Prediction of Human Disease Genes by Analysis of Conserved Coexpression

Ala U.[1], *Piro R.M.*[1], Grassi E.[1], Damasco C.[1], Silengo L.[1], Oti M.[2], Provero P.[1] and Di Cunto F.[1]

[1]Molecular Biotechnology Center, Department of Genetics, Biology and Biochemistry, University of Turin, Italy

[2]Department of Human Genetics and Centre for Molecular and Biomolecular Informatics, University Medical Centre Nijmegen, The Netherlands

rosario.piro@unito.it

Many phenotypes mapped on the genome by linkage analysis are not yet associated to any validated disease gene (850 OMIM entries for phenotypes with unknown molecular basis had at least one associated disease locus on July 2nd, 2007). The identification of disease genes within disease-associated loci is a very demanding task even in the post-genomic era because orphan loci may typically contain hundreds of positional candidates.

Most published methods for disease gene prediction or prioritization rely on accurate gene annotation information (e.g. Gene Ontology) or mine PubMed/MEDLINE abstracts to infer relations between genes and phenotypes, thus being strongly biased towards well-characterized genes and tending to overlook genes about which little is known.

We present a **method [1] that exploits microarray gene expression data and a quantitative measure for similarity between human phenotypes to identify best candidates** among the positional candidates for a given disease as those that show significant coexpression with genes already known to be involved in similar phenotypes. Since the method uses a notion of similarity among phenotypes it **can also be applied to phenotypes of so far unknown molecular basis** (as long as they show similarity to other phenotypes of known molecular basis). Also, avoiding information on gene annotation and previous research is **potentially much less biased towards consolidated knowledge** although current microarray platforms still have their limitations and hence do not allow the evaluation of all positional candidates.

Since microarray data can be very noisy we focus on **coexpression that is evolutionary conserved** between human and mouse and therefore is more likely to be biologically meaningful. For this purpose we construct a human-mouse conserved coexpression network and verify its biological meaning and applicability to disease gene prediction by analyzing the prevalence of Gene Ontology terms, known interactions between human proteins and similar OMIM phenotypes within the networks coexpression clusters.

Our results demonstrate that conserved coexpression, even at the human-mouse phylogenetic distance, represents a very strong criterion to predict disease-relevant relationships among human genes. We propose **high-probability candidates for 81 OMIM loci characterized by unknown molecular basis**.

(NON-EXHAUSTIVE) COMPARISON OF METHODS FOR DISEASE GENE PREDICTION/PRIORITIZATION:

- ***Methods based on functional annotation*** (e.g. [2], [3]): represent the most straightforward approach for candidate prioritization; *overlook non-annotated candidate genes* [3][4]; it is not always evident how the annotated functions of the candidates relate to the disease phenotype.
- ***Methods based on protein-protein interactions*** (e.g. [5]): due to high-throughput data less biased towards already consolidated knowledge, but *not exhaustive* because very close functional relationships between genes and proteins are possible in the absence of direct molecular binding.
- ***Methods based on microarray (co-)expression data*** (the one presented here): potentially unbiased high-throughput data; no previous knowledge about candidate genes required; no direct protein-protein interactions required to infer functional relationship; but: microarray expression data is known to be noisy and co-expression does not imply a functional relationship between genes => we *need a filter for identifying biologically relevant co-expression*
- ***Methods based on multiple sources*** (e.g. [6] based on microarray expression data plus functional annotation): inherit strength and weaknesses of both methods.

CONSERVED CO-EXPRESSION NETWORK:

Phylogenetic conservation as a very strong criterion to **identify functionally relevant coexpression** links between genes [7][8]: significant coexpression that is phylogenetically conserved is likely due to selective advantage, suggesting a functional relation

Stanford MicroArray Database

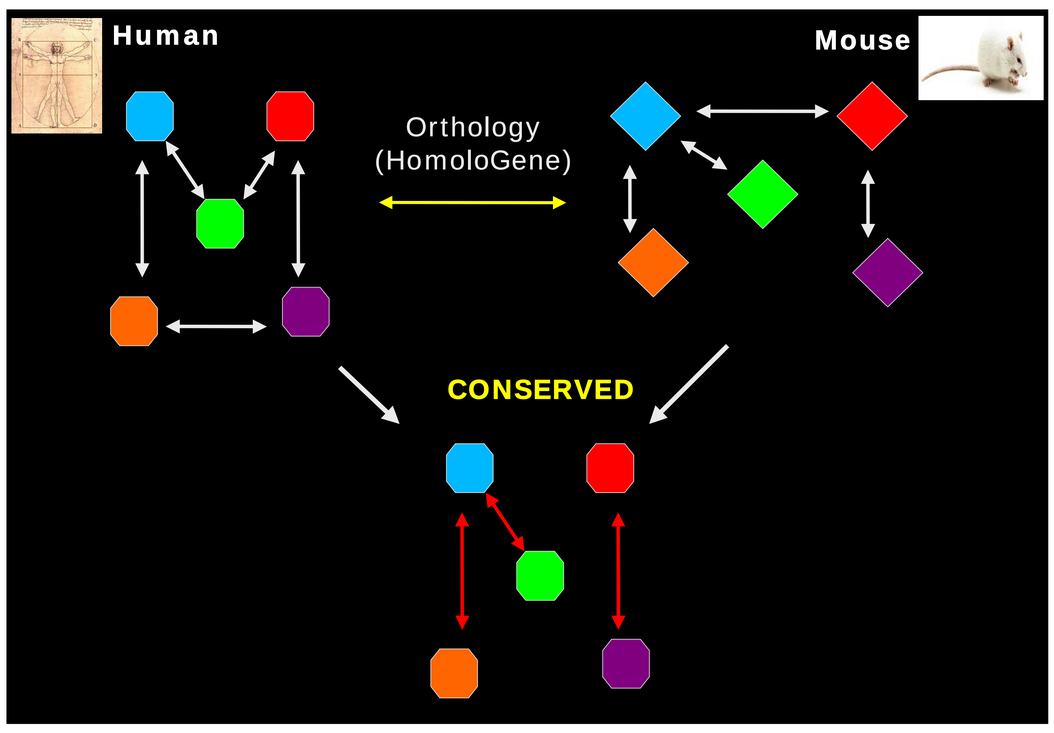
cDNA microarrays from *tumor cell lines*;
4192 experiments for human and 467 for mouse

OR

Entrez GEO DataSets

Affymetrix microarray data from 65 *normal* human tissues
(Roth et al. [9] and 61 *normal* mouse tissues (Su et al. [10])

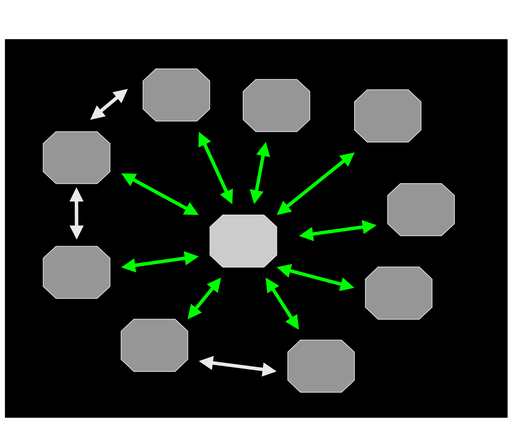
- Single-species co-expression networks (human and mouse; distinct networks for cDNA and Affymetrix):
  a link is established if gene A is among the 1% of most co-expressed genes of gene B and vice versa
- Conserved co-expression networks (distinct networks for cDNA and Affymetrix):
  only links present in both the human and the mouse cDNA/Affymetrix network are retained

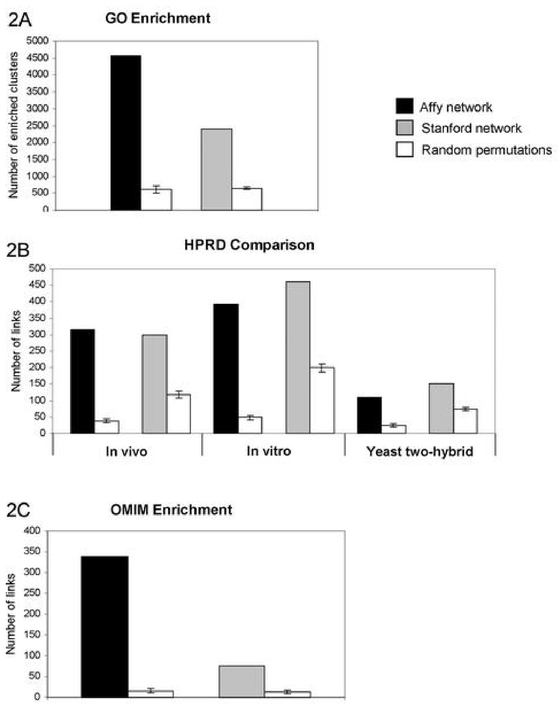


- Affymetrix network: 12,766 nodes (genes) with 155,403 links (conserved co-expression relationships)
- cDNA ("Stanford") network: 8,512 nodes with 56,397 links/edges

CONSERVED CO-EXPRESSION CLUSTERS (CCCs):

each CCC consists of a given gene (the center of the cluster) and all **next neighbors in the network** (co-expressed in both human and mouse)

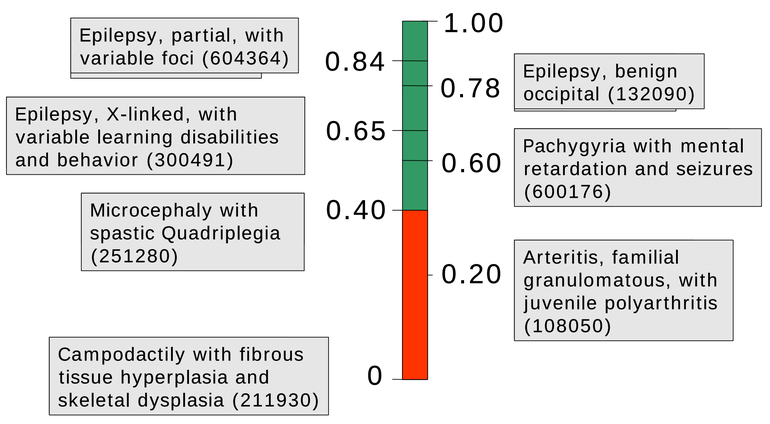




PHENOTYPE SIMILARITY: MIMMINER

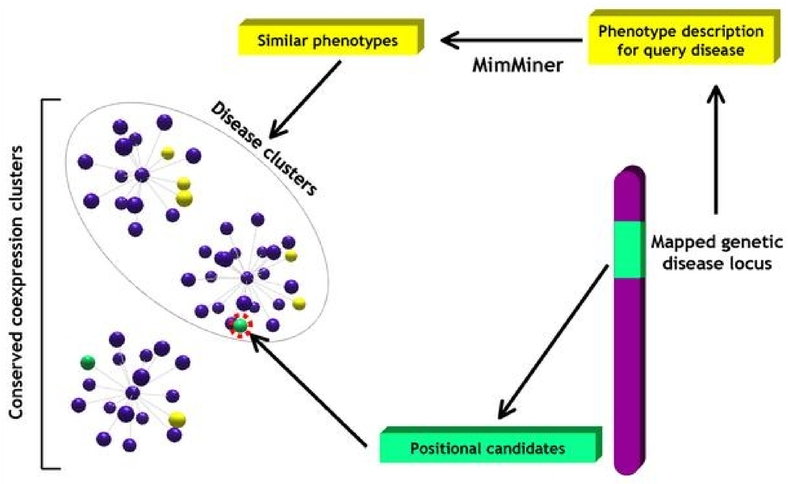
We use **MimMiner** as developed by van Driel et al. [11] to determine the similarity between two OMIM phenotype entries. MimMiner provides normalized scores for phenotype similarity (from 0 to 1); a threshold of 0.4 is used to denote similarity [11].

Example: **Epilepsy, partial, with pericentral spikes** (OMIM 607221)

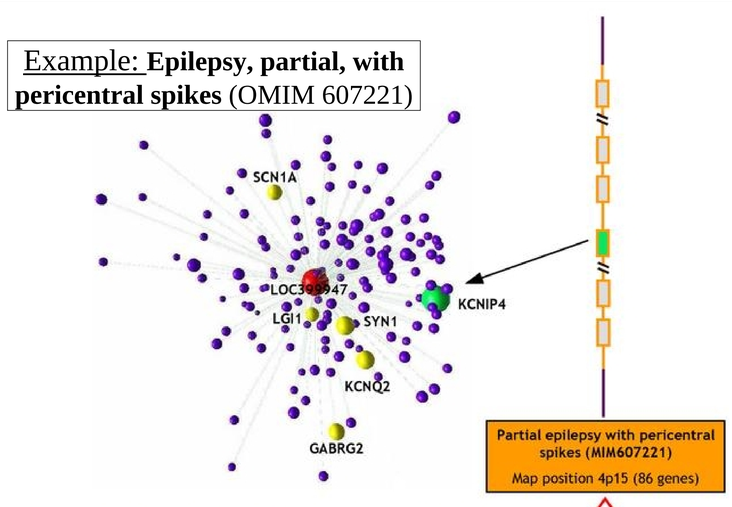


DISEASE GENE CANDIDATE SELECTION:

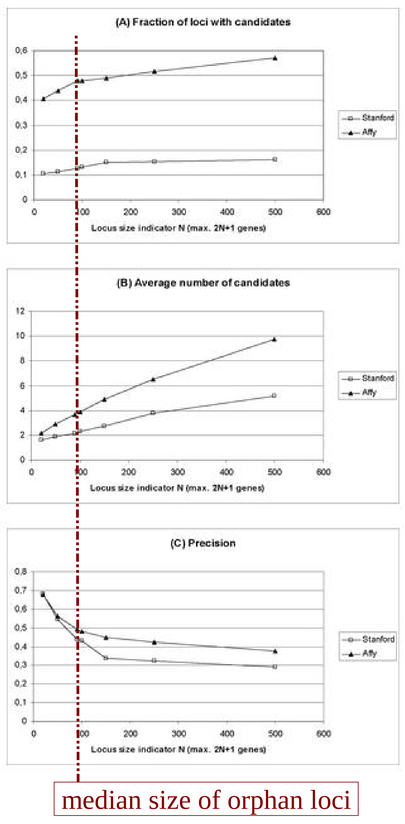
As **best candidates** among the genes in the disease-associated orphan loci we select those that appear in a CCC together with at least two genes known to be involved in similar phenotypes (i.e. they *show conserved co-expression with other genes that cause similar phenotypes*)



Example: **Epilepsy, partial, with pericentral spikes** (OMIM 607221)



Leave-one-out test results:



median size of orphan loci

Our predictions for 81 loci associated to phenotypes with unknown molecular basis can be found in [1].

REFERENCES:

1. Ala U, Piro RM, Grassi E, Damasco C, Silengo L, et al. (2008) Prediction of Human Disease Genes by Human-Mouse Conserved Coexpression Analysis. PLoS Comput Biol 4(3): e1000043.
2. Franke L, et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78: 1011–1025.
3. Turner FS, Clutterbuck DR, Semple CA (2003) POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol 4: R75.
4. Perez-Iratxeta C, Bork P, Andrade MA (2002) Association of genes to genetically inherited diseases using data mining. Nat Genet 31: 316–319.
5. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25: 309–316.
6. Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, et al. (2006) TOM: a web-based integrated approach for identification of candidate disease genes. Nucleic Acids Res 34: W285–292.
7. Pellegrino M, et al. (2004) CLOE: identification of putative functional relationships among genes by comparison of expression profiles between two species. BMC Bioinformatics 5: 179.
8. Stuart JM, Segal E, Koller D, Kim SK (2003) A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science 302: 249–255.
9. Roth RB, Hevezi P, Lee J, Willhite D, Lechner SM, et al. (2006) Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. Neurogenetics 7: 67–80.
10. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062–6067.
11. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA (2006) A text-mining analysis of the human phenome. Eur J Hum Genet 14: 535–542.