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 Zhang,4, Xingyi Guo,5,6, Alexander Platt,7, Jingjing Wu,1, Qingrun Zhang,1,2,3#

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, Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA.

6, Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA.

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

*: joint first authors

#: Correspondence should be address to Q.Z.: qingrun.zhang@ucalgary.ca

Keywords: Cumulative Linkage Disequilibrium, Gene-gene Interaction, Rare Genetic Variants, 3D Chromatin Interaction, Statistical Instability

ABSTRACT

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, which reflects more biological interactions between variants, in addition to lack of recombination. We derived the theoretical variance of cLD using delta methods to demonstrate its higher stability than LD for rare variants. This property is also verified by bootstrapped simulations using real data. 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.

INTRODUCTION

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 (Slatkin 2008). 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) (Weissbrod et al. 2020). To this end, LD is critically important in analyzing the fine resolution of genotype-phenotype association mapping (Flint-Garcia et al. 2003) and forming polygenic risk scores (Amariuta et al. 2020). 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 favored by selection (Gregersen et al. 2006). 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 (**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 (Auton et al. 2015).

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 (Li and Leal 2008) 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**).

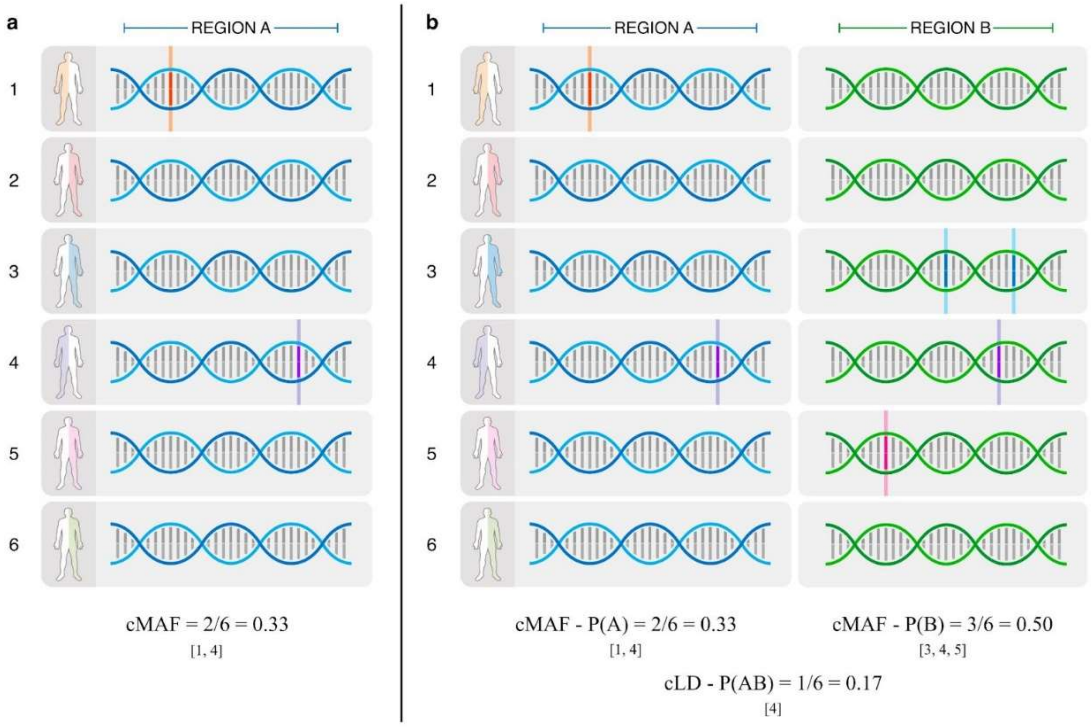
Building on the idea of cMAF and the essence of LD, we developed a statistic, cumulative Linkage Disequilibrium (cLD) to capture the aggregated correlation between two sets of rare variants (**Methods; Fig. 1b**).

We thoroughly tested the property of cLD. First, using both theoretical closed-form derivation and bootstrapped simulations (**Methods**), it is verified that cLD enjoys way lower variance than the standard LD when applied to rare variants, evidencing cLD's higher stability (**Fig. 2**). We then applied cLD to four scenarios in genetic analysis (**Methods**), discovering additional knowledge that have not been reported (or attempted but negatively reported) using standard LD (**Figs. 3 – 6**).

RESULTS

The intuitive idea of defining cLD. In the similar vein of definition of cMAF, we define cLD below. 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**). Mathematical details are spelt out in **Methods** and **Supplementary Materials 1.1 & 1.2**).

Figure 1. Illustration of the idea of a) cMAF and b) cLD. An example to show the calculation of cLD, inspired by cMAF. **a)** Out of six haplotypes, there are two [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 haplotypes [3, 4, 5] who have mutations in region B and the cMAF P(B) for region B is 3/6 = 0.50. If one considers 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.

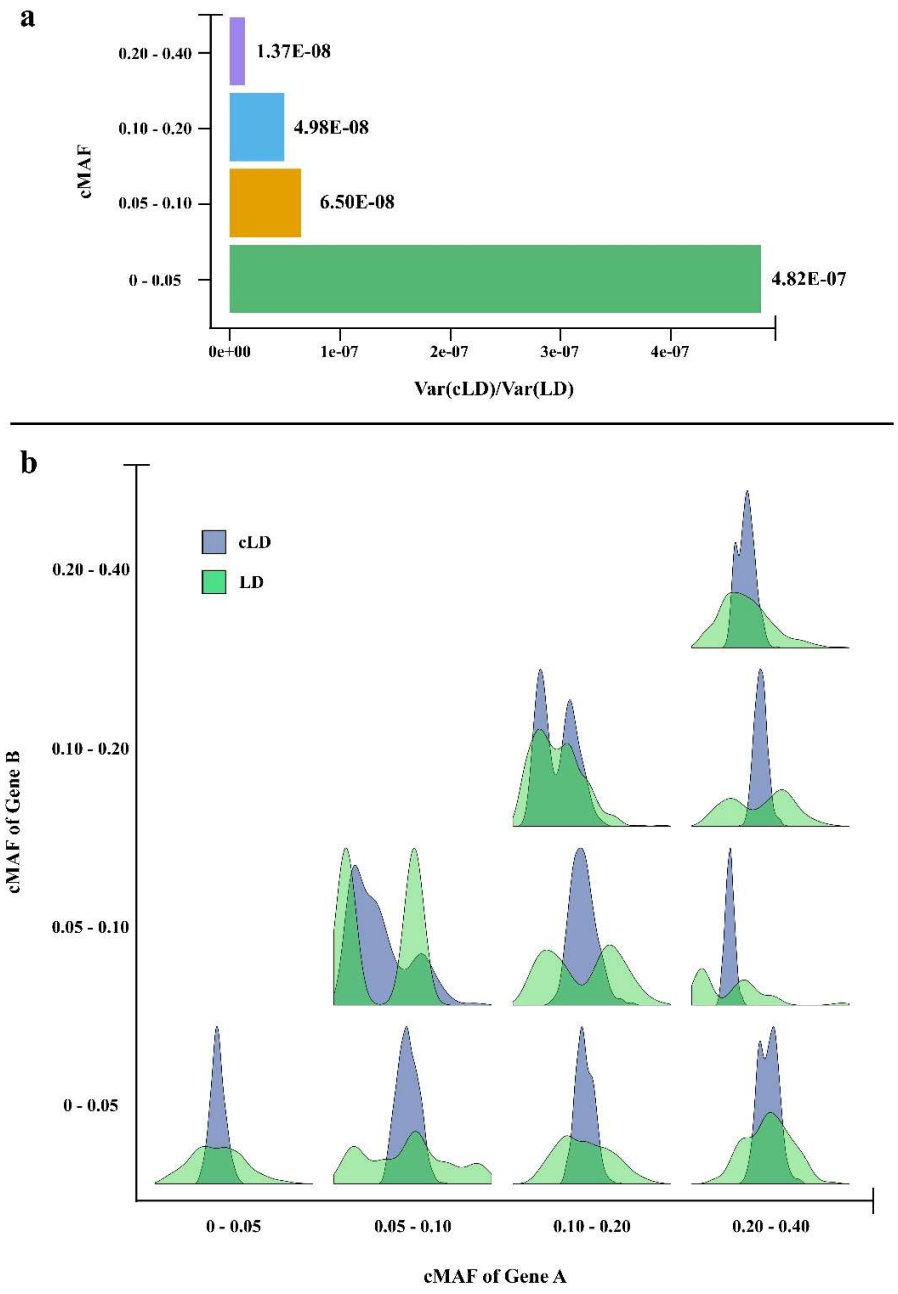


High stability of cLD in contrast to standard LD. Both LD and cLD could be used to capture the correlation between two sets of rare variants. However, these two measures differ in the aspect of stability. Intuitively, as cMAF is always higher than MAF, cLD's variance (reflecting its instability) should 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 (Lehmann Springer) (**Methods; Supplementary Materials 2.1 & 2.2**). by plugging in the allele frequencies calculated using the 1000

Genomes Project data (Auton et al. 2015) (**Supplementary Materials 2.3**), we observed that the variance of cLD is at least six 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. S2.1a & S2.2a**). Additionally, following the conventional statistical procedure of bootstrapping to empirically estimate stability, we sub-sampled half of each population sample 1,000 times to form bootstrapped distributions for both cLD and LD (**Methods; Supplementary Materials 2.4**). The subsampling showed that cLD exhibits much slimmer bootstrapped distributions than LD across all cMAF bins and all three ethnic groups (**Fig. 2b, Supplementary Figs. S2.1b & S2.2b**), further confirming the greater stability of cLD compared to traditional measures of LD.

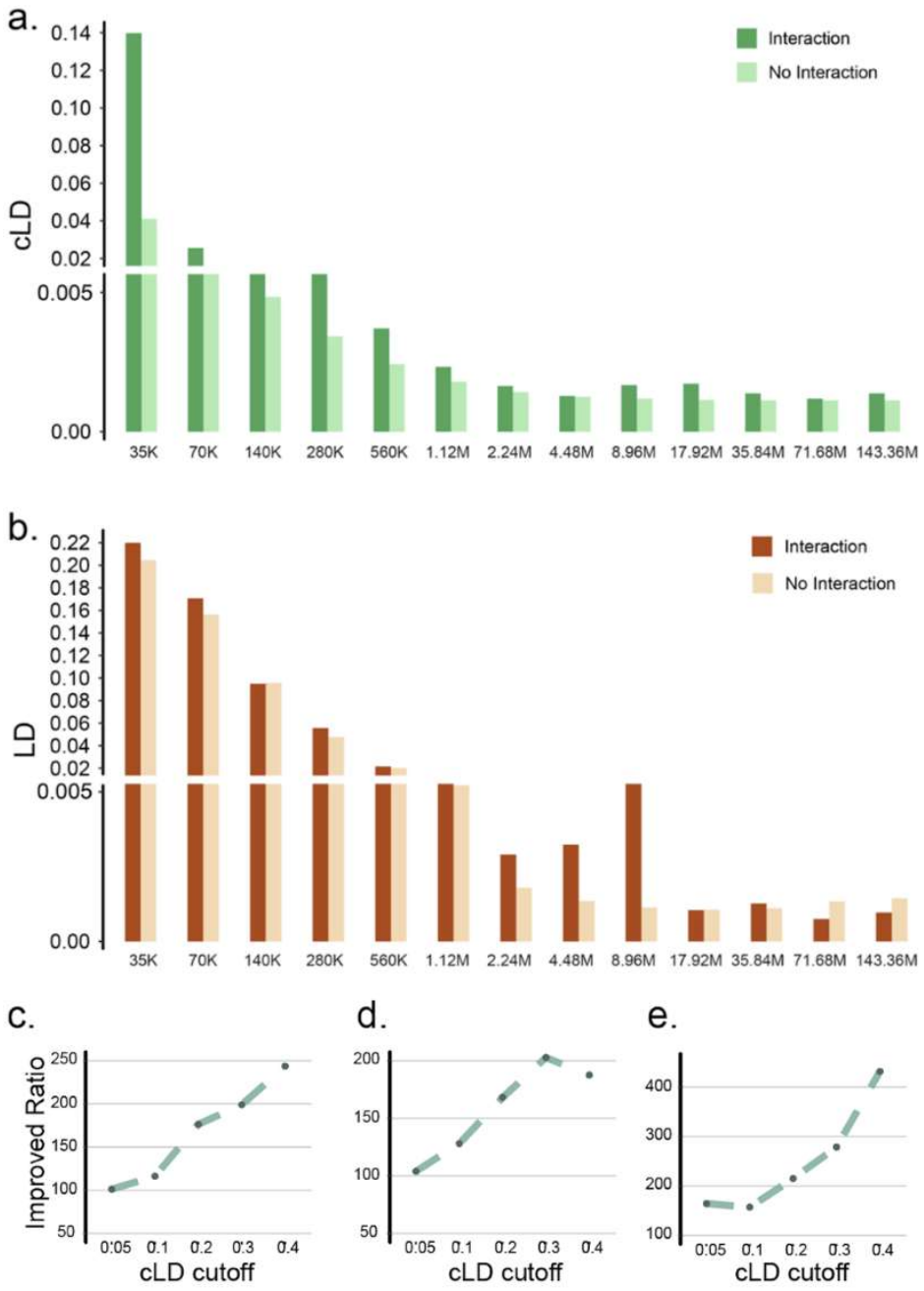
cLD reveals linkage disequilibrium between 3D contact regions where standard LD fails. A distinct advantage of cLD over LD is the ability to reveal linkage disequilibrium between 3D contact regions. By aggregating information from multiple independent mutations, cLD is sensitive to subtle interactions poorly reflected by LD (which can only account for two at a time). As such, cLD captures more biological interactions in addition to traditional LD that focuses more on the lack of recombination. Interactions within the 3D structure of genomes is one place where this difference allows for insight from cLD where LD-based methods fail. The availability of high-throughput experimental technologies that can assess chromatin conformation such as Hi-C (Rajarajan et al. 2018; Akbarian et al. 2015) allows researchers to analyze genetic regions that are in close contact in 3D spatial structure. There was a widely disseminated expectation that the 3D genomic interaction in the form of chromatin

Figure 2. Stability of cLD and LD revealed by closed-form variance calculation and bootstrapped distributions. a) The gene pairs were split into four different bins based on the cMAF values, i.e., <0.05, 0.05 - 0.10, 0.10 - 0.20, and 0.20 - 0.40 (y-axis). The x-axis is the ratio between the variances of cLD and LD, i.e., Var(cLD)/Var(LD). **b)** Probability density distribution of cLD and LD from bootstrapped samples. Results from the European population are shown. See **Supplementary Figs. S2.1 & S2.2** for other populations.



contact may leave a footprint in the form of genetic LD (Joiret et al. 2019). Motivated by such expectation, Whalen and Pollard calculated the standard LD based on common variants (MAF>0.05) in 1000 Genomes Project data (Auton et al. 2015) and reported negative results stating that genetic LD map is not overlapping with the 3D contact map (Whalen and Pollard 2019). However, by reanalyzing the 1000 Genomes sequencing data and Hi-C data (Akbarian et al. 2015; Rajarajan et al. 2018) in the developing brain using cLD on rare variants (**Methods; Supplementary Materials 3.1 & 3.2**), we revealed that the 3D chromatin interactions did leave genetic footprints in the form of higher cLD in pairs of genes that are in the adjacent Hi-C regions (**Fig 3a; Supplementary Fig. S3.1**). To assess the statistical significance of the enrichment of cLD in 3D contact regions, we conducted Mantel-Haenszel and Fisher exact tests (**Supplementary Materials 3.4**), both of which are highly significant (P-value < 1.0E-50; **Supplementary Tables S3.2 & S3.6, Supplementary Materials 3.4.1**). As Whalen & Pollard's work (Whalen and Pollard 2019) is not at the resolution of genes, we re-calculated standard LD using common variants based on gene pairs (**Supplementary Materials 3.2**), which shows a subtle effect (**Fig. 3b, Supplementary Fig. S3.2**) but still not statistically significant with Mantel-Haenszel and Fisher exact tests (P-value =0.999**; Supplementary Tables S3.3 & S3.4; Supplementary Materials 3.4.1**). Additionally, we checked the ratio between the number of pairs of genes within the 3D contact regions and the number of pairs outside the 3D contact regions as a function of their cLD cut-off. More specifically, we prespecified a cLD value cutoff and only counted the gene pairs with cLD value higher than this cutoff; then we separated the number of genes within or outside 3D contact regions and calculated their ratios (**Supplementary Materials 3.5**). Indeed, we found that the ratios are significantly larger than 1.0 and increase as the cLD cutoffs increase (**Fig 3c,d,e, Supplementary Table S3.7)**. Taking together, 3D

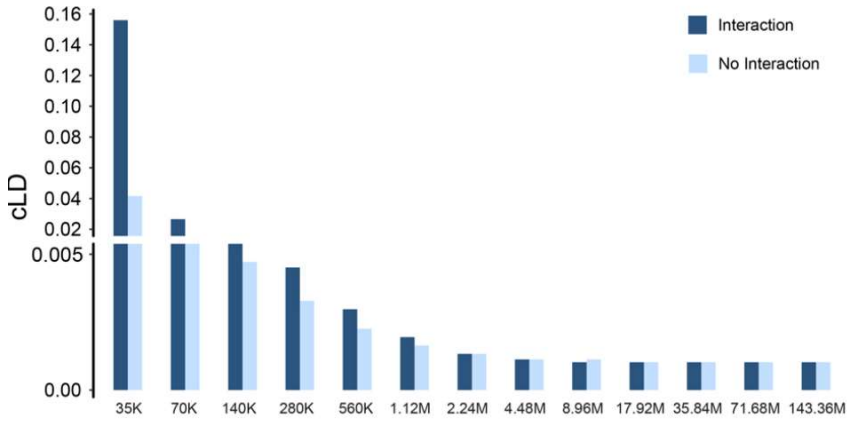
Figure 3. Enrichment of cLD among pairs of genes in chromatin contact regions. **a)** The comparisons of cLD values between the 3D chromatin interaction regions and non-interaction regions among 13 different distance groups in the European population. (Other populations are shown in **Supplementary Fig. S3.1**) The confidence intervals for these bars are presented in **Supplementary Table S3.1**. **b)** The same comparisons using standard LD in the European population. (Other populations are shown in **Supplementary Fig. S3.2**) **c-e)** 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 above which the gene pairs are counted. **c)** European population. **d)** African population. **e)** East Asian population.



interactions clearly overlap with genetic interactions; and cLD is instrumental in observing this whereas standard LD fails.

cLD is enriched in known interacting genes. To demonstrate that gene-gene interactions leave footprints in rare genetic mutations regardless of their physical positions we computed the distribution of cLD among interacting pairs genes reported in Reactome (Fabregat et al. 2018) and BioGRID (Stark et al. 2006), MINT (Orchard 2012) and IntAct (Orchard et al. 2014) (**Methods; Supplementary Materials 3.3**). We compared this distribution against a null distribution formed by all pairs of genes. Indeed, the comparisons led to the expected result: for gene pairs separated by any physical distance within 2MB, cLD is elevated in interacting gene pairs (**Fig. 4; Supplementary Fig. S3.3**). Again, the Mantel-Haenszel and Fisher exact tests confirm that the differences are significant (P-value < 1.0E-20; **Supplementary Table S3.5; Supplementary Materials 3.4.2**).

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, as depicted in **Supplementary Fig. S3.3**)

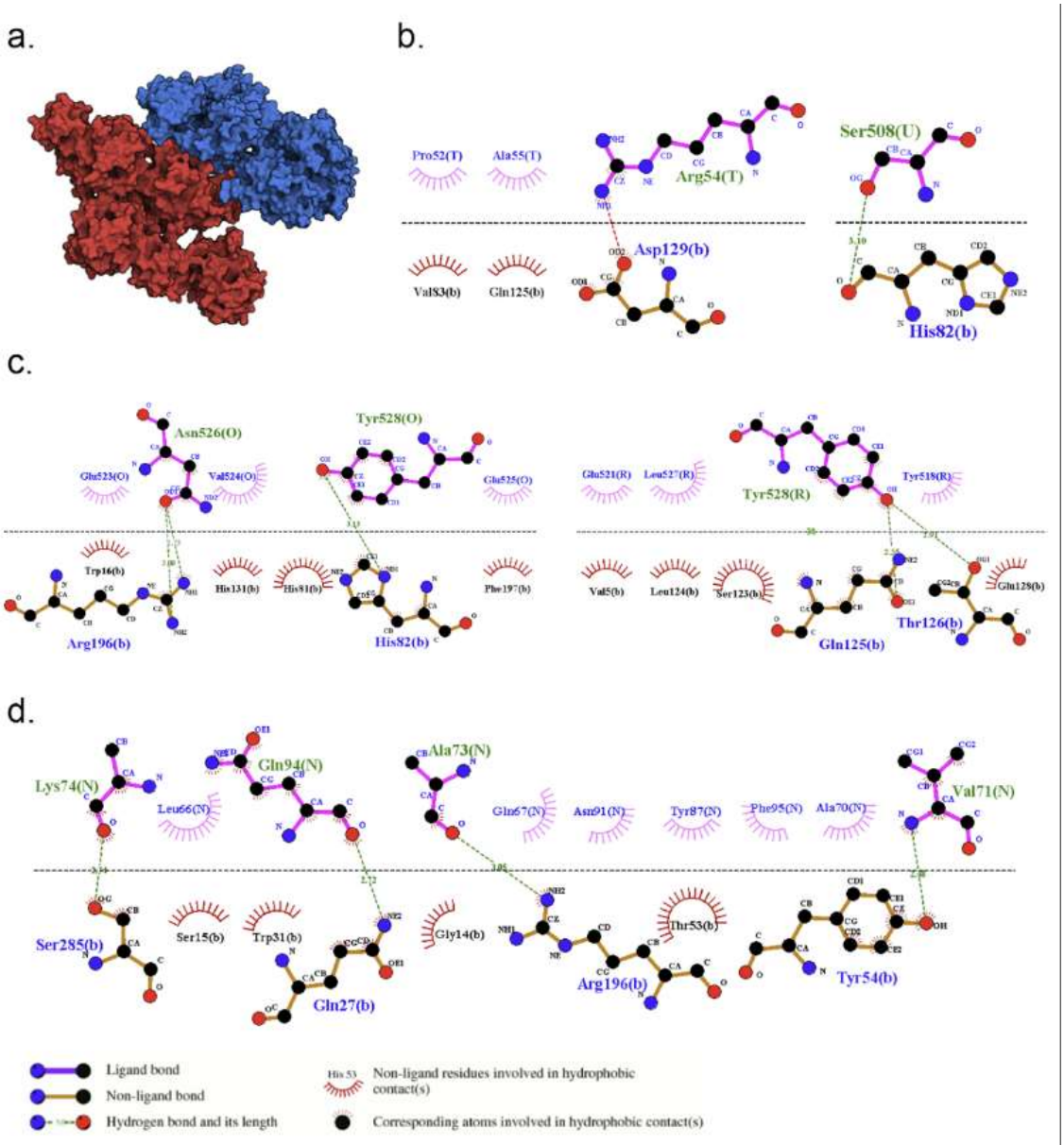


193 ***cLD identified novel pairs of likely interacting proteins***. To examine the novel gene
194 pairs with higher cLD values have the receptor-ligand interactions of their translated
195 proteins, we performed protein-docking analysis to obtain the evidence. Looking at all
196 pairs of genes, we observed several pairs without prior evidence of interaction with
197 extraordinarily high cLD, such as between genes *MEMO1* and *DPY30* (encoding
198 proteins 3BCZ and 4RIQ, respectively) with a cLD of 0.86. We conducted protein
199 docking analysis for the genes of large cLD values (top 0.01% among all gene pairs)
200 with cMAF > 0.05 and existing IDs in PDB, however, not reported in any interaction
201 databases (**Methods; Supplementary Materials 4.1; Supplementary Table S4.1**).
202 These criteria lead to 19 pairs of genes for protein-docking. We found multiple lines of
203 evidence of the interaction at protein level for five pairs (**Supplementary Table S4.2**) in
204 terms of both binding affinity and interacting residues (**Fig. 5a-d; Supplementary Figs.**
205 **S4.1 - S4.4**).

206

207 ***Differences in cLD distinguish cases/controls in Autism exome data***. In the context
208 of case/control association studies, cLD can be used to identify pairs of genes whose
209 interactions may be responsible for human diseases. Using data from the *Autism*
210 *Spectrum Disorders* (ASD) whole exome sequencing dataset (Satterstrom et al. 2020),
211 we calculated cLD values for all pairs of genes, separately conducted for the populations
212 of cases and controls (**Methods; Supplementary Materials 5.1 & 5.2**). The difference
213 in cLD for a pair of genes conditional on case/control status, defined as ***ΔcLD***, is
214 indicative of an interaction that is non-random associating with disease status. We
215 collected gene pairs with high ΔcLD and checked their annotation and enrichment in
216 existing databases. Using a hypergeometric test, we analyzed the enrichment among

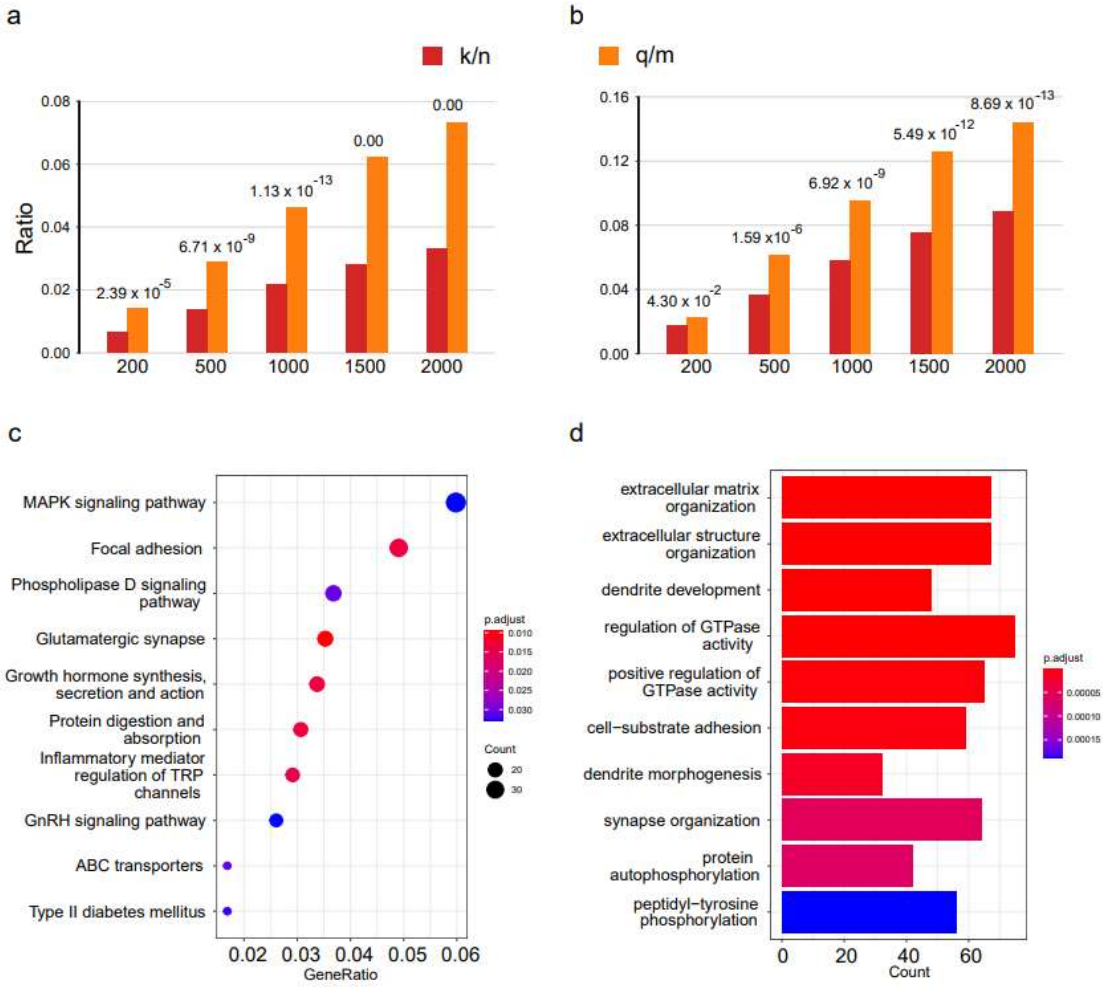
Figure 5: Protein docking interaction between 3BCZ and 4RIQ revealed by cLD (=0.86) with a binding affinity of -341.21 kJ/mol. **a)** Structure of 3BCZ (red) and 4RIO (blue) protein-protein complex. **b-d)** 2D representation of closest interacting residues around the protein-protein interaction interfaces, 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.



high-ΔcLD genes for ASD genes reported by DisGeNet (Piñero et al. 2017), an established general database for diseases and SFARI (Abrahams et al. 2013), a gold-

standard database focusing on ASD (**Supplementary Materials 5.3**). The genes included in the pairs with high ΔcLD scores are highly enriched in both the Autism related genes in DisGeNet (**Fig. 6a**) and SFARI (**Fig. 6b**). Gene Ontology (Ashburner et al. 2000) and pathways (KEGG) (Kanehisa and Goto 2000; Kanehisa et al. 2009) enrichment analysis for the high ΔcLD genes (**Methods; Supplementary Table S5.2; Supplementary Materials 5.4**) also showed sensible biological functions and pathways (**Fig. 6c,d**) that are well supported by the literature (**Supplementary Materials 5.4**) (Ashburner et al. 2000; Kanehisa and Goto 2000; Kanehisa et al. 2009; Yu et al. 2012; Rojas 2014; Hannelius et al. 2005; Richler et al. 2006; O ' Roak et al. 2012; Fung and Hardan 2015; Sato et al. 2012; Berkel et al. 2010; Durand et al. 2007; Wei et al. 2021; Ye et al. 2011; Betancur et al. 2009; Lin et al. 2016). 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 S5.1**). Moreover, among the rest of six genes, we also identified indirect links of two, *RAB11A* and *IDUA* with ASD (**Supplementary Materials 5.3**).

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) as well as the number of selected genes dividing the total number of all known 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 using different cutoffs, [200, 500, …, 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.



DISCUSSION

LD is a critical concept applicable to many types of genetic analyses. In this work, we have defined cLD, a new statistic addressing the association between genetic regions using rare genetic variants. In contrast to the previous attempts to utilize LD between multiple variants focusing on dominant haplotypes (Zan et al. 2018) or joint distributions (Turkmen and Lin 2017), cLD emphasizes biological interactions. Additionally, previously researchers have proposed composite linkage (Hamilton and Cole 2004; Zaykin 2004),

which addresses the property of variances and its normalization, however, does not incorporate rare variants.

By both closed-form derivations and statistical simulations, we proved the stability of cLD in contrast to the high instability of standard LD (when applied to rare variants). The stability and the focus on biological interaction allows cLD to capture additional information from the distributions of many variants segregating in a population at low frequencies within particular regions of a genome. Indeed, by applying cLD to real data, we observed interesting overlapping pattern of 3D interactions and genetic interactions that have been negatively reported by using standard LD. We also successfully analyzed protein docking and association mapping, providing two broadly impactable use-cases of cLD. With its demonstrated power in identifying gene and protein interactions, cLD might offer an essential tool to analyze biological interactions and their evolution using rare genetic variants.

METHODS

Definition of LD and cLD:

The definition of LD between two bi-allelic loci relies on the calculation 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:

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

An alternative definition of LD is D', which has a different way of normalization. In this paper, we used r^2 as the representative. Because LD involves P_A and P_B in the denominator, it is highly instable when P_A or P_B are 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. In the similar vein of definition of cMAF, the idea of cLD is illustrated in **Fig. 1b**. 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 $S_{1i}, S_{1i}, \dots, S_{mi}$ to denote the allele of the *s*-th SNP $(s = 1, 2, \dots, m)$ in the *i*-th individual $(i = 1, 2, \dots, n)$. Similarly, for gene B, we use $\{K_{1i}, K_{2i}, \dots, K_{ri}\}$ to denote the allele of the *k*-th SNP $(k = 1, 2, \dots, r)$ in the *i*-th individual $(i = 1, 2, \dots, n)$. 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:

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

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

Where $I(.)$ is the indicator function. P_{AB} is then defined as the proportion of individual haplotypes with a minor allele in both regions:

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

Following the convention of LD, we define the r^2 version of cLD:

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

The more rigorous mathematical descriptions and the definition of D' version is provided in **Supplementary Materials 1.1 & 1.2**.

Derivation of theoretical variance of cLD in contrast to LD

To obtain the theoretical variance of cLD and LD, we derived their asymptotic distributions. The details are in **Supplementary Materials 2.1 & 2.2**. The gist of our approach is summarized in the following three steps:

First, we rewrote the formula of cLD and LD in terms of counts to use multinomial random variables. In the definition, we used 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}) $(i = 1, 2, \dots, n)$ can take possible values (1,1), (0,1), (1,0) and (0,0). Using O_1 to O_4 to denote the count of the 4 possible pairs in two variants, the distribution of $\boldsymbol{O} = (O_1, O_2, O_3, O_4)$ is $\boldsymbol{O} \sim multinom(n; \boldsymbol{p})$ with $\boldsymbol{p} = (p_1, p_2, p_3, p_4)$ represents the population probability. The LD between the *u*-th and *v*-th variants can be re-written as:

$$LD_{(u,v)} = \frac{(O_1 O_4 - O_2 O_3)^2}{(O_1+O_2)(O_1+O_3)(O_2+O_4)(O_3+O_4)}.$$

Similarly, we followed the same strategy of using multinomial random variables to describe cLD as below:

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}) $(i = 1, 2, \dots, n)$ 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 $\boldsymbol{M} = (M_1, M_2, M_3, M_4)$ is $\boldsymbol{M} \sim multinom(n; q)$ with $\boldsymbol{q} = (q_1, q_2, q_3, q_4)$ 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 used the central limit theorem (CLT) to derive the asymptotic multivariate normal distribution. In the LD case, with the population mean $\boldsymbol{p} = (p_1, p_2, p_3, p_4)$, we can write the covariance matrix as

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

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

In the cLD case, with the population mean $\boldsymbol{q} = (q_1, q_2, q_3, q_4)$, 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 (Lehmann Springer) we have $\sqrt{n}\left(\frac{\boldsymbol{M}}{n} - \boldsymbol{q}\right) \xrightarrow{L} N(\boldsymbol{0}, \mathbf{Q})$.

Third, as the cLD and LD are functions of random variables, we applied the multivariate Delta method (Lehmann Springer) to derive the distribution of cLD and LD. In the LD case, suppose the Jacobian matrix of $LD(\boldsymbol{O}/n)$ is $\boldsymbol{J_{LD}} = \left[\frac{\partial LD(\boldsymbol{O}/n)}{\partial \boldsymbol{O}}\right]|_{\boldsymbol{O}=n\boldsymbol{p}}$. Then the asymptotic distribution of $LD(\boldsymbol{O}/n)$ is $LD(\boldsymbol{O}/n) - LD(\boldsymbol{p}) \sim AN\left(0, n\boldsymbol{J_{LD}}\boldsymbol{\Sigma}\boldsymbol{J_{LD}^{\top}}\right)$, where '*AN*' stands for asymptotic normal.

In the cLD case, suppose the Jacobian matrix of $cLD(\boldsymbol{M}/n)$ is $\boldsymbol{J_{cLD}} = \left[\frac{\partial cLD(\boldsymbol{M}/n)}{\partial \boldsymbol{M}}\right]|_{\boldsymbol{M}=n\boldsymbol{q}}$. Then the asymptotic distribution of $cLD(\boldsymbol{M}/n)$ is $cLD(\boldsymbol{M}/n) - cLD(\boldsymbol{q}) \sim AN\left(0, n\boldsymbol{J_{cLD}}\boldsymbol{Q}\boldsymbol{J_{cLD}^{\top}}\right)$.

Genotype data used for the calculations

The 1000 Genomes Variant Call Data were used to validate the properties of cLD. In particular, the phased (i.e., haploid instead of diploid) variant call data of the Phase 3 of the 1000 Genomes dataset was obtained through The European Bioinformatics Institute's dedicated FTP server (Fairley et al. 2020).

Assessing the instability of LD and cLD using bootstrapped distributions

To use bootstrapped samples to quantify instability, we randomly sampled half of the haplotypes in three main 1000 Genomes Project populations (EUR, AFR, or EAS), and calculated the average cLD and average LD over the gene pairs within cMAF bins and repeated this procedure 1,000 times. Based on these bootstrapped cLD and LD values we formed bootstrapped distributions for cLD and LD respectively (with appropriate re-scaling described in **Supplementary Materials 2.3**). More specifically, we randomly

sampled 1,000 genes and assessed their pairwise LD and cLD in stratified cMAF bins (**Supplementary Materials 2.4**) using half of the haplotypes in the given population (AFR, EAS or EUR). These randomly drawn subsamples (each with half of the individuals in the original population) form bootstrapped samples. We define the LD of a gene pair as the average value of LD over all rare SNV pairs within that gene pair. In each iteration, we calculate the average cLD over the gene pairs in each bin (**Supplementary Materials 2.4**).

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

To revisit a previously negatively reported relationship between 3D interaction regions and genetic linkage disequilibrium (Whalen and Pollard 2019) , we calculated both cLD and LD in a Hi-C assessment in the developing brain (Li et al.), which has 27,982 brain-specific paired 3D-interacting regions, measured from neurons derived from human induced pluripotent stem cells (hiPSCs).

Again, the 1000 Genomes Project data were used. We first calculated the distance between the genes in each pair and separate the gene pairs into 13 distance groups (**Supplementary Materials 3.1**). After stratifying all gene pairs into distance groups, within each distance group, we calculated cLD between all gene pairs and further split them into two categories: the ones that are located in 3D interaction regions (assessed by Hi-C experiments) and the ones that are located in non-3D interaction regions. The gene pairs with exactly one gene in an interaction region were discarded. Finally, the average cLD values over gene pairs within interaction and non-interaction regions were used to conduct the comparison, quantified by two two-sample tests, namely Mantel-Haenszel and Fisher exact tests (**Supplementary Materials 3.4**).

392 The procedure of calculating standard LD mirrors the one used above for cLD using the
393 same distance groups and 3D-interaction vs non-interaction categories. As standard LD
394 is defined by individual variants (not by genes), the following averaging steps were
395 taken. For each gene pair in the 3D interaction regions, we randomly chose 2,000 rare
396 variant pairs from it to calculate their LD values. For each selected rare variant pair, we
397 calculated its distance and then, among the gene pairs without 3D interactions, we
398 randomly selected another rare variant pair with the same or closest possible distance
399 (**Supplementary Materials 3.2**). As a result, we achieved 2,000 randomly selected
400 variant pairs from gene pairs without interaction that were matched up with the 2,000
401 variant pairs from gene pairs with interaction. The average values of the 2,000 variant-
402 pairs were deemed as the LD between the gene pair.

403

404 ***Calculation of cLD and LD for gene pairs in gene-gene interaction databases***

405 Four frequently used interaction databases, Biogrid (Stark et al. 2006), Reactome
406 (Fabregat et al. 2018), MINT (Orchard 2012) and Intact (Orchard et al. 2014) were
407 aggregated as the source of gene-gene interactions (**Supplementary Materials 3.3**).
408 The related datasets were downloaded from their corresponding websites and the IDs
409 were matched using standard gene models (gencode v17). To quantify the distance
410 between genes, only data for the gene pairs within the same chromosomes were used.
411 Calculation of cLD and LD follows the same procedure as described for the 3D-
412 interaction analysis, and the two-sample tests (Mantel-Haenszel and Fisher exact tests)
413 were used to quantify the significant levels (**Supplementary Materials 3.4**).

414

Protein docking analysis

We used protein docking to validate the novel gene-gene interactions predicted by unexpected high cLD values. HDOCKlite-v1.1 (Yan et al. 2020, 2017) was employed for conducting the protein-protein docking analysis between the cLD prioritized protein pairs (**Supplementary Materials 4**). The protein's crystal structure was obtained from the Protein Data Bank (Berman et al. 2000) and further validated (Perera et al. 2021) (**Supplementary Materials 4.1**). The output file of the docked complex was visualized by PyMOL 2.5.1 (Delano), and the 2D plot of the protein-protein binding region was analyzed and deduced using LigPlot+ v.2.2 (Laskowski and Swindells 2011) (**Supplementary Materials 4.2**).

ΔcLD genes, their functional annotation, and pathway enrichment

Calculation of cLD-differential gene pairs. To explore the use of cLD in distinguishing cases and controls in a typical association study, we calculated cLD using the whole exome sequencing data to study Autism Spectrum Disorder (ASD) (Satterstrom et al. 2020) [dbGaP ID: phs000298.v4.p3]. We first calculated cLD values for each gene pair for cases and controls groups separately. Then, we calculated the absolute differences between the cLD values in case and control groups for each gene pair, which was called ΔcLD. These absolute differences were sorted from largest to smallest. The top ranked genes pairs were collected and called cLD-differential gene pairs, or ΔcLD genes (**Supplementary Materials 5.2 & 5.3**).

Functional annotation and pathway enrichment. Based on their ΔcLD values, we selected the top 200, 500, 1,000, 1,500 and 2,000 cLD-differential gene pairs (i.e., ΔcLD genes) and used the genes sets for the downstream functional annotations. We utilized two different databases, Simons Foundation Autism Research Initiative (SFARI) (Abrahams et al. 2013) and DisGeNet (Piñero et al. 2017) as the gold-standard because they are frequently used in the field of ASD studies and general disease gene queries, respectively. We used the hypergeometric distribution probability to assess the p-value of the significance of enrichment of the cLD-differential genes against the background of gold-standard genes (**Supplementary Materials 5.4**). Additionally, using the top 2,000 cLD-differential gene pairs, we conducted GO enrichment (Ashburner et al. 2000) and KEGG pathway analysis (Kanehisa et al. 2009).

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

Data and Code Availability:

The codes calculating cLD and conducting all the analyses in this work are publicly available at our GitHub: https://github.com/QingrunZhangLab/cLD

The 1000 Genome Variant Call Data used in this study could be downloaded from http://ftp.1000genomes.ebi.ac.uk. The complete variant call dataset was found using the webpage (Announcements | 1000 Genomes (internationalgenome.org)) (This is a sub-

page maintained by the 1000 Genome webpage) and downloaded from (Index of /vol1/ftp/release/20130502/ (ebi.ac.uk)).

The 3D Hi-C dataset is available in the Synapse database (https://www.synapse.org/) with Synapse ID: syn12979149.

The Protein Data Bank: https://www.rcsb.org/.

The DisGeNet Database: https://www.disgenet.org/

The SFARI Database: https://www.sfari.org/resource/sfari-gene/

The HDOCK protein docking software: http://hdock.phys.hust.edu.cn/

Competing Interest Statement: The authors declare no competing interests.

Acknowledgments. 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.W is supported by Alberta Graduate Excellence Scholarship; 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).

REFERENCES

Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, Menashe I, Wadkins T, Banerjee-Basu S, Packer A. 2013. SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism* **4**.

Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, Jaffe AE, Pinto D, Dracheva S, Geschwind DH, et al. 2015. The PsychENCODE project. *Nat Neurosci* **18**: 1707–1712.

Amariuta T, Ishigaki K, Sugishita H, Ohta T, Koido M, Dey KK, Matsuda K, Murakami Y, Price AL, Kawakami E, et al. 2020. Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nat Genet* **52**: 1346–1354.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. *Gene Ontology: tool for the unification of biology The Gene Ontology Consortium**. http://www.flybase.bio.indiana.edu.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Berkel S, Marshall CR, Weiss B, Howe J, Roeth R, Moog U, Endris V, Roberts W, Szatmari P, Pinto D, et al. 2010. Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. *Nat Genet* **42**: 489–491.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. *The Protein Data Bank*. http://www.rcsb.org/pdb/status.html.

Betancur C, Sakurai T, Buxbaum JD. 2009. The emerging role of synaptic cell-adhesion pathways in the pathogenesis of autism spectrum disorders. *Trends Neurosci* **32**: 402–412.

Delano WL. *PyMOL: An Open-Source Molecular Graphics Tool*.

Durand CM, Betancur C, Boeckers TM, Bockmann J, Chaste P, Fauchereau F, Nygren G, Rastam M, Gillberg IC, Anckarsäter H, et al. 2007. Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nat Genet* **39**: 25–27.

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, et al. 2018. The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**: D649–D655.

Fairley S, Lowy-Gallego E, Perry E, Flicek P. 2020. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res* **48**: D941–D947.

514 Flint-Garcia SA, Thornsberry JM, Edward IV SB. 2003. Structure of Linkage Disequilibrium in
515 Plants. *Annu Rev Plant Biol* **54**: 357–374.

516 Fung LK, Hardan AY. 2015. Developing Medications Targeting Glutamatergic Dysfunction in
517 Autism: Progress to Date. *CNS Drugs* **29**: 453–463.

518 Gregersen JW, Kranc KR, Ke X, Svendsen P, Madsen LS, Thomsen AR, Cardon LR, Bell JI, Fugger L.
519 2006. Functional epistasis on a common MHC haplotype associated with multiple sclerosis.
520 *Nature* **443**: 574–577.

521 Hamilton DC, Cole DEC. 2004. Standardizing a composite measure of linkage disequilibrium. *Ann*
522 *Hum Genet* **68**: 234–239.

523 Hannelius U, Lindgren CM, Melén E, Malmberg A, von Dobeln U, Kere J. 2005. Phenylketonuria
524 screening registry as a resource for population genetic studies. *J Med Genet* **42**.

525 Joiret M, Mahachie John JM, Gusareva ES, van Steen K. 2019. Confounding of linkage
526 disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData*
527 *Min* **12**.

528 Kanehisa M, Goto S. 2000. *KEGG: Kyoto Encyclopedia of Genes and Genomes*.
529 http://www.genome.ad.jp/kegg/.

530 Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. 2009. KEGG for representation and
531 analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**.

532 Laskowski RA, Swindells MB. 2011. LigPlot+: Multiple ligand-protein interaction diagrams for
533 drug discovery. *J Chem Inf Model* **51**: 2778–2786.

534 Lehmann Springer EL. *Elements of Large-Sample Theory*.

535 Li B, Leal SM. 2008. Methods for Detecting Associations with Rare Variants for Common
536 Diseases: Application to Analysis of Sequence Data. *Am J Hum Genet* **83**: 311–321.

537 Li Q, Cao C, Perera D, He J, Chen X, Azeem F, Howe A, Au B, Yan J, Long Q. Statistical model
538 integrating interactions into genotype-phenotype association mapping: an application to
539 reveal 3D-genetic basis underlying Autism. https://doi.org/10.1101/2020.07.27.222364.

540 Lin YC, Frei JA, Kilander MBC, Shen W, Blatt GJ. 2016. A subset of autism-associated genes
541 regulate the structural stability of neurons. *Front Cell Neurosci* **10**.

542 O ' Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, et al.
543 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo
544 mutations. *Nature* **485**: 246–250.

545 Orchard S. 2012. Molecular interaction databases. *Proteomics* **12**: 1656–1662.

546 Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G,
547 Chen C, Del-Toro N, et al. 2014. The MIntAct project - IntAct as a common curation
548 platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**.

549 Perera DDBD, Perera KML, Peiris DC. 2021. A novel in silico benchmarked pipeline capable of
550 complete protein analysis: A possible tool for potential drug discovery. *Biology (Basel)* **10**.

551 Piñero J, Bravo Á, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-
552 García J, Sanz F, Furlong LI. 2017. DisGeNET: A comprehensive platform integrating
553 information on human disease-associated genes and variants. *Nucleic Acids Res* **45**: D833–
554 D839.

555 Rajarajan P, Borrman T, Liao W, Schrode N, Flaherty E, Casiño C, Powell S, Yashaswini C, LaMarca
556 EA, Kassim B, et al. 2018. Neuron-specific signatures in the chromosomal connectome
557 associated with schizophrenia risk. *Science (1979)* **362**.

558 Richler E, Reichert JG, Buxbaum JD, Mcinnes LA. 2006. *Autism and ultraconserved non-coding
559 sequence on chromosome 7q*. Lippincott Williams & Wilkins
560 http://www.cse.ucsc.edu/Bjill/ultra.html.

561 Rojas DC. 2014. The role of glutamate and its receptors in autism and the use of glutamate
562 receptor antagonists in treatment. *J Neural Transm* **121**: 891–905.

563 Sato D, Lionel AC, Leblond CS, Prasad A, Pinto D, Walker S, O'Connor I, Russell C, Drmic IE,
564 Hamdan FF, et al. 2012. SHANK1 deletions in males with autism spectrum disorder. *Am J
565 Hum Genet* **90**: 879–887.

566 Satterstrom FK, Kosmicki JA, Wang J, Breen MS, de Rubeis S, An JY, Peng M, Collins R, Grove J,
567 Klei L, et al. 2020. Large-Scale Exome Sequencing Study Implicates Both Developmental
568 and Functional Changes in the Neurobiology of Autism. *Cell* **180**: 568-584.e23.

569 Slatkin M. 2008. Linkage disequilibrium - Understanding the evolutionary past and mapping the
570 medical future. *Nat Rev Genet* **9**: 477–485.

571 Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: a general
572 repository for interaction datasets. *Nucleic Acids Res* **34**.

573 Turkmen A, Lin S. 2017. Are rare variants really independent? *Genet Epidemiol* **41**: 363–371.

574 Wei H, Zhu Y, Wang T, Zhang X, Zhang K, Zhang Z. 2021. Genetic risk factors for autism-spectrum
575 disorders: a systematic review based on systematic reviews and meta-analysis. *J Neural
576 Transm* **128**: 717–734.

577 Weissbrod O, Hormozdiari F, Benner C, Cui R, Ulirsch J, Gazal S, Schoech AP, van de Geijn B,
578 Reshef Y, Márquez-Luna C, et al. 2020. Functionally informed fine-mapping and polygenic
579 localization of complex trait heritability. *Nat Genet* **52**: 1355–1363.

580 Whalen S, Pollard KS. 2019. Most chromatin interactions are not in linkage disequilibrium.
581 *Genome Res* **29**: 334–343.

582 Yan Y, Tao H, He J, Huang SY. 2020. The HDOCK server for integrated protein–protein docking.
583 *Nat Protoc* **15**: 1829–1852.

584 Yan Y, Zhang D, Zhou P, Li B, Huang SY. 2017. HDOCK: A web server for protein-protein and
585 protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res* **45**: W365–W373.

586 Ye H, Liu J, Wu JY. 2011. Cell adhesion molecules and their involvement in autism spectrum
587 disorder. *Neurosignals* **18**: 62–71.

588 Yu G, Wang LG, Han Y, He QY. 2012. ClusterProfiler: An R package for comparing biological
589 themes among gene clusters. *OMICS* **16**: 284–287.

590 Zan Y, Forsberg SKG, Carlborg Ö. 2018. On the relationship between high-order linkage
591 disequilibrium and epistasis. *G3: Genes, Genomes, Genetics* **8**: 2817–2824.

592 Zaykin D v. 2004. Bounds and normalization of the composite linkage disequilibrium coefficient.
593 *Genet Epidemiol* **27**: 252–257.

594
595