

# Interactome and gene expression based analyzes for translational medicine

Emre Güney, PhD

Hospital del Mar Research Institute (**IMIM**)  
& Pompeu Fabra University(**UPF**)

*MSc on Omics Data Analysis - Bioinformatics Applications*  
January 17<sup>th</sup>, 2018



Institut Hospital del Mar  
d'Investigacions Mèdiques



RESEARCH  
PROGRAMME  
ON BIOMEDICAL  
INFORMATICS



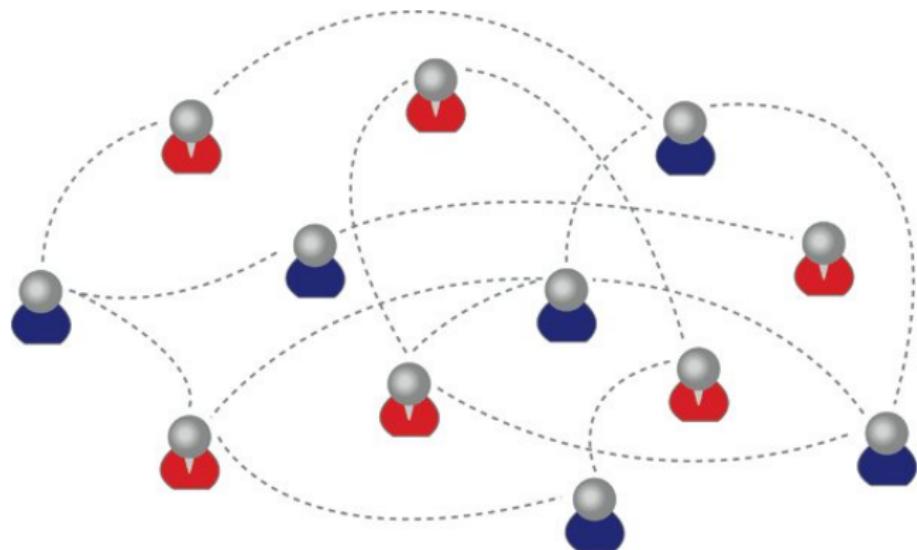
Facebook friends

~85%

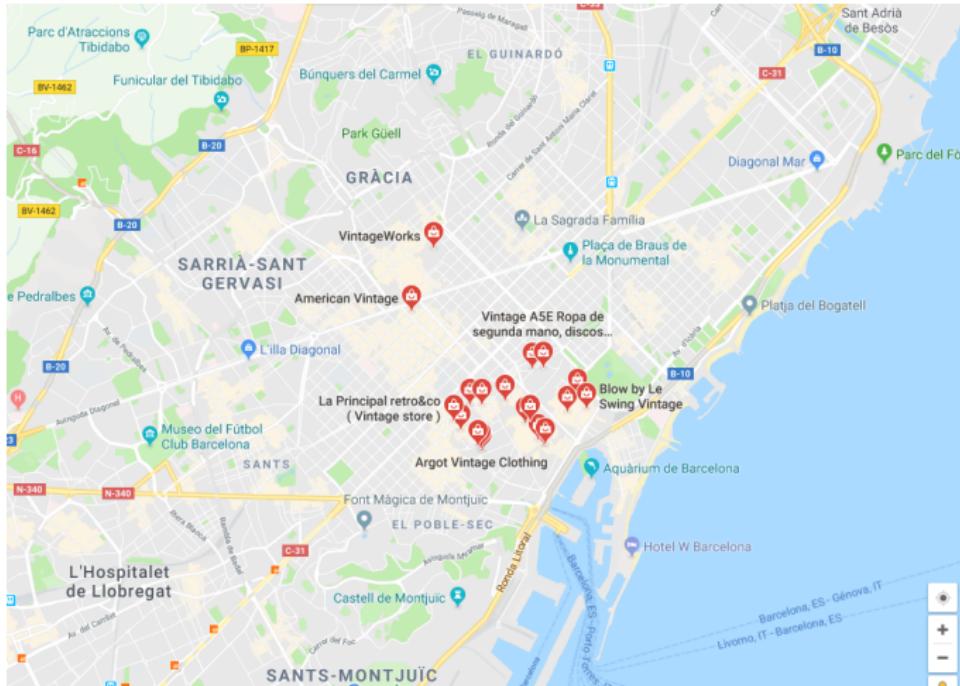
Percentage of your Facebook friends from the **same country**

(Ugander et al., 2011, arXiv:1111.4503)

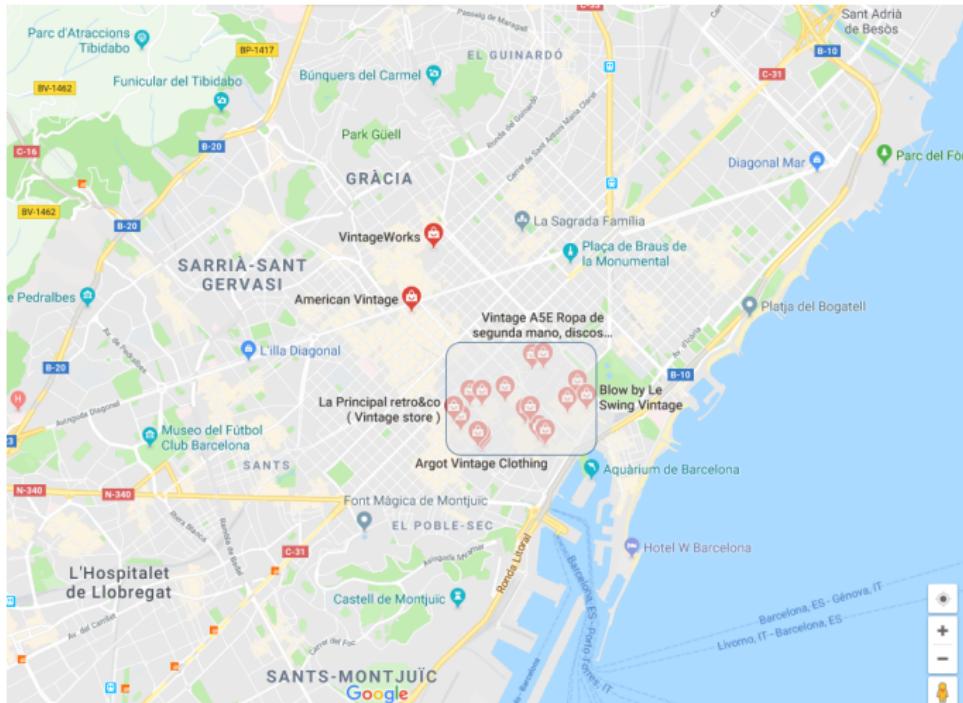
~85%



# Clustering & Guilt-by-association

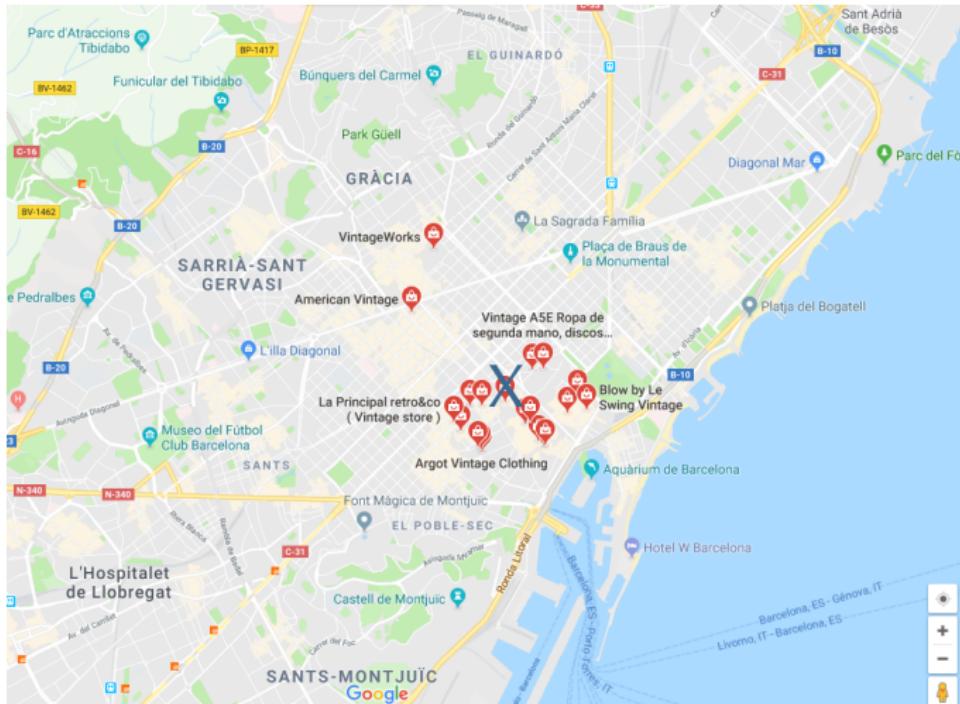


# Clustering & Guilt-by-association



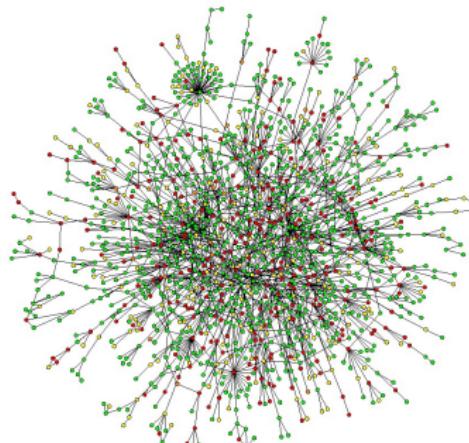
clustering (Vintage shops are located in the center)

# Clustering & Guilt-by-association



guilt-by-association (X is a vintage shop)

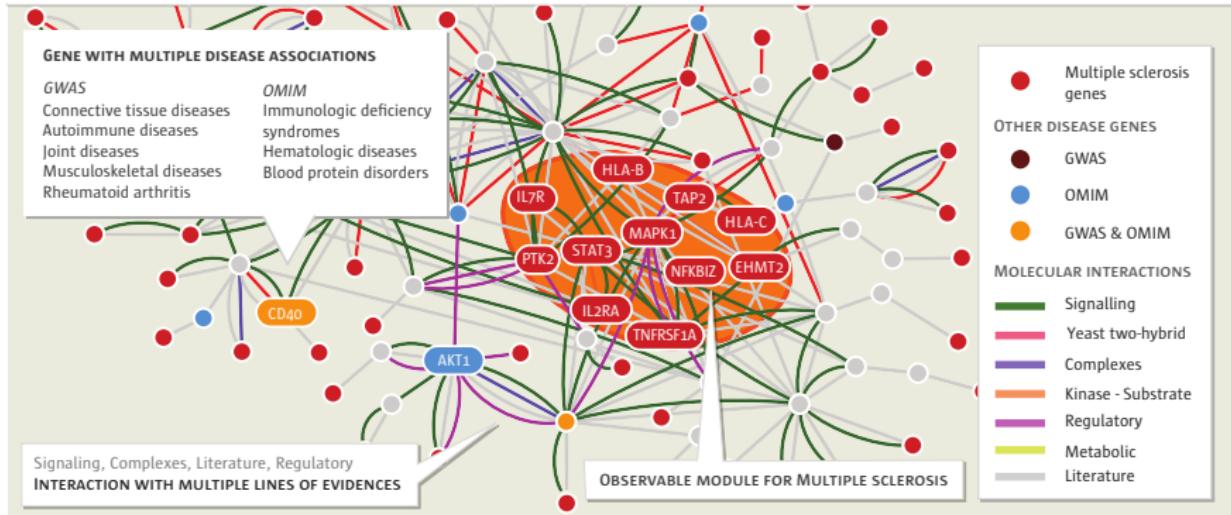
## Interactome: The cellular map



(Barabási and Oltvai, 2004, Nat Rev Genet)

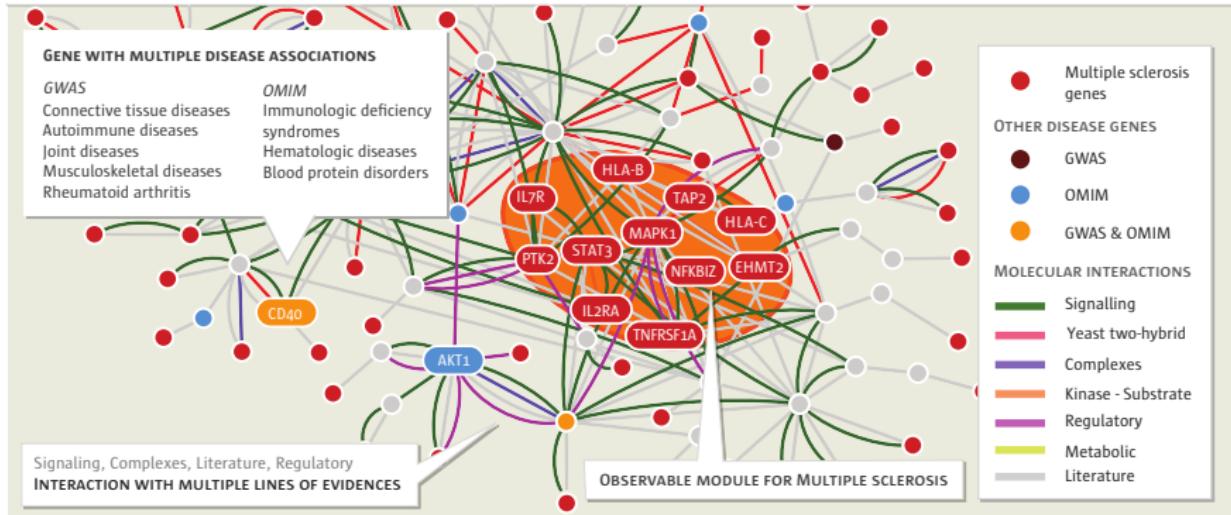
- Proteins “talk to each other” by physically interacting with each other
- These interactions are essential for performing biological processes
- The network of interactions between proteins: **Interactome**

# Guilt-by-association in the interactome



(Menche et al., 2015, Science)

# Guilt-by-association in the interactome



(Menche et al., 2015, Science)

How to generate and analyze the interactome?

*BIANA: Biological Interaction And Network Analysis*

SOFTWARE

Open Access

# Biana: a software framework for compiling biological interactions and analyzing networks

Javier Garcia-Garcia, Emre Guney, Ramon Aragues, Joan Planas-Iglesias, Baldo Oliva\*

## Abstract

**Background:** The analysis and usage of biological data is hindered by the spread of information across multiple repositories and the difficulties posed by different nomenclature systems and storage formats. In particular, there is an important need for data unification in the study and use of protein-protein interactions. Without good integration strategies, it is difficult to analyze the whole set of available data and its properties.

**Results:** We introduce BIANA (Biologic Interactions and Network Analysis), a tool for biological information integration and network management. BIANA is a Python framework designed to achieve two major goals: i) the integration of multiple sources of biological information, including biological entities and their relationships, and ii) the management of biological information as a network where entities are nodes and relationships are edges. Moreover, BIANA uses properties of proteins and genes to infer latent biomolecular relationships by transferring edges to entities sharing similar properties. BIANA is also provided as a plugin for Cytoscape, which allows users to visualize and interactively manage the data. A web interface to BIANA providing basic functionalities is also available. The software can be downloaded under GNU GPL license from <http://sbi.imim.es/web/BIANA.php>.

**Conclusions:** BIANA's approach to data unification solves many of the nomenclature issues common to systems dealing with biological data. BIANA can easily be extended to handle new specific data repositories and new specific data types. The unification protocol allows BIANA to be a flexible tool suitable for different user requirements: non-expert users can use a suggested unification protocol while expert users can define their own specific unification rules.

## Protein Interaction Data

Protein-protein interaction (PPI) data is spread across various repositories (similar to the data on social interactions that is spread across Facebook, Twitter, Instagram, etc.)

DIP

MIPS



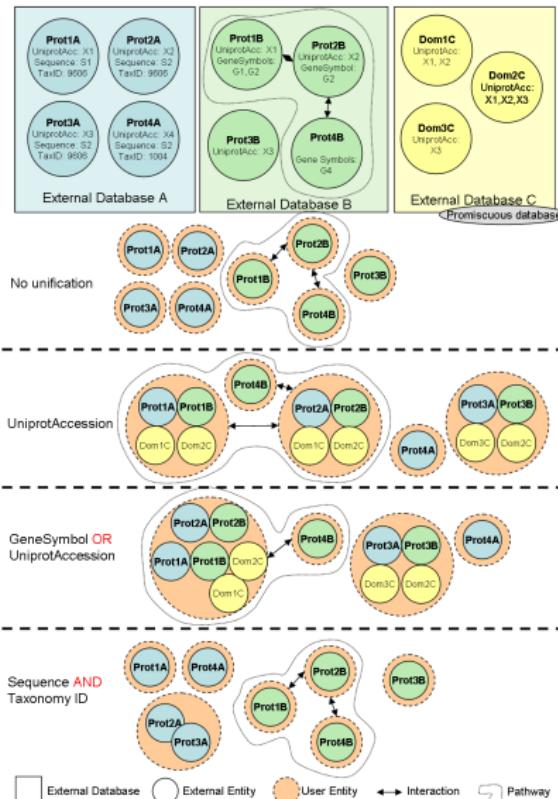
*IntAct*

STRING

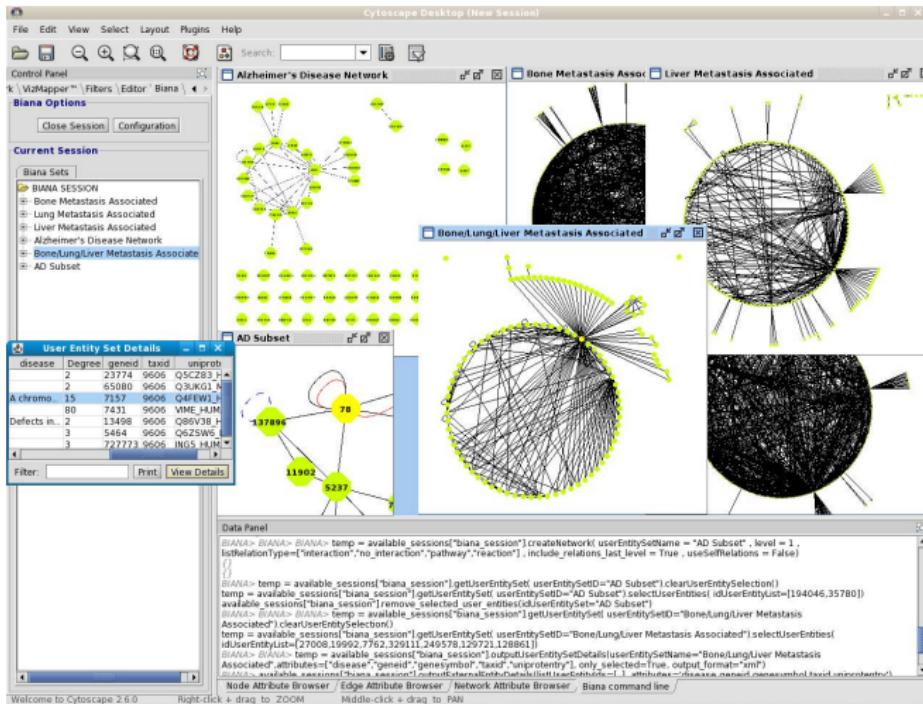
MINT

BioGrid

# Integrating Protein Interaction Data



# Analyzing Protein Interaction Data



## PPICoin

A “new” pseudo-crypto currency for computational biologists

### Rules for block generation and transaction verification

- Start a new block with the gene (protein X) associated to *“Nemaline myopathy 3”* in DisGeNET
- Add a protein Y interacting with X to the block such that it interacts with at least one other protein Z
- If there are multiple candidate proteins Y, add the one which has the smallest ENTREZ gene id
- Write that protein Y on a paper (along with your name) and pass it to the person on your left
- When you receive the paper, verify the correctness of previous transaction and repeat the steps above to generate a new one

## Analyzing Protein Interaction Data – Hands On!

Given  $G(V, E)$  the interactome represented as a graph ( $V$ : proteins,  $E$ : interactions) and  $k$  is node degree (number of the node's neighbors).

PPICoin:

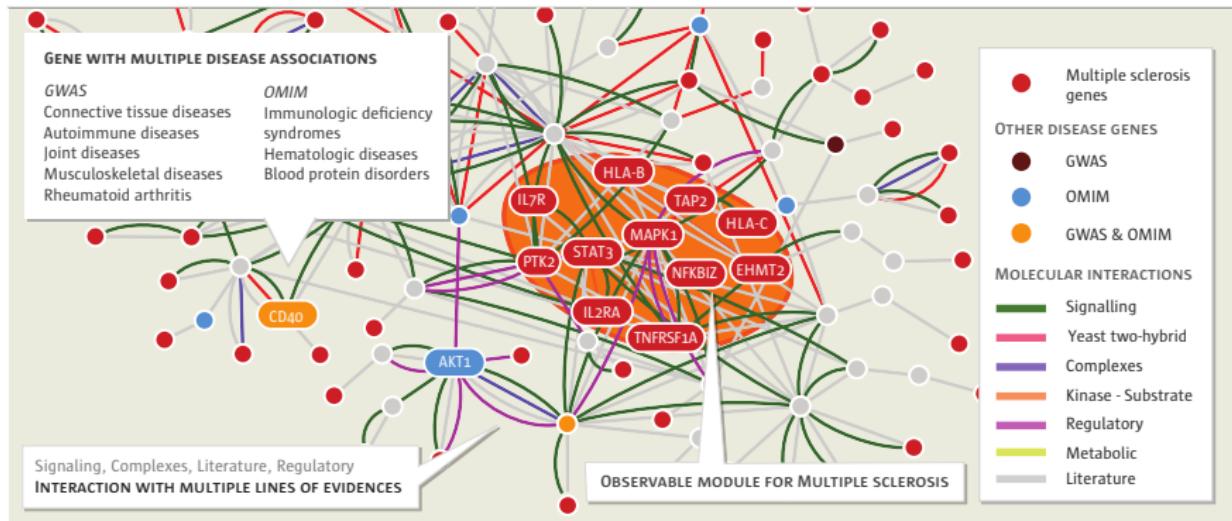
|      |       |     |     |
|------|-------|-----|-----|
| X    | Y     | Z   | ... |
| Emre | Alice | Bob | ... |

Subject to:

- $X, Y \in V$
- $Candidates(Y) = \{y : (X, y) \in E, k_y > 1\}$
- $Y = \operatorname{argmin}_{Geneid}(Candidates(Y))$

| Resource                         | URL  |
|----------------------------------|--|
| <b>Disease-gene information</b>  |  |
| DisGeNET                         | <a href="http://www.disgenet.org">www.disgenet.org</a>                             |
| <b>Interactome generation</b>    |  |
| BIANA (web server)               | <a href="http://sbi.imim.es/BIANA.php">sbi.imim.es/BIANA.php</a>                   |
| BIANA generated files            | <a href="http://emreguney.net/doc/ms_omics.zip">emreguney.net/doc/ms_omics.zip</a> |
| <b>Interactome visualization</b> |  |
| Cytoscape                        | <a href="http://www.cytoscape.org">www.cytoscape.org</a>                           |

# Guilt-by-association in the interactome

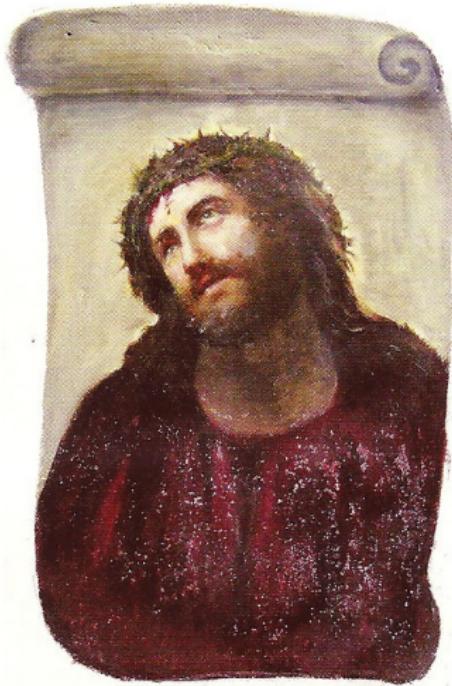


(Menche et al., 2015, Science)

How to use interactome neighborhood for extracting biologically meaningful information?

A similar problem in a different domain: Image recovery

guilt-by-association



A similar problem in a different domain: Image recovery

guilt-by-association



*GUILD: Genes Underlying Inheritance Linked Disorders*

# Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization

Emre Guney, Baldo Oliva\*

Structural Bioinformatics Group (GRIB), Universitat Pompeu Fabra, Barcelona Research Park of Biomedicine (PRBB), Barcelona, Catalonia, Spain

## Abstract

Complex genetic disorders often involve products of multiple genes acting cooperatively. Hence, the pathophenotype is the outcome of the perturbations in the underlying pathways, where gene products cooperate through various mechanisms such as protein-protein interactions. Pinpointing the decisive elements of such disease pathways is still challenging. Over the last years, computational approaches exploiting interaction network topology have been successfully applied to prioritize individual genes involved in diseases. Although linkage intervals provide a list of disease-gene candidates, recent genome-wide studies demonstrate that genes not associated with any known linkage interval may also contribute to the disease phenotype. Network based prioritization methods help highlighting such associations. Still, there is a need for robust methods that capture the interplay among disease-associated genes mediated by the topology of the network. Here, we propose a genome-wide network-based prioritization framework named GUILD. This framework implements four network-based disease-gene prioritization algorithms. We analyze the performance of these algorithms in dozens of disease phenotypes. The algorithms in GUILD are compared to state-of-the-art network topology based algorithms for prioritization of genes. As a proof of principle, we investigate top-ranking genes in Alzheimer's disease (AD), diabetes and AIDS using disease-gene associations from various sources. We show that GUILD is able to significantly highlight disease-gene associations that are not used *a priori*. Our findings suggest that GUILD helps to identify genes implicated in the pathology of human disorders independent of the loci associated with the disorders.

**Citation:** Guney E, Oliva B (2012) Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization. PLoS ONE 7(9): e43557. doi:10.1371/journal.pone.0043557

**Editor:** Narcis Fernandez-Fuentes, Aberystwyth University, United Kingdom

**Received** March 19, 2012; **Accepted** July 23, 2012; **Published** September 21, 2012

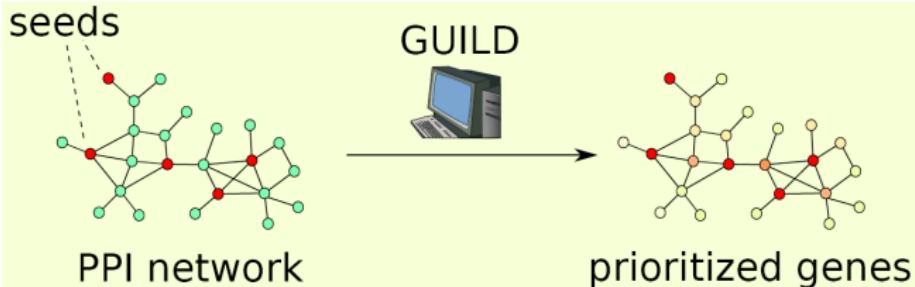
**Copyright:** © 2012 Guney, Oliva. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** Departament d'Educació i Universitats de la Generalitat de Catalunya i del Fons Social Europeu (Department of Education and Universities of the Generalitat of Catalonia and the European Social Fund), Spanish Ministry of Science and Innovation (MICINN), FEDER (Fonds Européen de Développement Régional) BIO2008-0205, BIO2011-22568, PSE-0100000-2007, and PSE-0100000-2009; and by EU grant EraSysbio+ (SHIPREC) Euroinvestigación (EUI2009-04018). 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.

\* E-mail: baldo.oliva@upf.edu

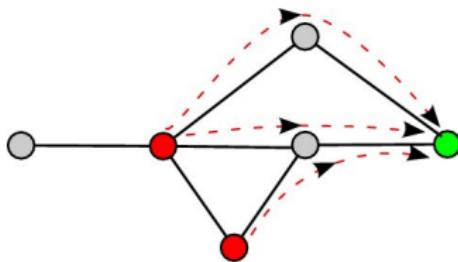
## GUILD framework



- NetScore
- NetZcore
- NetShort
- NetCombo
- Functional Flow (*Nabieva et al., 2005*)
- Random walk with restart (*Kohler et al., 2008*)
- PageRank with priors (*Chen et al., 2009*)
- Network propagation (*Vanunu et al., 2011*)

## NetScore

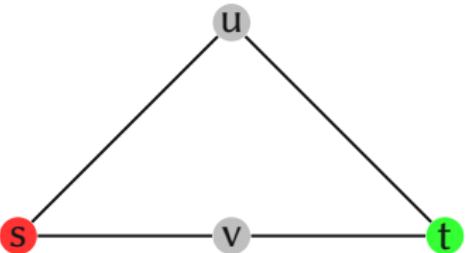
- considers multiple shortest paths in-between the two nodes



$$message = \left\{ \begin{array}{l} \textit{source} \\ \textit{timestamp} \\ \textit{path\_weight} \end{array} \right\}$$

|          |
|----------|
| <b>u</b> |
| $m_u$    |
| :        |
| $m_v$    |
| :        |

# NetScore



```

Input:  $G = (V, E)$  graph with score property  $\text{score}: V \rightarrow [0, 1]$ ,  

       weight property  $\text{weight}: E \rightarrow [0, 1]$ ,  $n\text{Repetition}$ ,  $n\text{Iteration}$ .  

Output:  $G = (V, E)$  graph with score property  $\text{score}': V \rightarrow [0, 1]$ .  

for  $i=1$  to  $n\text{Repetition}$  do  

    /* Initialize message arrays  

    foreach  $u \in V$  do  

         $m.\text{source} \leftarrow u$  // source node id  

         $m.\text{timestamp} \leftarrow 0$  // the iteration when received  

         $m.\text{path.weight} \leftarrow 1$  // weights of the traveled path  

         $messages(u) \leftarrow \{m\}$   

    end  

    for  $j=1$  to  $n\text{Iteration}$  do  

        /* Digest messages  

        foreach  $u \in V$  do  

            foreach  $v \in \{(u, v) \in E\}$  do  

                foreach  $m \in messages(v)$  do  

                    /* Do not accept messages from the same node  

                     received during previous iterations */  

                    if AcceptMessage( $messages(u)$ ,  $m$ ) then  

                         $m.\text{timestamp} \leftarrow j$   

                         $m.\text{path.weight} \leftarrow m.\text{path.weight} * \text{weight}(u, v)$   

                         $messages(u) \leftarrow messages(u) \cup m$   

                end  

            end  

        end  

        /* Update node scores  

        foreach  $u \in V$  do  

            foreach  $m \in messages(u)$  do  

                 $score(u) \leftarrow score(u) + m.\text{path.weight} *$   

                 $score(m.\text{source})$   

            end  

             $score(u) \leftarrow score(u) / \|messages(u)\|$   

        end  

    end

```

```

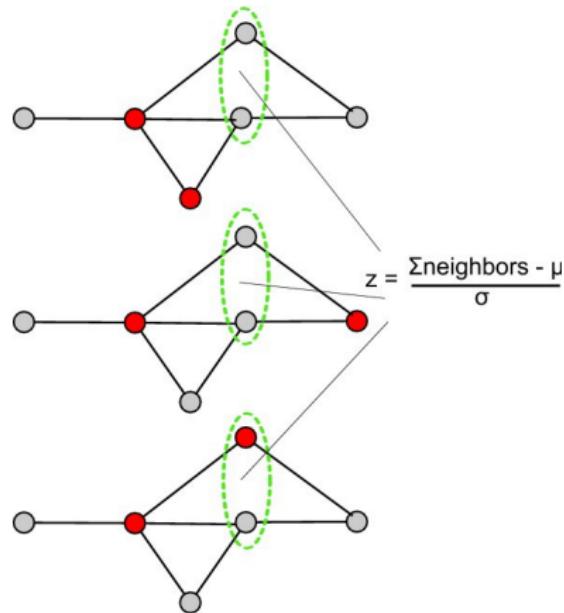
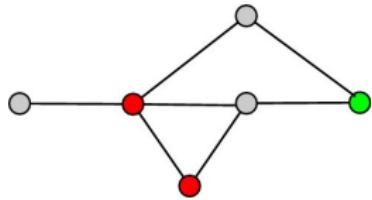
end
end

```

|             | s        | u       | v        | t        |
|-------------|----------|---------|----------|----------|
| initially   | $m_s^0$  | $m_u^0$ | $m_v^0$  | $m_t^0$  |
| iteration 1 | $m_u^1$  | $m_s^1$ | $m_u^1$  | $m_u^1$  |
| iteration 2 | $2m_t^2$ | $m_t^1$ | $2m_u^2$ | $2m_s^2$ |

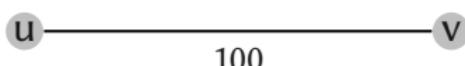
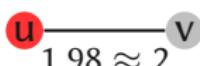
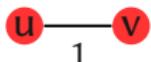
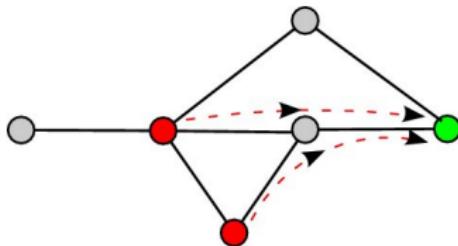
## NetZcore

- checks the “significance” of the neighborhood configuration



## NetShort

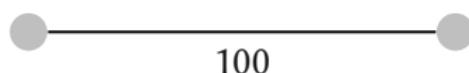
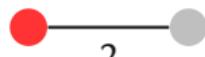
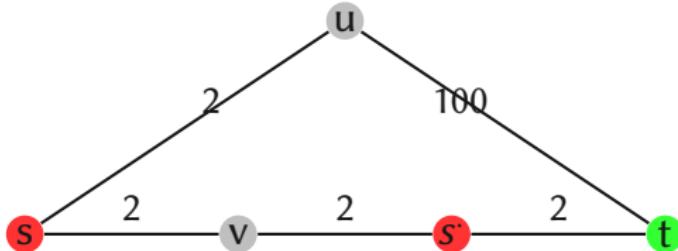
- incorporates “disease-relevance” of a path
- the more seeds the path has, the shorter it is



$$score(u) = \begin{cases} 1, & \text{if } u \text{ is seed} \\ 0.01, & \text{otherwise} \end{cases}$$

$$weight(u, v) = \frac{1}{(score(u) + score(v))/2}$$

## NetShort



$$path\_length(s, t)_{\{s, u, t\}} = 2 + 100 = 102$$

$$path\_length(s, t)_{\{s, v, s', t\}} = 2 + 2 + 2 = 6$$

## NetCombo

- combines the scores of the three methods
- calculates a z-score for each node using the score distribution of each method
- averages the z-score from all three algorithms for each node

$$score_u^{NetCombo} = \frac{1}{3} * \sum_{method \in \{NetScore, NetZcore, NetCombo\}} \frac{score_u^{method} - \mu^{method}}{\sigma^{method}}$$

## GUID accuracy

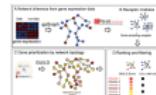
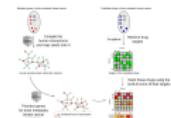
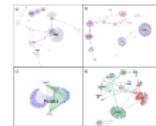
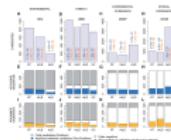
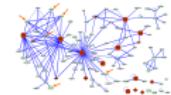
| Data Set | Metric | NetScore     | NetZcore | NetShort | NetCombo     | Func. Flow | PageRank | Random Walk | Network Prop. |
|----------|--------|--------------|----------|----------|--------------|------------|----------|-------------|---------------|
| OMIM     | AUC    | 67.49        | 62.99    | 65.63    | <b>72.09</b> | 58.55      | 57.03    | 55.36       | 65.97         |
|          | Sens.  | 20.69        | 19.62    | 15.41    | 21.46        | 22.31      | 10.76    | 14.64       | <b>23.24</b>  |
| Goh      | AUC    | <b>67.32</b> | 61.45    | 55.36    | 67.08        | 54.78      | 52.39    | 49.35       | 54.74         |
|          | Sens.  | <b>11.61</b> | 11.05    | 4.88     | 11.34        | 6.22       | 4.00     | 5.69        | 8.66          |
| Chen     | AUC    | 75.92        | 72.80    | 63.11    | <b>78.41</b> | 63.56      | 65.30    | 61.78       | 69.07         |
|          | Sens.  | <b>18.89</b> | 12.84    | 9.06     | 17.51        | 12.43      | 6.00     | 9.64        | 15.30         |

Improvement over existing methods is significant

- for NetCombo in all three datasets
- for NetScore in Goh and Chen datasets

# Several applications of GUILD

- implicating genes in AD (*Guney and Oliva, 2012, PLoS ONE*)
- extending apoptosis pathway (*Planas et al., 2012, OMICS: A Jour Int Bio*)
- prioritizing genes in bone metastatic breast cancer (*Santana-Codina et al., 2013, Mol & Cel Prot*)
- identifying subnetworks driving brain and lung metastasis in breast cancer (*Engin et al., 2013, PLoS ONE*)
- analysis of functional diversity of disease genes (*Guney and Oliva, 2014, PLoS ONE*)
- drug repurposing in brain metastatic breast cancer (*Martinez-Aranda et al., 2015, Oncotarget*)
- prioritization of regulatory modules in insulin secretion (*Hänelmann et al., 2015, Islets*)



*GUILDify: GUILD web server*

Systems biology

Advance Access publication February 14, 2014

## GUIDify: a web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms

Emre Guney<sup>†</sup>, Javier Garcia-Garcia and Baldo Oliva\*

Departament de Ciències Experimentals i de la Salut, Structural Bioinformatics Group (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona, 08003 Catalonia, Spain

Associate Editor: Alfonso Valencia

### ABSTRACT

**Summary:** Determining genetic factors underlying various phenotypes is hindered by the involvement of multiple genes acting cooperatively. Over the past years, disease–gene prioritization has been central to identify genes implicated in human disorders. Special attention has been paid on using physical interactions between the proteins encoded by the genes to link them with diseases. Such methods exploit the guilt-by-association principle in the protein interaction network to uncover novel disease–gene associations. These methods rely on the proximity of a gene in the network to the genes associated with a phenotype and require a set of initial associations. Here, we present GUIDify, an easy-to-use web server for the phenotypic characterization of genes. GUIDify offers a prioritization approach based on the protein–protein interaction network where the initial phenotype–gene associations are retrieved via free text search on biological databases. GUIDify web server does not restrict the prioritization to any predefined phenotype, supports multiple species and accepts user-specified genes. It also prioritizes drugs based on the ranking of their targets, unleashing opportunities for repurposing drugs for novel therapies.

**Availability and implementation:** Available online at <http://sbi.imim.es/GUIDify.php>

**Contact:** baldo.oliva@upf.edu

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

Methods using protein–protein interactions (PPIs) exploit the ‘guilt-by-association’ principle over the network topology to uncover new disease–gene associations. The guilt-by-association principle suggests that the genes whose products (proteins) interact with the products of known disease genes are more likely to be disease genes (Aerts *et al.*, 2006; Lage *et al.*, 2007). Recently, we proposed three novel algorithms for genome-wide prioritization of disease genes using PPI networks and showed that a consensus method combining these algorithms improved the prioritization (Guney and Oliva, 2012) when using the disease–gene associations in Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al.*, 2005). Combined with genomics and proteomics data, the method has been successfully used to identify a gene driving metastasis to bone in breast cancer (Santana-Codina *et al.*, 2013).

Available network-based prioritization tools use either disease–gene annotations from OMIM database (Gottlieb *et al.*, 2011; Kohler *et al.*, 2008) or a set of genes provided by the user (Chen *et al.*, 2009; Kacprowski *et al.*, 2013; Warde-Farley *et al.*, 2010) as initial associations (seed genes). These tools typically output the prioritization for a set of candidate genes; genes lying under a given genomic interval, a set of user-provided genes or top ranking genes (several hundred at most). Furthermore, some of these tools are accessible only through Cytoscape (Saito *et al.*, 2012) as a plugin (Gottlieb *et al.*, 2011; Kacprowski *et al.*, 2013).

# GUILDify

The screenshot shows the GUILDify Web Server interface. At the top, there's a navigation bar with links for Home, Documentation, GUILD, BIANA, and SBI Group. Below the navigation bar, the main content area has two search input fields: one for 'alzheimer' and another for 'Homo sapiens'. Between these fields is a button labeled 'Search in BIANA Knowledge base'. Below the search fields, there's a section titled 'Try it with the following examples:' followed by a list of search options: [Keywords] [Keywords (AND)] [Keywords (OR)] [Genes]. A detailed explanatory text follows, describing the network-topology based prioritization algorithm used by GUILD to score relevance of gene products. It mentions the BIANA knowledge base, the compilation of data from various sources, and the creation of a species-specific interaction network. At the bottom of the interface, it says 'Powered by PyriNet'.

- free text search on UniProt, OMIM, DisGeNET databases and GO annotations for a given species
- matching genes are retrieved and used as seeds
- the PPI network is compiled from DIP, HPRD, IntAct, MINT, MPact, BioGRID, BIND
- tissue-specific PPI networks for human using HPA and Tissues databases

# GUILDify

## GUILDify Web Server - Matching Proteins

35 BIANA entries are found for the query **alzheimer** in *Homo sapiens*. 2 of these entries have no interactions (not in the network).

GUILDify! using selected entries below [ Options ]

Options

NetScore (Repetition: 3 ▾ Iteration: 2 ▾)  
 NetZcore (Iteration: 5 ▾)  
 NetShort

[ Hide ]

### BIANA entries in the network

[ Select All / None ]

| Keep                                | Gene ID | UniProt ID | Gene Symbol | Description  |
|-------------------------------------|---------|------------|-------------|--|
| <input checked="" type="checkbox"/> | 2       | P01023     | A2M         | <a href="#">[alzheimer disease, susceptibility to]</a> , 104300 (3) <a href="#">[omim]</a><br>May function as a general inhibitor of the histone deacetylase HDAC1. Binding to the pocket region of RB1 may displace HDAC1 from RB1/E2F complexes, leading to activation of E2F-target genes and cell cycle progression. Conversely, displacement of HDAC1 from SP1 bound to the CDKN1A promoter leads to increased expression of this CDK inhibitor and blocks cell cycle progression. Also antagonizes PAWR mediated induction of aberrant amyloid peptide production in <a href="#">alzheimer</a> disease (presenile and senile dementia), although the molecular basis for this phenomenon has not been described to date. <a href="#">[swissprot]</a>   |
| <input checked="" type="checkbox"/> | 26574   | Q9NY61     | AATF        | <a href="#">[alzheimer disease, susceptibility to]</a> , 104300 (3) <a href="#">[omim]</a><br>Cleaves aggrecan, a cartilage proteoglycan, and may be involved in its turnover. May play an important role in   |
| <input checked="" type="checkbox"/> | 1636    | P12821     | ACE         | <a href="#">[alzheimer disease, susceptibility to]</a> , 104300 (3) <a href="#">[omim]</a><br>The destruction of aggrecan in arthritic diseases. Could also be a critical factor in the exacerbation of neurodegeneration in <a href="#">alzheimer</a> disease. Cleaves aggrecan at the 392-Glu-[Ala-393] site. <a href="#">[swissprot]</a>  |
| <input checked="" type="checkbox"/> | 323     | Q92B70     | APBB2       | <a href="#">[alzheimer disease, late-onset]</a> , 104300 (3) <a href="#">[omim]</a><br><a href="#">[alzheimer disease]</a> <a href="#">[swissprot]</a><br>Defects in APOE are a cause of hyperlipoproteinemia type 3 (HLP3) [MIM:107741]; also known as familial dysbetaipoproteinemia. Individuals with HLP3 are clinically characterized by xanthomas, yellowish lipid deposits in the palmar crease, or less specific on tendons and on elbows. The disorder rarely manifests before the third decade in men. In women, it is usually expressed only after the menopause. The vast majority of the patients are homozygous for APOE*2 alleles. More severe cases of HLP3 have also been observed in individuals heterozygous for rare APOE variants. The influence of APOE on lipid levels is often suggested to have major implications for the risk of coronary artery disease (CAD). Individuals carrying the common APOE*4 variant are at higher risk of CAD. Genetic variations in APOE are associated with <a href="#">alzheimer</a> disease type 2 (AD2) [MIM:104310]. It is a late-onset neurodegenerative disorder characterized by progressive dementia, loss of cognitive abilities, and deposition of fibrillar amyloid proteins as intraneuronal neurofibrillary tangles, extracellular amyloid plaques and vascular amyloid deposits. The major constituent of these plaques is the neurotoxic amyloid-beta-APP 40-42 peptide (s), derived proteolytically from the transmembrane precursor protein APP by sequential secretase processing. The cytosolic C-terminal fragments (CTFs) and the caspase-cleaved products such as C31 derived from APP, are also implicated in neuronal death. |
| <input checked="" type="checkbox"/> | 348     | P02649     | APOE        | Note: The APOE*4 allele is genetically associated with the common late onset familial and sporadic forms of <a href="#">alzheimer</a> disease. Risk for AD increased from 20% to 90% and mean age at onset decreased from 84 to 68 years with increasing number of APOE*4 alleles in 42 families with late onset AD. Thus APOE*4 gene dose is a major risk factor for late onset AD and, in these families, homozygosity for APOE*4 was virtually sufficient to cause AD by age 80. The mechanism by which APOE*4 participates in pathogenesis is not known. Defects in  |

# GUILDFy

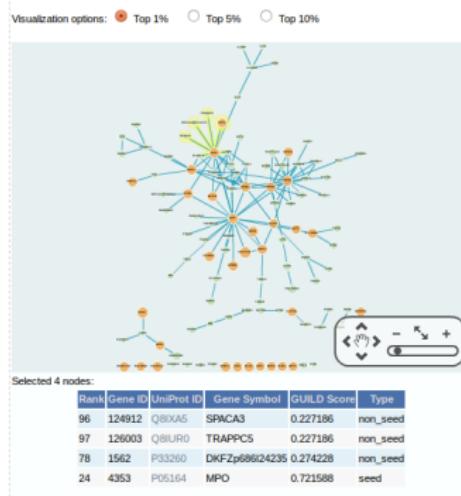
## GUILDFy Web Server - Results

Calculated scores for proteins in the BIANA Homo sapiens interactome:

[ Download all scores ] [ Download seed proteins ] [ Download interactome ]

Proteins 1 - 20 of 11238 ||<< << >> >>||

| Rank | Gene ID | UniProt ID | Gene Symbol  | GUILDFy Score |
|------|---------|------------|--------------|---------------|
| 1    | 84670   | Q9BX50     | COL25A1      | 1.0000        |
| 2    | 27123   | Q9UBLU     | seed:DKK2    | 0.9398        |
| 3    | 348     | P02649     | seed:APOE    | 0.8481        |
| 4    | 6620    | Q16143     | seed:SNCB    | 0.8435        |
| 5    | 10531   | Q5JRX0     | seed:PITRM1  | 0.8373        |
| 6    | 5664    | P49810     | seed:PSEN2   | 0.8123        |
| 7    | 5663    | P49768     | seed:PSEN1   | 0.7905        |
| 8    | 6653    | Q92673     | seed:SORL1   | 0.7861        |
| 9    | 1795    | Q8ZD9      | seed:DOCK3   | 0.7858        |
| 10   | 1636    | P12821     | seed:ACE     | 0.7848        |
| 11   | 5328    | P00749     | seed:PLAU    | 0.7824        |
| 12   | 351     | P05067     | seed:APP     | 0.7774        |
| 13   | 323     | Q92870     | seed:APBB2   | 0.7756        |
| 14   | 9507    | O75173     | seed:ADAMTS4 | 0.7754        |
| 15   | 4846    | P29474     | seed:NOS3    | 0.7717        |
| 16   | -       | Q8IVG9     | seed:MT-RNR2 | 0.7705        |
| 17   | 2       | P01023     | seed:A2M     | 0.7599        |
| 18   | 642     | Q13867     | seed:BLMH    | 0.7556        |
| 19   | 81035   | Q5KU26     | seed:COLEC12 | 0.7523        |
| 20   | 6622    | P37840     | seed:SNC     | 0.7518        |



### Relationship between Alzheimer's disease and type 2 diabetes

Investigate the common genetic factors involved in both diseases

| Resource                                 | URL  |
|--|--|
| <b><i>Disease-gene information</i></b>   |  |
| DisGeNET                                 | <a href="http://www.disgenet.org">www.disgenet.org</a>                   |
| <b><i>Interactome-based analysis</i></b> |  |
| GUILDify web server                      | <a href="http://sbi.imim.es/GUILDify2.php">sbi.imim.es/GUILDify2.php</a> |
| <b><i>Scripting</i></b>                  |  |
| Your favorite tool                       | bash / python / R / perl / ...   |

# Patient-level heterogeneity

## IMPRECISION MEDICINE

For every person they do help (blue), the ten highest-grossing drugs in the United States fail to improve the conditions of between 3 and 24 people (red).

**1. ABILIFY (aripiprazole)**  
Schizophrenia



**2. NEXIUM (esomeprazole)**  
Heartburn



**3. HUMIRA (adalimumab)**  
Arthritis



**4. CRESTOR (rosuvastatin)**  
High cholesterol



**5. CYMBALTA (duloxetine)**  
Depression



**6. ADVAIR DISKUS (fluticasone propionate)**  
Asthma



**7. ENBREL (etanercept)**  
Psoriasis



**8. REMICADE (infliximab)**  
Crohn's disease



**9. COPAXONE (glatiramer acetate)**  
Multiple sclerosis



**10. NEULASTA (pegfilgrastim)**  
Neutropenia



Based on published number needed to treat (NNT) figures. For a full list of references, see Supplementary Information at [go.nature.com/4dr78f](http://go.nature.com/4dr78f).

Schork, 2015, Nature

*PEPPER: PErsonalized exPression ProfilER*

## ARTICLE

## OPEN

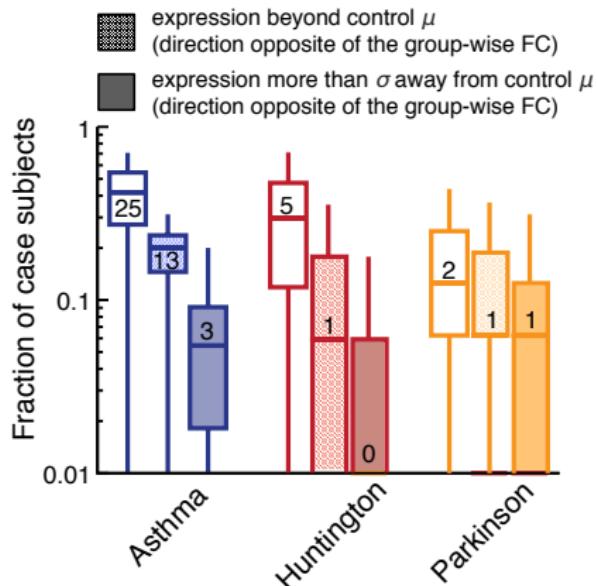
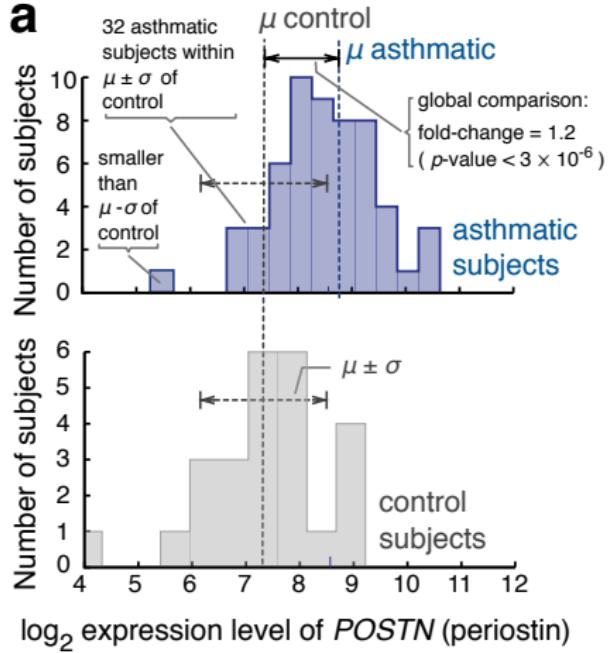
# Integrating personalized gene expression profiles into predictive disease-associated gene pools

Jörg Menche<sup>1,2,3</sup>, Emre Guney<sup>1,4</sup>, Amitabh Sharma<sup>1,4,5</sup>, Patrick J. Branigan<sup>6</sup>, Matthew J. Loza<sup>6</sup>, Frédéric Baribaud<sup>6</sup>, Radu Dobrin<sup>6</sup> and Albert-László Barabási<sup>1,2,4,5</sup>

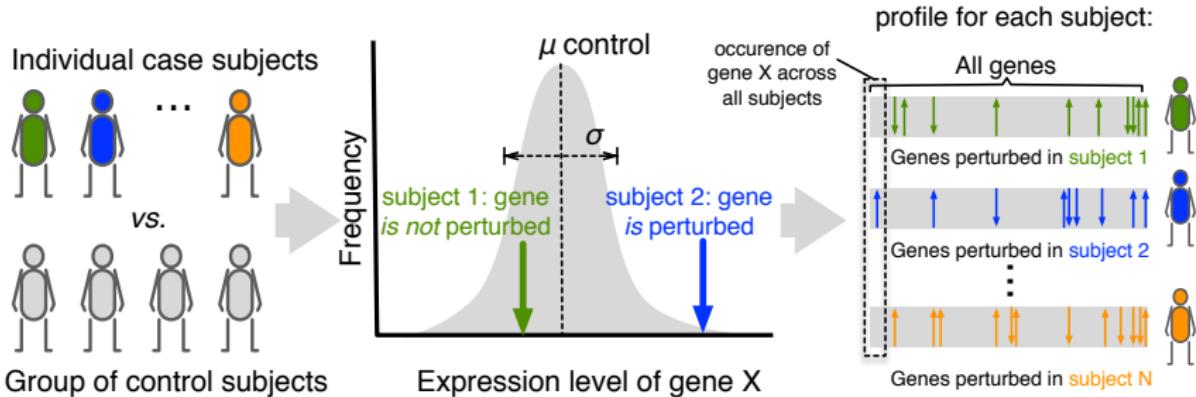
Gene expression data are routinely used to identify genes that *on average* exhibit different expression levels between a case and a control group. Yet, very few of such differentially expressed genes are detectably perturbed in individual patients. Here, we develop a framework to construct *personalized* perturbation profiles for individual subjects, identifying the set of genes that are significantly perturbed in each individual. This allows us to characterize the heterogeneity of the molecular manifestations of complex diseases by quantifying the expression-level similarities and differences among patients with the same phenotype. We show that despite the high heterogeneity of the individual perturbation profiles, patients with asthma, Parkinson and Huntington's disease share a broadpool of sporadically disease-associated genes, and that individuals with statistically significant overlap with this pool have a 80–100% chance of being diagnosed with the disease. The developed framework opens up the possibility to apply gene expression data in the context of precision medicine, with important implications for biomarker identification, drug development, diagnosis and treatment.

*npj Systems Biology and Applications* (2017)3:10; doi:10.1038/s41540-017-0009-0

# Group-wise differentially expressed genes do not capture transcriptomic heterogeneity

**a**

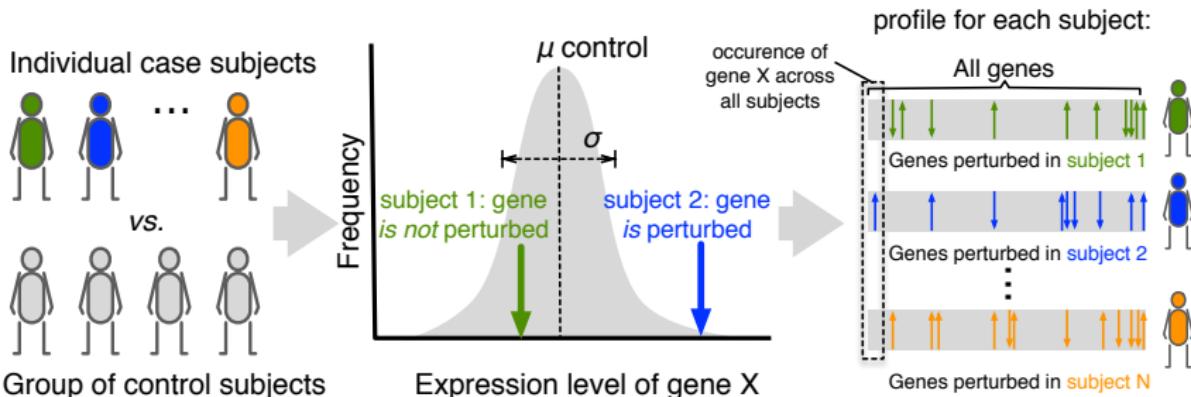
# PeeP: PErsonalized Expression Profile



|         | Samples |   |     |   |
|---------|---------|---|-----|---|
| Case    | x       | x | ... | x |
| Control | c       | c | c   | c |

$$z(\text{gene in } \textcolor{blue}{x}) = \frac{\text{expression}_{\textcolor{blue}{x}}(\text{gene}) - \mu_c(\text{gene})}{\sigma_c(\text{gene})}$$

# PeeP: PErsonalized Expression Profile



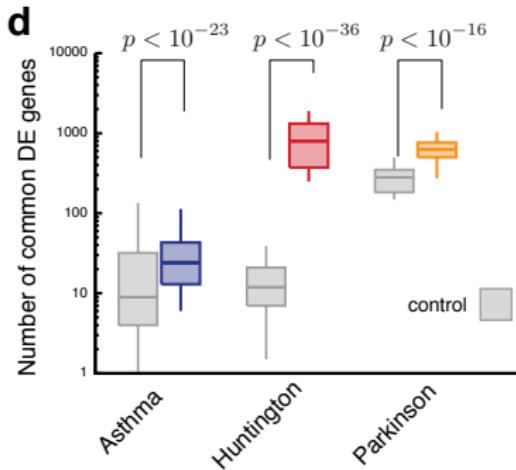
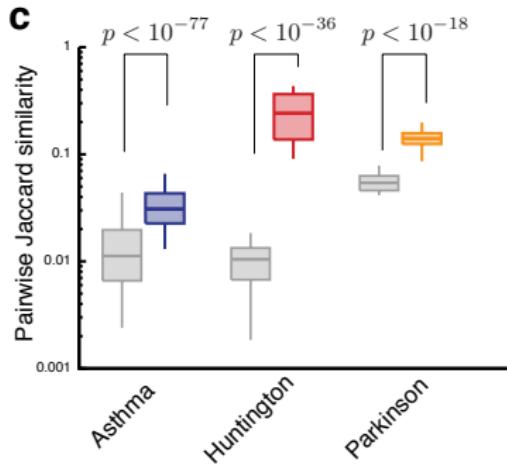
|         | Samples |   |     |   |
|---------|---------|---|-----|---|
| Case    | x       | x | ... | x |
| Control | c       | c | c   | c |

$$z(\text{gene in } \textcolor{blue}{x}) = \frac{\text{expression}_{\textcolor{blue}{x}}(\text{gene}) - \mu_c(\text{gene})}{\sigma_c(\text{gene})}$$

$$\text{PeeP}(\textcolor{blue}{x}) : \forall \text{gene } |z(\text{gene in } \textcolor{blue}{x})| > z_{threshold}$$

[ Genes that are significantly perturbed in each individual ]

## Quantifying the heterogeneity using PeePs



The overlap between PeePs of two individuals with the same disease

- is low (< 30%), suggesting high heterogeneity at the transcription level
- is higher than the overlap between the PeePs of healthy subjects

## Quantifying the heterogeneity using PeePs – Hands On!

### Patient-level gene signatures in Parkinson disease

Identify genes significantly perturbed across patients

| Resource                                 | URL   |
|--|---|
| <b><i>Gene expression data</i></b>       | NCBI GEO - GSE7621 <a href="http://ncbi.nlm.nih.gov/geo">ncbi.nlm.nih.gov/geo</a>               |
| <b><i>Interactome-based analysis</i></b> | PEPPER R package <a href="http://emreguney.net/doc/PEPPER.tgz">emreguney.net/doc/PEPPER.tgz</a> |

## Concluding remarks

- Biological processes involve coordinated activity of proteins that physically interact with each other
- Interactome, the network of interactions between proteins, provides a framework for understanding and characterizing biological processes
- Network medicine is an emerging field that aims to use interactome-based analyzes to identify genes associated to diseases and develop novel therapeutics
- Several tools available for the generation, visualization and analysis of the interactome as well as transcriptomic heterogeneity across patients