*Bioinformatics*, 32(12), 2016, 1761–1765
doi: 10.1093/bioinformatics/btw054
Advance Access Publication Date: 9 February 2016
Discovery Note



Systems biology

# Characterization of clinical signs in the human interactome

**Monica Chagoyen* and Florencio Pazos**

Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), Darwin 3, Madrid 28049, Spain

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on 2 November 2015; revised on 5 January 2016; accepted on 22 January 2016

## Abstract

**Motivation**: Many diseases are related by shared associated molecules and pathways, exhibiting comorbidities and common phenotypes, an indication of the continuous nature of the human pathological landscape. Although it is continuous, this landscape is always partitioned into discrete diseases when studied at the molecular level. Clinical signs are also important phenotypic descriptors that can reveal the molecular mechanisms that underlie pathological states, but have seldom been the subject of systemic research. Here, we quantify the modular nature of the clinical signs associated with genetic diseases in the human interactome.

**Results**: We found that clinical signs are reflected as modules at the molecular network level, to at least to the same extent as diseases. They can thus serve as a valid complementary partition of the human pathological landscape, with implications for etiology research, diagnosis and treatment.

**Contact**: monica.chagoyen@cnb.csic.es

**Supplementary information**: Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Disease-associated phenotypes, also known as clinical signs, are commonly used in clinical practice to describe pathological manifestations. These signs help in the differential diagnosis and treatment of a patient, enabling identification of a particular disease. Diseases can thus be regarded as a special and important type of pathological phenotype.

Most work on the molecular basis of pathologies has focused on diseases or syndromes, instead of clinical signs. This is the case, e.g. in the determination of the genetic variations associated with a given disease (Civelek and Lusis, 2014). Some exceptions include the analysis of quantitative traits (Plomin *et al.*, 2009) (such as height, body mass index, erythrocyte count) or certain common clinical manifestations (alopecia, hearing loss, cardiac arrhythmias).

Diseases are the basic pathological units in the systemic approaches to human pathologies from a network perspective (Furlong, 2013; Piro, 2012). Nevertheless, clinical signs can also be valid, useful descriptors of the human pathological landscape, currently neglected by its traditional partitioning into distinct diseases. If that is the case, partitioning the human pathological landscape into clinical signs should result in a clear link between the molecular and physiological levels.

Clinical signs are thought to be reflected at the molecular network level, possibly to an even greater extent than diseases. There are numerous examples of diseases or groups of diseases with similar phenotypes caused by functionally related genes (Oti and Brunner, 2007). Previous studies observed a correlation between disease similarity as measured by common phenotypes/clinical signs and their similarity due to common molecular mechanisms (such as shared genes, protein interactions, etc.) (van Driel *et al.*, 2006; Zhou *et al.*, 2014). In a recent study, Menche *et al.* (2015) provided evidence for the disease module hypothesis, demonstrating that genes associated with different diseases appear as distinct modules in the still incomplete human interactome. While most of the diseases studied appeared as separate modules, some overlap.

Here, we quantify the extent to which clinical signs, not diseases, are reflected in the human interactome. Due to the lack of data on the genetic basis of most clinical signs, we built a dataset of gene-clinical sign relationships through the genetic diseases they are linked to. As most genetic diseases are monogenic, it is reasonable to

assume that these genes are pleiotropic and make an important contribution to the clinical sign observed. We found that clinical signs are modular. This dataset also allowed us to establish relationships between distinct clinical signs and to compare these relationships with those obtained from shared genes. We find significant overlap in neighborhoods between clinical signs and between signs and complex diseases.

## 2 Methods

### 2.1 Mapping clinical signs on the interactome

We compiled clinical signs and their disease associations from the Human Phenotype Ontology (HPO) (Kohler *et al.*, 2014). Clinical sign–gene associations were inferred from the phenotype–disease relationships defined in HPO, as well as the disease-gene linkages available at OMIM (Online Mendelian Inheritance in Man) (McKusick, 2007) and Orphanet (Orphadata: Free access data from Orphanet © INSERM 1997. Available on http://www.orphadata.org). In this way, a gene was linked with a phenotype if it is associated with at least one disease with the observed phenotype. Gene–phenotype associations were further transferred to the parent terms of a given phenotype in the HPO hierarchy.

We studied only those phenotypes under the 'Phenotypic abnormality' category with at least 25 associated genes in our dataset. This threshold was established in a previous study of diseases (Menche *et al.*, 2015) to account for observable modules in the current incomplete interactome. We further filtered out redundant phenotypes. Two phenotypes were considered identical if they were associated with exactly the same set of genes. A total of 1474 clinical signs (belonging to 'Phenotypic abnormality' in HPO) were associated with at least 25 genes. These correspond to 1433 distinct signs (non-redundant gene sets), which we analyzed.

Genes associated with clinical signs were mapped to the corresponding proteins in the human interactome obtained from the supplementary data of Menche *et al.* (2015). This interactome integrates several types of relationships, including literature (88 349), protein–protein physical interactions (28 653), protein complexes (31 276), regulatory interactions (1335), enzyme-coupled interactions (5325), kinase-substrate pairs (6066) and signaling interactions (32 706).

### 2.2 Measuring localization

In the interactome, each clinical sign is represented as a subnetwork. We used two measures to quantify the modular nature of each subnetwork, as defined for the analysis of diseases (Menche *et al.*, 2015):

- Module size (Si), calculated as the size of the largest connected component.
- Inner distance (di): for each node in the subnetwork, we calculated the shortest distance to another subnetwork node to obtain a distance distribution $P(d)$. We then calculated di as the mean value of $P(d)$.

For both measures, we calculated their statistical significance following the mathematical formalism described for the analysis of diseases (Menche *et al.*, 2015). Briefly, Si and di values were compared with those obtained in a null model. Null models were based on random assignments of *ni* proteins in the interactome to a clinical sign of size *ni* (genes). Statistical significance with random expectation was calculated using 10, 000 randomizations.

For module sizes (Si), z-score statistics was used to assess significance:

$$z-score = \frac{Si - \langle Si^{rand} \rangle}{\sigma(S^{rand})} \quad (1)$$

where $\langle Si^{rand} \rangle$ and $\sigma(Si^{rand})$ correspond to the mean value and standard deviation of the random expectation $P^{rand}(Si)$.

For inner distances (di), we calculated the $P$ value with the Mann–Whitney $U$ test, comparing $P(d)$ with the random distribution $P^{rand}(d)$.

Significance thresholds (Si z-score $\geq$ 1.6 and di p-val $<$ 0.05) were established as in Menche *et al.* (2015).

### 2.3 Measuring relationships

We calculated the network-based separation *Sab* of two clinical signs *a* and *b* as defined in Menche *et al.* (2015):

$$Sab = \langle d_{ab} \rangle - \frac{\langle d_{aa} \rangle + \langle d_{bb} \rangle}{2} \quad (2)$$

where $\langle d_{aa} \rangle$ and $\langle d_{bb} \rangle$ are the mean shortest distances within *a* and *b* clinical signs, respectively, and $\langle d_{ab} \rangle$ is the mean shortest distance between *a* and *b*.

We assessed the significance of all pairs, comparing *Sab* to the network separation obtained for a null model (by random assignment of proteins to two subnetworks). We performed 1000 randomizations and calculated the z-score. A significance threshold z-score $<-1.65$ was established for overlapping pairs ($Sab < 0$), as in Menche *et al.* (2015).

To quantify the relationships between clinical signs and complex diseases, we obtained gene associations to complex diseases from the DisGeNet (Pinero *et al.*, 2015). The network-based separation *Sab* of a clinical sign *a* and a disease *b* was then calculated as above (Eq. 2).

## 3 Results

### 3.1 Clinical signs localization

We analyzed clinical signs as defined in the HPO, which provides a thorough ontology of clinical signs as well as their association to genetic diseases as defined in curated databases (such as OMIM and Orphanet). These curated data allowed us to establish gene associations to 1433 distinct signs that represent non-redundant gene sets.

To assess whether clinical signs (HPO phenotypes) are modular in the human interactome, we measured the size (Si) of the largest connected component formed by their associated genes in the interactome. We also measured the inner distance (di), defined as the mean distance to the closest phenotype protein.

Based on expected deviation of size (Si) and inner distance (di) from random, we found a total of 1154 significantly modular clinical signs (Si z-score $\geq$1.6 and di p-val $<$ 0.05). Most signs (80.53%) are therefore modular according to both measures. Considering only one measure, 1169 signs were significant according to size (Si z-score $\geq$ 1.6) and 1356 according to distance (di p-val $<$ 0.05), with module size thus being the most stringent criterion. Detailed results for all phenotypes analyzed are found in Supplementary Table S1.

As an example, Figure 1 shows the genes associated with the '*café-au lait* spot' phenotype (HPO:0000957) and their interactome context. This is a highly modular phenotype according to our measures (Si z-score = 32.16, di p-val = 2.09e−35). This module includes many genes related to signaling pathways and cancer, as well

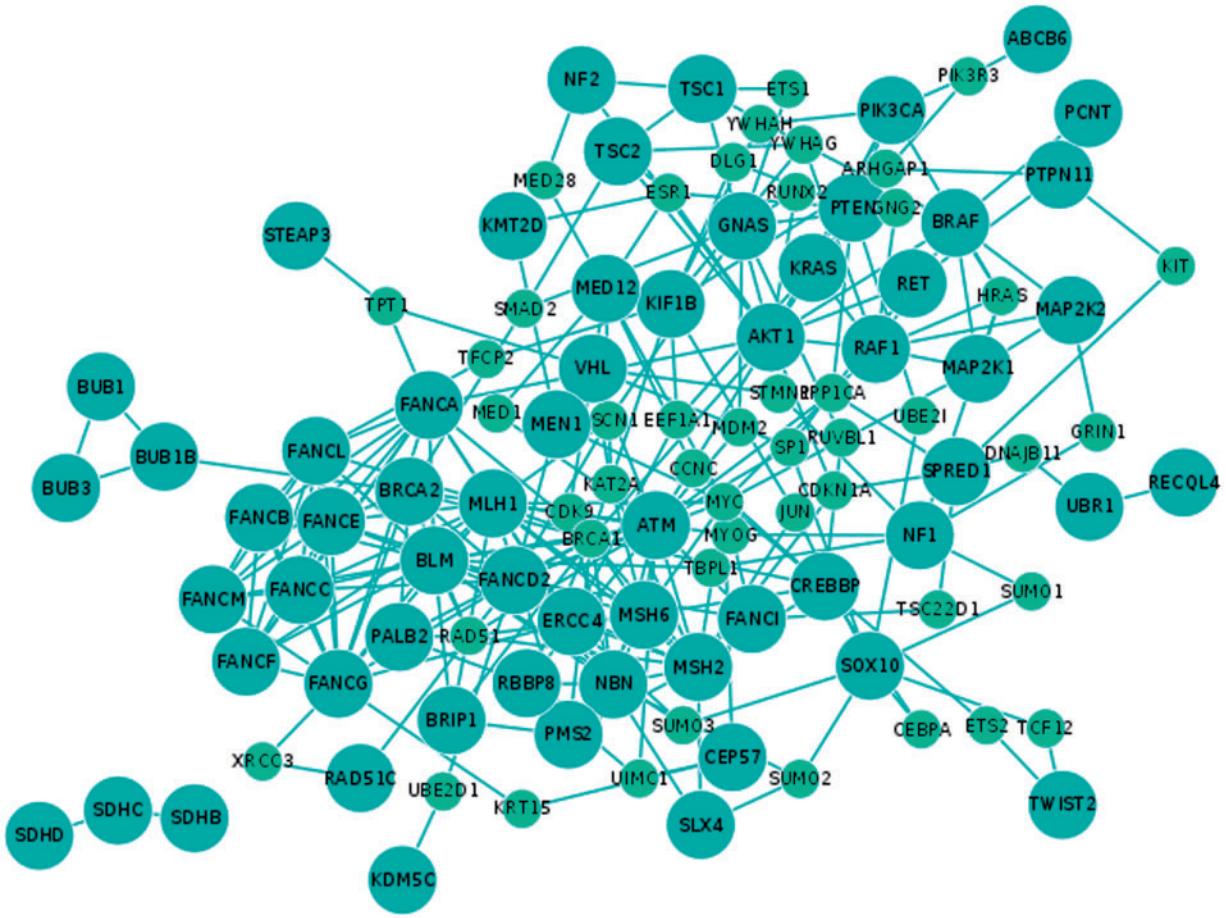


**Fig. 1.** Example of a modular clinical sign: *café-au-lait* spots. Genes associated with diseases that report this sign are mapped in the human interactome. The module consists of 61 protein products (large dark circles), with a largest connected component (Si) of 44 proteins, and inner distance (di) of 1.20. Intermediate nodes that link otherwise isolated protein products (those with inner shortest path length $> 1$) are shown as small light circles.

as other known complexes such as BUB1, 1B and 3 (involved in the mitotic checkpoint), the FAN complex (a nuclear pore complex related to Fanconi anemia) and succinate dehydrogenase (not connected with the rest). An enrichment analysis of this gene set reports 'nucleus', 'DNA repair' and various signaling pathways as the keywords enriched in the set (not shown). *Café-au-lait* spots are in many cases due to abnormal melanocyte proliferation and differentiation. Consequently, the modular nature of CSs can be used, e.g. to extend the set of genes associated to them (as previously done for diseases) so as to obtain more information on their molecular basis.

Only 60 clinical signs (4.19%) are not modular according to any measure. This might be due network incompleteness in the specific regions that affect these clinical signs, or to the fact that some of these signs are not clearly reflected at the molecular level.

### 3.2 Clinical signs relationships

We explored the degree of overlap between clinical signs in the interactome by calculating the network-based separation (*Sab*) of two clinical signs, *a* and *b*, and their statistical significance. A pair of clinical signs overlap if $Sab < 0$.

Due to the way we constructed the clinical sign–gene associations, we anticipated a large degree of overlap between signs. For example, all HPO phenotypes associated with a disease will share genes (those associated with the disease). Due to the HPO hierarchy, a parent phenotype will also contain all the genes associated with its children phenotypes; building relationships among HPO phenotypes based on shared genes will thus report all those linkages. More interesting are the potential relationships between clinical signs defined only by network neighborhoods. We therefore focused on the relationships between the most specific modular clinical signs in our set (comprising 510 HPO phenotypes), among which there are no parent–child relationships in HPO.

A total of 22,162 pairs (17.11%) showed significant overlap (z-score $<-1.65$ and $Sab<0$) (see Supplementary Table S2), while the remaining overlapping pairs (11 946, 9.22%) were not statistically significant. Most pairs (73 850, 57.01%) are significantly non-overlapping or non-overlapping but not significant (21 582, 16.66%).

Supplementary Table 2 lists the pairs of clinical signs related in the interactome (significant overlap). The table is sorted by the z-score of the *Sab* parameter and the number of shared genes, to pinpoint the most interesting cases of clinical signs that do not share associated genes but are close in the interactome. This is the case of the relationship between 'wormian bones' (small extra bones in the cranium, HP:0002645) and 'abnormality of intracranial pressure' (HP:0012640), supported by eight interactions and with no shared genes. These two symptoms are not related in the HPO hierarchy; whereas 'wormian bones' belongs to the 'abnormality of skeletal system' branch, 'abnormality of intracranial pressure' is classified within 'abnormality of the nervous system'.

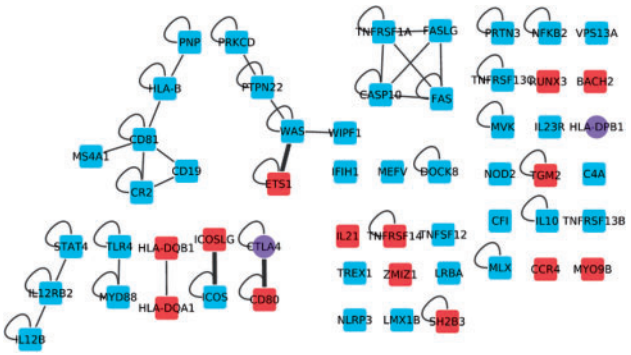


**Fig. 2.** Example of a significantly overlapping relationship between a clinical sign and a complex disease. Celiac disease (dark/red) and vasculitis (light/blue) have two genes in common (circles) and three interactions (wide lines) (Color version of this figure is available at *Bioinformatics* online.)

### 3.3 Clinical signs–complex disease relationships

To identify network-based relationships between clinical signs and complex diseases, we first obtained diseases and their associated genes from DisGeNet (Pinero *et al.*, 2015), an integrated resource that compiles information from several public databases and the literature. Supplementary Table S3 lists the relationships between diseases and clinical signs extracted from the interactome. Again, the most interesting cases are those relationships supported by interactions that share very few genes or no genes at all. An example is 'celiac disease' (umls:C0007570), which is linked to a number of symptoms such as those involving lung problems. A literature and web search showed that such relationships were indeed reported in specific cases (Tarlo *et al.*, 1981). According to our data, celiac disease is also related to the 'vasculitis' clinical sign (inflammation of blood vessels, HP:0002633). This relationship was also reported for specific cases (Caproni *et al.*, 2012; Holdstock and Oleesky, 1970; Meyers *et al.*, 1981); Figure 2 shows the set of genes associated with both features and their interactions. Celiac disease and vasculitis have two genes in common (Fig. 2, purple) and three interactions (wide lines). The Wiskott–Aldrich syndrome (WAS) protein is involved in transduction of membrane receptor signals to the actin cytoskeleton, and the inducible T-cell costimulator (ICOS) is a cell receptor involved in enhancing T-cell responses to antigens. Both are related to vasculitis, and interact respectively with the ETS1 transcription factor (involved in cell proliferation, death and cancer) and the ICOSLG (ICOS ligand, a signaling molecule that enhances T-cell activity). The third interaction is between CD80 (another membrane receptor involved in T-cell proliferation), associated with celiac disease and CTLA4 (also a membrane receptor involved in T-cell inhibition), which is associated with both celiac disease and vasculitis. These interactions, together with the other shared gene (HLA-DPB1, a member of the MHC-II complex involved in antigen presentation), can provide a molecular explanation for the relationship between celiac disease and vasculitis, reported in particular cases, and possibly a way to target this disease symptom in certain patients.

## 4 Discussion

Here, we analyzed the organization of clinical signs associated with genetic diseases in the human interactome, and found that the clinical manifestations of these diseases form modules in the interactome. The size (Si) and compactness (di) of the modules differ due to the current coverage of the known interactome and to the nature of the clinical signs, but in most cases (80.53%) Si and di deviate from random expectation with the most stringent criteria.

We applied the same mathematical formalism developed to estimate the modular nature of a set of complex diseases and groups of diseases (Menche *et al.*, 2015). In their study, Menche *et al.* analyzed diseases and their relationships, and showed that diseases that overlap in the interactome have more similar clinical signs that those that do not. We found that some clinical sign modules overlap with disease modules, and that some of these overlaps are statistically significant.

Our results show that not only genes associated with complex diseases are modular in the interactome, as demonstrated previously, but also genes associated with clinical signs. This offers new alternatives to the study of pathological states of common diseases and their association to genetic determinants, especially for the development of personalized medicine, as it will allow better understanding of a given disease manifestation in a given patient (with a specific set of signs and symptoms) and its relationship with the genetic variations found.

As the human pathological landscape is partitioned into distinct clinical signs as well as into diseases, both can serve as targets for the study of pathological states using molecular network concepts. For diseases with no existing or immediately foreseeable treatments, one could devise ways of targeting some clinical signs by looking at their associated genes and network modules. As the human pathological landscape is a continuum, it can and should be explored beyond traditional disease definitions and classifications.

In conclusion, our results show that clinical signs associated with genetic diseases appear as modules in the interactome, and some show significant overlap. As such, they can be used to identify molecular mechanisms that could explain the manifestations of more complex diseases. To quote Oti *et al.* '*. . . we have traditionally split the human phenome into discrete entities called diseases, instead of searching for common pathogenetic mechanisms that link diseases together [. . . ] However, there is more to the human phenome than a list of diseases*'. Partitioning the human pathological landscape into other (molecularly supported) entities could help to approach diagnosis and treatment of disease states from a new perspective.

## Acknowledgements

We thank the members of the Computational Systems Biology Group (CNB-CSIC) for useful comments and suggestions. Special thanks to Drs. Francisco J. del Castillo (Hospital Ramón y Cajal) and Nuria Malats (CNIO) for interesting discussions. We thank C. Mark for editorial assistance.

## Funding

Spanish Ministry for Economy and Competitiveness [BIO2010-22109].

*Conflict of Interest*: none declared.

## References

Caproni,M. *et al.* (2012) Celiac disease and dermatologic manifestations: many skin clue to unfold gluten-sensitive enteropathy. *Gastroenterol. Res. Pract.*, **2012**, 952753.

Civelek,M. and Lusis,A.J. (2014) Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.*, **15**, 34–48.

Furlong,L.I. (2013) Human diseases through the lens of network biology. *Trends Genet.*, **29**, 150–159.

Holdstock,D.J. and Oleesky,S. (1970) Vasculitis in coeliac diseases. *Br. Med. J.*, **4**, 369.

Kohler,S. *et al*. (2014) *The human phenotype ontology project: linking molecular biology and disease through phenotype data*. *Nucleic Acids Res*., **42**, D966–D974.

McKusick,V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Human Genet*., **80**, 588–604.

Menche,J. *et al*. (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.

Meyers,S. *et al*. (1981) Cutaneous vasculitis complicating coeliac disease. *Gut*, **22**, 61–64.

Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet*., **71**, 1–11.

Oti,M. *et al*. (2008) Phenome connections. *Trends Genet*., **24**, 103–106.

Pinero,J. *et al*. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**, bav028.

Piro,R.M. (2012) Network medicine: linking disorders. *Human Genet*., **131**, 1811–1820.

Plomin,R. *et al*. (2009) Common disorders are quantitative traits. *Nat. Rev. Genet*., **10**, 872–878.

Tarlo,S.M. *et al*. (1981) Association between celiac disease and lung disease. *Chest*, **80**, 715–718.

van Driel,M.A. *et al*. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet*., **14**, 535–542.

Zhou,X. *et al*. (2014) Human symptoms-disease network. *Nat. Commun*., **5**, 4212.