TITLE: Landscape of pleiotropic proteins causing human disease: structural and system biology insights

Authors: Sirawit Ittisoponpisan, Eman Alhuzimi, Michael J. E. Sternberg, Alessia David

Structural Bioinformatics Group, Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK

S.I. was supported by a scholarship from the Royal Thai Government; E.A. was supported by a scholarship from King Abdulaziz City for Science and Technology; A.D. was supported by the Medical Research Council UK (grant MR/K021613/1 to A.D.) and the Wellcome Trust (grant WT/104955/Z/14/Z to M.J.E.S.).

Corresponding author

Dr Alessia David MD, PHD

Structural Bioinformatics Group,

Department of Life Sciences,

Imperial College London,

London, SW7 2AZ, UK

Email: alessia.david09@imperial.ac.uk

ABSTRACT

Pleiotropy is the phenomenon by which the same gene can result in multiple phenotypes. Pleiotropic proteins are emerging as important contributors to rare and common disorders. Nevertheless, little is known on the mechanisms underlying pleiotropy and the characteristic of pleiotropic proteins.

We analysed disease-causing proteins reported in UniProt and observed that 12% are pleiotropic (variants in the same protein cause more than one disease). Pleiotropic proteins were enriched in deleterious and rare variants, but not in common variants. Pleiotropic proteins were more likely to be involved in the pathogenesis of neoplasms, neurological and circulatory diseases, and congenital malformations, whereas non-pleiotropic proteins in endocrine and metabolic disorders. Pleiotropic proteins were more essential and had a higher number of interacting partners compared to non-pleiotropic proteins. Significantly more pleiotropic than non-pleiotropic proteins contained at least one intrinsically long disordered region ($p<0.001$). Deleterious variants occurring in structurally disordered regions were more commonly found in pleiotropic, rather than non-pleiotropic proteins.

In conclusion, pleiotropic proteins are an important contributor to human disease. They represent a biologically different class of proteins compared to non-pleiotropic proteins and a better understanding of their characteristics and genetic variants, can greatly aid in the interpretation of genetic studies and drug design.

Key words:

Pleiotropy, genetic variants, human disease, disordered protein region

INTRODUCTION

Pleiotropy is the phenomenon by which the same gene can result in multiple phenotypes. The 'human disease gene network' developed by Goh et al. unveiled a highly shared genetic background between different rare human diseases (Goh et al. 2007). Genetic overlap has also been demonstrated between different common disorders (Solovieff et al. 2013; Ligthart et al. 2015). A recent analysis of genome wide associations studies (GWAS) reported in the GWAS catalogue showed that 16.9% of genes and 4.6% of genetic variants have pleiotropic effects (Sivakumaran et al. 2011). The widespread presence of pleiotropic genes poses a formidable challenge when analysing and prioritizing genes and genetic variations responsible for disease. Furthermore, in-depth understanding of pleiotropic proteins and their mechanisms of action is crucial in drug design, in view of the high risk of an off-target effect posed by these proteins.

Pleiotropy is a widespread phenomenon across all organisms and is a central feature of evolution (Orr 2000; Wagner et al. 2008; Le Nagard et al. 2011) and the development of complex organisms from a relatively limited set of genes (Hodgkin 1998; Wagner and Zhang 2011). Although the phenomenon of pleiotropy has been known for over 100 years (Stearns 2010), the mechanisms by which a single gene can affect multiple traits is still far from being fully understood. Analysis of the properties of pleiotropic protein interaction networks associated with human disease has shown that, similarly to essential proteins, pleiotropic proteins are central in protein-protein interaction networks and have more interactors compared to non-pleiotropic proteins (Zou et al. 2008; Chavali et al. 2010). At molecular level, several mechanisms have been proposed to explain the pleiotropic effect of a gene and its protein product (Hodgkin 1998). Of these, two appear to prevail: 1) the presence, on the same protein, of different domains with different functions (type 1 pleiotropy) and 2) the utilization of the same protein in multiple biological processes (type 2 pleiotropy) (Wagner and Zhang 2011).

It has been shown that genetic variants with a pleiotropic effect are more likely to be located in a coding rather than non-coding region (Sivakumaran et al. 2011). This suggests that analysis of the

three-dimensional structure of pleiotropic proteins and their disease-associated genetic variants may provide insight into the mechanisms underlying pleiotropy. In particular, long structurally disordered regions — protein regions that lack a fixed three-dimensional structure — are compelling candidates to explain pleiotropy. The lack of a fixed 3D structure allows the protein amino acid sequence to adopt different three-dimensional conformations, thus allowing for flexibility and interaction with multiple partners in response to different stimuli and the environment (Dunker et al. 2005; Haynes et al. 2006). Although protein disorder is a well-recognized phenomenon (Dunker et al. 2008; Oldfield and Dunker 2014), its impact on human disease has just began to be recognized (Uversky et al. 2008; Vacic et al. 2012). More importantly, its relation to pleiotropy remains to be analysed.

In this study, we examined the distribution of single amino acid variations (SAVs) in pleiotropic and non-pleiotropic proteins, focusing on the absence of a fixed three-dimensional structure and we analysed the impact of deleterious SAVs occurring in pleiotropic and non-pleiotropic proteins on human disease. Our results show that: pleiotropic proteins are enriched in disease-causing and rare variants, whereas common variants are more likely to be found in non-pleiotropic proteins; b) although deleterious variants occur mainly in structurally ordered regions, a deleterious SAV occurring in a disordered region is more common in a pleiotropic rather than non-pleiotropic protein; c) pleiotropic and non-pleiotropic proteins cause different classes of disorders: neoplasms, neurological and circulatory diseases and congenital malformations are more likely to be caused by pleiotropic rather than non-pleiotropic proteins, whereas endocrine and metabolic disorders are more likely caused by a non-pleiotropic protein.

MATERIALS AND METHODS

Construction of the dataset

We surveyed genotype-phenotype associations reported in UniProt (Humsavar database, release: 2014_08) (Ongoing and future developments at the Universal Protein Resource 2011) for 12,543 proteins. SAVs were divided according to the Humsavar classification in: '*disease_causing*', when associated with disease and 'non-disease_causing' (corresponding to the term '*polymorphisms*' in Humsavar database) if no association with disease was present. SAVs reported as "*unclassified*" were not included in the analysis.

Non-disease_causing SAVs were further divided according to their global minor allele frequency (MAF) in: 1) '*rare variants*' (MAF <0.01) and 2) '*common variants*' (MAF ≥0.01). The '*common variants*' dataset was enriched by adding SAVs with MAF ≥0.01 reported in the EXAC database (Lek et al. 2016). MAF data for SAVs not listed in EXAC were extracted from Ensembl, using the Biomart data-mining tool (Smedley et al. 2015).

Each disease associated to the proteins in our dataset was classified according to the International Classification of Diseases tenth revision (ICD-10) (Brämer 1988), the standard medical classification of diseases used for clinical and epidemiological purposes by the World Health Organization (WHO).

A protein was classified as '*pleiotropic'* when associated to at least two disorders affecting different physiological systems (e.g. cardiovascular and respiratory systems) or two different conditions belonging to the same physiological system according to the ICD-10 classification (e.g. diabetes type I and obesity both under the 'Endocrine and Metabolic' system). A protein was classified as '*non-pleiotropic*' when associated with only one disorder (i.e. the same identifier in the Online Mendelian Inheritance in Men (OMIM) database).

Since UniProt only reports phenotypes caused by SAVs, diseases caused by truncating variants (i.e. nonsense and frameshift) or canonical splice junction variants are not annotated, thus potentially leading to the erroneous classification of a protein into the non-pleiotropic set. In order to overcome this potential misclassification, we screened proteins in the non-pleiotropic dataset against the morbid map in the OMIM database, which reports protein-disease associations regardless of the type of genetic variant causing them. Proteins reported to be associated with more than one disease in OMIM but not UniProt, were excluded from the non-pleiotropic protein dataset but not re-assigned to the pleiotropic dataset. This was done because our aim was to examine the distribution of SAVs on protein structure. The dataset of investigated proteins and variants is available from http://www.sbg.bio.ic.ac.uk/pleiotropydb

Prediction of intrinsically disordered regions

An 'intrinsically disordered region' was defined by the presence of at least 50 consecutive disordered amino acids. Protein disorder was predicted using IUPred (Dosztányi et al. 2005) and the default score 0.5. IUPred predicts disordered regions by calculating the inter-residue interaction energy. Since intrinsically disordered domains can be less than 50 residues long, a threshold of 30 consecutive disordered residues was also used. Moreover, all results were replicated using a second disorder prediction program, Disopred3, which is trained to recognized missing coordinates from high resolution X-ray structures (Ward et al. 2004; Jones and Cozzetto 2015). IUPred and Disopred3 were chosen among others because they are state of the art (Monastyrskyy et al. 2014), freely downloadable and sensitive to long disordered regions.

Essential proteins

Proteins were classified as "essential" or "non-essential" according to whether their mouse ortholog

produced a non-viable phenotype. The Mouse Genome Informatics (MGI) database (Bult et al. 2016) was used to retrieve mouse genes that produce a lethal phenotype. A total of 3440 mouse genes were classified as essential and 3,303 of these genes could be mapped to a human ortholog. Since some human essential genes have non-essential mouse orthologs (Liao and Zhang 2008), the Online Gene Essentiality database (OGEE) (Chen et al. 2012) was used to provide alternative assignment of essential genes results. The OGEE database includes data for 1528 essential human genes obtained from genome-wide experiments (Silva et al. 2008).

Pathways, protein interactors and Gene Ontology (GO) terms

Pathways data were extracted from the Reactome database (Milacic et al. 2012). GO terms for biological processes, molecular function and cellular component were retrieved from the GO database (Gene Ontology Consortium 2015). Protein interactors were retrieved from BioGRID database (version 3.4.131) (Chatr-Aryamontri et al. 2015).

Statistics

The χ^2 test was used to compare observed and expected frequencies for categorical values. Comparison of medians was performed using the Mann-Whitney-Wilcoxon test. The preference of a SAV to be located in *i* rather than *j* was expressed as the odds ratio (OR_{ij}), calculated as follows:

$$OR_{ij} = x_i/(1- x_i) \,/\, x_j /(1- x_j)$$

for x_i = number of SAVs in region *i* / total number of residues in region *i*.

The p value for the difference between two ORs (OR_1 and OR_2) was calculated from the z score:

$$z= \delta/ SE(\delta)$$

where δ is the difference between ORs calculated as: $\delta= |OR_1| - |OR_2|$, and SE(δ) is the standard error for δ calculated as: $SE(\delta) = \sqrt{SE_1^2 + SE_2^2}$

The hypergeometric test was used to calculate the enrichment in GO terms and classes of diseases. Results were considered statistically significant if p value < 0.05. The Benjamini-Hochberg correction (Benjamini and Hochberg 1995) was applied to control for the error rate arising from multiple testing and results were considered significant if a corrected two-sided p value was < 0.05.

RESULTS

Protein dataset

We surveyed genotype-phenotype associations reported in the UniProt database for 12,543 protein-coding genes. 257 proteins (12% of all proteins associated with at least one disease-causing SAV reported in UniProt) were classified as "pleiotropic", since they were associated with at least two different disorders affecting the same or different physiological systems. 1345 protein-coding genes were classified as "non-pleiotropic", since they were associated with only one disease. All the results presented hereafter were generated from the analysis of this dataset comprising 257 pleiotropic proteins.

It is worth noting, that, an additional 155 protein-coding genes (7.3% of disease-causing proteins reported in UniProt) were associated with at least two similar disorders (identified by different OMIM identifiers), which may differ in terms of degree of severity or response to treatment, e.g. Crigler-Najjar syndrome type I (MIM:218800) and type II (MIM:606785) (a hereditary hyperbilirubinemia caused by genetic variants in UGT1A1). These 155 protein-coding genes did not fulfil our conservative definition of pleiotropy, however they could still be considered pleiotropic when using a less stringent criterion to define pleiotropism. For this reason, these 155 proteins were not included among the 1345 non-pleiotropic. Moreover, for completeness we present in Supp_Mat the results obtained by including these 155 protein-coding genes in the pleiotropic dataset (total number of pleiotropic protein-coding genes presented in Supp_Mat= 412). The dataset of investigated proteins and variants is available from http://www.sbg.bio.ic.ac.uk/pleiotropydb

Characteristics of pleiotropic and non-pleiotropic proteins

The median length of proteins encoded by pleiotropic protein-coding genes (pleiotropic proteins) was 570 (range 110-4967) compared to 509 (range 51-6885) for proteins encoded by non-

pleiotropic protein-coding genes (non-pleiotropic proteins, $p < 0.001$, Mann-Whitney two-tailed test; data for extended dataset of 412 proteins available in Supp. Table S1).

Next, we analyzed the distribution of essential proteins in our dataset by mapping 3,303 human orthologs of mouse essential genes to our protein dataset and found that 59.1% (152) pleiotropic proteins were essential compared to 35.9% (483) non-pleiotropic proteins ($p<0.001$, χ^2 test, Figure 1 panel A and Supp. Table S2A). Since not all human essential genes have non-essential mouse orthologs (Liao and Zhang 2008), we confirmed our results using the OGEE database (20.2% of pleiotropic proteins were classified as essential compared to 10.5% of non-pleiotropic proteins, $p < 0.001$, χ^2 test, Figure 1 panel B and Supp. Table S2B).

We determined the degree of protein-protein interaction (defined as the number of partners a protein interacts with) for each protein in the dataset. The median number of interactors for pleiotropic proteins was 11 (range 0-1980) compared to 6 (range 0-1187) for non-pleiotropic proteins ($p < 0.001$, two tailed Mann-Whitney test; Figure 2 and Supp. Table S3).

These results, which show that pleiotropic proteins are on average longer, more essential and have more interactors compared to non-pleiotropic proteins, suggest that these two groups comprise of proteins with tendencies to different biological properties.

Differences in the distribution of disease-causing SAVs and non-disease_causing SAVs across pleiotropic and non-pleiotropic proteins.

We mapped 24,070 disease-causing SAVs and 37,947 additional variants with no known disease association (non-disease_causing SAVs) annotated in Humsavar to our protein dataset (results are presented in Table 1 and Supp. Tables S4A and S4B). Our final non-redundant dataset contained 15,222 disease-causing SAVs (6,100 in pleiotropic and 9,122 in non-pleiotropic proteins) and 4913 non-disease_causing SAVs (1292 in pleiotropic proteins and 3,621 in non-pleiotropic proteins).

Pleiotropic proteins had a ratio of 'disease_causing' to 'non-disease_causing' SAVs almost double that of non-pleiotropic proteins (4.7 versus 2.5). Moreover, deleterious SAVs were more likely to occur in pleiotropic proteins compared to non-pleiotropic proteins (OR 2.74, p <0.001). Surprisingly, SAVs annotated as non-disease_causing in the Humsavar database, were also more likely to occur in pleiotropic than non-pleiotropic proteins (OR 1.46, p < 0.001). We, therefore, examined the distribution of common SAVs (MAF ≥ 0.01) identified through exome sequencing studies in 60,706 unrelated individuals and reported in the EXAC database. We found that common SAVs were significantly less likely to be in pleiotropic compared to non-pleiotropic proteins (OR 0.73, p <0.001, Table 1 and Supp. Tables S4A and S4B). This raised the hypothesis that MAF is an effect modifier for non-disease_causing SAVs distribution. When MAF was taken into account, common variants reported in Humsavar (356 SAVs in pleiotropic and 1650 in non-pleiotropic) were less likely to be found in pleiotropic proteins (OR 0.88, p<0.03, although this was just a trend when the larger dataset of 412 pleiotropic proteins was analysed), whereas rare SAVs (320 SAVs in pleiotropic and 917 in non-pleiotropic) were more likely to be found in pleiotropic proteins (OR 1.43 p<0.001). For the remaining 1670 non-disease_causing SAVs reported in Humsavar (616 in pleiotropic and 1054 in non-pleiotropic proteins) no MAF data were available. These SAVs were more likely to be found in pleiotropic rather than non-pleiotropic proteins (OR 2.40, p<0.001). As these rare variants and SAVs with no MAF show a distribution similar to that of disease-causing SAVs, this raises the question of whether many of these variants are actually mildly damaging.

Ordered versus disordered protein structure

Significantly more pleiotropic (132 proteins, 51.4%) than non-pleiotropic proteins (516 proteins, 38.4%) contained at least one intrinsically long disordered region of over 50 residues in length (p<0.001, χ^2 test). This difference was maintained even when lowering our cut-off to define a long

intrinsically disordered region from 50 to 30 residues, and when utilizing the disorder prediction program DISOPRED3 (Supp. Table S5).

Distribution of SAVs in relation to protein disorder and pleiotropy

We examined the distribution of disease_causing and non-disease_causing SAVs in ordered and disordered regions of pleiotropic and non-pleiotropic proteins (results for the 412 pleiotropic dataset are presented in Supp. Table S6) and found that deleterious SAVs were more likely to occur in ordered rather than disordered regions in both pleiotropic (OR 1.71, $p< 0.001$) and non-pleiotropic proteins (OR 3.51, $p<0.001$). Nevertheless the OR for a deleterious SAV occurring in a disordered region was significantly greater for pleiotropic compared to non-pleiotropic proteins (pleiotropic proteins OR 0.58, 95%C.I. 0.54-0.63, $p<0.001$; non-pleiotropic OR 0.29, 95%C.I. 0.26-0.31 $p< 0.001$; z score for the difference between ORs 11.6, $p< 0.001$). Common variants from EXAC were similarly more likely to occur in a disordered rather than an ordered region, in both pleiotropic (OR 1.55, $p< 0.001$) and non-pleiotropic proteins (OR 1.32 $p< 0.001$; z score for difference between ORs 1.47, $p=0.07$). When examining common SAVs reported in the Humsavar database, we confirmed that these are more likely to occur in disordered rather than ordered regions, although this was just a trend for pleiotropic proteins (non-pleiotropic proteins OR 1.29, $p< 0.001$, pleiotropic proteins OR 1.22, $p=0.11$). The lack of statistical significance in the pleiotropic set is likely to be due to the small number of common SAVs analysed (80 in disordered and 276 in ordered regions).

When rare variants were examined, their location was not influenced by the absence of a fixed tertiary structure, neither in pleiotropic (ordered vs disordered regions: OR 0.80, $p=0.09$) or in non-pleiotropic proteins (OR 0.96, $p=0.63$). Finally a very weak correlation was found between the length of the disordered protein sequence and the degree of pleiotropism, defined as the number of different disorders caused by the same protein ($r=0.14$, $p=0.02$).

Pleiotropic and non-pleiotropic proteins cause different classes of diseases

We examined the distribution of diseases caused by pleiotropic and non-pleiotropic proteins. Congenital malformations, neoplasms, circulatory and neurological diseases were significantly ($p<0.01$) more likely to be caused by pleiotropic rather than non-pleiotropic proteins, whereas the reverse was true for endocrine, metabolic, blood and immune disorders (Figure 3).

Pleiotropy and biological pathways

We mapped 986 out of 1602 pleiotropic and non-pleiotropic proteins in our dataset on pathways annotated in Reactome. Although the majority of proteins were mapped to only one (602 proteins) or two pathways (238 proteins), a significantly greater number of pleiotropic proteins shared multiple (>2) pathways compared to non-pleiotropic proteins (χ^2 test $p<0.001$, Figure 4). Moreover, there was an enrichment of pleiotropic proteins in "signal transduction", "developmental biology", "muscle contraction", "extracellular matrix organization", "membrane trafficking", "binding and uptake of ligands by scavenger receptors" and "haemostasis" pathways, while enrichment in non-pleiotropic proteins was found in the "metabolic pathway". When GO terms were examined, pleiotropic proteins were mainly annotated as involved in "cell aggregation", "receptor regulatory activity" and "collagen formation".

Case studies: Vinculin and Homeobox pleiotropic proteins

The following case studies exemplify how protein structural analysis can aid in the identification of the molecular mechanisms underlying the pleiotropic effect of proteins.

Vinculin (VCL, UniProt ID: P18206, HGNC:12665), a binding protein involved in cell-matrix and cell-cell adhesion, is a key component of heart and brain development (Borgon et al. 2004; Carisey

and Ballestrem 2011). It is characterized by the presence of two main domains, vinculin's head (Vh) and vinculin's tail (Vt), which can form intermolecular interactions. In its inactive conformation, most of vinculin's binding sites are hidden and not available for protein-protein interaction (Ziegler et al. 2006). In the activated state, Vh and Vt domains interact with different proteins: the Vh domain interacts with talin, α-catenin and α-actinin and the Vt domain interacts with actin, paxillin, and phosphatidylinositol (4,5)-biphosphate (PIP_2) (Borgon et al. 2004; Carisey and Ballestrem 2011). Two diseases have been associated with *Vinculin* mutations: dilated cardiomyopathy (MIM:611407), which is caused by p.L954del and p.R975W, and familial hypertrophic cardiomyopathy (MIM:613255), which is caused by p.L277M. We analysed the structure of vinculin and showed that residues L954 and R975 are located in the Vt domain. In particular, R975 is part of a positively charged patch on the surface of vinculin and of its alternative spliced form metavinculin and is likely to represent an interaction site. We predict that the R to W substitution alters this charged patch (Figure 5, panel A), thus affecting protein-protein interaction. This hypothesis is supported by *in vitro* experiments, which demonstrate that p.R975W results in a lack of interaction between vinculin and actin (Olson et al. 2002). Moreover, R975W could also disrupt an intrachain interaction between R975 and D505, as previously proposed (Rangarajan et al. 2010). Residue L227 is part of the Vh domain and its substitution with methionine has been described in patients with hypertrophic cardiomyopathy (Vasile et al. 2006). This residue is partially buried and is part of an alpha helix. Although replacement with M is predicted to have a minimal effect on the overall structure of vinculin, the occurrence of a steric clash with the neighbor F212, cannot be excluded.

The homeobox protein (OTX2, UniProt ID: P32243, HGNC:8522) is a transcription factor which plays a crucial role in eye and visual cortex development (Beby and Lamonerie 2013). Two diseases are associated with mutations on this protein: severe ocular malformations (MIM: 610125) caused by p.R89G and p.R90S (Ragge et al. 2005), and anterior pituitary hypoplasia with multiple hormone deficiency (MIM:613986), caused by p.N225S (Diaczok et al. 2008). The homeobox

protein is characterized by two domains: the Homeobox (Homeobox) and the TF-Otx domain (Figure 5, panel B). The Homeobox domain is formed by three alpha helices and interacts with DNA. In particular, the side chains of amino acids located on one of these helices, called recognition helices, make extensive interactions with the DNA major groove. Invariant residues R89 and R90 are part of the recognition helix (Billeter et al. 1993) and their substitution is likely to disrupt the formation of the protein-DNA complex. Residue N225 is located in the TF-Otx domain, thus outside the DNA binding domain. Its substitution with Serine has been shown *in vitro* to result in a mutant homeobox protein that can still bind to target genes but has a dominant negative inhibitory effect on HESX1 gene expression (Diaczok et al. 2008), resulting in abnormal development of the anterior pituitary and, hence, deficiency of all pituitary hormones.

In both these case studies, SAVs in different domains resulted in different diseases, which is in agreement with a previous report that mutation pairs on different protein interacting domains are more likely to cause different disorders than those on the same protein interacting domain (Wang et al. 2012).

DISCUSSION

Pleiotropic proteins are emerging as important contributors to human disease. We demonstrated that pleiotropic proteins are enriched in deleterious SAVs and rare genetic variations and contribute more than non-pleiotropic proteins to certain classes of diseases, such as neoplasias, neurological and congenital disorders.

We analysed proteins causing human disease annotated in the UniProt database and showed that 12% of protein-coding genes harbouring deleterious SAVs are pleiotropic. Since we used a stringent criterion to define pleiotropy and only considered protein-coding genes harbouring amino acid substitutions, it is likely that pleiotropy is an even more widespread phenomenon. A recent systematic review of genetic variations associated with common diseases demonstrated that 17% of genes identified by GWAS are responsible for multiple phenotypes (Sivakumaran et al. 2011). Pleiotropy poses a formidable challenge when analysing GWAS data for assessing the dependent or independent association between two phenotypes, or when comparing the enrichment in deleterious variants between patients and controls for establishing the pathogenicity of a gene in relation to a specific phenotype. Identification and characterization of pleiotropic genes and their genetic variations is, therefore, of paramount importance.

In our study, proteins in the pleiotropic and non-pleiotropic datasets differed in terms of their biological properties, with enrichment in essential proteins and a higher number of protein interactors in the pleiotropic compared to the non-pleiotropic dataset. This is consistent with previous studies in humans and yeast (Chavali et al. 2010) and supports the hypothesis that pleiotropic proteins have a tendency to have a different biological role compared to non-pleiotropic proteins. Studies in *C. elegans* have shown that pleiotropic proteins have an important role in early embryological life and potentially bridge different protein pathways (Zou et al. 2008). Enrichment in essential proteins in our pleiotropic dataset suggests that pleiotropic proteins could also play a similar role in human embryonic life.

Our analysis demonstrated an enrichment in pleiotropic proteins in neoplasias, neurological, circulatory and congenital disorders, which are among the leading causes of morbidity and mortality in developed countries (Muka et al. 2015). Conversely, endocrine and metabolic disorders were most commonly caused by non-pleiotropic proteins. These results are in agreement with the human disease network developed by Goh et al., which shows that the degree of shared genetic background is not similar for all diseases, but is high for neoplasias and low for endocrine related disorders (Goh et al. 2007). Consistent with our results are recent reports of a genetic association between lung and other cancer types from the PAGE and TRICL consortia (Park et al. 2014) and between colorectal cancer and other cancers from the PAGE and GECCO and CCFR consortia (Cheng et al. 2014). In the case of neurological disorders, there is clinical and genetic evidence of an overlap between conditions, such as frontotemporal dementia and amyotrophic lateral sclerosis (ALS) (Lillo and Hodges 2009; Geser et al. 2010) or ALS, bone dysfunction and myopathy (Guerreiro et al. 2014) while there is an inverse association between neurological disorders, such as Parkinson and Alzheimer, and the incidence of cancer (Driver 2014) in the elderly. Moreover, there is increasing evidence of a shared genetic background between embryological development and cancer (Becker et al. 2012) and evidence of an association between congenital malformations and an increased risk of cancer, especially in young children (Bjørge et al. 2008; Sun et al. 2014). Although additional studies are needed to clarify the biological mechanisms behind the enrichment of pleiotropic proteins in particular classes of diseases, our findings can help identify a subset of candidate genes for GWAS and exome sequencing studies in cancer and circulatory and neurological disorders.

GWAS have shown that the metabolic syndrome as well as immune-mediated disorders (Sivakumaran et al. 2011; Kraja et al. 2014) are associated with pleiotropic genes, which appears to contradict our results. Two important aspects are worth noticing: firstly, in our study pleiotropic genes were identified by analysing disease phenotypes caused by rare variants in coding regions, resulting in amino acid substitutions. On the contrary, GWAS identify pleiotropic genes by

analysing phenotypes associated with common variants, often located in the non-coding region and affecting gene expression or methylation (Gratten and Visscher 2016). A comprehensive study looking at phenotypes associated with rare and common variants is likely to result in a higher number of pleiotropic genes associated with disease compared to our method. Secondly, the disease classification used in our study differs from that used in the above-cited GWAS (Sivakumaran et al. 2011; Kraja et al. 2014). The metabolic syndrome is characterized by several biochemical and clinical conditions (Kaur 2014) many of which, such as hypertension, an hypercoagulable state, chronic inflammation and chronic stress are not classified as "endocrine and metabolic disorders" in the ICD-10. We adhered to the ICD-10 classification and chose not to manually define new classes (e.g. for the metabolic syndrome). Moreover, we treated endocrine and metabolic disorders as a single class. Similarly, we chose not to manually merge all immune-mediated disorders into a new class. For this reason, conditions such as type I diabetes or Crohn disease are classified in our study as "metabolic" and "gastrointestinal", respectively, as was the case in other studies (Goh et al. 2007), whereas these are classified as "immune-mediated conditions" in GWAS analyses linking pleiotropism to immune disorders (Sivakumaran et al. 2011).

We showed that pleiotropic proteins share multiple pathways. This can — at least in part — explain how different genetic variants in the same pleiotropic protein-coding gene can affect several phenotypic traits and result in similar or opposite disease risks for different disorders. Moreover, we demonstrated an enrichment of pleiotropic proteins in a broad range of core biological pathways, such as signal transduction, developmental biology and extracellular matrix organization, which is in agreement with the enrichment of pleiotropic proteins in diseases, such as neoplasias and congenital disorders. Although our findings could be the result of a better characterization of pleiotropic proteins compared to non-pleiotropic proteins, it is noteworthy that genes involved in signalling pathways, such as the Notch signalling pathway, are involved in embryogenesis, cancer (Reichrath and Reichrath 2012), cardiovascular (Rusanescu et al. 2008) and neurological disorders (Lathia et al. 2008).

In our dataset, disease-causing SAVs and rare variants were enriched in pleiotropic proteins, whereas common variants were more likely to be found in non-pleiotropic proteins. This suggests that pleiotropic proteins are less likely to tolerate genetic variations and questions arise on the nature — neutral, moderately deleterious or disease-risk protective — of rare variants observed in these proteins. This becomes particularly important when interpreting the pathogenicity of different genetic variants harboured by the same gene in patients and controls sequenced as part of individual projects or large sequencing efforts, such as the 100K Genomes project.

Although deleterious SAVs were enriched in structurally ordered regions, deleterious variants occurring in structurally disordered regions were more likely to be in a pleiotropic rather than non-pleiotropic protein. Disordered proteins play an important role in cell biology, as they participate in signalling, regulation process, DNA-protein and protein-protein interaction (Xie et al. 2007) (Tompa 2005). The lack of a fixed 3D structure allows disordered regions to assume different conformations, thus, having the ability to interact with multiple partners (Haynes et al. 2006) (Dunker et al. 2005). The evidence that structure disorder is a common feature in hub proteins (proteins with multiple interactors) (Haynes et al. 2006) prompted us to evaluate its role in the distribution of deleterious SAVs in pleiotropic and non-pleiotropic proteins. Moreover, protein disorder has recently been recognised as an important contributor to disease (Dunker et al. 2008; Vacic et al. 2012) and there is evidence that disordered proteins are associated with cancer (Iakoucheva et al. 2002) and neurological disorders, such as Alzheimer and Parkinson disease (Uversky 2008). Interestingly, the deleterious effect of SAVs occurring in disordered regions is underestimated by variant prediction programs, such as SIFT (Mort et al. 2010), possibly because of the low sequence conservation, which characterizes these regions (Vacic et al. 2012). Our results suggest that knowledge of the nature of a protein — pleiotropic versus non-pleiotropic — and of the protein 3D structure — ordered or disordered — could help establish the benign or deleterious nature of a genetic variant. The presence of a genetic variant in a long disordered region of a pleiotropic protein should, in fact, prompt investigation of the effect of such a variant using

additional methods, such as ordered-to-disordered transition upon amino acid change (Vacic et al. 2012).

One important potential limitation of our study is that pleiotropic proteins are likely to be better characterised compared to non-pleiotropic proteins, thus potentially biasing some of our findings, such as the higher number of protein interactors in the pleiotropic set. Nevertheless, other results, such as protein essentiality and protein disorder are less likely to be biased in favour of extensively characterized proteins and build a consistent picture which support the hypothesis that pleiotropic proteins are biologically different from non-pleiotropic proteins.

In conclusion, this study provides a better understanding of pleiotropic proteins and their genetic variants, which could greatly aid in the analysis of genetic studies and drug design. In particular, this study provides evidence that pleiotropic proteins are enriched in deleterious SAVs and rare variants and significantly contribute to cancer, neurological, circulatory and congenital disorders. This emphasizes the importance of acknowledging the phenomenon of pleiotropy when establishing the contribution of genetic variants to human health. The results of this study can guide the selection of candidate genes for classes of disorders enriched in pleiotropic proteins.

Acknowledgements

We would like to thank Dr S Islam for the technical support.

Competing interests

The authors declare that they have no competing interests.

REFERENCES

Beby F, Lamonerie T. 2013. The homeobox gene Otx2 in development and disease. Exp Eye Res 111:9–16.

Becker D, Sfakianakis I, Krupp M, Staib F, Gerhold-Ay A, Victor A, Binder H, Blettner M, Maass T, Thorgeirsson S, Galle PR, Teufel A. 2012. Genetic signatures shared in embryonic liver development and liver cancer define prognostically relevant subgroups in HCC. Mol Cancer 11:55.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57:289–300.

Billeter M, Qian YQ, Otting G, Müller M, Gehring W, Wüthrich K. 1993. Determination of the nuclear magnetic resonance solution structure of an Antennapedia homeodomain-DNA complex. J Mol Biol 234:1084–1093.

Bjørge T, Cnattingius S, Lie RT, Tretli S, Engeland A. 2008. Cancer risk in children with birth defects and in their families: a population based cohort study of 5.2 million children from Norway and Sweden. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol 17:500–506.

Borgon RA, Vonrhein C, Bricogne G, Bois PRJ, Izard T. 2004. Crystal structure of human vinculin. Struct Lond Engl 1993 12:1189–1197.

Brämer GR. 1988. International statistical classification of diseases and related health problems. Tenth revision. World Health Stat Q Rapp Trimest Stat Sanit Mond 41:32–36.

Bult CJ, Eppig JT, Blake JA, Kadin JA, Richardson JE, Mouse Genome Database Group. 2016. Mouse genome database 2016. Nucleic Acids Res 44:D840-847.

Carisey A, Ballestrem C. 2011. Vinculin, an adapter protein in control of cell adhesion signalling. Eur J Cell Biol 90:157–163.

Chatr-Aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, et al. 2015. The BioGRID interaction database: 2015 update. Nucleic Acids Res 43:D470-478.

Chavali S, Barrenas F, Kanduri K, Benson M. 2010. Network properties of human disease genes with pleiotropic effects. BMC Syst Biol 4:78.

Chen W-H, Minguez P, Lercher MJ, Bork P. 2012. OGEE: an online gene essentiality database. Nucleic Acids Res 40:D901-906.

Cheng I, Kocarnik JM, Dumitrescu L, Lindor NM, Chang-Claude J, Avery CL, Caberto CP, Love S-A, Slattery ML, Chan AT, Baron JA, Hindorff LA, et al. 2014. Pleiotropic effects of genetic risk variants for other cancers on colorectal cancer risk: PAGE, GECCO and CCFR consortia. Gut 63:800–807.

Diaczok D, Romero C, Zunich J, Marshall I, Radovick S. 2008. A novel dominant negative mutation of OTX2 associated with combined pituitary hormone deficiency. J Clin Endocrinol Metab 93:4351–4359.

Dosztányi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinforma Oxf Engl 21:3433–3434.

Driver JA. 2014. Inverse association between cancer and neurodegenerative disease: review of the epidemiologic and biological evidence. Biogerontology 15:547–557.

Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J 272:5129–5148.

Dunker AK, Silman I, Uversky VN, Sussman JL. 2008. Function and structure of inherently disordered proteins. Curr Opin Struct Biol 18:756–764.

Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. Nucleic Acids Res 43:D1049-1056.

Geser F, Lee VM-Y, Trojanowski JQ. 2010. Amyotrophic lateral sclerosis and frontotemporal lobar degeneration: a spectrum of TDP-43 proteinopathies. Neuropathol Off J Jpn Soc Neuropathol 30:103–112.

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. 2007. The human disease network. Proc Natl Acad Sci U S A 104:8685–8690.

Gratten J, Visscher PM. 2016. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. Genome Med 8:78.

Guerreiro R, Brás J, Hardy J, Singleton A. 2014. Next generation sequencing techniques in neurological diseases: redefining clinical and molecular associations. Hum Mol Genet ddu203.

Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM. 2006. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2:e100.

Hodgkin J. 1998. Seven types of pleiotropy. Int J Dev Biol 42:501–505.

Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK. 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 323:573–584.

Jo S, Vargyas M, Vasko-Szedlar J, Roux B, Im W. 2008. PBEQ-Solver for online visualization of electrostatic potential of biomolecules. Nucleic Acids Res 36:W270-275.

Jones DT, Cozzetto D. 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinforma Oxf Engl 31:857–863.

Kaur J. 2014. A comprehensive review on metabolic syndrome. Cardiol Res Pract 2014:943162.

Kraja AT, Chasman DI, North KE, Reiner AP, Yanek LR, Kilpeläinen TO, Smith JA, Dehghan A, Dupuis J, Johnson AD, Feitosa MF, Tekola-Ayele F, et al. 2014. Pleiotropic genes for metabolic syndrome and inflammation. Mol Genet Metab 112:317–338.

Lathia JD, Mattson MP, Cheng A. 2008. Notch: from neural development to neurological disorders. J Neurochem 107:1471–1481.

Le Nagard H, Chao L, Tenaillon O. 2011. The emergence of complexity and restricted pleiotropy in adapting networks. BMC Evol Biol 11:326.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291.

Liao B-Y, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. Proc Natl Acad Sci U S A 105:6987–6992.

Ligthart S, Vries PS de, Uitterlinden AG, Hofman A, CHARGE Inflammation working group, Franco OH, Chasman DI, Dehghan A. 2015. Pleiotropy among common genetic loci identified for cardiometabolic disorders and C-reactive protein. PloS One 10:e0118859.

Lillo P, Hodges JR. 2009. Frontotemporal dementia and motor neurone disease: overlapping clinic-pathological disorders. J Clin Neurosci Off J Neurosurg Soc Australas 16:1131–1135.

Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, D'Eustachio P, Stein L. 2012. Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers 4:1180–1211.

Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K. 2014. Assessment of protein disorder region predictions in CASP10. Proteins 82 Suppl 2:127–137.

Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters BJ, Sathyesh R, Li B, Sun Y, Xue B, Shah NH, et al. 2010. In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. Hum Mutat 31:335–346.

Muka T, Imo D, Jaspers L, Colpani V, Chaker L, Lee SJ van der, Mendis S, Chowdhury R, Bramer WM, Falla A, Pazoki R, Franco OH. 2015. The global impact of non-communicable diseases on healthcare spending and national income: a systematic review. Eur J Epidemiol 30:251–277.

Oldfield CJ, Dunker AK. 2014. Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem 83:553–584.

Olson TM, Illenberger S, Kishimoto NY, Huttelmaier S, Keating MT, Jockusch BM. 2002. Metavinculin mutations alter actin interaction in dilated cardiomyopathy. Circulation 105:431–437.

Ongoing and future developments at the Universal Protein Resource. 2011. Nucleic Acids Res 39:D214–D219.

Orr HA. 2000. Adaptation and the cost of complexity. Evol Int J Org Evol 54:13–20.

Park SL, Fesinmeyer MD, Timofeeva M, Caberto CP, Kocarnik JM, Han Y, Love S-A, Young A, Dumitrescu L, Lin Y, Goodloe R, Wilkens LR, et al. 2014. Pleiotropic associations of risk variants identified for other cancers with lung cancer risk: the PAGE and TRICL consortia. J Natl Cancer Inst 106:dju061.

Ragge NK, Brown AG, Poloschek CM, Lorenz B, Henderson RA, Clarke MP, Russell-Eggitt I, Fielder A, Gerrelli D, Martinez-Barbera JP, Ruddle P, Hurst J, et al. 2005. Heterozygous mutations of OTX2 cause severe ocular malformations. Am J Hum Genet 76:1008–1022.

Rangarajan ES, Lee JH, Yogesha SD, Izard T. 2010. A helix replacement mechanism directs metavinculin functions. PloS One 5:e10679.

Reichrath J, Reichrath S. 2012. Notch Signaling in Embryology and Cancer. Springer Science & Business Media. 346 p.

Rusanescu G, Weissleder R, Aikawa E. 2008. Notch signaling in cardiovascular disease and calcification. Curr Cardiol Rev 4:148–156.

Silva JM, Marran K, Parker JS, Silva J, Golding M, Schlabach MR, Elledge SJ, Hannon GJ, Chang K. 2008. Profiling essential genes in human mammary cells by multiplex RNAi screening. Science 319:617–620.

Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, Rudan I, McKeigue P, Wilson JF, Campbell H. 2011. Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet 89:607–618.

Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G, Bardou P, Beck T, et al. 2015. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res 43:W589–W598.

Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. 2013. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet 14:483–495.

Stearns FW. 2010. One hundred years of pleiotropy: a retrospective. Genetics 186:767–773.

Sun Y, Overvad K, Olsen J. 2014. Cancer risks in children with congenital malformations in the nervous and circulatory system-A population based cohort study. Cancer Epidemiol 38:393–400.

Tompa P. 2005. The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett 579:3346–3354.

Uversky VN. 2008. Amyloidogenesis of natively unfolded proteins. Curr Alzheimer Res 5:260–287.

Uversky VN, Oldfield CJ, Dunker AK. 2008. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys 37:215–246.

Vacic V, Markwick PRL, Oldfield CJ, Zhao X, Haynes C, Uversky VN, Iakoucheva LM. 2012. Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. PLoS Comput Biol 8:e1002709.

Vasile VC, Ommen SR, Edwards WD, Ackerman MJ. 2006. A missense mutation in a ubiquitously expressed protein, vinculin, confers susceptibility to hypertrophic cardiomyopathy. Biochem Biophys Res Commun 345:998–1003.

Wagner GP, Kenney-Hunt JP, Pavlicev M, Peck JR, Waxman D, Cheverud JM. 2008. Pleiotropic scaling of gene effects and the "cost of complexity." Nature 452:470–472.

Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. Nat Rev Genet 12:204–213.

Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. 2012. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol 30:159–164.

Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. 2004. The DISOPRED server for the prediction of protein disorder. Bioinforma Oxf Engl 20:2138–2139.

Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z. 2007. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteome Res 6:1882–1898.

Ziegler WH, Liddington RC, Critchley DR. 2006. The structure and regulation of vinculin. Trends Cell Biol 16:453–460.

Zou L, Sriswasdi S, Ross B, Missiuro PV, Liu J, Ge H. 2008. Systematic analysis of pleiotropy in C. elegans early embryogenesis. PLoS Comput Biol 4:e1000003.

FIGURE LEGENDS

Figure 1. Percentage of essential proteins among pleiotropic and non-pleiotropic proteins. In panel A, proteins are classified as "essential" if their mouse orthologs produce a non-viable phenotype; in panel B, proteins are classified as essential if they appear in the OGEE database.

Figure 2 Number of protein interactors for pleiotropic and non-pleiotropic proteins. The box plots depict median, lower (25%, Q1) and upper quartiles (75%, Q3). Whisker represents 25% – 1.5×the interquartile range (IQR) and 75% + 1.5×IQR. **, $p<0.01$.

Figure 3 Pleiotropic proteins and classes of diseases. The enrichment of disease-causing pleiotropic and non-pleiotropic proteins according to different classes of disease defined using the ICD-10 classification is presented.

Other, remaining ICD-10 classes. **, $P<0.01$.

Figure 4 Pleiotropic proteins and biological pathways. Percentage of pleiotropic and non-pleiotropic proteins which are present in less or more than two biological pathways.

Figure 5: Structural analysis of the deleterious variants occurring in two pleiotropic proteins: Vinculin and Homeobox. Panel A: Vinculin protein. Electrostatic map shows the presence of a large positively charged patch on the surface of Vt domain (the electrostatic potential was calculated using PBEQ (Jo et al. 2008) program , which computes the protein electrostatic potential by solving the Poisson–Boltzmann equation). Panel B: Homeobox protein (OTX2). The homeobox

domain alpha helices on OTX2 are presented in green. Residues R89 and R90 are presented in blue. DNA is presented in purple.

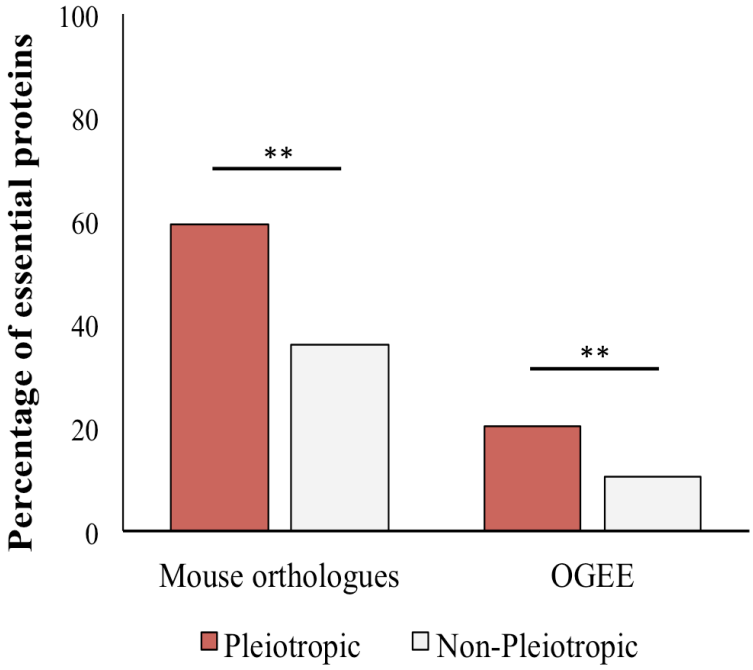
Protein structures were visualised using the Pymol visualisation program (http://www.pymol.org/).

Table 1. Odds ratio (OR) for disease-causing and non-disease-causing SAVs in pleiotropic and non-pleiotropic proteins.

Pleiotropic versus non-pleiotropic proteins	OR	95% CI	p value
Disease-causing SAVs	2.74	2.65-2.83	<0.001
Non-disease-causing SAVs (Humsavar - all)	1.46	1.37-1.56	<0.001
Non-disease-causing SAVs (EXAC - common)	0.73	0.66-0.80	<0.001
Non-disease-causing SAVs (Humsavar - common)	0.88	0.79-0.99	0.03
Non-disease-causing SAVs (Humsavar - rare)	1.43	1.26-1.62	<0.001
Non-disease-causing SAVs (Humsavar - no MAF)	2.41	2.17-2.65	<0.001

Non-disease-causing SAVs are defined as 'common' when MAF ≥ 0.01 and 'rare' when MAF < 0.01. 95% CI, 95% confidence intervals.

Figure 1

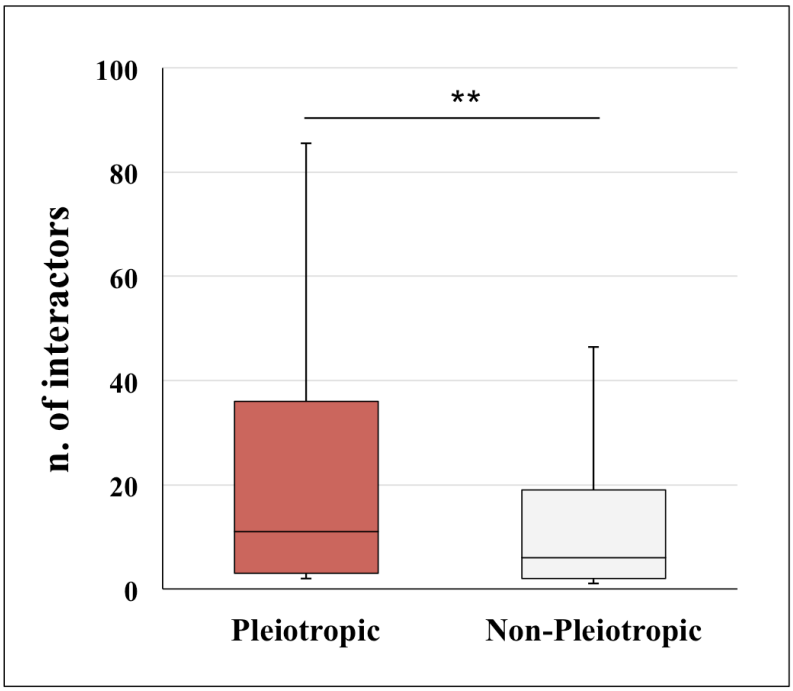

Percentage of essential proteins
100
80
60
40
20
0
**
**
Mouse orthologues
OGEE
Pleiotropic
Non-Pleiotropic

1

2

3

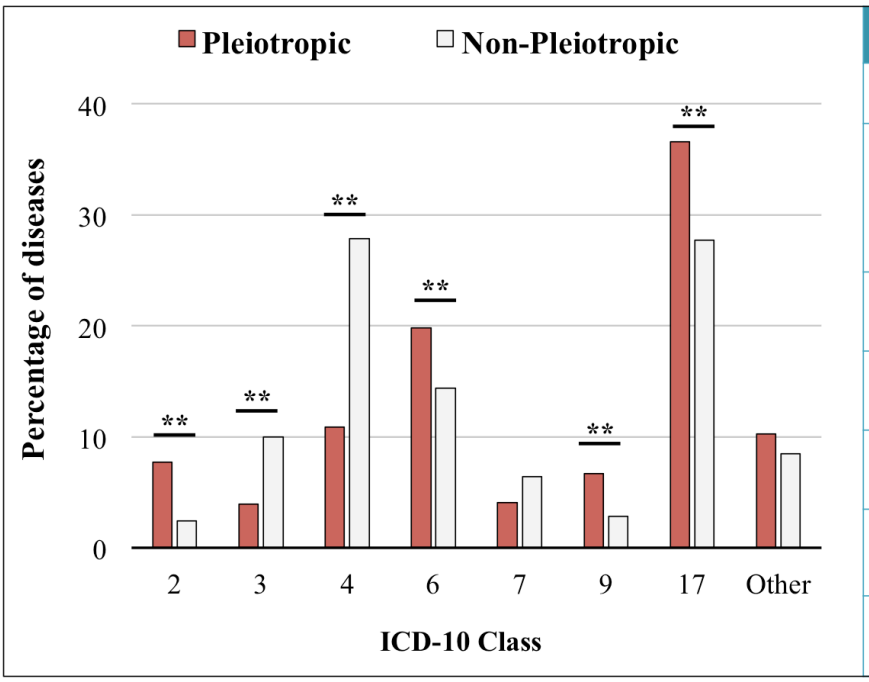
Figure 2


100
80
60
40
20
0
n. of interactors
**
Pleiotropic
Non-Pleiotropic

1

2

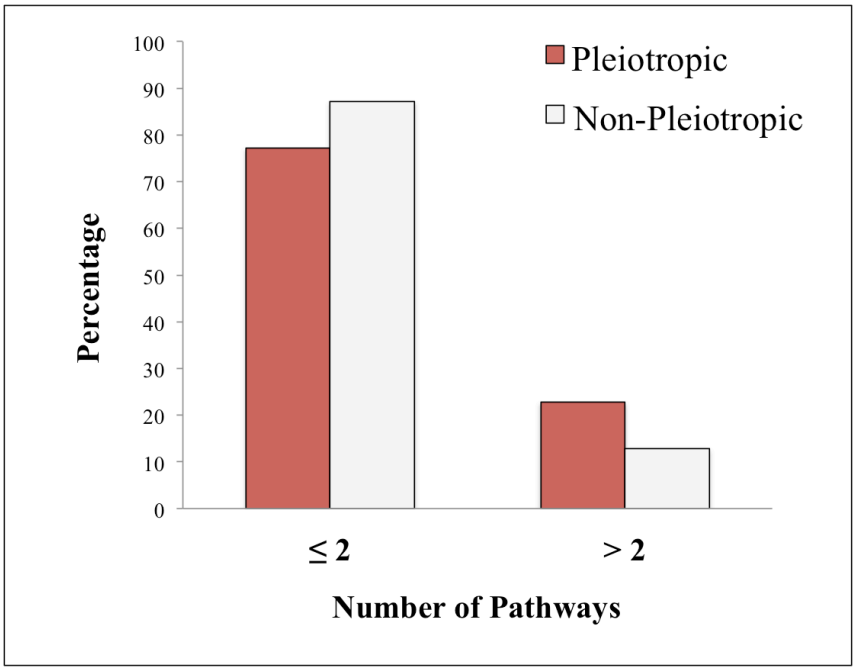
Figure 3



Class	Disease type
2	Neoplasms
3	Diseases of the blood and certain disorders involving the immune mechanism
4	Endocrine and metabolic diseases
6	Diseases of the nervous system
7	Diseases of the eye and adnexa
9	Diseases of the circulatory system
17	Congenital malformations

1

Figure 4

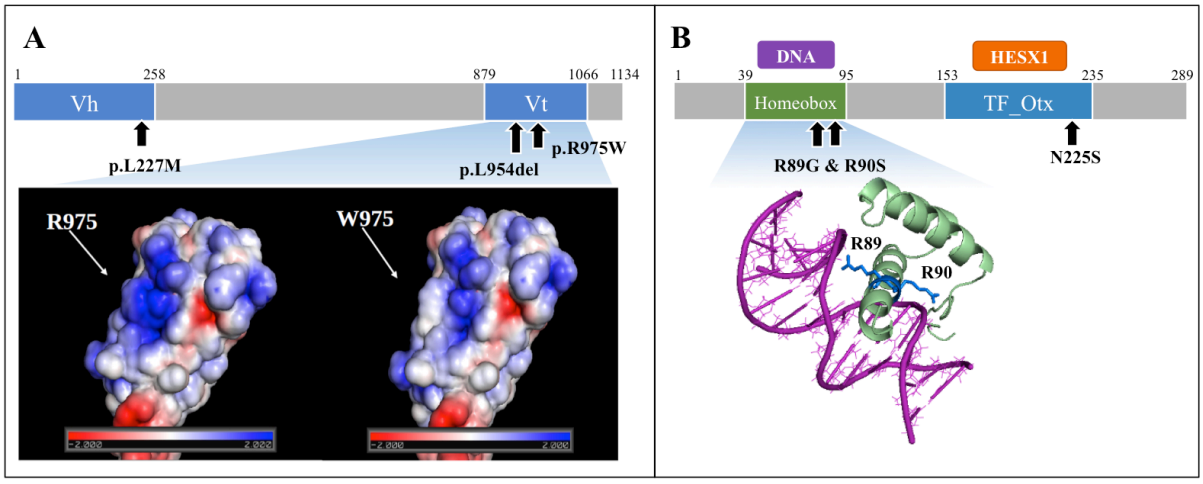


2

3

4

Figure 5


A
1
258
879
1066
1134
Vh
Vt
p.L227M
p.L954del
p.R975W
R975
W975
B
DNA
HESX1
1
39
95
153
235
289
Homeobox
TF_Otx
R89G & R90S
N225S
R89
R90

1

2

3

Landscape of pleiotropic proteins causing human disease: structural and system biology insights

Sirawit Ittisoponpisan, Eman Alhuzimi, Michael J. E. Sternberg, Alessia David

Supp. Table S1 Protein length analysed in pleiotropic proteins and non-pleiotropic proteins. Data for the 257 pleiotropic dataset are presented in the main text.

Protein length	Pleiotropic (257)	Pleiotropic (412)	Non-Pleiotropic (1345)
Min	110	79	51
Max	4967	5202	6885
Mean	878.14	890.96	686.99
Median	570	592.5	509
Q1	390	397.5	338
Q3	1218	1108	777

Supp. Table S2A Essentials proteins in the pleiotropic and non-pleiotropic dataset. Essential proteins are classified by mapping 3,303 human orthologs of mouse essential genes to the pleiotropic and non-pleiotropic datasets.

	Pleiotropic proteins (412)	Non-pleiotropic proteins (1345)	*p* value
Essential (n.)	228	483	<0.001
Essential (%)	55.3	35.9	

P-value was calculated using χ^2 test.

Supp. Table S2B Essentials proteins in the pleiotropic and non-pleiotropic dataset Essentials proteins are classified using the human Online Gene Essentiality (OGEE) database.

	Pleiotropic proteins (412)	**Non-pleiotropic proteins (1345)**	***p value***
Essential (n.)	69	141	<0.001
Essential (%)	16.6	10.5	

P-value was calculated using χ^2 test.

Supp. Table S3 Protein-protein interactions in pleiotropic and non-pleiotropic proteins. Protein-protein interactions data were extracted from BioGRID database. Highlighted in grey are the results presented in the main text.

	Pleiotropic (257)	**Pleiotropic (412)**	**Non-pleiotropic (1345)**
Min	0	0	0
Max	1980	1980	1187
Mean	40.16	32.56	18.88
Median	11	9	6
Q1	3	2	2
Q3	36	30	19
STD	136.47	112.75	48.98

Supp. Table S4A Distribution of disease-causing and non-disease-causing SAVs in 257 pleiotropic and non-pleiotropic proteins.

Non-disease-causing are defined as 'common' if MAF ≥ 0.01 and 'rare' if MAF < 0.01.

Disease_causing SAVs	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**	***p* value**
pleiotropic (257 proteins)	6100	219581	225681	2.74	2.65 - 2.83	< 0.0001
non-pleiotropic	9122	914880	924002			
EXAC - common SAVs	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**	***p* value**
pleiotropic (257 proteins)	499	225182	225681	0.73	0.66 - 0.80	< 0.0001
non-pleiotropic	2811	921191	924002			
Uniprot - all SAVs	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**	***p* value**
pleiotropic (257 proteins)	1292	224389	225681	1.46	1.37 - 1.56	< 0.0001
non-pleiotropic	3621	920381	924002			
Uniprot - common SAVs	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**	***p* value**
pleiotropic (257 proteins)	356	225325	225681	0.88	0.79 - 0.99	0.03
non-pleiotropic	1650	922352	924002			
Uniprot - rare SAVs	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**	***p* value**
pleiotropic (257 proteins)	320	225361	225681	1.43	1.26 - 1.62	< 0.0001
non-pleiotropic	917	923085	924002			
Uniprot - no MAF	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**	***p* value**
pleiotropic (257 proteins)	616	225065	225681	2.40	2.17 - 2.65	< 0.0001
non-pleiotropic	1054	922948	924002			

Total, the total number of residues; OR, odds ratio; 95%CI, 95% Confidence intervals.

Supp. Table S4B Distribution of disease-causing and non-disease-causing SAVs in 412 pleiotropic and non-pleiotropic proteins.

Non-disease-causing SAVs are defined as 'common' if MAF ≥ 0.01 and 'rare' if MAF < 0.01.

Disease_causing SAVs	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**		**p value**
pleiotropic (412 proteins)	10154	356922	367076	2.85	2.77 -	2.94	< 0.0001
non-pleiotropic	9122	914880	924002				
EXAC - common SAVs	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**		**p value**
pleiotropic (412 proteins)	962	366114	367076	0.86	0.80 -	0.93	< 0.0001
non-pleiotropic	2811	921191	924002				
Uniprot - all SAVs	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**		**p value**
pleiotropic (412 proteins)	2055	365021	367076	1.43	1.36 -	1.51	< 0.0001
non-pleiotropic	3621	920381	924002				
Uniprot - common SAVs	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**		**p value**
pleiotropic (412 proteins)	634	366442	367076	0.97	0.88 -	1.06	0.4752
non-pleiotropic	1650	922352	924002				
Uniprot - rare SAVs	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**		**p value**
pleiotropic (412 proteins)	531	366545	367076	1.46	1.31 -	1.62	< 0.0001
non-pleiotropic	917	923085	924002				
Uniprot - no MAF	**SAVs**	**not SAVs**	**Total**	**OR**	**95% CI**		**p value**
pleiotropic (412 proteins)	890	366186	367076	2.13	1.95 -	2.33	< 0.0001
non-pleiotropic	1054	922948	924002				

Total, the total number of residues; OR, odds ratio; 95%CI, 95% Confidence intervals.

Supp. Table S5 Number of pleiotropic and non-pleiotropic proteins with at least one predicted long disordered region.

Data are presented as number (%) for the 257 proteins pleiotropic dataset and for the 412 proteins pleiotropic dataset. Results are calculated using IUPred and DISOPRED3 and two different cut-offs: ≥30 and ≥50 consecutive disordered amino acids. The *p* value was calculated using the χ^2 test. Highlighted in grey are the results presented in the main text.

Prediction tool	Cut-off	Pleiotropic (257)	Non-pleiotropic (1345)	*p* value
IUPred	30	148 (57.6%)	594 (44.2%)	<0.001
IUPred	50	132 (51.4%)	516 (38.4%)	<0.001
DISOPRED3	30	183 (71.2%)	843 (62.7%)	<0.001
DISOPRED3	50	143 (55.6%)	555 (41.3%)	<0.001

Prediction tool	Cut-off	Pleiotropic (412)	Non-pleiotropic (1345)	*p* value
IUPred	30	231 (56.1%)	594 (44.2%)	<0.001
IUPred	50	209 (50.7%)	516 (38.4%)	<0.001
DISOPRED3	30	287 (69.7%)	843 (62.7%)	0.009
DISOPRED3	50	227 (55.1%)	555 (41.3%)	<0.001

Supp. Table S6 Distribution of disease-causing and non-disease-causing SAVs between ordered and disordered regions.

Results are presented for the three datasets: 257 pleiotropic proteins, 412 pleiotropic proteins and 1345 non-pleiotropic proteins.

257 pleiotropic proteins			
Order versus Disorder	**OR**	**95% CI**	**p value**
Disease-causing SAVs	1.71	1.58-1.85	< 0.001
EXAC common variants	0.65	0.53-0.79	< 0.001
Humsavar all non-disease-causing SAVs	0.95	0.83-1.09	0.47
- Humsavar common SAVs	0.82	0.64-1.05	0.11
- Humsavar rare SAVs	0.80	0.62-1.04	0.09

412 pleiotropic proteins			
Order versus Disorder	**OR**	**95% CI**	**p value**
Disease-causing SAVs	1.87	1.76-1.99	< 0.001
EXAC common variants	0.66	0.57-0.77	<0.001
Humsavar all non-disease-causing SAVs	0.93	0.83-1.04	0.18
- Humsavar common SAVs	0.80	0.66-0.96	0.02
- Humsavar rare SAVs	0.87	0.70-1.07	0.91

1345 non-pleiotropic proteins			
Order versus Disorder	**OR**	**95% CI**	**p value**
Disease-causing SAVs	3.51	3.19-3.85	< 0.001
EXAC common variants	0.76	0.69-0.83	< 0.001
Humsavar all non-disease-causing SAVs	0.87	0.80-0.95	0.002
- Humsavar common SAVs	0.77	0.68-0.87	<0.001
- Humsavar rare SAVs	0.96	0.81-1.14	0.63