PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Known allosteric proteins have central roles in genetic disease

György Abrusán[1]*, **David B. Ascher**[2,3,4], **Michael Inouye**[1,5,6,7,8,9]

1 Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, School of Medicine, University of Cambridge, Cambridge, United Kingdom, **2** Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, **3** Structural Biology and Bioinformatics, Department of Biochemistry, Bio21 Institute, University of Melbourne, Melbourne, Australia, **4** Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Australia, **5** Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Australia, **6** British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom, **7** British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, United Kingdom, **8** Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, United Kingdom, **9** The Alan Turing Institute, London, United Kingdom

* gyabr12@gmail.com

Citation: Abrusán G, Ascher DB, Inouye M (2022) Known allosteric proteins have central roles in genetic disease. PLoS Comput Biol 18(2): e1009806. https://doi.org/10.1371/journal.pcbi.1009806

Editor: Turkan Haliloglu, Bogazici University, TURKEY

Received: July 30, 2021

Accepted: January 5, 2022

Published: February 9, 2022

Data Availability Statement: The following public databases were used in the study: Allosteric Database (http://mdl.shsmu.edu.cn/ASD/), ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), OMIM (https://www.omim.org/), UniProt (https://www.uniprot.org/diseases/), DrugBank (https://go.drugbank.com/), The Open Biological and Biomedical Ontology Foundry (http://www.obofoundry.org/ontology/doid.html), Gene Ontology, (http://geneontology.org/docs/downloads), eggNOG Database (http://eggnog5.embl.de/#/app/home), comorbidity network (http://nlp.case.edu/public/data/FAERS_comb/), GWAS

Abstract

Allostery is a form of protein regulation, where ligands that bind sites located apart from the active site can modify the activity of the protein. The molecular mechanisms of allostery have been extensively studied, because allosteric sites are less conserved than active sites, and drugs targeting them are more specific than drugs binding the active sites. Here we quantify the importance of allostery in genetic disease. We show that 1) known allosteric proteins are central in disease networks, contribute to genetic disease and comorbidities much more than non-allosteric proteins, and there is an association between being allosteric and involvement in disease; 2) they are enriched in many major disease types like hematopoietic diseases, cardiovascular diseases, cancers, diabetes, or diseases of the central nervous system; 3) variants from cancer genome-wide association studies are enriched near allosteric proteins, indicating their importance to polygenic traits; and 4) the importance of allosteric proteins in disease is due, at least partly, to their central positions in protein-protein interaction networks, and less due to their dynamical properties.

Author summary

Allostery is a form of protein regulation, that enables the cell to modulate the activity of certain proteins. Research on allostery usually focuses on the analysis of individual proteins(complexes), to identify the mechanisms of allostery within them. However, proteins perform their functions in the networked environment of the cell, yet, a global, systems-level analysis of the role of allostery in genetic disease is currently missing. In this work, the authors examine the role of allosteric proteins in genetic disease, and show that allosteric proteins contribute to disease much more than non-allosteric proteins. The analysis shows that allosteric proteins are involved in hundreds of diseases, and are most enriched

catalog (https://www.ebi.ac.uk/gwas/), IntAct (https://www.ebi.ac.uk/intact/home), BioGrid (https://thebiogrid.org/), gnomAD (https://gnomad.broadinstitute.org/downloads), Appris database (https://apprisws.bioinfo.cnio.es/landing_page/). We used the 2014.2 version of HGMD; HGMD can be acquired from Qiagen (https://www.qiagen.com/us/products/discovery-and-translational-research/next-generation-sequencing/informatics-and-data/interpretation-content-databases/hgmd/), individual genes can be accessed at (http://www.hgmd.cf.ac.uk/ac/index.php). The supplementary data and code are available at Zenodo (https://zenodo.org/record/7723277; DOI: 10.5281/zenodo.7723277).

Funding: G.A. and M.I were supported by the Cambridge-Baker Systems Genomics Initiative. M. I. was supported by the Munz Chair of Cardiovascular Prediction and Prevention. This work was supported by core funding from the: British Heart Foundation (RG/13/13/30194; RG/18/13/33946), BHF Cambridge Centre of Research Excellence (RE/13/6/30180), and NIHR Cambridge Biomedical Research Centre (BRC-1215-20014)*. It was also supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. Additionally, this work was supported by the NHMRC GNT1174405 grant for D.A. and in part by the Victorian Government's Operational Infrastructure Support Program. The authors declare no competing interests. *The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

diseases of the hematopoietic system, (cardio)vascular diseases, and cancers. The findings also indicate that the central role of allosteric proteins in disease is less due to their dynamical properties, but, at least partly, due to their central positions in protein-protein interaction networks. Thus, evolutionary pressure for fine-tuned regulation of network hubs is likely to be a significant factor in the evolution of allostery.

Introduction

Allostery is a form of regulation of protein activity, that enables the cell to fine tune the spatial and temporal activity of certain proteins. The most permissive definitions of allostery simply require that a perturbation at one site of a protein (mutation, modification, or binding) results in a dynamical or topological change at a distant site, and in this respect, most dynamical proteins are likely to be allosteric [1–3]. However, the number of proteins that are regulated by allostery while performing their normal biological functions is much more limited, and the mechanism of allostery in proteins where allosteric regulation has evolved for millions of years is likely to be very different from proteins for which an artificial allosteric effector could be designed, but otherwise are not known to display allosteric behaviour in nature. Thus, allosteric proteins are frequently defined more restrictively ("classic" definition of allostery), as proteins that use allostery while performing their normal biological function, which usually involves protein complexes, switching between distinct conformational states, and information transfer between distant ligand binding sites [4,5].

Allosteric proteins typically have multiple ligand-binding sites: an orthosteric site, which is the binding site responsible for the main biological function of the protein, like catalysis; and one or more allosteric sites, which can be seen as "switches" that turn the activity of the protein on or off, by inducing conformational changes [4,6–8]. The actual mechanisms of allosteric signal transduction can be diverse, but in most proteins it can be explained by shifts between populations of different conformations, that are modified or induced by ligands binding at critical sites of the protein [5,9,10]. Allostery and the identification of allosteric binding sites have received considerable attention from the drug discovery community [11–13] for several reasons. First, some of the most frequent drug targets like GPCRs or protein kinases are frequently allosteric [14,15]. Second, the orthosteric binding sites of proteins are frequently highly conserved, which makes it difficult to design drugs that target these sites and at the same time are specific, without significant off-target effects. However, allosteric sites are much less conserved [16], and their high structural diversity offers the possibility of designing drugs that target allosteric proteins in a more specific way.

The fraction of proteins regulated by allostery while performing their normal biological function is unclear; kinases or GPCRs represent only a minority, and the number of known allosteric proteins [17] probably significantly underestimate their real number in the human proteome. Typically, allosteric proteins form protein complexes (see [18] for an overview and citations therein), and are particularly common in complexes where ligands connect several proteins in the complex [18] (with so called "polydesmic" ligands [19]). Most research on allostery has focused on the analysis of individual proteins or protein complexes, typically to identify the mechanism of allostery and allosteric pathways within them, or to find novel druggable allosteric sites. However, proteins perform their functions in the networked environment of the cell, e.g. in metabolic networks, signalling networks, gene regulatory and protein interaction networks amongst others. Nussinov and colleagues suggested that allostery should be analysed at the level of cellular networks [20,21], because many key proteins in cellular signalling

pathways (e.g. receptors) are allosteric [11,22,23], as allostery is well suited to propagate signals. However, a global, systems-level analysis of allostery and its role in genetic disease is currently missing, although a large-scale investigation of somatic cancer mutations found that they are enriched near allosteric sites [24].

In this work, we examine the role of allosteric proteins in genetic disease, focusing on proteins where allostery has been characterised experimentally. We show that known allosteric proteins are much more likely to cause genetic disease than non-allosteric proteins, even if proteins involved in signaling (including kinases and GPCRs) are excluded from the data, and are central in disease networks. Our analysis shows that allosteric proteins contribute to hundreds of diseases, and are most common in those of the hematopoietic system, (cardio)vascular diseases, and cancers. The analysis of cancer GWAS data indicates that their variants are also enriched near allosteric proteins, indicating that their contribution is enriched for polygenic traits. Surprisingly, we find that the central role of allosteric proteins in disease is not so much caused by their dynamical properties directly, but, at least partly, by their central positions in protein-protein interaction networks. This suggests that evolutionary pressure for fine-tuned regulation of network hubs is likely to contribute to the evolution of allostery, and that allostery is not so much the cause, but rather the consequence of the centrality of these proteins in disease and PPI networks.

Results

Data summary

We compiled a list of 6170 proteins which are associated with genetic disease, i.e. harbouring pathogenic mutations from ClinVar [25], OMIM [26], Uniprot and HGMD [27] (Methods). As the disease annotations and names in different databases are not identical, we used the human disease ontology (doid.obo file, Methods), to standardise disease nomenclature and excluded proteins that could not be mapped to a disease term present in the disease ontology. In total, there were 5050 proteins associated with at least one disease ontology term which were available for downstream analyses. We used the Allosteric Database (ASD, v4.10) [17] to obtain the list of known allosteric proteins. Of the 835 allosteric proteins in humans, 450 were associated with disease, and 380 are associated with at least one disease term present in the disease ontology.

Allosteric proteins are enriched in many major disease types, including hematopoietic and vascular diseases, and cancers

We performed disease ontology and gene ontology analyses, to examine whether there are diseases where mutations in allosteric proteins are particularly common. We identified the full list of disease ontology terms that map to each protein (diseases and their parental terms), and performed an ontology analysis similar to a gene ontology (GO) enrichment analysis. We found that allosteric proteins contribute to a large number of genetic diseases and conditions (S1 Table): they are significantly enriched in 214 disease ontology terms (S1 Table, $p < 0.05$, FDR), which include many of the major disease types: cardiovascular diseases, cancers, diabetes, rheumatoid arthritis and central nervous diseases (Fig 1A). The diseases where allosteric proteins are the most significantly enriched ($p < 0.005$ after Bonferroni correction, Fig 1B) are hematopoietic diseases (DOID:74) a diverse set of cancers, and vascular disease (DOID:178). Proteins associated with these diseases are characterised by a 2- to 5-fold enrichment among allosteric proteins compared to all disease proteins (S1 Table).

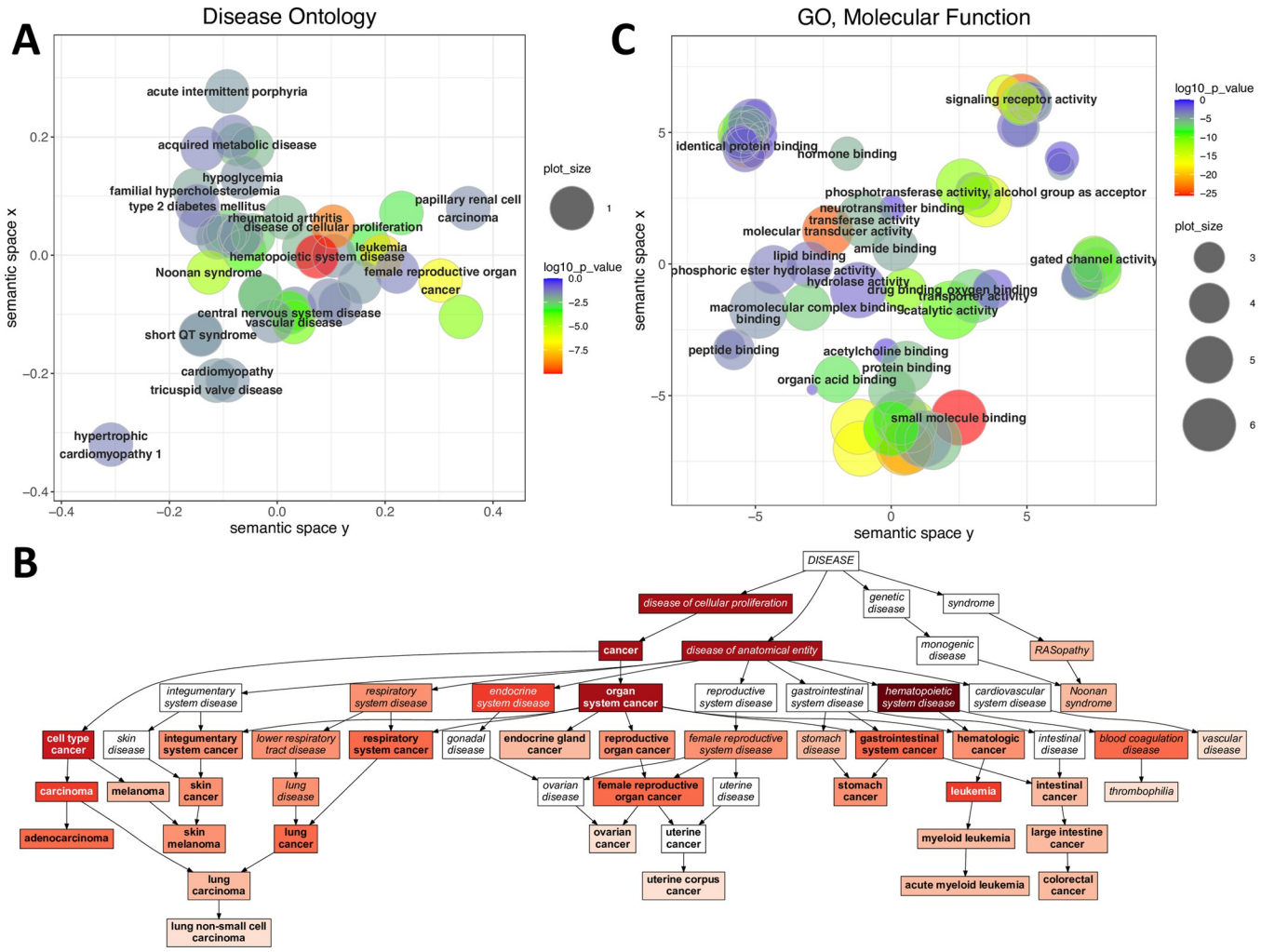

Fig 1. Enrichment of allosteric proteins in disease ontologies and gene ontologies. A) Summary of the disease ontology terms where allosteric proteins are significantly enriched. The terms were filtered for redundancy and plotted using their semantic similarities, the ones highlighted were selected manually from the filtered list of terms. Allostery is overrepresented in diverse diseases, which include many of the major disease types like cardiovascular diseases, metabolic diseases (e.g. hypercholesterolemia, diabetes) central nervous system diseases, and cancers. **B)** Graph of the disease ontology terms where allosteric proteins are most significantly enriched. The intensity of red corresponds to significance, only terms with $p < 0.005$ (Bonferroni-correction) are plotted. Allosteric proteins are most significantly associated with diseases of the hematopoietic system, cancers and vascular disease. **C)** GO analysis of significantly enriched Molecular Function terms. Related terms were grouped and visualized with REVIGO using their sematic similarity. The list of proteins in each term, and p-values are provided in S2 Table. The analysis shows that allosteric proteins are involved in diverse functions, and, unlike disease ontology terms, the enriched GO terms are not dominated by a few categories.

https://doi.org/10.1371/journal.pcbi.1009806.g001

To identify which functions of allosteric proteins contribute most to cancers, hematopoietic and vascular diseases, we performed a Molecular Function (MF) GO enrichment analysis on the proteins used in disease ontology analyses, and next a gene overlap analysis (see Methods) to test for overlaps between the lists of allosteric proteins of the significantly enriched disease ontology and MF GO terms. MF GO terms where allosteric proteins were overrepresented were not dominated by a few categories but have diverse functions (Fig 1C, $p < 0.05$ after Bonferroni correction, see S2 Table for full results and exact p-values). The gene overlap analysis identified several disease and MF gene otology terms with a significant overlap between cancer related terms and MF GO terms of protein kinases (S1 Fig, $p < 0.05$ after Bonferroni correction), consistent with the known roles of kinases in carcinogenesis [28,29]. However, in the

case of allosteric proteins of hematopoietic and vascular diseases no comparably strong overlap with proteins of MF GO terms was observed.

Allosteric proteins cause more diseases than non-allosteric ones, and are central in disease-protein networks

Next, we examined whether allosteric proteins were more likely to be involved in diseases than non-allosteric ones. As the frequency of allostery and the topology of allosteric pathways can vary substantially among different protein complexes [18], we performed both a pooled analysis of all allosteric proteins and assessed whether the patterns depend on the quaternary structure of proteins. Quaternary structure (i.e. heteromer, homomer, monomer) was assigned using the structures of the Protein Data Bank (PDB), and could be assigned to only 53% of proteins with disease annotations. The diseases associated with each protein were determined as previously, using the disease ontology terms, with the difference that the parental terms of diseases were not used, only disease terms that are independent from each other (see Methods). The analysis shows that allosteric proteins are more likely (2-fold, when all proteins are included) to cause disease than non-allosteric proteins (Fig 2A), and were also associated with significantly more diseases per protein (Fig 2B). This pattern is somewhat more pronounced for cancers than for non-cancerous diseases, but they were not qualitatively different (S2 Fig). We found the number of diseases associated with allosteric vs non-allosteric proteins to be different also when stratified by quaternary structure; it is strongest in heteromers, and not significant in homomers (Fig 2B).

To examine the importance of allostery at the systems level, we constructed a disease-protein network, using the approach of Goh et al.[30] (see Methods). Each protein is a node in the network, nodes are connected with an edge if both proteins are associated with the same disease, and the weight of the edge is defined by the number of diseases. The resulting network (Fig 2F) has one large connected component with 3553 proteins, several small clusters, and 1068 isolated nodes. The analysis of the parameters of the nodes indicates that allosteric proteins had significantly higher betweenness centrality than non-allosteric proteins (except for proteins that form homomers, Figs 2C and S3). Betweenness centrality measures the number of shortest paths in the network that pass through a particular node, and nodes with high betweenness centrality typically connect clusters and control most of the information flow in a network [31]. High betweenness centrality can result from several different biological processes, for example due to being bottlenecks in metabolic networks, being a transcription factor, or addition of posttranslational modifications, but frequently it is associated with signaling. However, the pattern does not change qualitatively when all kinases and GPCRs, or all proteins involved in signal transduction are excluded from the analysis (S4 Fig). Allosteric binding sites are sometimes identified during screening for novel drug binding sites, raising the possibility that the central role of allosteric proteins in the disease network is due to being a drug target rather than due to being allosteric. Thus, we also examined whether a similar pattern is present within pharmacologically active drug targets. We found that allosteric proteins have significantly higher betweenness than non-allosteric proteins, both when only drug targets (Fig 2D) or only non-drug targets were included (Fig 2E), indicating that being a drug target is not sufficient to explain the observed pattern, even though drug targets are characterised with higher betweenness.

Homologs of allosteric proteins indicate an association between allostery and disease

The number of known allosteric proteins is incomplete, and well-studied proteins are overrepresented among them. To further test the possible causality between being allosteric and disease associated, we examined the properties of human proteins that are homologous to known

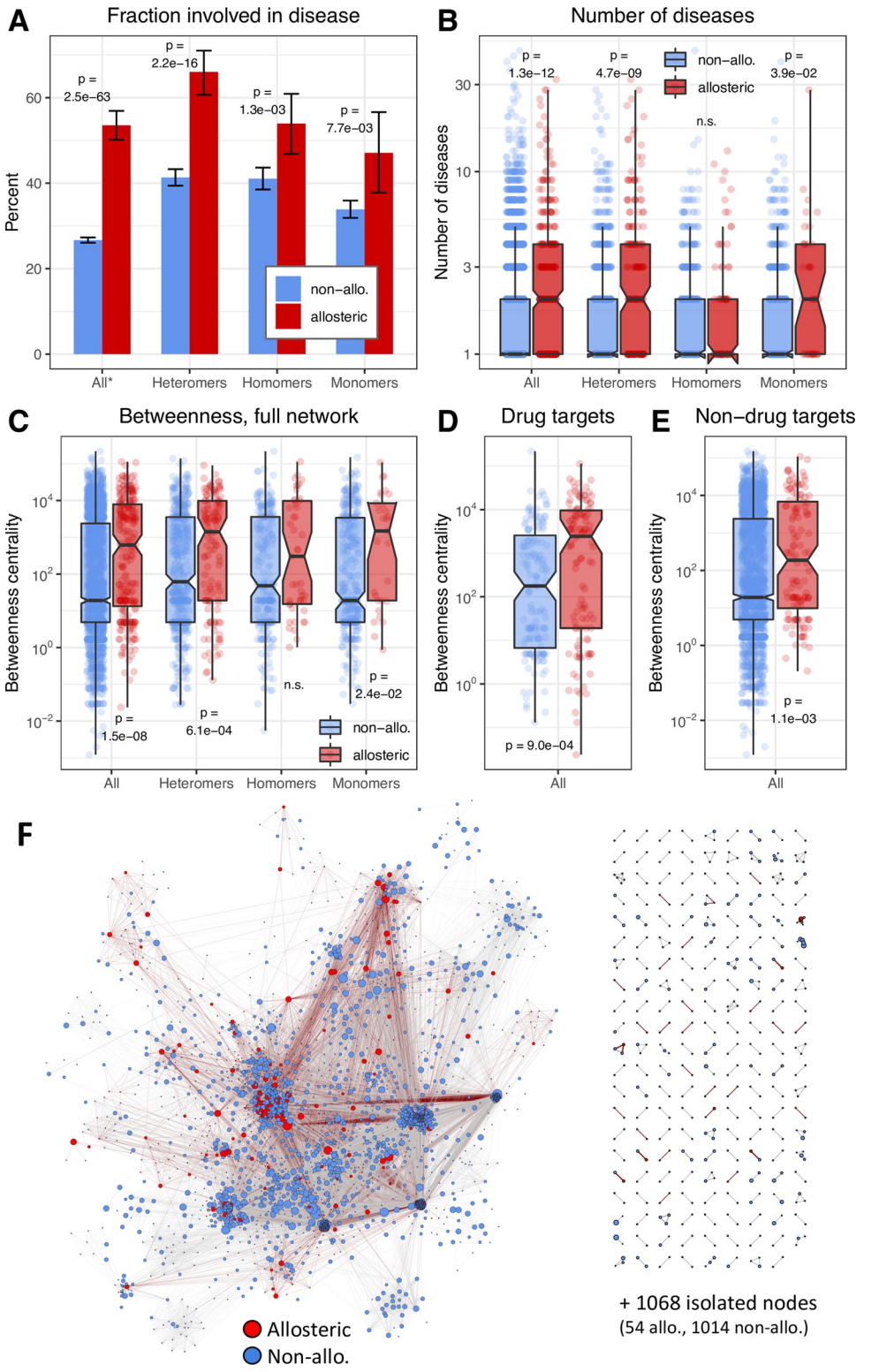

Fig 2. Allosteric proteins are overrepresented in disease, and are central in disease-protein networks. A) Allosteric proteins are significantly more frequently involved in disease than non-allosteric proteins, irrespectively of their quaternary structure. * In the 'All' category the number of non-allosteric proteins was defined as human SwissProt proteins minus all human allosteric proteins. Note that the 'All' category includes also the proteins where quaternary structure is not known. **B)** Allosteric proteins (except homomers) are involved in significantly more diseases than non-

allosteric ones. **C)** Allosteric proteins, particularly the ones forming heteromeric complexes have significantly higher betweenness centrality in the disease-network than non-allosteric ones. **D-E)** Drug-targets have generally higher betweenness centralities than non-drug target proteins, however, allosteric proteins have significantly higher betweenness than non-allosteric proteins in both groups. **F)** The disease-protein network. Allosteric proteins are represented by red nodes, non-allosteric ones by blue, the size of the nodes was calculated as 1 + Log(nr of diseases). The largest connected component was visualized with the OpenOrd algorithm of the Gephi platform.

https://doi.org/10.1371/journal.pcbi.1009806.g002

allosteric proteins, but currently are not known to be allosteric, and are not present in the allosteric database (thus in these proteins allostery—if exists—is not the result of study bias). A protein can have two basic types of homologs: orthologs and paralogs. Orthologs represent homology between different species, while paralogs are homologs within the same species, resulting from a duplication event. Orthology and paralogy have very different consequences for functional conservation: orthologs typically have similar functions in different species, frequently across large phylogenetic distances [32–34]. In contrast paralogs, due to being released from selective pressure after duplication, typically evolve and acquire new functions rapidly, and are one of the main source of evolutionary innovations [35–37]. Thus, if there is an association between disease and allostery, that predicts that orthologs of allosteric proteins will be more important in disease than their paralogs.

First, using the orthogroups of the eggNOG5 database [38] we examined whether the human orthologs of non-human allosteric proteins of mammals, metazoans, and eukaryotes are important in disease. We only used human proteins that are not known to be allosteric, however due to orthology, most of them are likely to be allosteric also in humans. Our results show that similarly to human allosteric proteins these proteins are enriched in disease, and have high betweenness in the disease network (Fig 3A and 3B). Next, we examined whether non-allosteric paralogs of human allosteric proteins (i.e. the inparalogs in mammalian, vertebrate, metazoan and eukaryotic orthogroups of eggNOG5) are important in disease. Paralogs are much less likely to have similar functions to their allosteric homologs then orthologs and our results are consistent with this expectation: paralogs that are not known to be allosteric are much less important in disease than allosteric proteins (Fig 3C and 3D).

Finally, we examined to what degree the age of duplications can influence the above results. Proteins that undergo duplications, especially recent ones, are expected to be less important than singletons [39,40], because their increased dosage is not very deleterious. Thus, the weak enrichment of paralogs in disease might also reflect the "unimportance" of their homologous allosteric proteins. We did find this effect (Fig 3E and 3F), however, except the mammalian orthogroup, duplicated allosteric proteins are still significantly enriched in disease compared to non-allosteric proteins.

Taken together, these results show that the homologs of allosteric proteins have the characteristics that are expected based on their evolutionary history, i.e. orthologs are important in disease, while paralogs are not, indicating that there is an association between allostery and disease. In addition, this pattern is not consistent with the predictions of study bias, in orthologs study bias predicts low importance in disease (however, in some cases allostery in non-human proteins is likely to be discovered due to their importance in humans), while in the case of human paralogs one would expect that due to being homologous to known allosteric proteins, they nevertheless receive significant attention from the research community.

Allosteric proteins are involved in more comorbidities than non-allosteric proteins

It has been demonstrated that genes involved in multiple diseases are also more likely to be responsible for comorbidities [41]. In addition, as clusters in disease networks usually

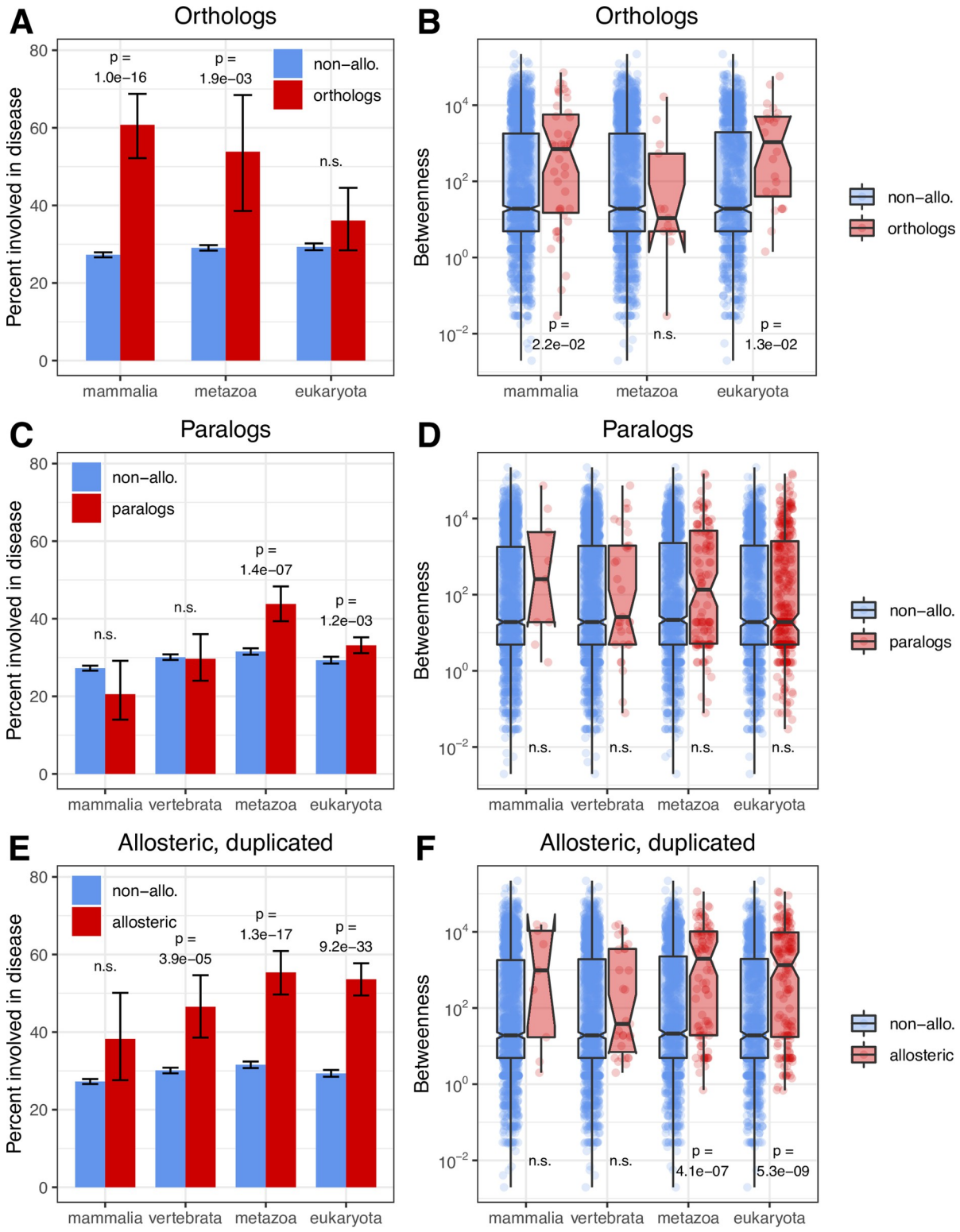

Fig 3. Human orthologs and paralogs of known allosteric proteins, and which are not present in the allosteric dataset have different disease properties. Orthology and paralogy have different consequences for functional conservation: orthologs usually have similar functions in different species, while paralogs, due to being released from selective pressure after duplication evolve and acquire new functions rapidly. **A-B)** Similarly to known human allosteric proteins, human orthologs of non-human allosteric proteins are enriched in disease, and have high betweenness centralities in the disease network. **C-D)** Paralogs of human allosteric proteins show no, or only weak importance in disease.

E-F) Young duplicated human allosteric proteins are less important in disease than older duplications. On all panels the analyses were performed for separate phylogenetic datasets of the eggNOG5 database, excluding orthogroups that have only members of the lower taxonomic group (i.e. orthogroups having only mammalian proteins were not used in the vertebrate or metazoan sets).

https://doi.org/10.1371/journal.pcbi.1009806.g003

correspond to proteins/genes of specific diseases [30], the nodes/proteins connecting them (i.e. with high betweenness centrality) are particularly good candidates for causing disease comorbidities [42]. To examine this, we analysed a recently published comorbidity network, that was constructed using 17 million cases of the FDA Adverse Event Reporting System [43], without the use of genetic information. For proteins involved in at least two diseases, a significantly higher fraction of allosteric proteins was associated with comorbid diseases or conditions (S5A Fig), and allosteric proteins are also involved in significantly more comorbidities (S5B Fig). However, the relationship between the number of diseases and number of comorbidities between them is not qualitatively different for the two protein types (S5C Fig), indicating that the higher number of comorbid conditions of allosteric proteins is primarily the consequence of the higher number of diseases they are involved in.

Common variants for cancers are enriched near allosteric proteins

The genetic architecture of diseases is typically grouped into two broad categories: Mendelian diseases caused by a single or small number of low frequency genetic variants with strong phenotypic effects; and complex (polygenic) diseases caused by a large number of common genetic variants with weak effects. (The two types are part of a continuum though [44].) The ClinVar and HGMD databases are focused on Mendelian diseases. To examine whether allosteric proteins also have a preferential contribution to complex diseases, we asked whether genome-wide association studies (GWAS) are more likely to identify loci near allosteric proteins than expected by chance. As our previous analyses indicated a consistent enrichment of allosteric proteins in diverse cancers (Fig 1), and a large number GWAS have been performed for multiple cancer types [45], we focused our analysis on GWAS of cancers.

To obtain a list of variants associated with cancers in GWAS, we used the GWAS Catalog [46] (v1.0.2). Variants identified by GWAS are typically not the causal variants of the trait being analysed, but are correlated with the causal variants through linkage disequilibrium. Nevertheless, the gene closest to the index SNP at a locus is usually the causal gene [47]; thus, for each associated variant, we identified a candidate causal gene—the nearest protein-coding gene with at least one exon closer than 100kb to the variant, then examined the enrichment of allostery in the candidate causal genes. We chose 100kb as the cutoff distance, because most long-range transcription factors are located closer than 100kb to their target gene [48], however our results were similar when a 50kb threshold was used. In addition, for every base in the genome, we identified the closest protein coding gene (within 100kb), and we identified the total number of bases that are located closest to proteins from three groups: allosteric proteins, 'genetic disease' proteins (as defined above in the Data summary, i.e. 6170 proteins, excluding allosteric ones), or 'other' proteins (which do not belong to either of the prior two groups). The frequencies of the bases proximal to the three protein groups were used to normalize the frequencies of variants identified by GWAS (see Methods).

Cancer GWAS are highly variable in terms of sample size (and thus statistical power) and the number of identified variants per study, even for the same cancer types. The majority of studies include 1,000–10,000 cancer cases, and identify less than 10 significant ($p < 10^{-8}$) variants (Fig 4A and 4B). Cancers are characterised by high genetic heterogeneity even when they originate in the same tissues [49,50], and GWAS with large numbers of cancer cases are typically available only for the most common cancer types. We performed three different analyses

which treat the underlying heterogeneity of the GWAS data differently. In the first analysis (Fig 4–4F), for every cancer type (mapped trait of GWAS catalog, see Methods) we used only the variants of the study with the highest number of cancer cases. In the second analysis (Fig 4G–4J), we compiled a single nonredundant list of genes from all available studies of the same cancer type, and enrichment was calculated in the pooled lists. In the third analysis (S6 Fig), we used the variants/genes of all studies which identified at least partially nonredundant lists of genes for every cancer type. In addition, in all three analyses, we used two different datasets: one that uses cancer GWAS studies irrespectively of the number of reported variants ("All"), and one that only use GWAS reporting less than 10 significant variants, which is less affected by the few commonest cancer types, and the included studies have less variability in statistical power to detect associations. The list of all GWAS used, along with their candidate causal genes is provided in S4 Table.

We found similar patterns of enrichment in all three analyses. Allosteric proteins show a significant, 2-fold enrichment compared to "other" proteins that are neither allosteric nor involved in disease (Figs 4C, 4G, and S6A), while 'genetic disease' proteins show a less pronounced, but still significant 1.5-fold enrichment (Figs 4C, 4G, and S6A). When using only the variants with the highest significance in each GWAS (20% with the lowest p-values), we found an even stronger, 3-fold enrichment near allosteric proteins compared to other proteins (Figs 4D, 4H, and S6B). In the analyses based on the studies reporting less than 10 variants, variants near allosteric proteins also show a high, 3-fold enrichment compared to other proteins of the genome, similarly to the most significant variants of the full dataset (Figs 4E, 4H, and S6C; note that a 3-fold enrichment means that 11.4% of GWAS variants is located close to allosteric proteins, while in the genome 3.8%). Taken together, these results indicate that, at least in cancers, genes of allosteric proteins are enriched as the nearest genes for common variants, and the effect is strongest among the genes near the variants with the highest significance.

Allosteric proteins are enriched for pathogenic mutations

The observed role of allosteric proteins in disease may be due to several factors. These include a generally greater functional importance of allosteric proteins, their distinct structural and dynamical characteristics, central positions in cellular networks, or research bias. To assess functional importance, we calculated the level of conservation for every disease associated protein using their mammalian orthologs, excluding primates (see Methods). Surprisingly, we found only a small, 2% difference between the conservation of allosteric (88.12%) and non-allosteric (85.99%) proteins (Fig 5A). We also performed the same comparison for kinases, to check whether allostery has an effect within a single protein family, and for drug target proteins, which are likely to be less variable in the research effort they receive. We found no significant difference in kinases (Fig 5B), but there is a significant effect (4% difference, $p = 7.47e\text{-}06$) in drug targets (Fig 5C).

Next, using ClinVar and HGMD mutations (see Methods) we examined whether there were consistent differences in the numbers of known pathogenic missense mutations of allosteric/non-allosteric proteins, and whether it influences the number of diseases they cause. Allosteric proteins have a significantly greater number of pathogenic mutations (2–3 fold) than non-allosteric ones, both when the entire dataset, kinases or drug targets are compared (Fig 5D, 5E and 5F). This may reflect true biological differences, e.g. allosteric proteins may be more vulnerable to mutation, but also could be influenced by research biases, if allosteric proteins receive consistently more attention from the research community (however, research intensity and the actual biological importance of a protein are strongly correlated). To estimate bias, we examined whether the human orthologs of mammalian allosteric proteins are

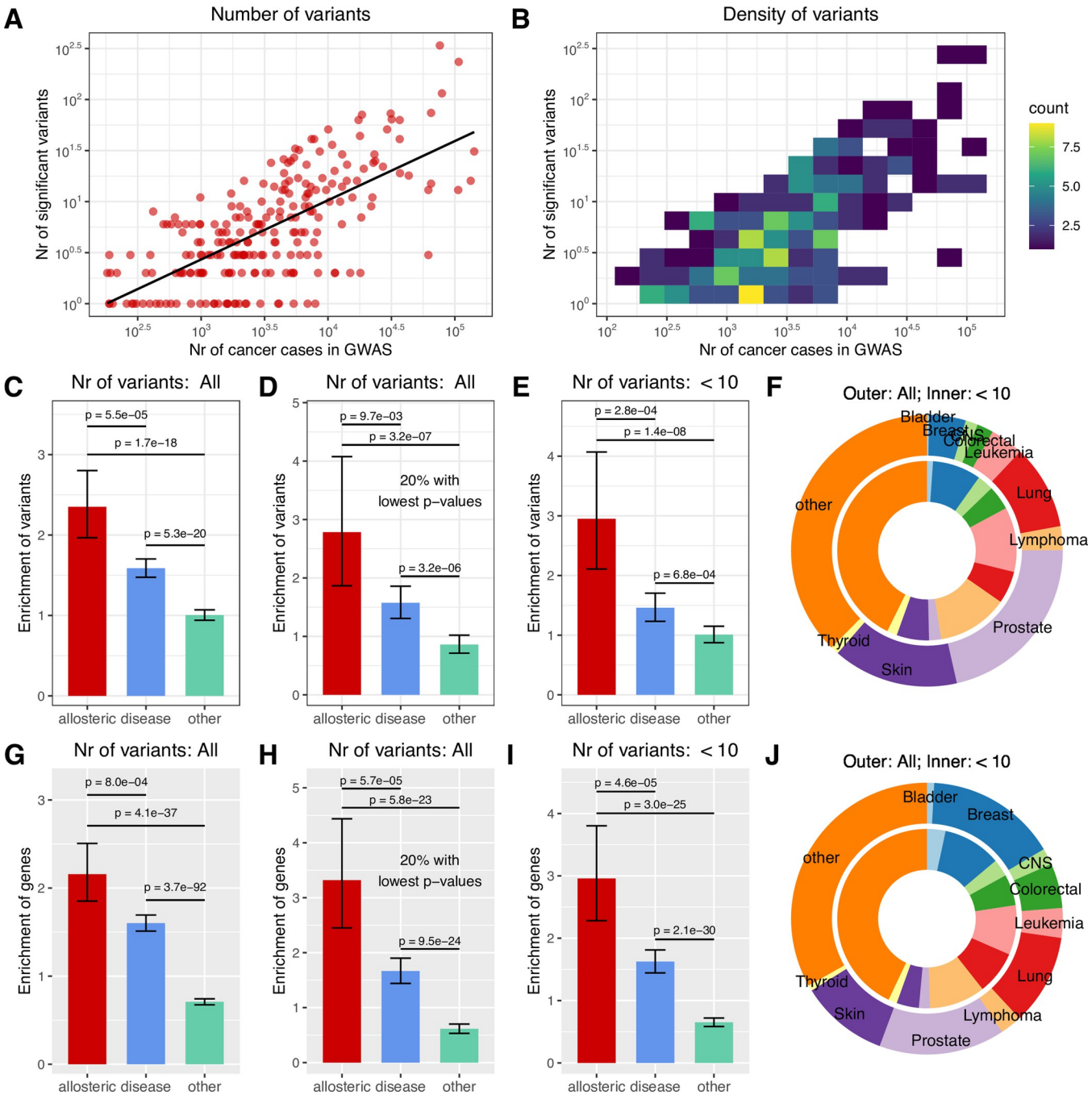

Fig 4. Variants of cancer GWAS are enriched near allosteric proteins. A-B) Relationship between the number of cancer cases and significant variants ($p < 10^{-8}$) in cancer GWA studies where at least one significant variant is located near a protein coding gene. The majority of GWAS have a moderate number of cases (and power), 10^3–10^4, and identify less than 10 significant variants. Enrichment near different protein types was calculated in two datasets; one using studies irrespectively of the number of significant variants ("All"), and one using GWA studies reporting less than 10 significant variants, which is influenced less by the largest studies of the most common cancer types, and has less variability in statistical power to detect significant associations. **C-F)** In the analysis which only uses the studies with the highest number of cancer cases, allosteric and (mostly Mendelian) disease associated proteins show a 2- and 1.5-fold enrichment compared to other (i.e. neither allosteric nor disease associated) proteins (C). Variants with the highest significance in each GWAS (20% with the lowest p-values) show an even more pronounced, 3-fold enrichment near allosteric proteins (D). In studies reporting less than 10 variants, the enrichment near allosteric proteins is also 3-fold, comparable to the pattern seen with the most significant variants of the full dataset (E). The distribution of main cancer types in the variants of the two GWAS sets (F). **G-J)** The analysis which uses a single nonredundant list of genes compiled from all studies of the same cancer type (mapped trait of GWAS Catalog) shows a similar degree of enrichment of allosteric and disease proteins among the proteins identified by GWAS.

https://doi.org/10.1371/journal.pcbi.1009806.g004

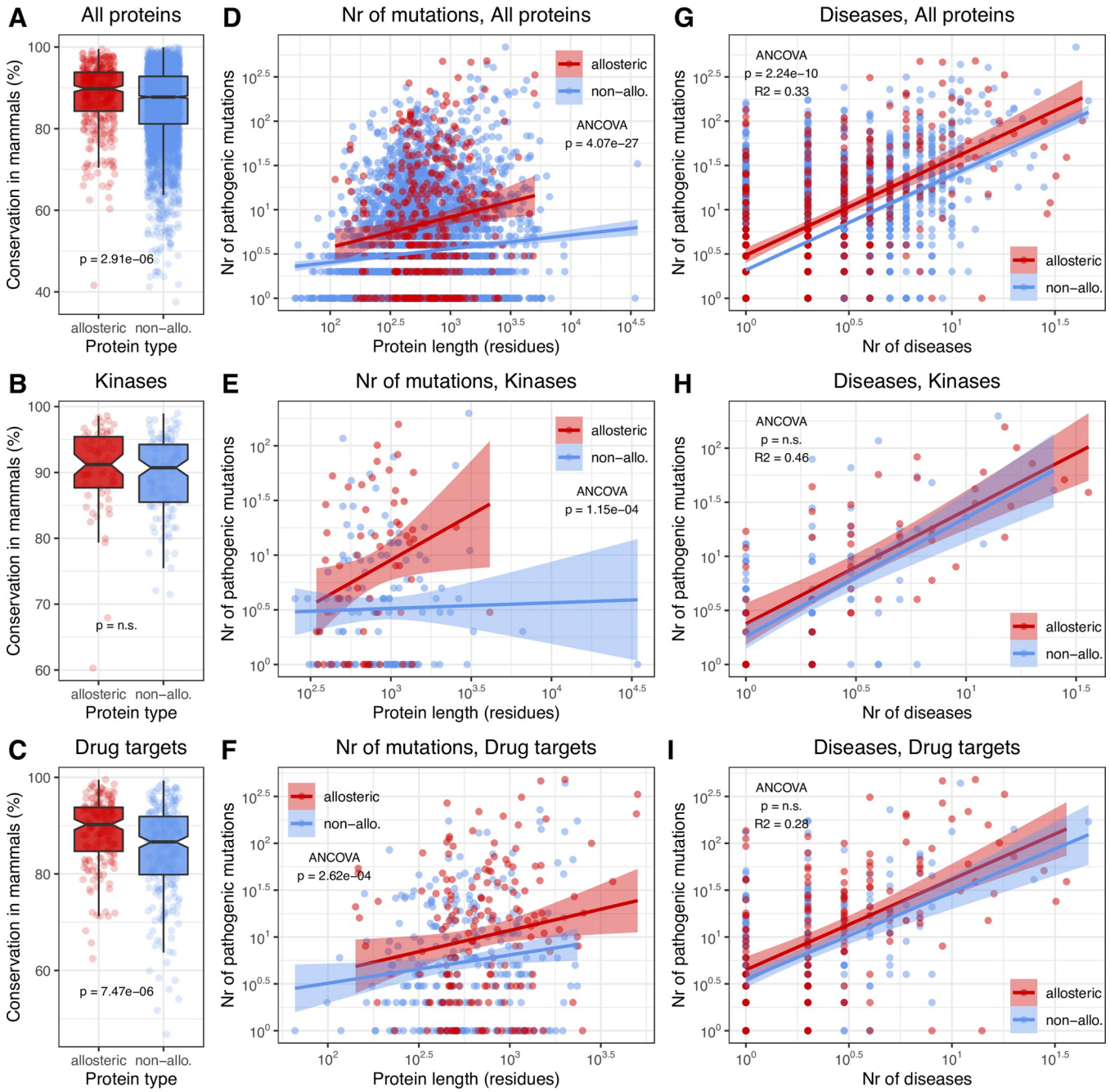

Fig 5. Conservation, mutation density, and disease associations of allosteric and non-allosteric proteins involved in disease. Kinases and drug targets are shown separately. **A-C)** Allosteric proteins are characterised by a somewhat higher conservation than non-allosteric proteins. **D-F)** Allosteric proteins are characterised by significantly higher numbers of pathogenic mutations than non-allosteric ones. **G-I)** Despite having more pathogenic mutations, the relationship between the number of diseases, and the number of pathogenic mutations is qualitatively similar for allosteric and non-allosteric proteins.

https://doi.org/10.1371/journal.pcbi.1009806.g005

characterised with a similarly higher number of pathogenic mutations as human allosteric proteins. We found that a similar pattern is present in orthologs as in known allosteric proteins (S7 Fig). We also examined whether the mutations of allosteric and non-allosteric proteins in ClinVar originate from same number of PubMed articles and found that there is a positive correlation between the number of diseases and number of articles reporting mutations (S6 Fig).

In addition, mutations of allosteric proteins are cited in significantly more research articles than non-allosteric ones (S8 Fig). However, the greater number of articles per disease does not translate to a similarly greater number pathogenic mutations per disease, the relationship between the number of diseases and the number of pathogenic mutations is the same in allosteric and non-allosteric proteins in all three sets (Fig 5G, 5H and 5I). The difference is not significant for kinases and drug targets (Fig 5H and 5I), and significant but explain <1% of variance for the entire dataset (Fig 5G, ANCOVA was performed on log-transformed values). These results, in conjunction with the evolutionary analyses and GWAS indicate that the high importance of allosteric proteins in disease and disease networks is not simply the by-product of research bias. Furthermore, the positive correlation between the number of pathogenic mutations and diseases (Fig 5G–5I) suggest that proteins associated with more diseases also have more vulnerabilities, possibly due to having more protein-protein interactions and interfaces.

Pathogenic mutations accumulate in regions of 3D structures that are important in dynamics

Allosteric proteins are characterised with multiple binding sites and allosteric pathways that connect them, and some pathogenic missense mutations are known to cause disease through allosteric effects [51]. Thus, the number of mutations that can interfere with the correct functioning of allosteric proteins might be greater than in the case of non-allosteric proteins, due to their dynamics, and we examined whether the dynamic nature of these proteins contributes to their centrality in disease. First, we tested whether there are topological differences between the folds of conserved Pfam domains in allosteric and non-allosteric proteins. We mapped non-covalent residue interactions of the domains to a 50 x 50 matrix (Fig 6A and 6B, see also Methods), where position 1 is the N-terminus of the domain, and 50 is the C-terminus. We found that compared to non-allosteric proteins, allosteric proteins are depleted of long-range residue interactions, i.e. interactions connecting residues that are distant in the protein sequence (Fig 6C and 6D). This indicates that protein domains in allosteric proteins are more flexible than in non-allosteric proteins, as long-range interactions generally stabilise proteins and result in more rigid folds.

Next, we examined whether disease associated mutations are enriched in the regions that are primarily responsible for the internal dynamics of proteins. Dynamic proteins can typically be partitioned into semi-rigid blocks of residues called "communities", which are characterised by correlated motions of residues [8,10,52]. The residues connecting the communities, particularly the ones characterised by high betweenness centrality in the residue-interaction network are the residues that control most of the motions of the proteins, including allosteric signal transduction (Fig 6J and 6K). We determined the community structure of the monomeric units of every disease-associated protein where a suitable PDB structure was available, using the STRESS tool [52,53] (Methods), and mapped the pathogenic mutations to the communities. We found a clear enrichment of pathogenic mutations in residues that connect communities, particularly the ones that take part in strong interactions, like H-bonds (Fig 6E and 6F). Surprisingly, the pattern is similar in allosteric and non-allosteric proteins, indicating that interference with the motions of proteins is important for pathogenesis in both cases.

Recent work in cancers have shown that somatic mutations show a clustering in the 3D structures of cancer genes, both in drivers and suppressors, and that this pattern can be used to identify them [54–57]. Kumar et al. [52] has reported that such clusters overlap with the dynamic communities of proteins, and driver genes are characterized by "hotspot communities" which are mutated much more frequently than others. We examined whether pathogenic

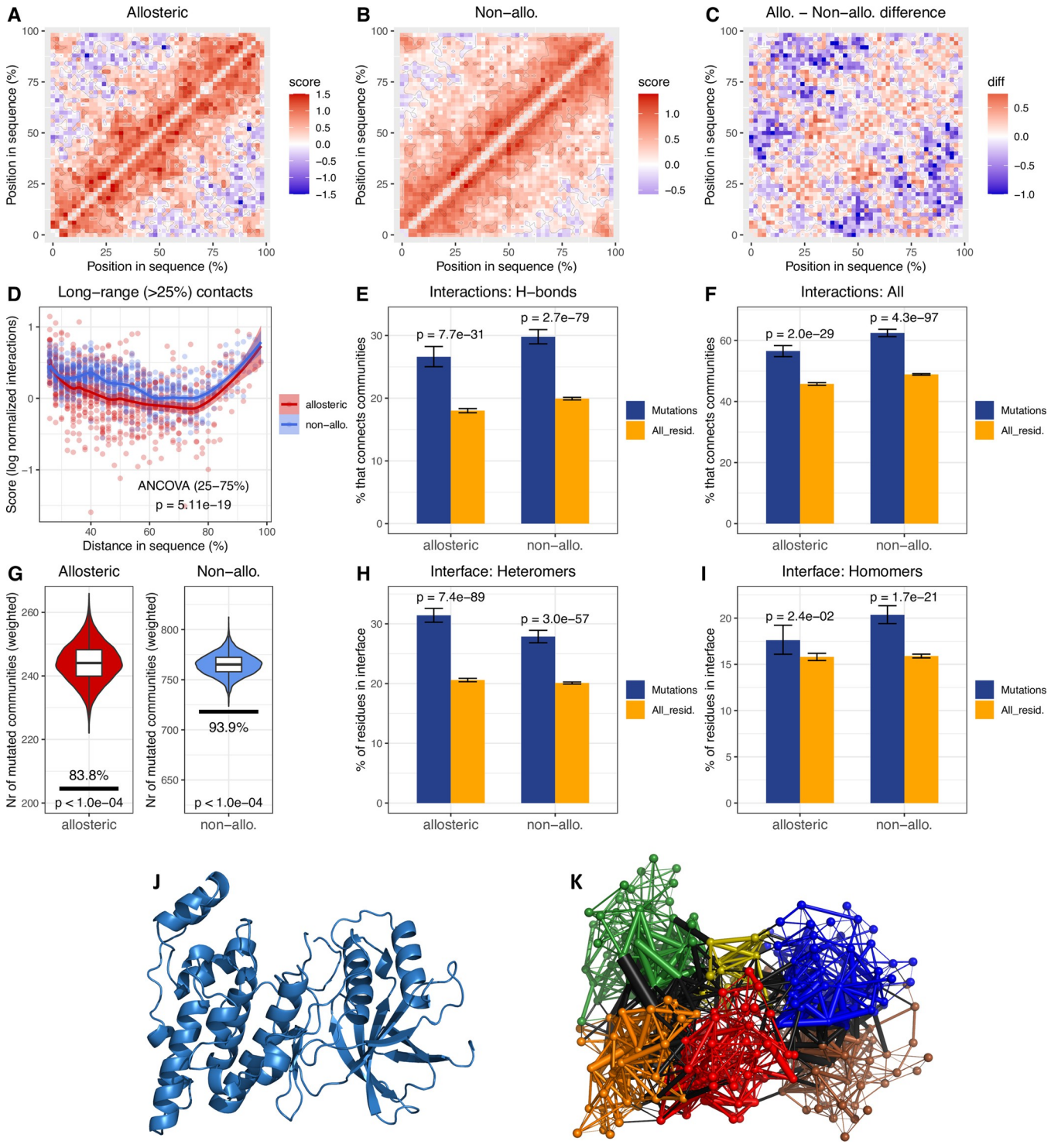

Fig 6. Structural and dynamical characteristics of pathogenic mutations in allosteric and non-allosteric proteins. A-B) Residue interaction matrix of Pfam domains of allosteric (A) and non-allosteric (B) proteins. **C-D)** The difference between the two matrices (panel C) shows that allosteric proteins have significantly fewer long-range interactions in their Pfam domains than non-allosteric proteins (panel D), and are likely to be more flexible. **E-F)** Disease associated mutations are significantly enriched in community interfaces, i.e. residues that interact with members of other communities, both in the case of allosteric and non-allosteric proteins. The enrichment is more pronounced for stronger interactions like H-bonds (panel E). **G)** The distribution of disease mutations across communities (horizontal bar) differs from the random expectation (violin plot) more in allosteric proteins than in non-allosteric ones, indicating that pathogenic mutations have a stronger effect in allosteric proteins. See also S9 Fig. **H)** Disease associated mutations are significantly enriched in the protein-protein interfaces of both allosteric and non-allosteric heteromers. **I)** A much less pronounced, but also significant enrichment is present in the

interfaces of homomers. **J)** The structure of Mitogen-activated protein kinase 8 (MAPK8, PDB ID: 4qtd). **K)** Community structure of MAPK8. Each community is represented by a different colour, residue-residue interactions between different communities (community interfaces) are indicated with black.

https://doi.org/10.1371/journal.pcbi.1009806.g006

mutations of Mendelian diseases (ClinVar/HGMD) also show similar biases across communities, with respect to allostery. While the pathogenic mutations we use are not a mixture of drivers and passengers, their degree of pathogenicity is variable [58], and we expected that similarly to passengers in cancers, the less pathogenic ones will be distributed more evenly across communities. Our results show that pathogenic mutations are significantly more clustered than the random expectation (i.e. are present in fewer communities than expected, Fig 6G) and that the effect is stronger in allosteric proteins than in non-allosteric ones (p = 9.37e-04, ANCOVA, Figs S9 and 6G). This suggest that mutations in allosteric proteins are more pathogenic than in non-allosteric ones.

Allosteric signal transduction can cross protein-protein interfaces of complexes; therefore, we also examined whether pathogenic mutations are distributed differently in the interfaces of known structures of allosteric and non-allosteric proteins. Our results indicate that, as it has been reported previously [59–61], pathogenic mutations are clearly enriched in interfaces (Fig 6H and 6I). The enrichment is more pronounced in heteromers (40–50%, Fig 6H) than in homomers (15–30%, Fig 6I), however the effect is not consistently stronger in allosteric proteins.

The importance of allostery in disease networks is partly explained by protein-protein interaction (PPI) networks

It has been shown that allosteric proteins have more connections in PPIs [24] than non-allosteric ones, thus the central role of allosteric proteins in disease may also be the result of their central positions in cellular networks. We examined this using two protein interaction datasets, IntAct, and the larger BioGrid. Using binary protein interactions, we constructed a PPI-network and determined the betweenness and degree centrality for every protein (Methods). The results show that, similar to the disease network, allosteric proteins in general, and particularly the ones forming heteromeric protein complexes, are characterised by significantly higher betweenness and degree centralities than non-allosteric disease proteins, both when IntAct (S10A and S10B Fig) or BioGrid (S10C and S10D Fig) were used to build the network. We found only a weak overall correlation between the centralities of the PPI and disease networks (Fig 7A, R = 0.063; and Fig 7B, R = 0.104), indicating that these two networks types have different topologies. However, allosteric proteins have consistently higher betweenness centralities than non-allosteric ones in both network types (Fig 7A and 7B), and are overrepresented among the proteins with the highest joint betweenness (both being above 1000, Fig 7C and 7D). This suggests that the more central position of allosteric proteins in PPI networks does contribute to the role of allostery in disease.

Hubs of PPI networks are frequently essential [31,62] (network centrality measures are even used to predict essentiality), therefore we examined whether the (at least monoallelic) loss-of-function in allosteric proteins is generally more deleterious than the loss-of-function for non-allosteric disease proteins. We used the distributions of premature stop codons (PSCs) of the gnomAD database [63] to determine the inactivation tolerance of every protein (see Methods). We found that allosteric proteins are characterised by a modest reduction in PSC density, and also by slightly lower mean allele frequencies of their PSCs compared to non-allosteric proteins (S11 Fig). However, the fraction of proteins that are intolerant to inactivation, i.e., have no detected PTCs in their sequence (and are most likely haploinsufficient or essential) is similar in both groups (9.7% vs. 8.7%, p = 0.561, test of proportions). Taken together,

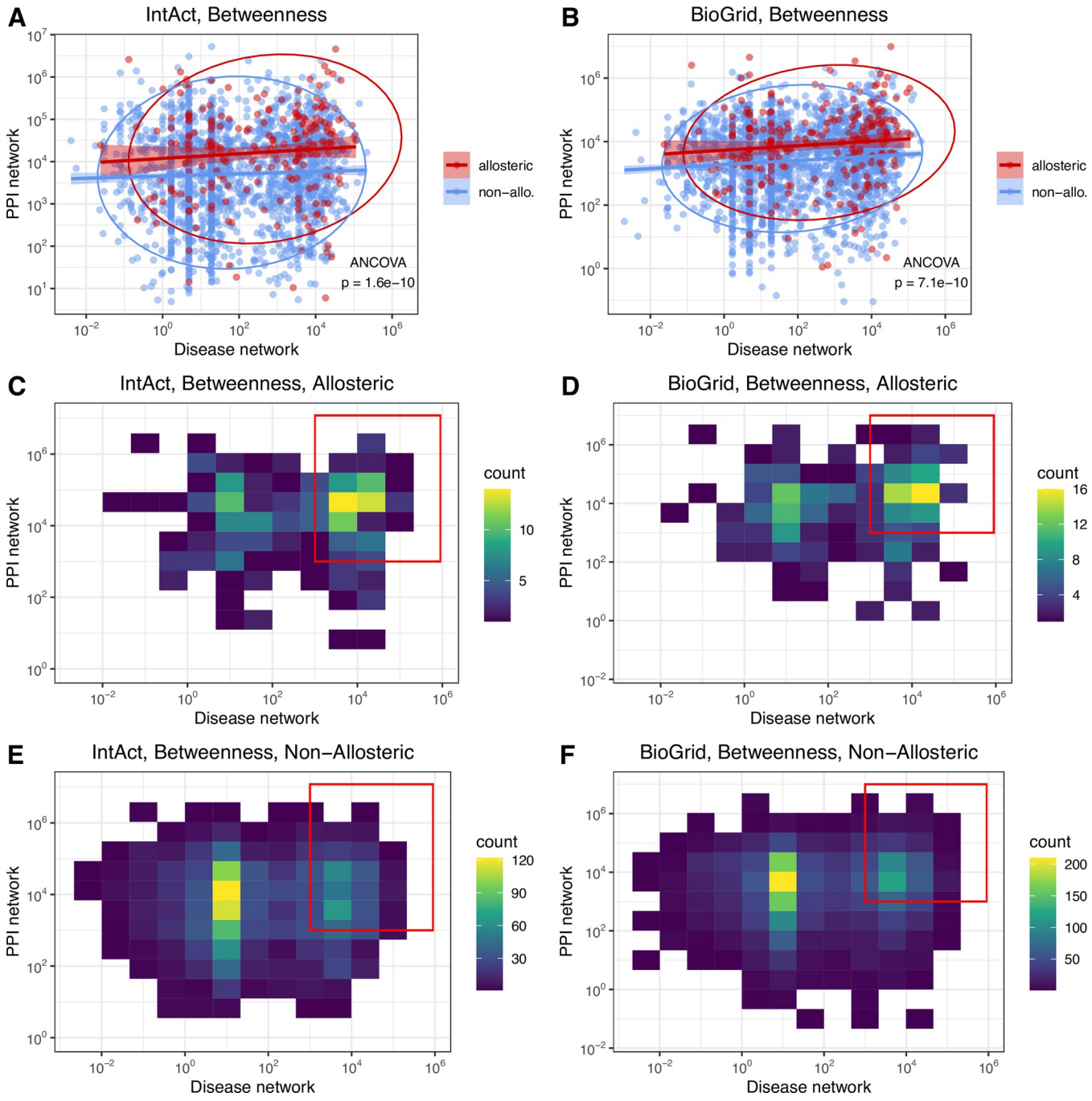

Fig 7. Allosteric proteins have higher betweenness both in PPI and disease networks than non-allosteric proteins, even though the overall correlation between the centralities of the two network types is weak. A-B) Correlations between PPI and disease protein network centralities. Proteins with 0 betweenness centrality in any of the two networks were excluded. **C-D)** Density plots of allosteric proteins. 43.5% (IntAct) and 39.9% (BioGrid) of proteins have betweenness centrality higher than 1000 in both networks (red rectangle). **E-F)** Density plots of non-allosteric proteins. 23.6% (IntAct) and 23.3% (BioGrid) of proteins have betweenness centrality higher than 1000 in both networks (red rectangle). Data ellipses on panel A and B were drawn with the stat_ellipse() function of ggplot2 (R), with default settings, ANCOVA was performed on log transformed data.

https://doi.org/10.1371/journal.pcbi.1009806.g007

these results indicate that inactivating allosteric proteins is somewhat more deleterious than inactivating non-allosteric disease proteins, but there are no significant differences in the essentiality of these two protein groups.

Discussion

Our results indicate that the currently known allosteric proteins are much more important in genetic diseases than other proteins: they cause disease more frequently, and are associated with more diseases than non-allosteric proteins, primarily in cancers, hematopoietic and (cardio)vascular diseases (Figs 1 and 2). In addition, they are central in disease-protein networks, are responsible for more disease comorbidities than non-allosteric proteins (Figs 2 and S5). The analyses of homologs of known allosteric proteins also indicate that there is an association between being allosteric and being involved in disease, and that the pattern is unlikely to be caused by study bias (Fig 3). Importantly, we also observed a clear enrichment of variants near allosteric proteins in cancer GWA studies (Fig 4), indicating that allosteric proteins also have higher than average contribution to complex disease.

The high importance of allosteric proteins in disease can be caused by at least two processes. First, as they have distinct dynamical properties [64] (i.e. signal transduction pathways that connect allosteric and orthosteric sites within the protein), they might have more residues that can make them inactive when mutated, thus they might simply be more vulnerable to mutations. The fact that allosteric proteins have more known pathogenic mutations and are somewhat more conserved (Fig 5) does support this. The analysis of the dynamical properties of the proteins in our dataset indicates that, as it has been suggested previously [64], mutations that alter dynamics are important contributors to disease, i.e. pathogenic mutations are overrepresented in residues important in transducing motions within proteins, and also in interfaces (Fig 6). However, except the more pronounced clustering of mutations in residue-communities (which does suggest that to some degree allostery itself contributes to disease), we found no dramatic differences between allosteric and non-allosteric proteins, suggesting that proteins involved in disease are generally dynamic, also in the non-allosteric group.

A second possibility is that known allosteric proteins are characterised by central positions in cellular networks, that result in higher-than-average contribution to disease and suggests that allostery evolves to regulate the most important nodes in the networks. Our finding, that allosteric proteins have high betweenness centralities both in disease and PPI networks do support this hypothesis (Figs 7 and S10), as well as their somewhat lower tolerance for loss-of-function (PSC) mutations (S11 Fig). It has been argued that allostery is common among signalling proteins [21,22], and signalling pathways are a subset of protein-protein interaction networks. Since signalling proteins are frequent contributors to disease, we also examined whether involvement in signaling is alone sufficient to explain the observed patterns. We found that removing all proteins involved in signaling from the data (or all kinases and GPCRs) does not change our results qualitatively (S4 Fig), therefore signaling alone cannot explain the pattern, and it is likely to be caused also by other or more general processes, for example evolutionary pressure for the emergence of fine-tuned regulation of nodes of high importance in cellular networks.

The number of studies that investigate the mechanistic aspects of allostery is large, however only a few have examined the biological factors that are responsible for its evolution. It has been suggested that the origins of allostery are rooted in protein evolvability itself and not necessarily in function [65], and that potentially allosteric residues and pathways are "prewired" in proteins in the form of networks of coevolving residues [66,67], which are conditionally neutral, and emerge due to adaptations to fluctuating conditions [65]. However, recently it has been demonstrated that allostery emerges particularly frequently in proteins which have a large number of homologs with conserved binding sites in the genome [18] (like kinases), suggesting that a major evolutionary force behind the emergence of allostery is preventing cross-reactivity among them. The finding that allostery is a characteristic of nodes with high

betweenness in disease and PPI networks indicates that network centrality, i.e., evolutionary pressure for regulating nodes critical in disease and PPIs is also a driving force behind its emergence.

Methods

Data sources

The list of known human allosteric proteins was downloaded from the Allosteric Database (v4.10, http://mdl.shsmu.edu.cn/ASD/), altogether 835 (from the total 1942). The list of proteins involved in disease was compiled from four sources: we used RefSeq transcripts from ClinVar [25] (https://www.ncbi.nlm.nih.gov/clinvar/) with mutations annotated as "pathogenic" or "likely pathogenic"; OMIM [26] genes (https://www.omim.org/); the list of disease proteins compiled by UniProt [68] (https://www.uniprot.org/diseases/); and RefSeq transcripts from HGMD[27] (version 2014.2) with missense mutations annotated as "disease causing mutation". The list of pharmacologically active drug target proteins was downloaded from DrugBank (https://go.drugbank.com/).

Ontology analyses

The human disease ontology [69] file (doid.obo) was downloaded from The Open Biological and Biomedical Ontology Foundry (http://www.obofoundry.org/ontology/doid.html). This ontology file was used to standardise disease nomenclature in all further analyses. For each protein of the full list of disease associated proteins, its MedGen identifiers, OMIM identifiers, and disease names were mapped to disease ontology IDs (DOIDs), provided by the doid.obo file. To reduce inconsistencies in disease names (HGMD does not provide MedGen or OMIM IDs, only disease names/descriptions), disease names were converted to lowercase, and commas were removed. Only those proteins were included in the ontology analyses where their MedGen/OMIM/disease names could be converted to a disease ontology ID, altogether 5045. Next, the full list of disease ontology IDs of every protein was determined by traversing the disease ontology hierarchy to the highest level "Disease" term, and a standard ID enrichment analysis (similar to a GO enrichment analysis) was performed with the GeneMerge tool [70]. In addition, we performed a GO enrichment analysis on the same set of proteins, using the mappings between UniProt proteins and GO Molecular Function terms, provided by the standard human gene ontology annotation file (goa_human.gaf), which was downloaded from http://geneontology.org/docs/downloads. The significance of the overlaps between the protein lists of enriched disease ontology and gene ontology terms was calculated using Fisher's exact test of the GeneOverlap R/Bioconductor package.

For visualisation and redundancy filtering of the significant molecular function GO terms, we used the REVIGO tool [71] with medium similarity setting. For disease ontology terms we used a methodology comparable to REVIGO, except that semantic similarities between the disease ontology terms were calculated with an edge-based method [72], which in the case of the disease ontology was more robust than node/information content based methods. We calculated the length of the shortest path between every possible pair of the enriched terms with igraph [73]. Next, we calculated a similarity metric for every pair using the path lengths: $sim = (length_{max} — length_{ij} + 1) / length_{max}$ where $length_{ij}$ is the number of nodes on the shortest path between nodes i and j, and $length_{max}$ is the length of the longest path. We filtered out the most redundant nodes by keeping only the more significant one of the node pairs with $sim > 0.9$. (In the case of similar significances, the term located higher in the ontology was kept.) Finally, we used multidimensional scaling (cmdscale function of R) to reduce the dimensionality of the matrix of the similarities, and plot them a in a two-dimensional semantic space.

Construction and visualization of disease-protein networks

Disease-protein networks were constructed using an approach outlined by Goh et al. [30] For every protein we determined the number of diseases it contributes to, and a network was constructed where the nodes are proteins, which are connected with an edge if they are associated with the same disease. The number of diseases shared by the two proteins were used as the weight of the edge. The list of diseases associated with each protein was determined with a comparable, but more restrictive method than used for DOID enrichment analysis. Similarly, using MedGen IDs, OMIM IDs, and disease names we identified the list of disease ontology IDs associated with each gene. For diseases names a "bag of words" approach was used, i.e. disease names containing the same words were treated as similar, irrespectively of word order. This corrects for common inconsistencies in disease names like "myopathy, congenital" and "congenital myopathy", in cases where MedGen or OMIM IDs were not present. In addition, in cases where proteins were associated with multiple disease ontology IDs, we filtered out non-independent ones, leaving in only the IDs at the lower levels of the ontology, by removing higher level terms that are located on the same path towards the top-level"disease" (DOID:4) term of the ontology. For example, if a protein is associated with the terms "myeloid leukemia" and "leukemia", then only "myeloid leukemia" was used.

The betweenness centrality of the network nodes (proteins) was determined using igraph [73] (v0.8.3). The largest connected component of the network was visualized using the OpenOrd algorithm of the Gephi tool (v0.92), smaller components of the network with the DrL layout algorithm of igraph.

Identification of quaternary structure of proteins

We used the first biounit of every PDB entry to define its quaternary structure. Proteins having multiple structures with different quaternary structures were classified as heteromers if at least one of their structures is a heteromer, homomers if they have no heteromeric structure and have at least one homomer, and momomers if all of their entries are monomeric. Since many PDB entries contain peptide ligands, we adopted a methodology of the BioLiP database [74], and we used a 30 amino-acid cutoff to identify peptide ligands; chains shorter than that were treated as ligands and not a separate protein chain. Homomers were defined as complexes of multiple identical proteins (with the same UniProt accession), heteromers as complexes of different proteins (two or more different Uniprot accessions), without using a sequence similarity threshold. PDB entries having topological variability between different biounits of the same entry, and biounits having interfaces absent in the asymmetric units were included in the analysis (e.g. homomeric biounits with monomeric asymmetric units).

Identification of orthology and paralogy

We used the orthogroups of the eggNOG database (v.5, http://eggnog5.embl.de/#/app/home) [38] in the analysis. Sequence identifiers of eggNOG were mapped to UniProt protein IDs using the e5.sequence_aliases.tsv file provided by eggNOG. From the 1942 proteins of ASD 1562 could be mapped to eggNOG. In the analysis of paralogs, we used four taxonomic datasets: mammals, vertebrates, metazoans and eukaryotes, while in the analysis of orthologs three: mammals, metazoans and eukaryotes, because the vertebrate dataset has too few non-human (and at the same time non-mammalian) allosteric proteins for a meaningful analysis. In each taxonomic dataset, we excluded orthogroups that have only proteins of the lower taxonomic level (thus, in the vertebrate dataset orthogroups that contain only mammalian proteins were not used). Human orthologs of non-human allosteric proteins were identified in orthogroups that do not have any human allosteric proteins, while (in)paralogs were identified in

orthogroups that do have a human allosteric protein, irrespectively whether they also have a non-human allosteric ortholog. The duplicated human allosteric protein set complements the paralogs, i.e. they include the allosteric proteins that have a non-allosteric paralog. For the non-allosteric human reference set we used all non-allosteric human proteins of the orthogroups included in each taxonomic level, and we excluded also all orthologs and paralogs of a given taxonomic level (i.e. in the analysis of metazoan orthologs, only non-mammalian orthologs were included as putatively allosteric, but from the non-allosteric reference set also the orthologs of mammalian allosteric proteins were excluded).

Calculating the number of comorbidities for proteins

We used only proteins that are involved in minimum two diseases (DOID terms, see above) in the disease-protein network. For every disease term we identified all their parental terms, up to the highest level "disease"term (DOID:4), using the disease ontology (doid.obo file). Next, we converted the disease ontology terms to MedGen IDs, using the mapping between DOID terms and MedGen/OMIM IDs of the disease ontology. OMIM IDs were converted to MedGen IDs using the MedGen/HPO/OMIM mapping file provided by MedGen (available at https://ftp.ncbi.nlm.nih.gov/pub/medgen/). The comorbidity network [43] based on FDA Adverse Event Reporting System (FAERS) was downloaded from http://nlp.case.edu/public/data/FAERS_comb/. The network contains 25217 edges between 1059 MedGen IDs describing diseases and conditions. The number of comorbidities for every protein was calculated by identifying every possible pairwise combination of their MedGen IDs, and counting the number of combinations that are present in the comorbidity network.

Calculating enrichment of GWAS variants

The GRCh38 release of the human genome, and the corresponding GTF file of genes were downloaded from Ensembl, release 101. Significant variants reported by all known cancer GWAS were downloaded from the GWAS catalog [46] (https://www.ebi.ac.uk/gwas/); in each study only variants with $p < 10^{-8}$ were used. The type of cancer was assigned as the "mapped trait" column of the associations file, after filtering out studies with disease descriptions that were not explicitly focusing on cancer risk genes, for example "Toxicity response to radiotherapy in prostate cancer". Additionally, we excluded all studies using targeted genotyping arrays (Oncoarray), because they are not powered to be genome-wide. If two cancer GWAS had exactly the same variants, we used only one study in the analysis, but overlaps between the variant sets of two studies were permitted. We identified the closest protein coding exon for every known base of the genome, and for every variant in the GWA studies. Bases and variants that were further than 100kb from the closest protein coding exon were not assigned to any gene. If the variants of two GWA studies of the same cancer type were located near the same set of genes (i.e. identified the same sets of candidate causal genes, or a subset of the other), then only one study was used (with the higher number genes, or cancer cases). Ensembl genes were mapped to Uniprot proteins with the ID mapping tool of Uniprot. The enrichment of variants that are located closest to a particular protein type (e.g., allosteric) was calculated as the fraction of all variants of the pooled GWA studies, divided by the fraction of bases in the genome (% in GWAS / % in the genome). Pooling data from several GWAS allows for the meaningful analysis of GWAS with a few variants (and the comparison with larger ones), because in these studies a single variant can result in large changes in enrichment: a single allosteric variant in a GWAS with five variants means 5.26x enrichment compared to the genomic average. In the gene enrichment analysis, a single list of genes was compiled for every cancer type from all of its GWA studies, and the enrichment of allosteric, disease and other proteins was calculated

compared to their frequencies in all protein coding genes of the genome. Significances of enrichment were calculated with the "metafor" R package using the heterogeneity test of the Mantel-Haenszel method (rma.mh).

Calculating conservation

We used the eggNOG (v5) orthology database [38] to calculate conservation. We downloaded the mammalian dataset, and calculated the level of conservation of human proteins using the multiple-alignments of the orthogroups. For every human protein in the alignment, we calculated their similarity with all non-primate proteins, as the fraction of the similar residues in the total number of aligned (i.e. non-gapped) residues in their pairwise alignment. Primates were excluded to ensure that all proteins in the alignment had approximately the same time to accumulate differences compared to their human ortholog, as most extant placental mammal groups diverged from primates in the mammalian radiation 70–80 mya [75]. The conservation of each human protein was defined as the average similarity with all other non-primate proteins of the alignment.

Identification of conserved domains and residue interactions

Raw conserved domains of disease associated proteins were identified with the hmmscan tool of HMMER 3.3[76], using Pfam-33.1 and an e-value cutoff (—domE) 0.001, and bitscore cutoff 22. When several domains mapped to the same region, we used only the most significant (typically the longest) hit. Next we mapped PDB structures to the identified Pfam domains with muscle [77], using structures with the highest resolution (and no worse than 3Å), and with the highest coverage, with a minimum of 90%; excluding domains shorter than 50 residues. Residue interactions within protein structures were calculated with the RINerator tool [78]. Residue interaction matrices were calculated as in [19] by dividing the length of each domain into 50 units. The interaction scores of each cell of the resulting 50 x 50 matrix were calculated by adding the number of residue-interactions of residue-pairs that fall into any given cell, using the *_nrint.ea files produced by RINerator. The number of residue-interactions was divided by the total number of interactions in the domain, to correct for differences in domain size. Additionally, the interaction scores of non-allosteric domains were normalized by the proportion of allosteric to non-allosteric proteins, to correct for the difference in the number Pfam domains in allosteric and non-allosteric proteins. In the final analyses Log transformed scores were used, to reduce the variation between cells.

Analysis of dynamics

For proteins associated with disease (allosteric and non-allosteric) we selected structures that cover at least 75% of the length of the protein, has the largest possible ligand (if it has a ligand), resolution is better than 3Å, and the structure has less than 12 chains. Altogether 563 proteins have such structures, 131 allosteric and 432 non-allosteric (see S5 Table). We used STRESS [53] to identify communities and critical residues in the structures. The MMTK-2.7.12 dependency of STRESS was installed with the install script by the lab of Pierre-Nicholas Roy at https://github.com/roygroup/mmtk_install. Before running STRESS the structures were preprocessed: if they were part of a protein complex than only the monomeric unit was used (a single chain with the disease protein), and we processed the structure with the DockPrep tool of Chimera [79], to add hydrogens, complete incomplete side chains, and remove residues with low occupancy when residues with alternative locations are present. In addition, similarly to [18] we modified STRESS to use the center-of-mass of residues instead of their C-α atoms, using the Bio3D R package [80] and in-house Perl scripts, as it was shown to significantly

improve performance both in molecular dynamics [81] and elastic-network based simulations [18].

Protein-protein interaction networks

IntAct [82] (10th Nov. 2020, https://www.ebi.ac.uk/intact/home) and BioGrid [83] (v.4.1.190, https://thebiogrid.org/) interaction datasets were downloaded from their corresponding websites. Only protein-protein interactions were used for both (annotated as "physical association" for IntAct and "physical" for BioGrid). Alltogether 139010 distinct interactions were identified in IntAct and 434534 in Biogrid; a weight 1 was assigned to all edges of the network. Betweenness and degree centralities of the network were calculated with igraph.

Determining loss-of-function intolerance

The gnomAD exome dataset [63] was downloaded from https://gnomad.broadinstitute.org/downloads. We discarded all variants from the dataset that did not pass all quality filters, or where the allele frequency was given as 0. For every Ensembl protein in the dataset we determined the number of PSCs that map to them (annotated as "stop_gained"), the length of the protein, and the average allele frequency of the PSCs per protein. When multiple SNPs in the same codon resulted in the same amino acid change (for example in the case of Tryptophan), the allele frequency of the amino acid change was determined as the sum of the allele frequencies of the variants. From the protein isoforms of the same genes, we used only the main isoform, as defined by the Appris database [84] (https://apprisws.bioinfo.cnio.es/landing_page/, Gencode19/Ensembl74 dataset). Ensembl proteins were mapped to Uniprot proteins by the ID mapping service of Uniprot.

Supporting information

S1 Fig. Matrix of disease ontology and molecular function GO terms, where their corresponding protein lists have significant overlap ($p < 0.05$, Bonferroni-correction). Most terms with significant overlaps are the result of associations between cancers and protein kinases.
(TIF)

S2 Fig. Allosteric proteins are associated with more diseases than non-allosteric ones (**A**), also when cancers (**B**), or non-cancerous diseases (**C**) are analysed separately. All statistical comparisons were made with Wilcoxon tests.
(TIF)

S3 Fig. The fraction of proteins with zero betweenness centrality is somewhat smaller in allosteric proteins than in non-allosteric ones.
(TIF)

S4 Fig. Excluding proteins annotated with "signal transduction" biological process GO term (and its lower-level child terms) does not change qualitatively the patterns observed on Fig 2: allosteric proteins are overrepresented in disease, and have central positions in disease networks.
(TIF)

S5 Fig. Allosteric proteins are more likely to cause comorbidities than non-allosteric ones. A) A significantly higher fraction of allosteric proteins is involved in known comorbidities than in non-allosteric proteins (p—tests of proportions). Only proteins associated with minimum two diseases were included. **B)** The number of comorbidities per protein is higher in

allosteric proteins (p—Wilcoxon test). **C)** The relationship between the number of diseases and number of comorbidities is not qualitatively different in the two protein types, indicating that primarily the higher number diseases per protein is responsible for the higher number of comorbidities in allosteric proteins.
(TIF)

S6 Fig. The analysis which uses all, at least partially nonredundant studies of the GWAS catalog shows similar enrichments as the analyses that use only the single study with the highest number of cases, or a single nonredundant gene list for every cancer type (Fig 3).
(TIF)

S7 Fig. Human orthologs of mammalian allosteric proteins have similarly high number of pathogenic mutations as human allosteric proteins (A), and the relationship between the number of mutations and number of diseases is also qualitatively similar (B).
(TIF)

S8 Fig. Allosteric proteins in ClinVar have more PubMed articles per disease than non-allosteric proteins. A) All proteins. **B)** Kinases **C)** Drug targets. The number of diseases was calculated using ClinVar only. The number of publications for each protein was calculated by summing all citation IDs of their "pathogenic" and "likely pathogenic" mutations, using the var_citations.txt file provided by ClinVar.
(TIF)

S9 Fig. Pathogenic mutations are not distributed uniformly across communities. A) In allosteric proteins the difference between the expected and observed number of mutated communities is significantly higher than in non-allosteric proteins (p = 9.37e-04, ANCOVA), indicating stronger pathogenicity of mutations in allosteric proteins. **B** and **C)** Location of pathogenic mutations and communities in a monomeric unit of the Transitional Endoplasmic Reticulum ATPase (PDB ID: 5ftj).
(TIF)

S10 Fig. Disease associated allosteric proteins have higher centralities in protein-protein interaction (PPI) networks than non-allosteric proteins. A) Betweenness centralities based on IntAct PPIs. Note that the 'All' category includes also the proteins where quaternary structure is not known (i.e. are not present in the PDB). **B)** Degree centralities (number of interactions) based on IntAct PPIs. **C-D)** Betweenness and degree centralities based on BioGrid PPIs.
(TIF)

S11 Fig. Premature stop codon (PSC) densities and their mean allele frequencies in disease proteins. Allosteric proteins are characterized by somewhat lower PSC densities and allele frequencies than non-allosteric ones.
(TIF)

S1 Table. Disease ontology.
(XLSX)

S2 Table. Molecular Function Ontology.
(XLSX)

S3 Table. Disease network.
(XLSX)

S4 Table. GWAS Studies.
(XLSX)

S5 Table. PDB structures.
(XLSX)

Acknowledgments

We thank Sergio Ruiz-Carmona and Loïc Lannelongue for critical reading of the manuscript.

Author Contributions

Conceptualization: György Abrusán.

Formal analysis: György Abrusán.

Investigation: György Abrusán.

Resources: David B. Ascher.

Software: György Abrusán.

Visualization: György Abrusán.

Writing – original draft: György Abrusán.

Writing – review & editing: György Abrusán, David B. Ascher, Michael Inouye.

References

1. Gunasekaran K, Ma B, Nussinov R. Is allostery an intrinsic property of all dynamic proteins? Proteins Struct Funct Bioinforma. 2004; 57: 433–443. https://doi.org/10.1002/prot.20232 PMID: 15382234
2. Guarnera E, Berezovsky IN. On the perturbation nature of allostery: sites, mutations, and signal modulation. Curr Opin Struct Biol. 2019; 56: 18–27. https://doi.org/10.1016/j.sbi.2018.10.008 PMID: 30439587
3. Sapienza PJ, Popov KI, Mowrey DD, Falk BT, Dokholyan NV, Lee AL. Inter-Active Site Communication Mediated by the Dimer Interface β-Sheet in the Half-the-Sites Enzyme, Thymidylate Synthase. Biochemistry. 2019; 58: 3302–3313. https://doi.org/10.1021/acs.biochem.9b00486 PMID: 31283187
4. Changeux J-P. 50 years of allosteric interactions: the twists and turns of the models. Nat Rev Mol Cell Biol. 2013; 14: 819–829. https://doi.org/10.1038/nrm3695 PMID: 24150612
5. Thirumalai D, Hyeon C, Zhuravlev PI, Lorimer GH. Symmetry, Rigidity, and Allosteric Signaling: From Monomeric Proteins to Molecular Machines. Chem Rev. 2019; 119: 6788–6821. https://doi.org/10.1021/acs.chemrev.8b00760 PMID: 31017391
6. Motlagh HN, Wrabl JO, Li J, Hilser VJ. The ensemble nature of allostery. Nature. 2014; 508: 331–339. https://doi.org/10.1038/nature13001 PMID: 24740064
7. Dokholyan NV. Controlling Allosteric Networks in Proteins. Chem Rev. 2016; 116: 6463–6487. https://doi.org/10.1021/acs.chemrev.5b00544 PMID: 26894745
8. Kornev AP, Taylor SS. Dynamics-Driven Allostery in Protein Kinases. Trends Biochem Sci. 2015; 40: 628–647. https://doi.org/10.1016/j.tibs.2015.09.002 PMID: 26481499
9. Tsai C-J, Nussinov R. A Unified View of "How Allostery Works." PLOS Comput Biol. 2014; 10: e1003394. https://doi.org/10.1371/journal.pcbi.1003394 PMID: 24516370
10. Guo J, Zhou H-X. Protein Allostery and Conformational Dynamics. Chem Rev. 2016; 116: 6503–6515. https://doi.org/10.1021/acs.chemrev.5b00590 PMID: 26876046
11. Nussinov R, Tsai C-J. Allostery in Disease and in Drug Discovery. Cell. 2013; 153: 293–305. https://doi.org/10.1016/j.cell.2013.03.034 PMID: 23582321
12. Zhang J, Nussinov R, editors. Protein Allostery in Drug Discovery. Springer Singapore; 2019. https://doi.org/10.1007/978-981-13-8719-7
13. Guarnera E, Berezovsky IN. Allosteric drugs and mutations: chances, challenges, and necessity. Curr Opin Struct Biol. 2020; 62: 149–157. https://doi.org/10.1016/j.sbi.2020.01.010 PMID: 32062398
14. Thal DM, Glukhova A, Sexton PM, Christopoulos A. Structural insights into G-protein-coupled receptor allostery. Nature. 2018; 559: 45–53. https://doi.org/10.1038/s41586-018-0259-z PMID: 29973731

15. Leroux AE, Biondi RM. Renaissance of Allostery to Disrupt Protein Kinase Interactions. Trends Biochem Sci. 2020; 45: 27–41. https://doi.org/10.1016/j.tibs.2019.09.007 PMID: 31690482

16. Yang J-S, Seo SW, Jang S, Jung GY, Kim S. Rational Engineering of Enzyme Allosteric Regulation through Sequence Evolution Analysis. PLOS Comput Biol. 2012; 8: e1002612. https://doi.org/10.1371/journal.pcbi.1002612 PMID: 22807670

17. Liu X, Lu S, Song K, Shen Q, Ni D, Li Q, et al. Unraveling allosteric landscapes of allosterome with ASD. Nucleic Acids Res. 2020; 48: D394–D401. https://doi.org/10.1093/nar/gkz958 PMID: 31665428

18. Abrusán G, Marsh JA. Ligand-Binding-Site Structure Shapes Allosteric Signal Transduction and the Evolution of Allostery in Protein Complexes. Mol Biol Evol. 2019; 36: 1711–1727. https://doi.org/10.1093/molbev/msz093 PMID: 31004156

19. Abrusán G, Marsh JA. Ligand Binding Site Structure Shapes Folding, Assembly and Degradation of Homomeric Protein Complexes. J Mol Biol. 2019; 431: 3871–3888. https://doi.org/10.1016/j.jmb.2019.07.014 PMID: 31306664

20. Nussinov R, Tsai C-J, Ma B. The Underappreciated Role of Allostery in the Cellular Network. Annu Rev Biophys. 2013; 42: 169–189. https://doi.org/10.1146/annurev-biophys-083012-130257 PMID: 23451894

21. Nussinov R, Tsai C-J, Liu J. Principles of Allosteric Interactions in Cell Signaling. J Am Chem Soc. 2014; 136: 17692–17701. https://doi.org/10.1021/ja510028c PMID: 25474128

22. Bu Z, Callaway DJE. Chapter 5—Proteins MOVE! Protein dynamics and long-range allostery in cell signaling. In: Donev R, editor. Advances in Protein Chemistry and Structural Biology. Academic Press; 2011. pp. 163–221. https://doi.org/10.1016/B978-0-12-381262-9.00005-7 PMID: 21570668

23. Shah NH, Kuriyan J. Understanding molecular mechanisms in cell signaling through natural and artificial sequence variation. Nat Struct Mol Biol. 2019; 26: 25–34. https://doi.org/10.1038/s41594-018-0175-9 PMID: 30598552

24. Shen Q, Cheng F, Song H, Lu W, Zhao J, An X, et al. Proteome-Scale Investigation of Protein Allosteric Regulation Perturbed by Somatic Mutations in 7,000 Cancer Genomes. Am J Hum Genet. 2017; 100: 5–20. https://doi.org/10.1016/j.ajhg.2016.09.020 PMID: 27939638

25. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res. 2020; 48: D835–D844. https://doi.org/10.1093/nar/gkz972 PMID: 31777943

26. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Res. 2019; 47: D1038–D1043. https://doi.org/10.1093/nar/gky1151 PMID: 30445645

27. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017; 136: 665–677. https://doi.org/10.1007/s00439-017-1779-6 PMID: 28349240

28. Torkamani A, Verkhivker G, Schork NJ. Cancer driver mutations in protein kinase genes. Cancer Lett. 2009; 281: 117–127. https://doi.org/10.1016/j.canlet.2008.11.008 PMID: 19081671

29. Bhullar KS, Lagarón NO, McGowan EM, Parmar I, Jha A, Hubbard BP, et al. Kinase-targeted cancer therapies: progress, challenges and future directions. Mol Cancer. 2018; 17: 48. https://doi.org/10.1186/s12943-018-0804-2 PMID: 29455673

30. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proc Natl Acad Sci. 2007; 104: 8685–8690. https://doi.org/10.1073/pnas.0701361104 PMID: 17502601

31. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. PLOS Comput Biol. 2007; 3: e59. https://doi.org/10.1371/journal.pcbi.0030059 PMID: 17447836

32. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005; 39: 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725 PMID: 16285863

33. Gabaldón T, Koonin EV. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 2013; 14: 360–366. https://doi.org/10.1038/nrg3456 PMID: 23552219

34. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016; 44: D286–D293. https://doi.org/10.1093/nar/gkv1248 PMID: 26582926

35. Rödelsperger C, Prabh N, Sommer RJ. New Gene Origin and Deep Taxon Phylogenomics: Opportunities and Challenges. Trends Genet. 2019; 35: 914–922. https://doi.org/10.1016/j.tig.2019.08.007 PMID: 31610892

36. Ohno S. Evolution by gene duplication. London, New York: Springer Verlag; 1970.

37. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 2010; 11: 97–108. https://doi.org/10.1038/nrg2689 PMID: 20051986

38. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019; 47: D309–D314. https://doi.org/10.1093/nar/gky1085 PMID: 30418610

39. Chen W-H, Trachana K, Lercher MJ, Bork P. Younger Genes Are Less Likely to Be Essential than Older Genes, and Duplicates Are Less Likely to Be Essential than Singletons of the Same Age. Mol Biol Evol. 2012; 29: 1703–1706. https://doi.org/10.1093/molbev/mss014 PMID: 22319151

40. He X, Zhang J. Higher Duplicability of Less Important Genes in Yeast Genomes. Mol Biol Evol. 2006; 23: 144–151. https://doi.org/10.1093/molbev/msj015 PMID: 16151181

41. Park J, Lee D-S, Christakis NA, Barabási A-L. The impact of cellular networks on disease comorbidity. Mol Syst Biol. 2009; 5: 262. https://doi.org/10.1038/msb.2009.16 PMID: 19357641

42. Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. Nat Rev Genet. 2016; 17: 615–629. https://doi.org/10.1038/nrg.2016.87 PMID: 27498692

43. Zheng C, Xu R. Large-scale mining disease comorbidity relationships from post-market drug adverse events surveillance data. BMC Bioinformatics. 2018; 19: 500. https://doi.org/10.1186/s12859-018-2468-8 PMID: 30591027

44. Katsanis N. The continuum of causality in human genetic disorders. Genome Biol. 2016; 17: 233. https://doi.org/10.1186/s13059-016-1107-9 PMID: 27855690

45. Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: current insights and future perspectives. Nat Rev Cancer. 2017; 17: 692–704. https://doi.org/10.1038/nrc.2017.82 PMID: 29026206

46. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019; 47: D1005–D1012. https://doi.org/10.1093/nar/gky1120 PMID: 30445434

47. Stacey D, Fauman EB, Ziemek D, Sun BB, Harshfield EL, Wood AM, et al. ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. Nucleic Acids Res. 2019; 47: e3–e3. https://doi.org/10.1093/nar/gky837 PMID: 30239796

48. Chen C-H, Zheng R, Tokheim C, Dong X, Fan J, Wan C, et al. Determinants of transcription factor regulatory range. Nat Commun. 2020; 11: 2472. https://doi.org/10.1038/s41467-020-16106-x PMID: 32424124

49. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature. 2013; 501: 355–364. https://doi.org/10.1038/nature12627 PMID: 24048068

50. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018; 15: 81–94. https://doi.org/10.1038/nrclinonc.2017.166 PMID: 29115304

51. Tee W-V, Guarnera E, Berezovsky IN. On the Allosteric Effect of nsSNPs and the Emerging Importance of Allosteric Polymorphism. J Mol Biol. 2019; 431: 3933–3942. https://doi.org/10.1016/j.jmb.2019.07.012 PMID: 31306666

52. Kumar S, Clarke D, Gerstein MB. Leveraging protein dynamics to identify cancer mutational hotspots using 3D structures. Proc Natl Acad Sci. 2019; 116: 18962–18970. https://doi.org/10.1073/pnas.1901156116 PMID: 31462496

53. Clarke D, Sethi A, Li S, Kumar S, Chang RWF, Chen J, et al. Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. Structure. 2016; 24: 826–837. https://doi.org/10.1016/j.str.2016.03.008 PMID: 27066750

54. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A. 2015; 112: E5486–5495. https://doi.org/10.1073/pnas.1516373112 PMID: 26392535

55. Buljan M, Blattmann P, Aebersold R, Boutros M. Systematic characterization of pan-cancer mutation clusters. Mol Syst Biol. 2018; 14: e7974. https://doi.org/10.15252/msb.20177974 PMID: 29572294

56. Niu B, Scott AD, Sengupta S, Bailey MH, Batra P, Ning J, et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat Genet. 2016; 48: 827–837. https://doi.org/10.1038/ng.3586 PMID: 27294619

57. Dincer C, Kaya T, Keskin O, Gursoy A, Tuncbag N. 3D spatial organization and network-guided comparison of mutation profiles in Glioblastoma reveals similarities across patients. PLOS Comput Biol. 2019; 15: e1006789. https://doi.org/10.1371/journal.pcbi.1006789 PMID: 31527881

58. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical

Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015; 17: 405–423. https://doi.org/10.1038/gim.2015.30 PMID: 25741868

59. David A, Razali R, Wass MN, Sternberg MJE. Protein–protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum Mutat. 2012; 33: 359–363. https://doi.org/10.1002/humu.21656 PMID: 22072597

60. Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, et al. Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. Cell. 2015; 161: 647–660. https://doi.org/10.1016/j.cell.2015.04.013 PMID: 25910212

61. Bergendahl LT, Gerasimavicius L, Miles J, Macdonald L, Wells JN, Welburn JPI, et al. The role of protein complexes in human genetic disease. Protein Sci. 2019; 28: 1400–1411. https://doi.org/10.1002/pro.3667 PMID: 31219644

62. Joy MP, Brock A, Ingber DE, Huang S. High-Betweenness Proteins in the Yeast Protein Interaction Network. In: Journal of Biomedicine and Biotechnology [Internet]. 2005 [cited 25 Dec 2020]. https://doi.org/10.1155/JBB.2005.96 PMID: 16046814

63. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581: 434–443. https://doi.org/10.1038/s41586-020-2308-7 PMID: 32461654

64. Campitelli P, Modi T, Kumar S, Ozkan SB. The Role of Conformational Dynamics and Allostery in Modulating Protein Evolution. Annu Rev Biophys. 2020; 49: null. https://doi.org/10.1146/annurev-biophys-052118-115517 PMID: 32075411

65. Raman AS, White KI, Ranganathan R. Origins of Allostery and Evolvability in Proteins: A Case Study. Cell. 2016; 166: 468–480. https://doi.org/10.1016/j.cell.2016.05.047 PMID: 27321669

66. Reynolds KA, McLaughlin RN, Ranganathan R. Hot Spots for Allosteric Regulation on Protein Surfaces. Cell. 2011; 147: 1564–1575. https://doi.org/10.1016/j.cell.2011.10.049 PMID: 22196731

67. Pincus D, Pandey JP, Feder ZA, Creixell P, Resnekov O, Reynolds KA. Engineering allosteric regulation in protein kinases. Sci Signal. 2018; 11. https://doi.org/10.1126/scisignal.aar3250 PMID: 30401787

68. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021; 49: D480–D489. https://doi.org/10.1093/nar/gkaa1100 PMID: 33237286

69. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015; 43: D1071–D1078. https://doi.org/10.1093/nar/gku1011 PMID: 25348409

70. Castillo-Davis CI, Hartl DL. GeneMerge—post-genomic analysis, data mining, and hypothesis testing. Bioinformatics. 2003; 19: 891–892. https://doi.org/10.1093/bioinformatics/btg114 PMID: 12724301

71. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLOS ONE. 2011; 6: e21800. https://doi.org/10.1371/journal.pone.0021800 PMID: 21789182

72. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic Similarity in Biomedical Ontologies. PLOS Comput Biol. 2009; 5: e1000443. https://doi.org/10.1371/journal.pcbi.1000443 PMID: 19649320

73. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst. 2006; 1695.

74. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Res. 2013; 41: D1096–D1103. https://doi.org/10.1093/nar/gks966 PMID: 23087378

75. Upham NS, Esselstyn JA, Jetz W. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLOS Biol. 2019; 17: e3000494. https://doi.org/10.1371/journal.pbio.3000494 PMID: 31800571

76. Eddy SR. Accelerated Profile HMM Searches. PLOS Comput Biol. 2011; 7: e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361

77. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32: 1792–1797. https://doi.org/10.1093/nar/gkh340 PMID: 15034147

78. Doncheva NT, Klein K, Domingues FS, Albrecht M. Analyzing and visualizing residue networks of protein structures. Trends Biochem Sci. 2011; 36: 179–182. https://doi.org/10.1016/j.tibs.2011.01.002 PMID: 21345680

79. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem. 2004; 25: 1605–1612. https://doi.org/10.1002/jcc.20084 PMID: 15264254

80. Grant BJ, Skjærven L, Yao X-Q. The Bio3D packages for structural bioinformatics. Protein Sci. n/a. https://doi.org/10.1002/pro.3923 PMID: 32734663

81. Vanwart AT, Eargle J, Luthey-Schulten Z, Amaro RE. Exploring residue component contributions to dynamical network models of allostery. J Chem Theory Comput. 2012; 8: 2949–2961. https://doi.org/10.1021/ct300377a PMID: 23139645

82. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014; 42: D358–D363. https://doi.org/10.1093/nar/gkt1115 PMID: 24234451

83. Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021; 30: 187–200. https://doi.org/10.1002/pro.3978 PMID: 33070389

84. Rodriguez JM, Rodriguez-Rivas J, Di Domenico T, Vázquez J, Valencia A, Tress ML. APPRIS 2017: principal isoforms for multiple gene sets. Nucleic Acids Res. 2018; 46: D213–D217. https://doi.org/10.1093/nar/gkx997 PMID: 29069475