cLD: Rare-variant disequilibrium between genomic regions identifies novel genomic interactions

Dinghao Wang,1*, Jingni He,2*, Deshan Perera,2*, Chen Cao,2, Pathum Kossinna,2, Qing Li,2, William Zhang4, Alexander Platt5, Jingjing Wu1, Qingrun Zhang1,2,3#

1, Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada

2, Department of Biochemistry and Molecular Biology, University of Calgary, Calgary, AB, Canada

3, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada

4, The Harker School, San Jose, CA, USA

5, Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA USA.

*: joint first authors

#: corresponding should be addressed: qingrun.zhang@ucalgary.ca

Introductory Paragraph:

Linkage disequilibrium (LD) is a fundamental concept in genetics; critical for studying genetic associations and molecular evolution. However, LD measurements are only reliable for common genetic variants, leaving low-frequency variants unanalyzed. In this work, we introduce cumulative LD (cLD), a stable statistic that captures the rare-variant LD between genetic regions and opens the door for furthering biological knowledge using rare genetic variants. In application, we find cLD reveals an increased genetic association between genes in 3D chromatin interactions, a phenomenon recently reported negatively by calculating standard LD between common variants. Additionally, we show that cLD is higher between gene pairs reported in interaction databases, identifies unreported protein-protein interactions, and reveals interacting genes distinguishing case/control samples in association studies.

Main

Linkage Disequilibrium (LD) is a fundamental concept in population genetics that statistically captures non-random associations between two genetic variants due to reasons such as lack of recombination or different age of mutations[1]. LD serves as a core component in genotype-phenotype association mapping, as a statistically significant genetic variant could be just a proxy in LD with the genuine causal variant(s)[2]. Also, LD is critically important in analyzing the fine resolution of genotype-phenotype association mapping[3] and forming polygenic risk[4]. Additionally, from the perspective of molecular evolution, LD values substantially higher than expected under neutrality may indicate interesting phenomena, e.g., interactions between loci that are favoured by selection[5]. As such, LD has been extensively utilized in evolutionary studies.

The calculation of LD involves the use of allele frequencies of the genetic variants in its denominator to normalize the statistic (**Online Methods**; **Supplementary Materials 1.1**) and therefore suffers from a high variance (instability) when allele frequencies are close to zero. As such, in practice, researchers only analyze common genetic variants with minor allele frequency (MAF) higher than a threshold (*e.g.*,0.05), excluding more than 90% of human genetic variants[6].

In the field of association mapping, researchers have developed multiple techniques to aggregate the associations of multiple rare variants with a phenotype into a single shared effect. One of the pioneering methods that is still popularly used[7] is synthesizing a cumulative allele frequency from multiple rare genetic variants in the same genetic region (*e.g.*, within a gene). The cumulative minor allele frequency (cMAF) is defined on a region containing multiple rare variants: an individual will be labelled as a “mutant” if it has at least one of the rare variants, and then the proportion of individuals in the sample that are labelled as mutants will be the cMAF for this region (**Fig. 1a**).

Building on the idea of cMAF and the essence of LD, we developed a statistic, cumulative Linkage Disequilibrium (cLD) to capture the aggregated correlation between two sets of rare variants. Specifically, for the traditional calculation of LD between two variants, g1 and g2 with minor alleles a and b respectively the essential part is the definition of individual MAF P(a) and P(b) and the frequency that a and b show up in the same haplotype, P(ab). For calculating cLD between two regions, A and B, we first use cMAF to define P(A) and P(B) (the proportion of individuals carrying a rare variant within regions A and B, respectively); and then P(AB), the proportion of individuals who have at least one rare variant in both regions A and B (**Fig. 1b; Online Methods; Supplementary Materials 1.1 & 1.2**).

As cMAF is always higher than MAF, cLD's variance (reflecting its instability) should be lower than LD's. We verify this intuition by deriving the closed-form of variance of both LD and cLD (denoted as Var(LD) and Var(cLD)) using multinomial distributions and their multivariate normal approximation as well as the multivariate Delta Method[8](**Online Methods; Supplementary Materials 2.1 & 2.2**). Using allele frequencies from the 1000 Genomes Project data[6] and the formula (**Supplementary Materials 2.3**), we see that the variance of cLD is orders of magnitudes smaller (i.e., more stable) than the alternative -- calculating LD directly on rare variants in all ethnic populations and all cMAF bins (**Fig. 2a; Supplementary Figs. 2.2a & 2.3a**). Additionally, following the conventional statistical procedure of bootstrapping to empirically estimate stability, we sub-sample half of each population sample 1,000 times to form bootstrapped distributions for both cLD and LD (**Online Methods; Supplementary Materials 2.4**). The subsampling shows that cLD exhibits a much slimmer bootstrapped distribution than LD across three ethnic groups (**Fig. 2b, Supplementary Figs. 2.2b & 2.3b**), further confirming the greater stability of cLD compared to traditional measures of LD.

By aggregating information from multiple independent mutations, cLD is sensitive to subtle interactions poorly reflected by LD (which can only account for two at a time). Interactions within the 3D structure of genomes is one place where this difference allows for insight from cLD where

84 LD-based methods fail. The availability of high-throughput experimental technologies that can
85 assess chromatin conformation such as Hi-C[9,10] allows researchers to analyze genetic regions that
86 are in close contact in 3D spatial structure. There was a widely disseminated expectation that the
87 3D genomic interaction in the form of chromatin contact may leave a footprint in the form of genetic
88 LD[11]. Motivated by such expectation, Whalen and Pollard calculated the standard LD based on
89 common variants (MAF>0.05) in 1000 Genomes Project data[6]and reported negative results stating
90 that genetic LD map is not overlapping with the 3D contact map[12]. However, by reanalyzing the
91 1000 Genomes sequencing data and Hi-C data[9,10] in the developing brain using cLD on rare
92 variants (**Online Methods; Supplementary Materials 3.1 & 3.2**), we revealed that the 3D
93 chromatin interactions did leave genetic footprints in the form of higher cLD in pairs of genes that
94 are in the adjacent Hi-C regions (**Fig 3a; Supplementary Fig. 3.1**). To assess the statistical
95 significance of the enrichment of cLD in 3D contact, we conducted Mantel-Haenszel Fisher exact
96 tests (**Supplementary Materials 3.4**), both of which are highly significant (P-value < 1.0E-50;
97 **Supplementary Table 3.2, Supplementary Materials 3.4.1**). As Whalen & Pollard's work[12] is not
98 at the resolution of genes, we re-calculated standard LD using common variants based on gene
99 pairs, which shows a subtle effect (**Fig. 3b, Supplementary Fig. 3.2**) but still not statistically
100 significant in (P-value =0.999**; Supplementary Table 3.3; Supplementary Materials 3.4.1**).
101 Additionally, we checked the ratio between the numbers of pairs of genes within the 3D contact
102 region and without as a function of cLD value (**Supplementary Materials 3.5**) and found that the
103 ratio is large and increases as the cLD cut-off increases (**Fig 3c,d,e, Supplementary Table 3.7)**.
104 Taking together, 3D interactions clearly overlap with genetic interactions and cLD outperforms LD
105 in observing this.
106
107 To demonstrate that gene-gene interactions leave footprints in rare genetic mutations regardless of
108 their physical position we computed the distribution of cLD among interacting pairs genes reported
109 in Reactome[13]and BioGRID[14], MINT[15]and IntAct[16] (**Online Methods; Supplementary Materials**
110 **3.3**). We compared this distribution against a null distribution formed by all pairs of genes. Indeed,
111 the comparisons led to the expected result: for gene pairs separated by any physical distance

112 within 2MB, cLD is elevated in interacting gene pairs(**Fig. 4; Supplementary Fig. 3.3**). Again, the
113 Mantel-Haenszel and Fisher exact tests confirm that the differences are significant (P-value <
114 1.0E-20; **Supplementary Table 3.5; Supplementary Materials 3.4.2**).
115
116 cLD is also effective at identifying novel pairs of likely interacting proteins. Looking at all pairs of
117 genes, we observed several pairs without prior evidence of interaction with extraordinarily high
118 cLD, such as between 3BCZ and 4RIQ (encoded for by genes *MEMO1* and *DPY30*, respectively)
119 with a cLD of 0.86. We conducted protein docking analysis for the 19 pairs of genes of large cLD
120 values (top 0.01% among all gene pairs) with cMAF >0.05 and existing IDs in PDB, however, not
121 reported in any databases (**Online Methods, Supplementary Materials 4.1, Supplementary**
122 **Table 4.1**). We found multiple lines of evidence of the interaction of five pairs (**Supplementary**
123 **Table 4.2**) in terms of both binding affinity and interacting residues (**Fig. 5a-d; Supplementary**
124 **Figs. 4.1-4.4**).
125
126 In the context of case/control association studies, cLD can be used to identify pairs of genes
127 whose interactions may be responsible for human disease. Using data from the Autism Spectrum
128 Disorders (ASD) whole exome sequencing dataset[17] we calculated cLD scores for all pairs of
129 genes, with separate scores for the populations of cases and controls (**Online Methods;**
130 **Supplementary Materials 5.1 & 5.2**). The difference in cLD for a pair of genes conditional on
131 case/control status (ΔcLD) is indicative of an interaction that is non-random associating with
132 disease status. Overall, using a hypergeometric test, we analyzed the enrichment among high-
133 ΔcLD genes for ASD sustainability genes reported by DisGeNet[18], an established general
134 database for diseases and SFARI[19], a gold-standard database focusing on ASD (**Supplementary**
135 **Materials 5.3**). The genes included in the pairs with the highest ΔcLD scores are highly enriched in
136 both the Autism related genes in DisGeNet (**Fig. 6a**) and SFARI (**Fig. 6b**). Gene Ontology[20] and
137 pathways (KEGG)[21,22]enrichment analysis for the high ΔcLD genes (**Online Methods;**
138 **Supplementary Materials 5.4** also showed sensible biological functions and pathways (**Figure**
139 **6c,d**) that are well supported by the literature (**Supplementary Materials 5.4**)[20–35]. By taking a

closer look of the 20 genes identified by the top 10 gene pairs with the highest ΔcLD values, found that 14 genes (70%) have been reported to be associated with ASD, including *DENND4A, EFCAB5, ABI2, RAPH1, MSTO1, DAP3, ARL13B, PRB2, PRB1, ZNF276, FANCA, ADAM7, SLC26A1* and *TUBB8* (**Supplementary Table 5.1**). Moreover, among the rest of six genes, we also identified indirect links of two, *RAB11A* and *IDUA* with ASD (**Supplementary Materials 5.3**).

LD is a broadly applicable concept applicable to many types of genetic analyses. cLD allows us to expand on this concept and capture additional information from the distributions of many variants segregating in a population at low frequencies within particular regions of a genome. In contrast to the previous attempts to utilize LD between multiple variants focusing on dominant haplotypes[36] or joint distributions[37], cLD emphasizes biological interactions. With it's demonstrated power in identifying gene and protein interactions, cLD might offer an essential tool to analyze biological interactions and their evolution.

Author contributions. Conceived the study: QZ. Supervised the study: JW and QZ. Analyzed real data: DW, JH, DP, PK, QL. Conducted mathematical derivation and statistical simulations: DW, WZ, JW. Provided comments: CC, AP. Wrote the paper: DW and QZ with major input from JH, DP, AP, and minor input from all authors.

Acknowledgement. Q.Z. is supported by NSERC Discovery Grant (RGPIN-2018-05147), University of Calgary VPR Catalyst grant and New Frontiers in Research Fund (NFRFE-2018-00748); J.W. is supported by NSERC Discovery Grant (RGPIN-2018-04328); A.P. is supported by NIH (R35 GM134957-01) and American Diabetes Association (Pathway to Stop Diabetes grant 1-19-VSN-02); D.P. is supported by Alberta Innovates Graduate Scholarship and Eyes High International Scholarship; J.H. is supported by CSC Scholarship. The computational infrastructure is funded by Canada Foundation for Innovation JELF grant (36605).

168

References:

1. Slatkin, M. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* vol. 9 477–485 (2008).

2. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature Genetics* **52**, 1355–1363 (2020).

3. Flint-Garcia, S. A., Thornsberry, J. M. & Edward IV, S. B. Structure of Linkage Disequilibrium in Plants. *Annual Review of Plant Biology* vol. 54 357–374 (2003).

4. Amariuta, T. *et al.* Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nature Genetics* **52**, 1346–1354 (2020).

5. Gregersen, J. W. *et al.* Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature* **443**, 574–577 (2006).

6. Auton, A. *et al.* A global reference for human genetic variation. *Nature* vol. 526 68–74 (2015).

7. Li, B. & Leal, S. M. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *American Journal of Human Genetics* **83**, 311–321 (2008).

8. Lehmann Springer, E. L. *Elements of Large-Sample Theory.*

9. Rajarajan, P. *et al.* Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* **362**, (2018).

10. Akbarian, S. *et al.* The PsychENCODE project. *Nature Neuroscience* vol. 18 1707–1712 (2015).

11. Joiret, M., Mahachie John, J. M., Gusareva, E. S. & van Steen, K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Mining* **12**, (2019).

12. Whalen, S. & Pollard, K. S. Most chromatin interactions are not in linkage disequilibrium. *Genome Research* **29**, 334–343 (2019).

13. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **46**, D649–D655 (2018).

14. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic acids research* **34**, (2006).

15. Orchard, S. Molecular interaction databases. *Proteomics* vol. 12 1656–1662 (2012).

16. Orchard, S. *et al.* The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* **42**, (2014).

17. 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).

18. Piñero, J. *et al.* DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* **45**, D833–D839 (2017).

19. Abrahams, B. S. *et al.* SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Molecular Autism* **4**, (2013).

20. Ashburner, M. *et al. Gene Ontology: tool for the unification of biology The Gene Ontology Consortium**. http://www.flybase.bio.indiana.edu (2000).

21. Kanehisa, M. & Goto, S. *KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research* vol. 28 http://www.genome.ad.jp/kegg/ (2000).

22. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research* **38**, (2009).

23. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology* **16**, 284–287 (2012).

24. Rojas, D. C. The role of glutamate and its receptors in autism and the use of glutamate receptor antagonists in treatment. *Journal of Neural Transmission* **121**, 891–905 (2014).

25. Hannelius, U. *et al.* Phenylketonuria screening registry as a resource for population genetic studies. *Journal of medical genetics* vol. 42 (2005).

26. Richler, E., Reichert, J. G., Buxbaum, J. D. & Mcinnes, L. A. *Autism and ultraconserved non-coding sequence on chromosome 7q. Psychiatric Genetics* vol. 16 http://www.cse.ucsc.edu/Bjill/ultra.html (2006).

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

28. Fung, L. K. & Hardan, A. Y. Developing Medications Targeting Glutamatergic Dysfunction in Autism: Progress to Date. *CNS Drugs* **29**, 453–463 (2015).

232 29. Sato, D. *et al.* SHANK1 deletions in males with autism spectrum disorder. *American Journal*
233 *of Human Genetics* **90**, 879–887 (2012).

234 30. Berkel, S. *et al.* Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum
235 disorder and mental retardation. *Nature Genetics* **42**, 489–491 (2010).

236 31. Durand, C. M. *et al.* Mutations in the gene encoding the synaptic scaffolding protein
237 SHANK3 are associated with autism spectrum disorders. *Nature Genetics* **39**, 25–27 (2007).

238 32. Wei, H. *et al.* Genetic risk factors for autism-spectrum disorders: a systematic review based
239 on systematic reviews and meta-analysis. *Journal of Neural Transmission* vol. 128 717–
240 734 (2021).

241 33. Ye, H., Liu, J. & Wu, J. Y. Cell adhesion molecules and their involvement in autism spectrum
242 disorder. *NeuroSignals* vol. 18 62–71 (2011).

243 34. Betancur, C., Sakurai, T. & Buxbaum, J. D. The emerging role of synaptic cell-adhesion
244 pathways in the pathogenesis of autism spectrum disorders. *Trends in Neurosciences* vol.
245 32 402–412 (2009).

246 35. Lin, Y. C., Frei, J. A., Kilander, M. B. C., Shen, W. & Blatt, G. J. A subset of autism-associated
247 genes regulate the structural stability of neurons. *Frontiers in Cellular Neuroscience* vol.
248 10 (2016).

249 36. Zan, Y., Forsberg, S. K. G. & Carlborg, Ö. On the relationship between high-order linkage
250 disequilibrium and epistasis. *G3: Genes, Genomes, Genetics* **8**, 2817–2824 (2018).

251 37. Turkmen, A. & Lin, S. Are rare variants really independent? *Genetic Epidemiology* **41**, 363–
252 371 (2017).

38. Yan, Y., Tao, H., He, J. & Huang, S. Y. The HDOCK server for integrated protein–protein docking. *Nature Protocols* **15**, 1829–1852 (2020).

39. Yan, Y., Zhang, D., Zhou, P., Li, B. & Huang, S. Y. HDOCK: A web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Research* **45**, W365–W373 (2017).

40. Berman, H. M. *et al. The Protein Data Bank*. *Nucleic Acids Research* vol. 28 http://www.rcsb.org/pdb/status.html (2000).

41. Perera, D. D. B. D., Perera, K. M. L. & Peiris, D. C. A novel in silico benchmarked pipeline capable of complete protein analysis: A possible tool for potential drug discovery. *Biology* **10**, (2021).

42. Delano, W. L. *PyMOL: An Open-Source Molecular Graphics Tool*.

43. Laskowski, R. A. & Swindells, M. B. LigPlot+: Multiple ligand-protein interaction diagrams for drug discovery. *Journal of Chemical Information and Modeling* **51**, 2778–2786 (2011).

Figures and Tables

Figure 1: **Illustrating the idea of a) cMAF and b) cLD.** An example to show the calculation of cLD. **a)** There are two individuals [1, 4] who have mutations in region A. Therefore, the cMAF P(A) for region A is 2/6 = 0.33. **b)** There are three individuals [3, 4, 5] who have mutations in region B and the cMAF P(B) for region B is 3/6 = 0.50. If we consider regions A and B together, there is one individual with mutations in both regions: [4]. Thus, the P(AB) is 1/6 = 0.17. Finally, by yielding P(A), P(B) and P(AB) into the standard formula of LD we have cLD = 0.375.

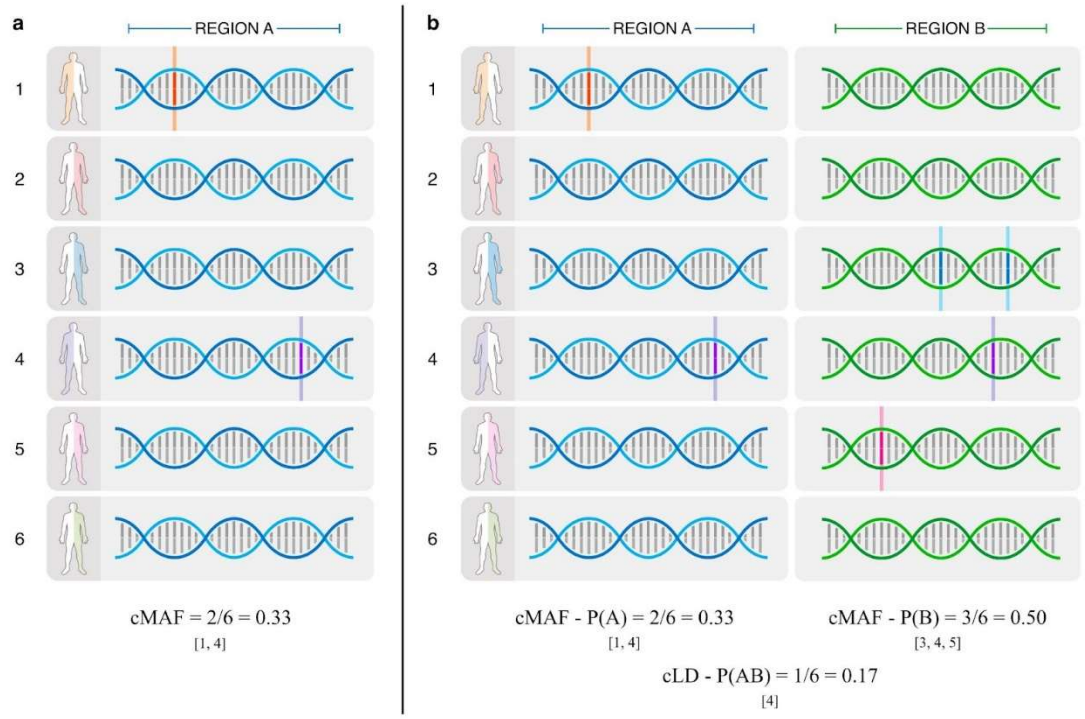


Figure 2: Stability of cLD and LD revealed by closed-form variance calculation and bootstrapped resampling. a) The gene pairs were split into four different bins based on the cMAF values, i.e., <0.05, 0.05 - 0.1, 0.1 - 0.2, and 0.2-0.4 (y-axis). The x-axis is the ratio between the variance of LD and cLD. **b)** Probability density distribution of cLD and LD from bootstrapped samples. Results from the European population are shown. See **Supplementary Materials** Section 2 for other populations.

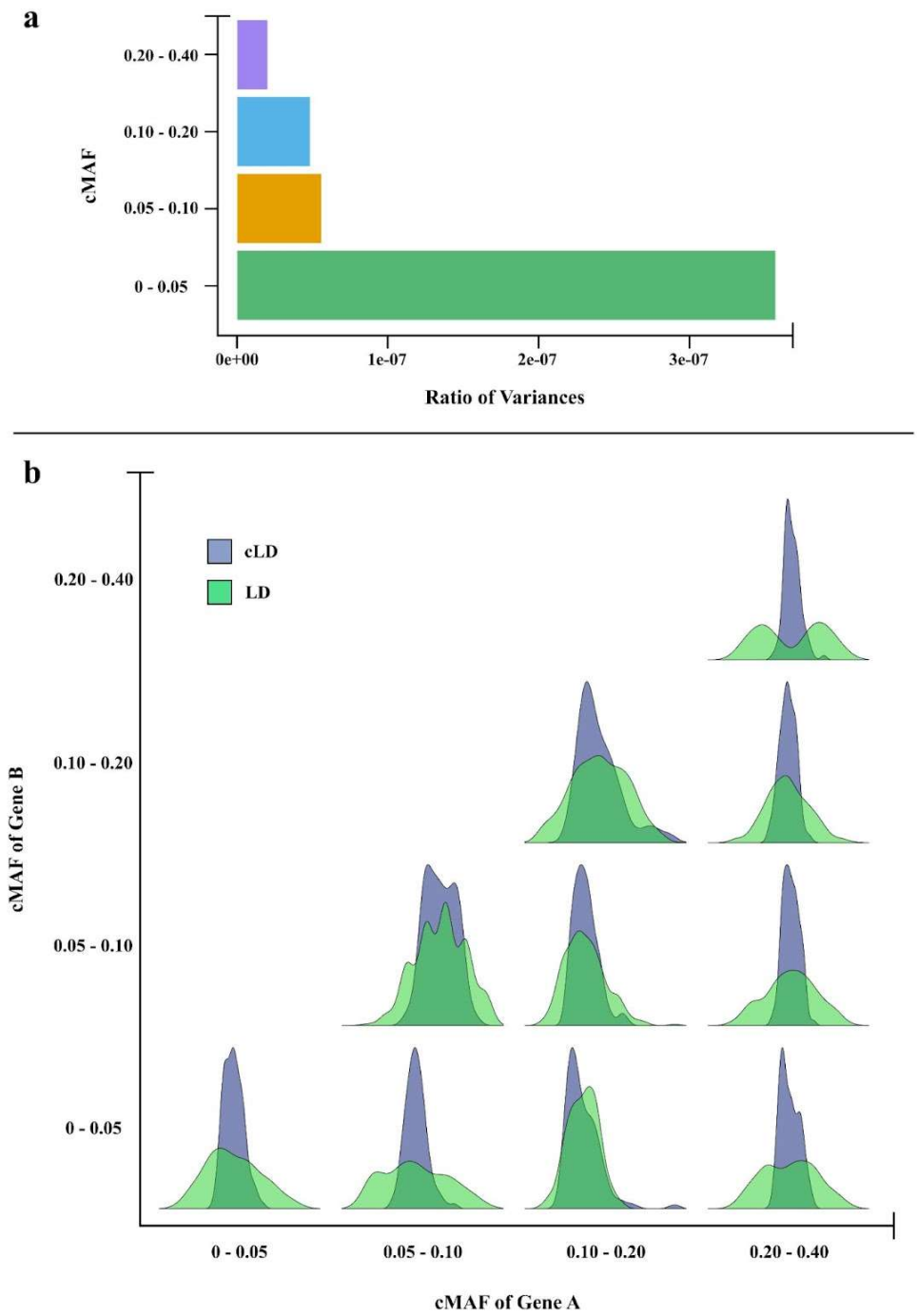


Figure 3: Enrichment of cLD among pairs of genes in chromatin contact regions. **a)** The comparisons of cLD values between the 3D chromatin interaction regions and non-interaction regions among 13 different distance groups in the European population. **b)** The same comparisons using LD values. **c-e)** The enrichment of the ratios between the number of gene pairs in 3D chromatin interaction regions against the number of gene pairs that are not in 3D regions. The x-axis is the cLD value cutoffs. **a)** European population. **b)** African population. **c)** East Asian population.

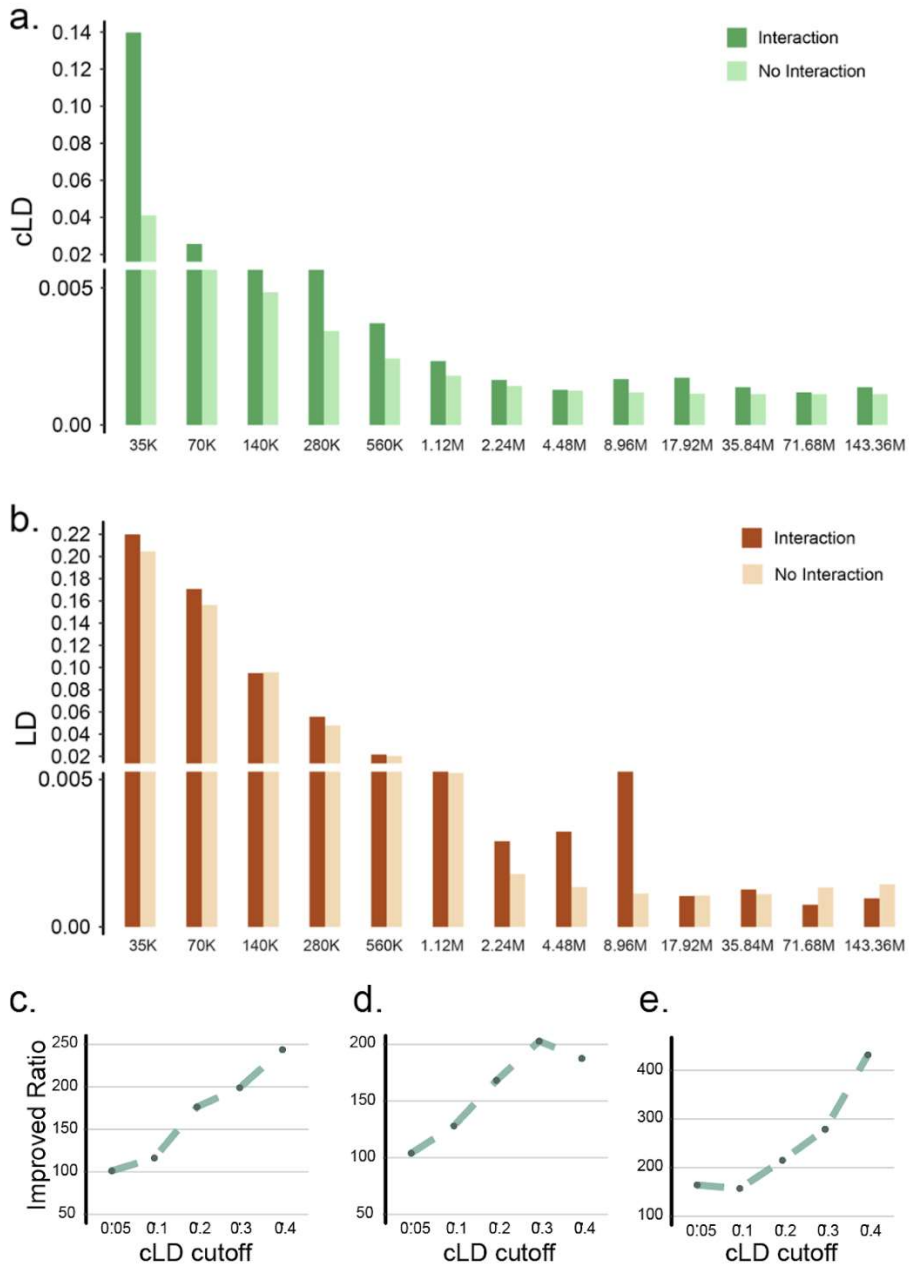


Figure 4: The comparisons of cLD values in European populations between gene pairs found in interaction databases and all pairs that are not in databases. Each bar represents the average of pairs with distance smaller than the value of its x-axis label but larger than the value of the previous x-axis label. (Other populations show the same trend, Supplementary Materials)

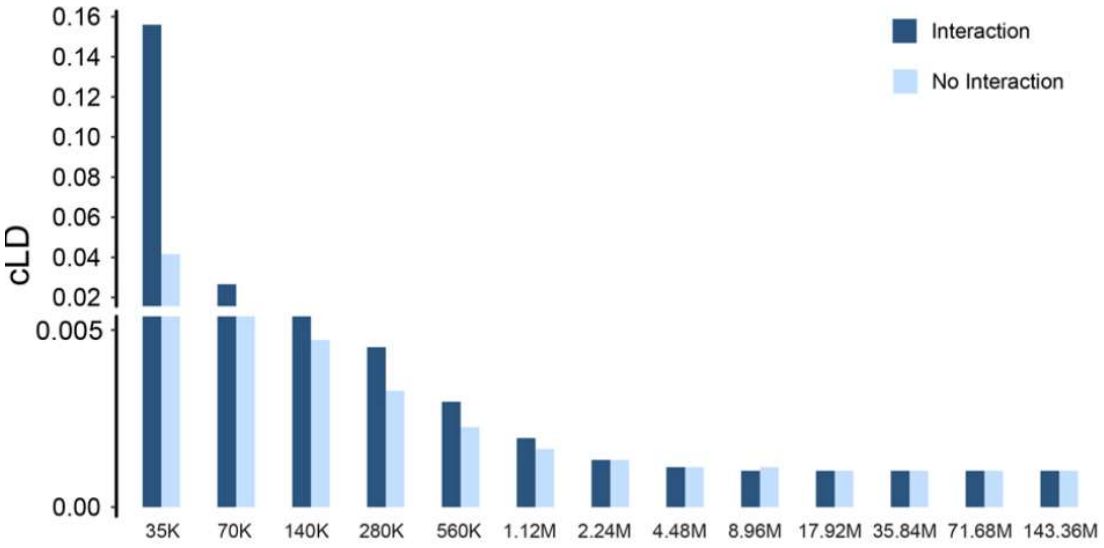


Figure 5: Protein docking interaction between 3BCZ and 4RIQ revealed by cLD (=0.86) with a binding affinity of -341.21 kJ/mol. **a)** Structure of 3BCZ (red) and 4RIO (blue) protein-protein complex. **b-d)** 2D representation of closest interacting residues around the protein-protein interaction interfaces, including multiple non-covalent bonds, for example, hydrogen bonds (green dotted line) and hydrophobic interactions (read and rose semi-circle with spikes). Residues for the 3BCZ are depicted in upper letters (T, U, O, R, N) and for the 4RIO are depicted in lower letters.

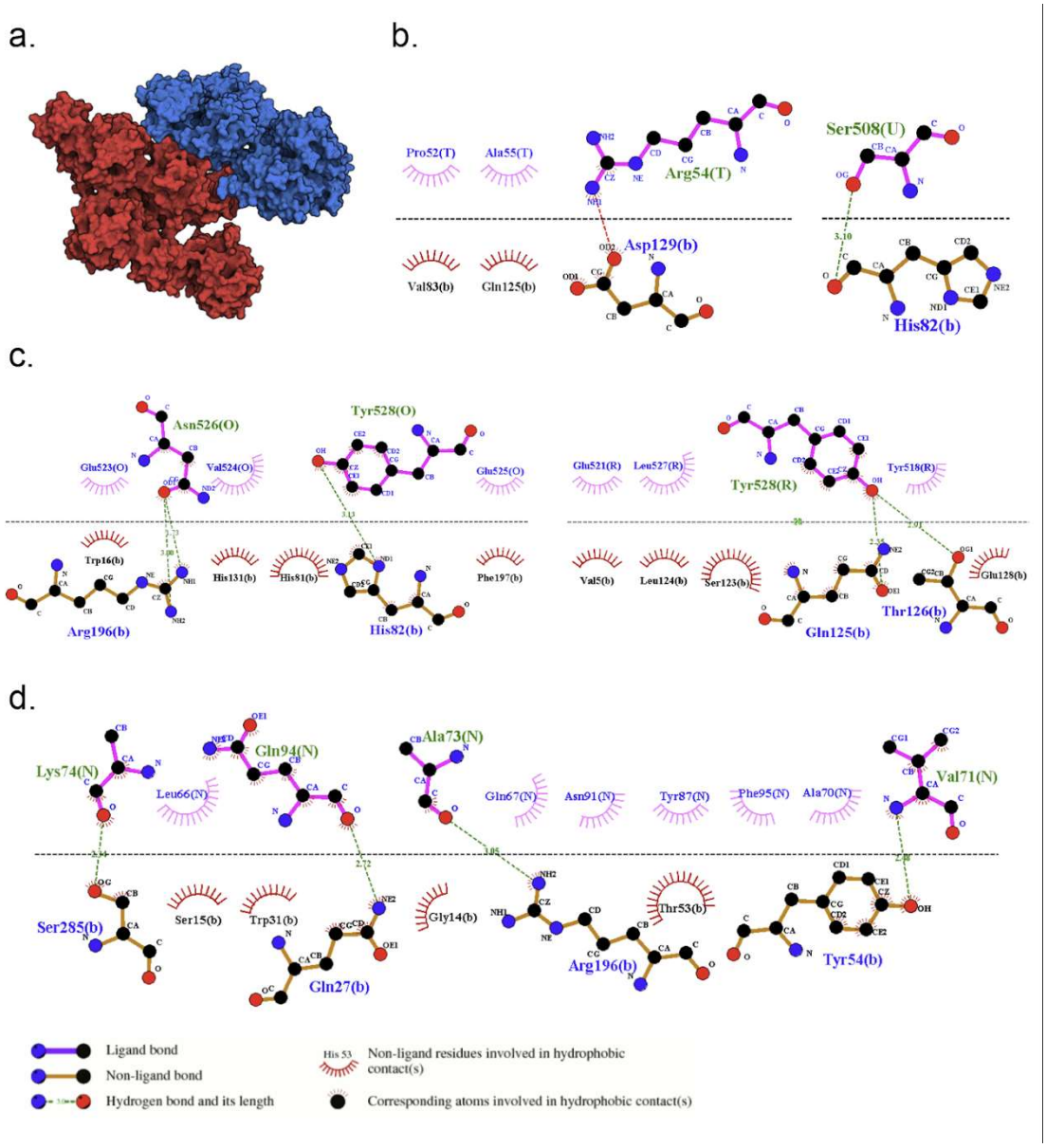
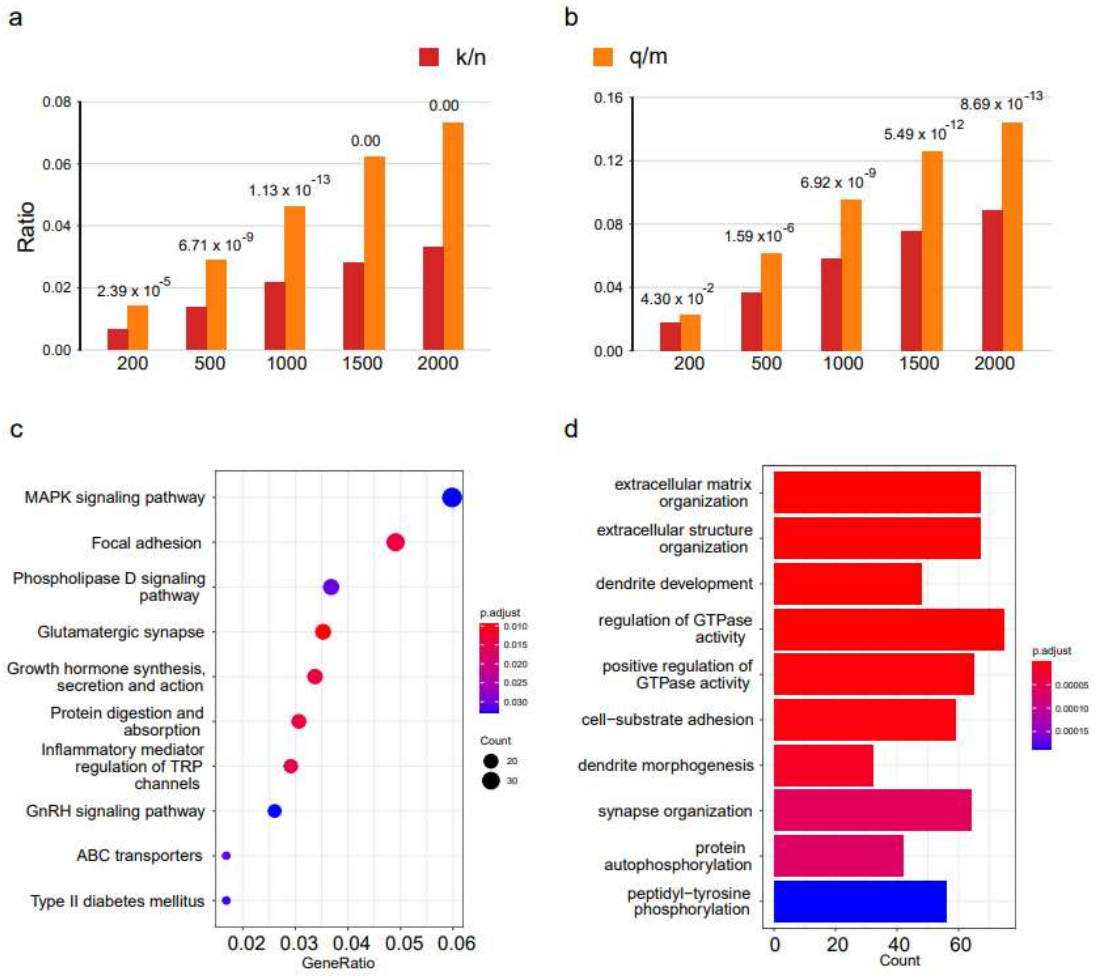


Figure 6: **ΔcLD gene pairs in case/control association mapping data: annotation of top genes and enrichment of pathways**. **a-b)** Group bar charts show the ratio between the number of selected genes being validated in the database dividing the number of genes in the database (q/m) and the number of selected genes dividing the total number of genes in the population minus m (k/n). The values on the top of each bar are the p-values of the hypergeometric distribution probability test. The x-axis indicated the top gene pairs from the top 200 to 2,000. **a)** DisGeNet database. **b**) SFARI database. **c)** a dot plot showing the top 10 KEGG pathways ranked by the GeneRatio values. The size of the balls indicates the number of the genes enriched and the color indicates the level of the enrichment (P-adjusted values). The GeneRatio is calculated as count/setSize. 'count' is the number of genes that belong to a given gene-set, while 'setSize' is the total number of genes in the gene-set. **d).** a bar plot showing the top 10 enriched biological processes ranked by p-values. The correlation is more significant as the red/blue ratio increases. The number on the x-axis indicates the number of genes that belong to a given gene set.



339

340

341

342

343

344

345

346

347

348

349

Online Methods (maximal 1,500 words)

Definition of LD and cLD:

The calculation of LD between two bi-allelic loci relies on the estimate of three key quantities: P_A, the allele frequency of an allele in locus A, P_B the allele frequency of an allele in locus B, and P_{AB}, the frequency of these two alleles of A and B showing up together. Then one can define the unnormalized disequilibrium statistic $D = P_{AB} - P_A P_B$. To rescale the statistic based on allele frequency, one can normalize D by dividing it by the allele frequency variances:

$$r^2 = \frac{D^2}{P_A(1-P_A)P_B(1-P_B)}$$

Another different definition is D' (that will not be the focus of this paper). As LD involves P_A and P_B in the denominator, it is highly instable when P_A or P_B are very close to zero, which means LD cannot be used if A or B are rare variants.

The cLD statistic is designed to handle the above problem by aggregating rare variants cumulatively. More specifically, here we look at two sets of variants in two genetic regions, e.g., two genes, again namely A and B. Assuming that there are m SNPs in gene A, and there are r SNPs in gene B. Also, we assume the sample size is n. Then, for gene A, we use $\{S_{1i}, S_{2i}, \ldots, S_{mi}\}$ to denote the allele of the s-th (s= 1, 2, …, m) SNP in the i-th individual ($i=1,2,\ldots,n$). Similarly, for gene B, we use $\{K_{1i}, K_{2i}, \ldots, K_{ri}\}$ to denote the allele of the k-th ($k= 1, 2, \ldots, r$) SNP in the i-th individual ($i=1,2,\ldots,n$). Note that S_{si} and K_{ki} is either 0 or 1. (0 denotes a major allele, whereas 1 denotes a minor allele).

Then we have the cMAF (P_A & P_B) defined below:

$$P_A = \frac{1}{n}\sum_{i=1}^{n} I\left(\sum_{s=1}^{m} S_{si} \geq 1\right)$$

$$P_B = \frac{1}{n}\sum_{i=1}^{n} I\left(\sum_{k=1}^{r} K_{ki} \geq 1\right)$$

$$P_{AB} = \frac{1}{n}\sum_{i=1}^{n} I\left(I\left(\sum_{k=1}^{R} K_{ki} \geq 1\right) + I\left(\sum_{s=1}^{m} S_{si} \geq 1\right) = 2\right)$$

Based on these values, we can calculate the r^2 version of cLD:

$$cLD = \frac{\left(P_{AB} - P_A P_B\right)^2}{P_A\left(1-P_A\right)P_B\left(1-P_B\right)}$$

The more rigorous mathematical descriptions and the definition of D' version are provided in Supplementary Materials.

Derivation of variances:

To obtain the variance of cLD and LD, we derived their asymptotic distributions. The details are in Supplementary Materials. The gist of our approach is summarized in the following three steps:

First, using multinomial random variables, we rewrite the formula of cLD and LD in terms of counts. In the definition, we use X_{ijk} to denote the allele of the *k-th* variant of the *j-th* gene for the *i-th* individual (haplotype) of. For a pair of variants, the *i*-th pair (X_{i1u},X_{i2v}) (i = 1,…,n) can take possible values (1,1), (0,1), (1,0) and (0,0). If we use O_1 to O_4 to denote the count of the 4 possible pairs in two variants, then the distribution of $\mathbf{O} = (O_1, O_2, O_3, O_4)$ is $\mathbf{O} \sim multinom\left(n; \mathbf{p}\right)$ with $\mathbf{p} = (p_1, p_2, p_3, p_4)$ represents the population probability. The LD between the *u*-th and *v*-th variants can be rewritten as $LD_{(u,v)} = \frac{\left(O_1O_4 - O_2O_3\right)^2}{\left(O_1+O_2\right)\left(O_1+O_3\right)\left(O_2+O_4\right)\left(O_3+O_4\right)}$. Similarly, we use the same strategy to using multinomial random variables to describe cLD as follows:

In analogy to the case of LD, we used X_{ij} to denote the allele of the *j-th* gene for the *i-th* individual (haplotype). For a pair of genes, the *i*-th pair (X_{i1},X_{i2}) (i = 1,…,n) can take possible values (1,1), (0,1), (1,0) and (0,0). Using M_1 to M_4 to denote the counts of the 4 possible pairs in two genes, then the distribution of $\mathbf{M} = (M_1, M_2, M_3, M_4)$ is $\mathbf{M} \sim multinom(n; q)$ with $\mathbf{q} = (q_1, q_2, q_3, q_4)$ represents the population probability. The cLD between a pair of genes could be rewritten as

$$cLD = \frac{(M_1 M_4 - M_2 M_3)^2}{(M_1 + M_2)(M_1 + M_3)(M_2 + M_4)(M_3 + M_4)}.$$

Second, we use the central limit theorem (CLT) to derive the asymptotic multivariate normal distribution. In the LD case, with the population mean $\mathbf{p} = (p_1, p_2, p_3, p_4)$, we can write the covariance matrix as

$$\Sigma = \begin{pmatrix} p_1 - p_1^2 & -p_1 p_2 & -p_1 p_3 & -p_1 p_4 \\ -p_2 p_1 & p_2 - p_2^2 & -p_2 p_3 & -p_2 p_4 \\ -p_3 p_1 & -p_3 p_2 & p_3 - p_3^2 & -p_3 p_4 \\ -p_4 p_1 & -p_4 p_2 & -p_4 p_3 & p_4 - p_4^2 \end{pmatrix}.$$

Then by the multivariate CLT[8] we have $\sqrt{n}\left(\frac{\mathbf{O}}{n} - \mathbf{p}\right) \xrightarrow{L} N(\mathbf{0}, \Sigma)$.

In the cLD case, with the population mean $\mathbf{q} = (q_1, q_2, q_3, q_4)$, we can write the covariance matrix as

$$\mathbf{Q} = \begin{pmatrix} q_1 - q_1^2 & -q_1 q_2 & -q_1 q_3 & -q_1 q_4 \\ -q_2 q_1 & q_2 - q_2^2 & -q_2 q_3 & -q_2 q_4 \\ -q_3 q_1 & -q_3 q_2 & q_3 - q_3^2 & -q_3 q_4 \\ -q_4 q_1 & -q_4 q_2 & -q_4 q_3 & q_4 - q_4^2 \end{pmatrix}.$$

Then by the multivariate CLT[8] we have $\sqrt{n}\left(\frac{\mathbf{M}}{n} - \mathbf{q}\right) \xrightarrow{L} N(\mathbf{0}, \mathbf{Q})$.

Third, as the cLD and LD are functions of random variables, we apply the multivariate Delta method[8] to derive the distribution of cLD and LD. In the LD case, suppose the Jacobian matrix of

418 $LD(\mathbf{O})$ is $\mathbf{J}_{LD} = [\frac{\partial LD(\mathbf{O})}{\partial \mathbf{O}}]|_{\mathbf{O}=n\mathbf{p}}$. Then the asymptotic distribution of $LD(\mathbf{O})$ is

419 $LD(\mathbf{O}/n) - LD(\mathbf{p}) \sim AN\left(0, n\mathbf{J}_{LD}\Sigma\mathbf{J}_{LD}^{T}\right)$, where '*AN*' stands for asymptotic normal.

420

421 In the cLD case, suppose the Jacobian matrix of $cLD(\mathbf{M})$ is $\mathbf{J}_{cLD} = \left[\frac{\partial cLD(\mathbf{M})}{\partial \mathbf{M}}\right]|_{\mathbf{M}=n\mathbf{q}}$. Then the

422 asymptotic distribution of $cLD(\mathbf{M})$ is $cLD(\mathbf{M}/n) - cLD(\mathbf{q}) \sim AN\left(0, n\mathbf{J}_{cLD}\mathbf{Q}\mathbf{J}_{cLD}^{T}\right)$

423

424 **Assessing the instability of LD and cLD using bootstrapped distributions**

425

426 We randomly sample half of the haplotypes in each population (EUR, AFR, or EAS), and
427 calculated the average cLD and average LD over the gene pairs within a cMAF group and repeat
428 this procedure 1,000 times. Based on these bootstrapped cLDs and LDs we form bootstrapped
429 distribution cLD and LD (with appropriate re-scaling (Supplementary Materials). More specifically,
430 we randomly sample 1000 genes from the 1000 Genomes Project data and assessed their
431 pairwise LD and cLD in stratified cMAF bins (Supplementary Materials) using half of the haplotypes
432 in the given population (AFR, EAS or EUR). These randomly drawn subsamples (each with half of
433 the individuals in the original population) form bootstrapped samples. We define the LD of a gene
434 pair as the average value of *LD(O/n)* over all rare SNV pairs within that gene pair. In each iteration,
435 we calculate the average cLD over the gene pairs in each group (Supplementary Materials). Then
436 using these average cLD and LDs, we plot the bootstrapped densities in **Fig. 2b** for the
437 visualization.

438

439 **Calculation of cLD and LD for gene pairs in 3D interaction regions.**

440

441 *1000 Genome Variant Call Data.* The variant call data of the Phase 3 analysis of the 1000 Genome
442 dataset was obtained through The European Bioinformatics Institute's dedicated FTP

443 (http://ftp.1000genomes.ebi.ac.uk) server. The complete variant call dataset was found using the
444 webpage (Announcements | 1000 Genomes (internationalgenome.org)) (This is a sub-page
445 maintained by the 1000 Genome webpage) and downloaded from (Index of
446 /vol1/ftp/release/20130502/ (ebi.ac.uk)).
447
448 *3D chromatin conformation assessed by Hi-C*.
449 Here, we utilized a Hi-C assessment in the developing brain, which has 27,982 brain-specific paired
450 3D-interacting regions, measured from neurons derived from human induced pluripotent stem cells
451 (hiPSCs) [39, 52]. This dataset is available in the Synapse database (https://www.synapse.org/) with
452 Synapse ID: syn12979149.
453
454 *cLD calculation*: We first calculated the distance between the genes in each pair and separate the
455 gene pairs into 13 distance groups (Supplementary Materials). After stratifying all gene pairs into
456 13 distance groups, within each distance group, we calculate cLD between all gene pairs and
457 further split them into two categories: the ones that are located in 3D interaction regions (assessed
458 by Hi-C experiments) and the ones that are located in non-3D interaction regions. The gene pairs
459 with one gene in interaction region will be discarded. Finally, the average cLD over gene pairs
460 within interaction and non-interaction regions will be used to generate the bar-charts in Fig. 3a.
461
462 *LD calculation*: In general, the procedure mirrors the one used above for cLD using distance
463 groups and 3D-interaction vs non-interaction categories. As LD is originally defined by individual
464 variants, not genes, the following averaging steps are taken. For each gene pair in the 3D
465 interaction regions, we randomly chose 2,000 rare variant pairs from it to calculate their LD values.
466 For each selected rare variant pair, we calculated its distance and then, among the gene pairs
467 without 3D interactions, we randomly selected another rare variant pair with the same or very
468 similar distance (Supplementary Materials). As a result, we achieved 2,000 randomly selected
469 variant pairs from gene pairs without interaction that are matched up with the 2,000 variant pairs

470 from gene pairs with interaction. The average LD within these two sets of 2,000 pairs will then
471 contribute to the bar charts in Fig 3b (Supplementary Materials).
472
473 **Calculation of cLD and LD for gene pairs in gene-gene interaction databases**
474
475 *Gene-gene interaction databases*. Four frequently used databases, Biogrid[14]
476 Reactome[13], MINT[15] and Intact[16] are aggregated as the source of gene-gene interactions
477 (Supplementary Materials). The related datasets were downloaded from their corresponding
478 websites and the IDs are matched using standard gene models (gencode v17). To quantify the
479 distance between genes, only data for the gene pairs within the same chromosomes were used.
480 Calculation of cLD and LD follows the same procedure as described in 3D-interaction case
481 (Supplementary Materials).
482
483 **Protein docking analysis**
484 HDOCKlite-v1.1[38,39] [PMID: 32269383] was employed for conducting the protein-protein docking
485 analysis between the protein pairs (Supplementary Materials). The protein's crystal structure was
486 obtained from the Protein Data Bank[40] (https://www.rcsb.org/) and validated[41] for the study
487 (Supplementary Materials). The output file of the docked complex was visualized with PyMOL
488 2.5.1[42], and the 2D plot of the protein-protein binding region was analyzed and deduced using
489 LigPlot+ v.2.2[43] (Supplementary Materials). These analyses are illustrated in Fig. 5.
490
491
492 **ΔcLD genes, their functional annotation, and pathway enrichment**
493
494 *Calculation of cLD-differential gene pairs*. We calculated cLD using the Autism Spectrum Disorder
495 (ASD)[17] [phs000298.v4.p3] whole exome sequencing dataset. We first calculate cLD values for
496 each gene pair for case and control groups separately. Then, we calculate the absolute differences

between the cLD values in case and control groups for each gene pair. These absolute differences were sorted from largest to smallest. The top ranked genes pairs are collected and called cLD-differential gene pairs (Supplementary Materials).

Functional annotation and pathway enrichment. We select the top 200, 500, 1,000, 1,500 and 2,000 cLD-differential gene pairs with the largest cLD differences, and include all genes within these top gene pairs as candidate sets for the downstream functional annotations. We make use two different databases, Simons Foundation Autism Research Initiative (SFARI)[19] and DisGeNet[18] (https://www.disgenet.org/) to be the gold-standard as they are frequently used in the field of ASD studies and general disease gene queries, respectively. We use the hypergeometric distribution probability to assess the p-value of the significance of enrichment of the cLD-differential genes against the gold-standard genes (Supplementary Materials). These results are illustrated in Fig. 6a,b. Additionally, we include all genes within the top 2,000 cLD-differential gene pairs and conduct GO enrichment[20] and KEGG pathway analysis[22] for these genes . These results are illustrated in Fig. 6c,d.