times 10^{-3}$, OR = 29.4). To further examine missense variants at this position, we examined another publication consisting of an independent set of 13,189 individuals with DNVs from whole-exome sequencing data from the SPARK autism cohort (42). In this cohort, there was one individual with a missense variant at the same amino acid position (17-28619916-G-A) resulting in an Arginine to Tryptophan change (1 allele / 26378 alleles, allele frequency = 3.79×10^{-5}) and this same variant is seen in 3 individuals in gnomAD (3 allele of 1613876 total alleles, allele frequency = 2.48×10^{-6}) in genome+exome. This allele is also enriched in NDDs (Fisher's Exact Test $p = 8.30 \times 10^{-4}$, OR = 15.2). In addition to the enrichment in NDDs, it is relevant to note that the Arginine is also highly conserved in several species (**Figure 8**).

There are few publications characterizing BLTP2 (KIAA0100) (61-71). None of these papers implicate BLTP2 in NDDs. However, what is known is that it is a member of the Bridge-Like Lipid Transport Protein family. These proteins are important for transfer of lipids and other members of the protein family have been implicated in neurodevelopmental and neurodegenerative disorders (72, 73). A recent preprint has also indicated a role for BLTP2 in the regulation of primary cilia (74); an area of molecular interest in neurodevelopmental disorders (75).

Comparison to Known Cancer Driver Genes

Estimates of the number of mutation-based cancer driver genes have varied over the years, but a commonly used source is the paper from the Cancer Genome Atlas (TCGA) (49). This study identified 299 driver genes, of which 259 were the result of consensus predictions of 26 computational tools. A limited number of these genes met the criteria for analysis in our study, given our threshold for independent missense mutations and limitations of our AlphaFold modeling described above (**Figure 7**). Of these genes, 14 were significantly clustered in NDDs and 11 were significantly clustered in cancers. There were 199 proteins that were not testable because there were not enough missense variants to run the test and 28 could not be tested because a structure could not be generated for the protein. There were 47 proteins that were tested and not significant.

Based on the analyses above, for the 299 known cancer proteins there were 25 (8.4%) with clustering based on CLUMP/3D-CLUMP analyses, and 227 (75.9%) were not testable either

because they did not have enough missense variants to perform the analyses or a protein structure could not be made, and 47 (15.7%) were not significant for clustering by any test.

62 potential proteome-wide significant genes in cancers not discovered by TCGA
Of the remaining genes, eleven in the Bailey list and 68 not in the list were found to have proteome-wide significant differential clustering in one of our cohorts (COSMIC-CNS, COSMIC-GI, TCGA-LUAD, TCGA-BRCA, or TCGA-PRAD) when compared to NDD variants based on CLUMP/3D-CLUMP analyses. Six of the 68 genes not on the Bailey list had been proposed as a driver in the literature, with an additional 11 proposed as prognostic or as a therapeutic target in a particular cancer type. Two genes were paralogs of drivers on the Bailey list.

Proteins Requiring Careful Consideration for Prognostics and Therapeutics
There were 220 known NDD proteins that we could test for missense clustering in NDDs or cancer (see above, **Figure 7**). Of these, 50 were proteome-wide significant for clustering in NDDs. However, there were 13 (5.9%) that were proteome-wide significant for clustering in cancer (AGAP2, ANK2, CLASP1, GNAS, KIDINS220, KMT2E, PIK3CA, PTEN, SCN1A, SRRM2, TCF20, TRIP12, UNC80). These proteins will need to be specially considered when thinking about functional, prognostic, and therapeutic aspects of the variants within them in different phenotypes. Likewise, there were 72 known cancer proteins that we could test for missense clustering in NDDs or cancer (see above, **Figure 7**). Of these, 11 were proteome-wide significant for clustering in cancer. However, there were 14 (19.4%) that were proteome-wide significant for clustering in NDDs (BRAF, CACNA1A, CREBBP, DDX3X, FBXW7, FGFR2, FGFR3, KIF1A, PPP2R1A, PTPN11, SETBP1, SETD2, SMAD4, SOS1). These proteins will also need to be specially considered when thinking about functional, prognostic, and therapeutic aspects of the variants within them in different phenotypes.

DISCUSSION

An outstanding question in the genomics of NDDs is why are several genes identified in NDDs also identified in cancer? In particular, genes identified in both are involved in molecular processes including chromatin remodeling and transcription (58). Other interesting observations include microcephaly and macrocephaly as a result of variation in some genes in NDDs. This has been compared to the cellular proliferation and differentiation related processes in cancer. Another area of interest has been in genome maintenance (57). Several hypotheses have been put forward for the overlap in genes (58).

In this study, we explore the hypothesis that genes with missense variants in NDDs and cancer have a different variant pattern in NDDs and in cancer. Three main options were considered at the protein-level including clustering of missense variants in NDDs and not in cancer, clustering of missense variants in cancer and not in NDDs, and no clustering of missense variants in NDDs or cancer. Since we focused on clustering of missense variation at the level of each protein, we utilized two strategies including examination of clustering on the 1D protein structure and clustering on the 3D protein structure. Our existing method CLUMP (52) was utilized to look for proteome-wide significance of clustering in 1D in a case-control design. We also developed the computational tool 3D-CLUMP that can perform proteome-wide significant case-control analysis

in three-dimensional protein structure space. This computational tool is open-source, available on GitHub (https://github.com/TNTurnerLab/3D-CLUMP), and will be beneficial to others wanting to test for clustering of variants on 3D protein structures in these and other phenotypes. As part of this work, we also generated AlphaFold structures for >4000 proteins with enough missense variation to be tested in our study. These structures were deposited in ModelArchive (https://www.modelarchive.org/doi/10.5452/ma-tur-clump) and will also be beneficial as a resource to the research community. We showed in this study in the comparison of CLUMP (1D) and 3D-CLUMP that using 3D structures boosts our power for discovery of proteins with clustering of missense variation.

Several outcomes of our study are novel and intriguing with regard to missense variants in NDDs and cancer. By comparing missense variants in NDDs to those in cancer, we identified proteins where there was significant clustering of missense variants in NDDs, proteins where there was significant clustering of missense variants in cancer, and proteins with no clustering in NDDs or cancer. This is an important discovery because it points to specific proteins where there are differences in the patterns of missense variation in the two phenotypes. This is another important resource for researchers studying the two phenotypes and will provide important information relevant in functional assessment of variation and in prognostics. While many of the genes we identified in NDDs have been identified in more broad searches looking at DNV enrichment of likely-gene disrupting variation and missense variation irrespective of clustering (36, 43), we did also identify one new proteome-wide significant protein (BLTP2) for NDDs. Overall, our work provides novel insights into patterns of missense variation in NDDs and cancer.

There are a few caveats to our study. One caveat is that there are some proteins where we do not have enough missense variants to perform the statistical test, and this is something that will be approachable with increased sample sizes in both NDDs and cancer. Another caveat is regarding the protein structures themselves. For some proteins, the protein isoform may not be supported by experimental evidence, the structure could not be made, or for some a high confidence structure could not be made. Finally, our statistical test does not currently identify when the proteins have clustered missense variants in both NDDs and cancer but the highly clustered regions are different in the two phenotypes. Potential future work would address each of these caveats to provide further insights into this question. One other interesting future direction is to explore the clustering of missense variants with regard to PPIs. We showed an enrichment of these and careful consideration of clusters at the interfaces of these PPIs would be useful.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (R01MH126933 to T.N.T., R00MH117165 to T.N.T., P50HD103525 to T.N.T. as a Member and Scientific Liaison in the Washington University in St. Louis Intellectual and Developmental Disabilities Research Center), the ITCR program at the National Cancer Institute (U24CA258393 to R.K.), the Simons Foundation (Award #734069 to T.N.T.), and funds from the Washington University in St. Louis McDonnell Center for Cellular and Molecular Neurobiology to T.N.T. Thank you also to Dan Western for his participation on the project during his rotation in the Turner Laboratory.

CRedIT AUTHOR STATEMENT REGARDING AUTHOR CONTRIBUTIONS

Jeffrey K. Ng: Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Yilin Chen:** Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Titilope M. Akinwe:** Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Hillary B. Heins:** Formal analysis, Writing - Review & Editing. **Elvisa Mehinovic:** Formal analysis, Writing - Review & Editing. **Yoonhoo Chang:** Formal analysis, Writing - Review & Editing. **Zachary Payne:** Formal analysis, Writing - Review & Editing. **Juana G. Manuel:** Visualization, Writing - Review & Editing. **Rachel Karchin:** Methodology, Software, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition. **Tychele N. Turner:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition.

DATA AND SOFTWARE AVAILABILITY STATEMENT

The AlphaFold structures for this project have been deposited in ModelArchive under accession number "ma-tur-clump" at https://www.modelarchive.org/doi/10.5452/ma-tur-clump. The code for this paper is available at:

CLUMP: https://github.com/KarchinLab/CLUMP
3D-CLUMP: https://github.com/TNTurnerLab/3D-CLUMP
AlphaFold Structure and 3D Protein Plot Generation:
https://github.com/TNTurnerLab/clustering_in_cancer_vs_ndd_paper

REFERENCES

1. Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, Mahajan M, Manaa D, Pawitan Y, Reichert J, Ripke S, Sandin S. Most genetic risk for autism resides with common variation. Nature genetics. 2014;46(8):881-5. doi: 10.1038/ng.3039. PubMed PMID: 25038753.

2. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, Raja A, Coe BP, Stessman HA, He ZX, Leal SM, Bernier R, Eichler EE. Excess of rare, inherited truncating mutations in autism. Nature genetics. 2015;47(6):582-8. Epub 2015/05/12. doi: 10.1038/ng.3303. PubMed PMID: 25961944; PMCID: PMC4449286.

3. Wilfert AB, Turner TN, Murali SC, Hsieh P, Sulovari A, Wang T, Coe BP, Guo H, Hoekzema K, Bakken TE, Winterkorn LH, Evani US, Byrska-Bishop M, Earl RK, Bernier RA, Zody MC, Eichler EE. Recent ultra-rare inherited variants implicate new autism candidate risk genes. Nature genetics. 2021;53(8):1125-34. Epub 2021/07/28. doi: 10.1038/s41588-021-00899-8. PubMed PMID: 34312540; PMCID: PMC8459613.

4. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimaki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M. Strong association of de novo copy number mutations with autism. Science (New York, NY). 2007;316(5823):445-9. Epub 2007/03/17. doi: 10.1126/science.1138659. PubMed PMID: 17363630; PMCID: Pmc2993504.

5. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, Thiruvahindrapduram B, Fiebig A, Schreiber S, Friedman J, Ketelaars CE, Vos YJ, Ficicioglu C, Kirkpatrick S, Nicolson R, Sloman L, Summers A, Gibbons CA, Teebi A, Chitayat D, Weksberg R, Thompson A, Vardy C, Crosbie V, Luscombe S, Baatjes R, Zwaigenbaum L, Roberts W, Fernandez B, Szatmari P, Scherer SW. Structural variation of chromosomes in autism spectrum disorder. American journal of human genetics. 2008;82(2):477-88. Epub 2008/02/07. doi: 10.1016/j.ajhg.2007.12.009. PubMed PMID: 18252227; PMCID: Pmc2426913.

6. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron. 2011;70(5):898-907. Epub 2011/06/11. doi: 10.1016/j.neuron.2011.05.021. PubMed PMID: 21658583; PMCID: Pmc3607702.

7. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH, Moreau MP, Gupta AR, Thomson SA, Mason CE, Bilguvar K, Celestino-Soper PB, Choi M, Crawford EL, Davis L, Wright NR, Dhodapkar RM, DiCola M, DiLullo NM, Fernandez TV, Fielding-Singh V, Fishman DO, Frahm S, Garagaloyan R, Goh GS, Kammela S, Klei L, Lowe JK, Lund SC, McGrew AD, Meyer KA, Moffat WJ, Murdoch JD, O'Roak BJ, Ober GT, Pottenger RS, Raubeson MJ, Song Y, Wang Q, Yaspan BL, Yu TW, Yurkiewicz IR, Beaudet AL, Cantor RM, Curland M, Grice DE, Gunel M, Lifton RP, Mane SM, Martin DM, Shaw CA, Sheldon M, Tischfield JA, Walsh CA, Morrow EM, Ledbetter DH, Fombonne E, Lord C, Martin CL, Brooks AI, Sutcliffe JS, Cook EH, Jr., Geschwind D, Roeder K, Devlin B, State MW. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron. 2011;70(5):863-85. Epub 2011/06/11. doi: 10.1016/j.neuron.2011.05.002. PubMed PMID: 21658581.

8. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, Rieder MJ, Nickerson DA, Bernier R, Fisher SE, Shendure J, Eichler EE.

Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nature genetics. 2011;43(6):585-9. Epub 2011/05/17. doi: 10.1038/ng.835. PubMed PMID: 21572417; PMCID: PMC3115696.

9. Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K, Buja A, Krieger A, Yoon S, Troge J, Rodgers L, Iossifov I, Wigler M. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron. 2011;70(5):886-97. Epub 2011/06/11. doi: 10.1016/j.neuron.2011.05.015. PubMed PMID: 21658582.

10. Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, Vu TH, Shafer N, Bernier R, Ferrero GB, Silengo M, Warren ST, Moreno CS, Fichera M, Romano C, Raskind WH, Eichler EE. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. PLoS genetics. 2011;7(11):e1002334. Epub 2011/11/22. doi: 10.1371/journal.pgen.1002334. PubMed PMID: 22102821; PMCID: Pmc3213131.

11. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, Bernier R, Shendure J, Eichler EE. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012;485(7397):246-50. Epub 2012/04/13. doi: 10.1038/nature10989. PubMed PMID: 22495309; PMCID: PMC3350576.

12. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, DePristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH, Jr., Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012;485(7397):242-5. Epub 2012/04/13. doi: 10.1038/nature11011. PubMed PMID: 22495311; PMCID: PMC3613847.

13. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, Meyer KA, Bilguvar K, Mane SM, Sestan N, Lifton RP, Gunel M, Roeder K, Geschwind DH, Devlin B, State MW. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237-41. Epub 2012/04/13. doi: 10.1038/nature10945. PubMed PMID: 22495306; PMCID: PMC3667984.

14. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A, Kendall J, Grabowska E, Ma B, Marks S, Rodgers L, Stepansky A, Troge J, Andrews P, Bekritsky M, Pradhan K, Ghiban E, Kramer M, Parla J, Demeter R, Fulton LL, Fulton RS, Magrini VJ, Ye K, Darnell JC, Darnell RB, Mardis ER, Wilson RK, Schatz MC, McCombie WR, Wigler M. De novo gene disruptions in children on the autistic spectrum. Neuron. 2012;74(2):285-99. Epub 2012/05/01. doi: 10.1016/j.neuron.2012.04.009. PubMed PMID: 22542183; PMCID: PMC3619976.

15. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W, Corominas R, Peoples A, Koren A, Gore A, Kang S, Lin GN, Estabillo J, Gadomski T, Singh B, Zhang K, Akshoomoff N, Corsello C, McCarroll S, Iakoucheva LM, Li Y, Wang J, Sebat J. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012;151(7):1431-42. Epub 2012/12/25. doi: 10.1016/j.cell.2012.11.019. PubMed PMID: 23260136; PMCID: Pmc3712641.

16. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, del Rosario M, Hoischen A, Scheffer H, de Vries BB, Brunner HG, Veltman JA, Vissers LE. Diagnostic exome sequencing in persons with severe intellectual disability. The New England journal of medicine. 2012;367(20):1921-9. Epub 2012/10/05. doi: 10.1056/NEJMoa1206524. PubMed PMID: 23033978.
17. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N, Dufke A, Cremer K, Hempel M, Horn D, Hoyer J, Joset P, Ropke A, Moog U, Riess A, Thiel CT, Tzschach A, Wiesener A, Wohlleber E, Zweier C, Ekici AB, Zink AM, Rump A, Meisinger C, Grallert H, Sticht H, Schenck A, Engels H, Rappold G, Schrock E, Wieacker P, Riess O, Meitinger T, Reis A, Strom TM. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet. 2012;380(9854):1674-82. Epub 2012/10/02. doi: 10.1016/s0140-6736(12)61480-9. PubMed PMID: 23020937.
18. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD, State MW, Devlin B, Roeder K. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS genetics. 2013;9(8):e1003671. Epub 2013/08/24. doi: 10.1371/journal.pgen.1003671. PubMed PMID: 23966865; PMCID: Pmc3744441.
19. Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, Epstein MP, Glauser T, Goldstein DB, Han Y, Heinzen EL, Hitomi Y, Howell KB, Johnson MR, Kuzniecky R, Lowenstein DH, Lu YF, Madou MR, Marson AG, Mefford HC, Esmaeeli Nieh S, O'Brien TJ, Ottman R, Petrovski S, Poduri A, Ruzzo EK, Scheffer IE, Sherr EH, Yuskaitis CJ, Abou-Khalil B, Alldredge BK, Bautista JF, Berkovic SF, Boro A, Cascino GD, Consalvo D, Crumrine P, Devinsky O, Dlugos D, Epstein MP, Fiol M, Fountain NB, French J, Friedman D, Geller EB, Glauser T, Glynn S, Haut SR, Hayward J, Helmers SL, Joshi S, Kanner A, Kirsch HE, Knowlton RC, Kossoff EH, Kuperman R, Kuzniecky R, Lowenstein DH, McGuire SM, Motika PV, Novotny EJ, Ottman R, Paolicchi JM, Parent JM, Park K, Poduri A, Scheffer IE, Shellhaas RA, Sherr EH, Shih JJ, Singh R, Sirven J, Smith MC, Sullivan J, Lin Thio L, Venkat A, Vining EP, Von Allmen GK, Weisenberg JL, Widdess-Walsh P, Winawer MR. De novo mutations in epileptic encephalopathies. Nature. 2013;501(7466):217-21. Epub 2013/08/13. doi: 10.1038/nature12439. PubMed PMID: 23934111; PMCID: PMC3773011.
20. Veeramah KR, Johnstone L, Karafet TM, Wolf D, Sprissler R, Salogiannis J, Barth-Maron A, Greenberg ME, Stuhlmann T, Weinert S, Jentsch TJ, Pazzi M, Restifo LL, Talwar D, Erickson RP, Hammer MF. Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. Epilepsia. 2013;54(7):1270-81. Epub 2013/05/08. doi: 10.1111/epi.12201. PubMed PMID: 23647072; PMCID: PMC3700577.
21. Jiang YH, Yuen RK, Jin X, Wang M, Chen N, Wu X, Ju J, Mei J, Shi Y, He M, Wang G, Liang J, Wang Z, Cao D, Carter MT, Chrysler C, Drmic IE, Howe JL, Lau L, Marshall CR, Merico D, Nalpathamkalam T, Thiruvahindrapuram B, Thompson A, Uddin M, Walker S, Luo J, Anagnostou E, Zwaigenbaum L, Ring RH, Wang J, Lajonchere C, Wang J, Shih A, Szatmari P, Yang H, Dawson G, Li Y, Scherer SW. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. American journal of human genetics. 2013;93(2):249-63. Epub 2013/07/16. doi: 10.1016/j.ajhg.2013.06.012. PubMed PMID: 23849776; PMCID: Pmc3738824.
22. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, Smith JD, Paeper B, Nickerson DA, Dea J, Dong S,

Gonzalez LE, Mandell JD, Mane SM, Murtha MT, Sullivan CA, Walker MF, Waqar Z, Wei L, Willsey AJ, Yamrom B, Lee Y-h, Grabowska E, Dalkic E, Wang Z, Marks S, Andrews P, Leotta A, Kendall J, Hakker I, Rosenbaum J, Ma B, Rodgers L, Troge J, Narzisi G, Yoon S, Schatz MC, Ye K, McCombie WR, Shendure J, Eichler EE, State MW, Wigler M. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014. doi: 10.1038/nature13908. PubMed PMID: 25363768; PMCID: PMC4313871

23. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, Kou Y, Liu L, Fromer M, Walker S, Singh T, Klei L, Kosmicki J, Shih-Chen F, Aleksic B, Biscaldi M, Bolton PF, Brownfeld JM, Cai J, Campbell NG, Carracedo A, Chahrour MH, Chiocchetti AG, Coon H, Crawford EL, Curran SR, Dawson G, Duketis E, Fernandez BA, Gallagher L, Geller E, Guter SJ, Hill RS, Ionita-Laza J, Jimenz Gonzalez P, Kilpinen H, Klauck SM, Kolevzon A, Lee I, Lei I, Lei J, Lehtimaki T, Lin CF, Ma'ayan A, Marshall CR, McInnes AL, Neale B, Owen MJ, Ozaki N, Parellada M, Parr JR, Purcell S, Puura K, Rajagopalan D, Rehnstrom K, Reichenberg A, Sabo A, Sachse M, Sanders SJ, Schafer C, Schulte-Ruther M, Skuse D, Stevens C, Szatmari P, Tammimies K, Valladares O, Voran A, Li-San W, Weiss LA, Willsey AJ, Yu TW, Yuen RK, Cook EH, Freitag CM, Gill M, Hultman CM, Lehner T, Palotie A, Schellenberg GD, Sklar P, State MW, Sutcliffe JS, Walsh CA, Scherer SW, Zwick ME, Barett JC, Cutler DJ, Roeder K, Devlin B, Daly MJ, Buxbaum JD. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515(7526):209-15. Epub 2014/11/05. doi: 10.1038/nature13772. PubMed PMID: 25363760.

24. Dong S, Walker MF, Carriero NJ, DiCola M, Willsey AJ, Ye AY, Waqar Z, Gonzalez LE, Overton JD, Frahm S, Keaney JF, 3rd, Teran NA, Dea J, Mandell JD, Hus Bal V, Sullivan CA, DiLullo NM, Khalil RO, Gockley J, Yuksel Z, Sertel SM, Ercan-Sencicek AG, Gupta AR, Mane SM, Sheldon M, Brooks AI, Roeder K, Devlin B, State MW, Wei L, Sanders SJ. De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. Cell reports. 2014;9(1):16-23. Epub 2014/10/07. doi: 10.1016/j.celrep.2014.08.068. PubMed PMID: 25284784; PMCID: Pmc4194132.

25. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A, Leach R, Klein R, Tearle R, Bo T, Pfundt R, Yntema HG, de Vries BB, Kleefstra T, Brunner HG, Vissers LE, Veltman JA. Genome sequencing identifies major causes of severe intellectual disability. Nature. 2014;511(7509):344-7. Epub 2014/06/05. doi: 10.1038/nature13394. PubMed PMID: 24896178.

26. Yuen RK, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, Chrysler C, Nalpathamkalam T, Pellecchia G, Liu Y, Gazzellone MJ, D'Abate L, Deneault E, Howe JL, Liu RS, Thompson A, Zarrei M, Uddin M, Marshall CR, Ring RH, Zwaigenbaum L, Ray PN, Weksberg R, Carter MT, Fernandez BA, Roberts W, Szatmari P, Scherer SW. Whole-genome sequencing of quartet families with autism spectrum disorder. Nature medicine. 2015;21(2):185-91. Epub 2015/01/27. doi: 10.1038/nm.3792. PubMed PMID: 25621899.

27. Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, Raja A, Baker C, Hoekzema K, Stessman HA, Zody MC, Nelson BJ, Huddleston J, Sandstrom R, Smith JD, Hanna D, Swanson JM, Faustman EM, Bamshad MJ, Stamatoyannopoulos J, Nickerson DA, McCallion AS, Darnell R, Eichler EE. Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. American journal of human genetics. 2016;98(1):58-74. Epub 2016/01/11. doi: 10.1016/j.ajhg.2015.11.023. PubMed PMID: 26749308; PMCID: PMC4716689.

28. Brandler WM, Antaki D, Gujral M, Noor A, Rosanio G, Chapman TR, Barrera DJ, Lin GN, Malhotra D, Watts AC, Wong LC, Estabillo JA, Gadomski TE, Hong O, Fajardo KV,

Bhandari A, Owen R, Baughn M, Yuan J, Solomon T, Moyzis AG, Maile MS, Sanders SJ, Reiner GE, Vaux KK, Strom CM, Zhang K, Muotri AR, Akshoomoff N, Leal SM, Pierce K, Courchesne E, Iakoucheva LM, Corsello C, Sebat J. Frequency and Complexity of De Novo Structural Mutation in Autism. American journal of human genetics. 2016;98(4):667-79. Epub 2016/03/29. doi: 10.1016/j.ajhg.2016.02.018. PubMed PMID: 27018473; PMCID: PMC4833290.

29. Yuen RK, Merico D, Cao H, Pellecchia G, Alipanahi B, Thiruvahindrapuram B, Tong X, Sun Y, Cao D, Zhang T, Wu X, Jin X, Zhou Z, Liu X, Nalpathamkalam T, Walker S, Howe JL, Wang Z, MacDonald JR, Chan A, D'Abate L, Deneault E, Siu MT, Tammimies K, Uddin M, Zarrei M, Wang M, Li Y, Wang J, Wang J, Yang H, Bookman M, Bingham J, Gross SS, Loy D, Pletcher M, Marshall CR, Anagnostou E, Zwaigenbaum L, Weksberg R, Fernandez BA, Roberts W, Szatmari P, Glazer D, Frey BJ, Ring RH, Xu X, Scherer SW. Genome-wide characteristics of de novo mutations in autism. NPJ genomic medicine. 2016;1:160271-1602710. Epub 2016/08/16. doi: 10.1038/npjgenmed.2016.27. PubMed PMID: 27525107; PMCID: PMC4980121.

30. Hashimoto R, Nakazawa T, Tsurusaki Y, Yasuda Y, Nagayasu K, Matsumura K, Kawashima H, Yamamori H, Fujimoto M, Ohi K, Umeda-Yano S, Fukunaga M, Fujino H, Kasai A, Hayata-Takano A, Shintani N, Takeda M, Matsumoto N, Hashimoto H. Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. Journal of human genetics. 2016;61(3):199-206. Epub 2015/11/20. doi: 10.1038/jhg.2015.141. PubMed PMID: 26582266; PMCID: PMC4819764.

31. DDD. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542(7642):433-8. Epub 2017/01/31. doi: 10.1038/nature21062. PubMed PMID: 28135719.

32. Yuen RK, Merico D, Bookman M, L Howe J, Thiruvahindrapuram B, Patel RV, Whitney J, Deflaux N, Bingham J, Wang Z, Pellecchia G, Buchanan JA, Walker S, Marshall CR, Uddin M, Zarrei M, Deneault E, D'Abate L, Chan AJS, Koyanagi S, Paton T, Pereira SL, Hoang N, Engchuan W, Higginbotham EJ, Ho K, Lamoureux S, Li W, MacDonald JR, Nalpathamkalam T, Sung WWL, Tsoi FJ, Wei J, Xu L, Tasse A-M, Kirby E, Van Etten W, Twigger S, Roberts W, Drmic I, Jilderda S, Modi BM, Kellam B, Szego M, Cytrynbaum C, Weksberg R, Zwaigenbaum L, Woodbury-Smith M, Brian J, Senman L, Iaboni A, Doyle-Thomas K, Thompson A, Chrysler C, Leef J, Savion-Lemieux T, Smith IM, Liu X, Nicolson R, Seifer V, Fedele A, Cook EH, Dager S, Estes A, Gallagher L, Malow BA, Parr JR, Spence SJ, Vorstman J, Frey BJ, Robinson JT, Strug LJ, Fernandez BA, Elsabbagh M, Carter MT, Hallmayer J, Knoppers BM, Anagnostou E, Szatmari P, Ring RH, Glazer D, Pletcher MT, Scherer SW. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nature neuroscience. 2017;20(4):602-11. doi: 10.1038/nn.4524.

33. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, Wright CF, Firth HV, FitzPatrick DR, Barrett JC, Hurles ME. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature. 2018;555(7698):611-6. Epub 2018/03/22. doi: 10.1038/nature25983. PubMed PMID: 29562236; PMCID: PMC5912909.

34. Guo H, Wang T, Wu H, Long M, Coe BP, Li H, Xun G, Ou J, Chen B, Duan G, Bai T, Zhao N, Shen Y, Li Y, Wang Y, Zhang Y, Baker C, Liu Y, Pang N, Huang L, Han L, Jia X, Liu C, Ni H, Yang X, Xia L, Chen J, Shen L, Li Y, Zhao R, Zhao W, Peng J, Pan Q, Long Z, Su W, Tan J, Du X, Ke X, Yao M, Hu Z, Zou X, Zhao J, Bernier RA, Eichler EE, Xia K. Inherited and multiple de novo mutations in autism/developmental delay risk genes suggest a multifactorial model. Molecular autism. 2018;9:64. Epub 2018/12/20. doi: 10.1186/s13229-018-0247-z. PubMed PMID: 30564305; PMCID: PMC6293633.

35. An JY, Lin K, Zhu L, Werling DM, Dong S, Brand H, Wang HZ, Zhao X, Schwartz GB, Collins RL, Currall BB, Dastmalchi C, Dea J, Duhn C, Gilson MC, Klei L, Liang L, Markenscoff-Papadimitriou E, Pochareddy S, Ahituv N, Buxbaum JD, Coon H, Daly MJ, Kim YS, Marth GT, Neale BM, Quinlan AR, Rubenstein JL, Sestan N, State MW, Willsey AJ, Talkowski ME, Devlin B, Roeder K, Sanders SJ. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science (New York, NY). 2018;362(6420). Epub 2018/12/14. doi: 10.1126/science.aat6576. PubMed PMID: 30545852; PMCID: PMC6432922.
36. Coe BP, Stessman HAF, Sulovari A, Geisheker MR, Bakken TE, Lake AM, Dougherty JD, Lein ES, Hormozdiari F, Bernier RA, Eichler EE. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nature genetics. 2019;51(1):106-16. Epub 2018/12/19. doi: 10.1038/s41588-018-0288-4. PubMed PMID: 30559488; PMCID: PMC6309590.
37. Turner TN, Wilfert AB, Bakken TE, Bernier RA, Pepper MR, Zhang Z, Torene RI, Retterer K, Eichler EE. Sex-Based Analysis of De Novo Variants in Neurodevelopmental Disorders. American journal of human genetics. 2019;105(6):1274-85. Epub 2019/12/02. doi: 10.1016/j.ajhg.2019.11.003. PubMed PMID: 31785789; PMCID: PMC6904808.
38. Ruzzo EK, Pérez-Cano L, Jung JY, Wang LK, Kashef-Haghighi D, Hartl C, Singh C, Xu J, Hoekstra JN, Leventhal O, Leppä VM, Gandal MJ, Paskov K, Stockham N, Polioudakis D, Lowe JK, Prober DA, Geschwind DH, Wall DP. Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. Cell. 2019;178(4):850-66.e26. Epub 2019/08/10. doi: 10.1016/j.cell.2019.07.015. PubMed PMID: 31398340; PMCID: PMC7102900.
39. Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, Fak JJ, Funk J, Yao K, Tajima Y, Packer A, Darnell RB, Troyanskaya OG. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nature genetics. 2019;51(6):973-80. Epub 2019/05/28. doi: 10.1038/s41588-019-0420-0. PubMed PMID: 31133750; PMCID: PMC6758908.
40. Wang W, Corominas R, Lin GN. De novo Mutations From Whole Exome Sequencing in Neurodevelopmental and Psychiatric Disorders: From Discovery to Application. Front Genet. 2019;10:258. Epub 2019/04/20. doi: 10.3389/fgene.2019.00258. PubMed PMID: 31001316; PMCID: PMC6456656.
41. Padhi EM, Hayeck TJ, Cheng Z, Chatterjee S, Mannion BJ, Byrska-Bishop M, Willems M, Pinson L, Redon S, Benech C, Uguen K, Audebert-Bellanger S, Le Marechal C, Férec C, Efthymiou S, Rahman F, Maqbool S, Maroofian R, Houlden H, Musunuri R, Narzisi G, Abhyankar A, Hunter RD, Akiyama J, Fries LE, Ng JK, Mehinovic E, Stong N, Allen AS, Dickel DE, Bernier RA, Gorkin DU, Pennacchio LA, Zody MC, Turner TN. Coding and noncoding variants in EBF3 are involved in HADDS and simplex autism. Hum Genomics. 2021;15(1):44. Epub 2021/07/15. doi: 10.1186/s40246-021-00342-3. PubMed PMID: 34256850; PMCID: PMC8278787.
42. Zhou X, Feliciano P, Shu C, Wang T, Astrovskaya I, Hall JB, Obiajulu JU, Wright JR, Murali SC, Xu SX, Brueggeman L, Thomas TR, Marchenko O, Fleisch C, Barns SD, Snyder LG, Han B, Chang TS, Turner TN, Harvey WT, Nishida A, O'Roak BJ, Geschwind DH, Michaelson JJ, Volfovsky N, Eichler EE, Shen Y, Chung WK. Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nature genetics. 2022;54(9):1305-19. Epub 2022/08/19. doi: 10.1038/s41588-022-01148-2. PubMed PMID: 35982159.
43. Kaplanis J, Samocha KE, Wiel L, Zhang Z, Arvai KJ, Eberhardt RY, Gallone G, Lelieveld SH, Martin HC, McRae JF, Short PJ, Torene RI, de Boer E, Danecek P, Gardner EJ, Huang N,

Lord J, Martincorena I, Pfundt R, Reijnders MRF, Yeung A, Yntema HG, Vissers L, Juusola J, Wright CF, Brunner HG, Firth HV, FitzPatrick DR, Barrett JC, Hurles ME, Gilissen C, Retterer K. Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature. 2020;586(7831):757-62. Epub 2020/10/16. doi: 10.1038/s41586-020-2832-5. PubMed PMID: 33057194; PMCID: PMC7116826.

44. Ware JS, Samocha KE, Homsy J, Daly MJ. Interpreting de novo variation in human disease using denovolyzeR. Current protocols in human genetics. 2015;87:7.25.1-15. Epub 2015/10/07. doi: 10.1002/0471142905.hg0725s87. PubMed PMID: 26439716; PMCID: PMC4606471.

45. Bernier R, Golzio C, Xiong B, Stessman HA, Coe BP, Penn O, Witherspoon K, Gerdts J, Baker C, Vulto-van Silfhout AT, Schuurs-Hoeijmakers JH, Fichera M, Bosco P, Buono S, Alberti A, Failla P, Peeters H, Steyaert J, Vissers LE, Francescatto L, Mefford HC, Rosenfeld JA, Bakken T, O'Roak BJ, Pawlus M, Moon R, Shendure J, Amaral DG, Lein E, Rankin J, Romano C, de Vries BB, Katsanis N, Eichler EE. Disruptive CHD8 mutations define a subtype of autism early in development. Cell. 2014;158(2):263-76. Epub 2014/07/08. doi: 10.1016/j.cell.2014.06.017. PubMed PMID: 24998929; PMCID: PMC4136921.

46. Zhou T, Enyeart PJ, Wilke CO. Detecting clusters of mutations. PloS one. 2008;3(11):e3765. Epub 2008/11/20. doi: 10.1371/journal.pone.0003765. PubMed PMID: 19018282; PMCID: PMC2582452.

47. Ye J, Pavlicek A, Lunney EA, Rejto PA, Teng CH. Statistical method on nonrandom clustering with application to somatic mutations in cancer. BMC bioinformatics. 2010;11:11. Epub 2010/01/08. doi: 10.1186/1471-2105-11-11. PubMed PMID: 20053295; PMCID: PMC2822753.

48. Gao J, Chang MT, Johnsen HC, Gao SP, Sylvester BE, Sumer SO, Zhang H, Solit DB, Taylor BS, Schultz N, Sander C. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome medicine. 2017;9(1):4. Epub 2017/01/25. doi: 10.1186/s13073-016-0393-x. PubMed PMID: 28115009; PMCID: PMC5260099.

49. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PK, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, Nagarajan N, Cortés-Ciriano I, Zhou DC, Liang WW, Hess JM, Yellapantula VD, Tamborero D, Gonzalez-Perez A, Suphavilai C, Ko JY, Khurana E, Park PJ, Van Allen EM, Liang H, Lawrence MS, Godzik A, Lopez-Bigas N, Stuart J, Wheeler D, Getz G, Chen K, Lazar AJ, Mills GB, Karchin R, Ding L. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;173(2):371-85.e18. Epub 2018/04/07. doi: 10.1016/j.cell.2018.02.060. PubMed PMID: 29625053; PMCID: PMC6029450.

50. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, Masica DL, Karchin R. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer research. 2016;76(13):3719-31. Epub 2016/05/20. doi: 10.1158/0008-5472.Can-15-3190. PubMed PMID: 27197156; PMCID: PMC4930736.

51. Hotspot Mutations Optimize the Oncogenic-Immunogenic Trade-off. Cancer Discov. 2022;12(7):Of19. Epub 2022/05/21. doi: 10.1158/2159-8290.Cd-rw2022-090. PubMed PMID: 35593588.

52. Turner TN, Douville C, Kim D, Stenson PD, Cooper DN, Chakravarti A, Karchin R. Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. Human molecular genetics. 2015;24(21):5995-6002. Epub 2015/08/08. doi: 10.1093/hmg/ddv309. PubMed PMID: 26246501.

53. Geisheker MR, Heymann G, Wang T, Coe BP, Turner TN, Stessman HAF, Hoekzema K, Kvarnung M, Shaw M, Friend K, Liebelt J, Barnett C, Thompson EM, Haan E, Guo H, Anderlid

BM, Nordgren A, Lindstrand A, Vandeweyer G, Alberti A, Avola E, Vinci M, Giusto S, Pramparo T, Pierce K, Nalabolu S, Michaelson JJ, Sedlacek Z, Santen GWE, Peeters H, Hakonarson H, Courchesne E, Romano C, Kooy RF, Bernier RA, Nordenskjold M, Gecz J, Xia K, Zweifel LS, Eichler EE. Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. Nature neuroscience. 2017;20(8):1043-51. Epub 2017/06/20. doi: 10.1038/nn.4589. PubMed PMID: 28628100; PMCID: PMC5539915.

54. Lelieveld SH, Wiel L, Venselaar H, Pfundt R, Vriend G, Veltman JA, Brunner HG, Vissers L, Gilissen C. Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. American journal of human genetics. 2017. Epub 2017/09/05. doi: 10.1016/j.ajhg.2017.08.004. PubMed PMID: 28867141.

55. Wiel L, Hampstead JE, Venselaar H, Vissers L, Brunner HG, Pfundt R, Vriend G, Veltman JA, Gilissen C. De novo mutation hotspots in homologous protein domains identify function-altering mutations in neurodevelopmental disorders. American journal of human genetics. 2023;110(1):92-104. Epub 2022/12/24. doi: 10.1016/j.ajhg.2022.12.001. PubMed PMID: 36563679; PMCID: PMC9892778.

56. Bewley AF, Akinwe TM, Turner TN, Gutmann DH. Neurofibromatosis-1 Gene Mutational Profiles Differ Between Syndromic Disease and Sporadic Cancers. Neurology Genetics. 2022;8(4):e200003. doi: 10.1212/nxg.0000000000200003.

57. Crawley JN, Heyer WD, LaSalle JM. Autism and Cancer Share Risk Genes, Pathways, and Drug Targets. Trends in genetics : TIG. 2016;32(3):139-46. Epub 2016/02/03. doi: 10.1016/j.tig.2016.01.001. PubMed PMID: 26830258; PMCID: PMC4769654.

58. Nussinov R, Tsai CJ, Jang H. How can same-gene mutations promote both cancer and developmental disorders? Sci Adv. 2022;8(2):eabm2059. Epub 2022/01/15. doi: 10.1126/sciadv.abm2059. PubMed PMID: 35030014; PMCID: PMC8759737.

59. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. Genome biology. 2016;17(1):122. Epub 2016/06/09. doi: 10.1186/s13059-016-0974-4. PubMed PMID: 27268795; PMCID: PMC4893825.

60. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9. Epub 2021/07/16. doi: 10.1038/s41586-021-03819-2. PubMed PMID: 34265844; PMCID: PMC8371605 have filed non-provisional patent applications 16/701,070 and PCT/EP2020/084238, and provisional patent applications 63/107,362, 63/118,917, 63/118,918, 63/118,921 and 63/118,919, each in the name of DeepMind Technologies Limited, each pending, relating to machine learning for predicting protein structures. The other authors declare no competing interests.

61. Cui H, Lan X, Lu S, Zhang F, Zhang W. Bioinformatic prediction and functional characterization of human KIAA0100 gene. J Pharm Anal. 2017;7(1):10-8. Epub 2018/02/07. doi: 10.1016/j.jpha.2016.09.003. PubMed PMID: 29404013; PMCID: PMC5686863.

62. Wang L, Jin W, Wu X, Liu Y, Gu W. Circ_0000520 interacts with miR-512-5p to upregulate KIAA0100 to promote malignant behaviors in lung cancer. Histol Histopathol. 2023;38(1):73-89. Epub 2022/07/23. doi: 10.14670/hh-18-498. PubMed PMID: 35866672.

63. Song J, Yang W, Shih Ie M, Zhang Z, Bai J. Identification of BCOX1, a novel gene overexpressed in breast cancer. Biochim Biophys Acta. 2006;1760(1):62-9. Epub 2005/11/18. doi: 10.1016/j.bbagen.2005.09.017. PubMed PMID: 16289875.
64. Zhong Z, Pannu V, Rosenow M, Stark A, Spetzler D. KIAA0100 Modulates Cancer Cell Aggression Behavior of MDA-MB-231 through Microtubule and Heat Shock Proteins. Cancers (Basel). 2018;10(6). Epub 2018/06/06. doi: 10.3390/cancers10060180. PubMed PMID: 29867023; PMCID: PMC6025110.
65. Cui H, Lan X, Lu S, Zhang F, Zhang W. Preparation of monoclonal antibody against human KIAA0100 protein and Northern blot analysis of human KIAA0100 gene. J Pharm Anal. 2017;7(3):190-5. Epub 2018/02/07. doi: 10.1016/j.jpha.2017.02.001. PubMed PMID: 29404037; PMCID: PMC5790689.
66. Ding YC, Song H, Adamson AW, Schmolze D, Hu D, Huntsman S, Steele L, Patrick CS, Tao S, Hernandez N, Adams CD, Fejerman L, Gardner K, Nápoles AM, Pérez-Stable EJ, Weitzel JN, Bengtsson H, Huang FW, Neuhausen SL, Ziv E. Profiling the Somatic Mutational Landscape of Breast Tumors from Hispanic/Latina Women Reveals Conserved and Unique Characteristics. Cancer research. 2023;83(15):2600-13. Epub 2023/05/05. doi: 10.1158/0008-5472.Can-22-2510. PubMed PMID: 37145128; PMCID: PMC10390863.
67. Sharma S, Bollinger KE, Kodeboyina SK, Zhi W, Patton J, Bai S, Edwards B, Ulrich L, Bogorad D, Sharma A. Proteomic Alterations in Aqueous Humor From Patients With Primary Open Angle Glaucoma. Invest Ophthalmol Vis Sci. 2018;59(6):2635-43. Epub 2018/05/31. doi: 10.1167/iovs.17-23434. PubMed PMID: 29847670; PMCID: PMC6733532.
68. Guo J, Wang M, Liu X. MicroRNA-195 suppresses tumor cell proliferation and metastasis by directly targeting BCOX1 in prostate carcinoma. J Exp Clin Cancer Res. 2015;34(1):91. Epub 2015/09/05. doi: 10.1186/s13046-015-0209-7. PubMed PMID: 26338045; PMCID: PMC4559360.
69. Liu T, Zhang XY, He XH, Geng JS, Liu Y, Kong DJ, Shi QY, Liu F, Wei W, Pang D. High levels of BCOX1 expression are associated with poor prognosis in patients with invasive ductal carcinomas of the breast. PloS one. 2014;9(1):e86952. Epub 2014/02/04. doi: 10.1371/journal.pone.0086952. PubMed PMID: 24489812; PMCID: PMC3904964.
70. Zhou FL, Zhang WG, Meng X, Chen G, Wang JL. [Bioinformatic analysis and identification for a novel antigen MLAA-22 in acute monocytic leukemia]. Zhongguo Shi Yan Xue Ye Xue Za Zhi. 2008;16(3):466-71. Epub 2008/06/14. PubMed PMID: 18549609.
71. Levine TP. Sequence Analysis and Structural Predictions of Lipid Transfer Bridges in the Repeating Beta Groove (RBG) Superfamily Reveal Past and Present Domain Variations Affecting Form, Function and Interactions of VPS13, ATG2, SHIP164, Hobbit and Tweek. Contact (Thousand Oaks). 2022;5:251525642211343. Epub 2022/12/27. doi: 10.1177/25152564221134328. PubMed PMID: 36571082; PMCID: PMC7613979.
72. Hanna M, Guillén-Samander A, De Camilli P. RBG Motif Bridge-Like Lipid Transport Proteins: Structure, Functions, and Open Questions. Annu Rev Cell Dev Biol. 2023;39:409-34. Epub 2023/07/05. doi: 10.1146/annurev-cellbio-120420-014634. PubMed PMID: 37406299.
73. Neuman SD, Levine TP, Bashirullah A. A novel superfamily of bridge-like lipid transfer proteins. Trends Cell Biol. 2022;32(11):962-74. Epub 2022/05/02. doi: 10.1016/j.tcb.2022.03.011. PubMed PMID: 35491307; PMCID: PMC9588498.
74. Parolek J, Burd CG. Bridge-like lipid transfer protein family member 2 suppresses ciliogenesis. bioRxiv. 2023:2023.12.07.570614. doi: 10.1101/2023.12.07.570614.

75. Karalis V, Donovan KE, Sahin M. Primary Cilia Dysfunction in Neurodevelopmental Disorders beyond Ciliopathies. J Dev Biol. 2022;10(4). Epub 2022/12/23. doi: 10.3390/jdb10040054. PubMed PMID: 36547476; PMCID: PMC9782889.