Tissue-specific conserved coexpression for disease gene prediction and hypotheses generation

***Piro R.M.*, Ala U., Molineris I., Grassi E., Damasco C., Bracco C., Provero P. and Di Cunto F.**
Molecular Biotechnology Center, Department of Genetics, Biology and Biochemistry, University of Turin, Italy
rosario.piro@unito.it

Even considering recent technological and methodological advances, such as next generation sequencing and genome-wide association studies, the **identification of genes involved in human hereditary disease** remains a demanding task that can be significantly aided by computational predictions. We discuss a method based on high-throughput microarray expression data that uses the **conservation of coexpression** as a powerful filter for biological significance and allows to specifically focus on tissue-specific relationships between disease and candidate genes.

The most important novelty of our approach, the **tissue-specificity** that we show to be highly complementary to multi-tissue coexpression, has allowed to identify **novel high-confidence candidates** for several genetic diseases. Moreover, disease gene prediction via tissue-specific conserved coexpression can additionally generate meaningful **hypotheses about functions of candidate genes** and biological processes underlying disease. Notably, these hypotheses can be important also for cases other than disease conditions, and have provided us with several promising novel candidates for pluripotency.

Through an explicit **integration of phenomics**, in particular the concept of phenotype similarity, we can apply our method also to disease phenotypes with so far unknown molecular basis. We present a **user-friendly web tool** for custom analysis, discuss the results obtained, and describe a case study that confirms USP9X as a particularly interesting candidate for **X-linked mental retardation**. The latter is also a proof of concept that our method can be efficiently integrated with deep sequencing to provide high-confidence candidate genes.

http://www.cbu.mbcunito.it/ts-coexp/

CONSERVED CO-EXPRESSION:

Phylogenetic conservation as a very strong criterion to **identify functionally relevant coexpression** links between genes [1][2]: significant coexpression that is phylogenetically conserved is likely due to selective advantage, suggesting a functional relation.
Conserved co-expression can be used for **disease gene prediction** [3][4].

TISSUE-SPECIFIC CO-EXPRESSION:

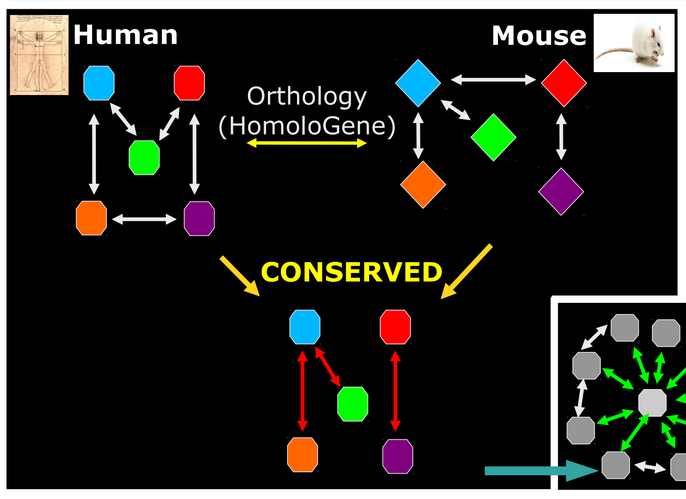
While many functionally relevant expression correlations can be extracted from heterogeneous (multi-tissue) microarray datasets, many others are likely to remain unnoticed, because they **specifically occur in one or few tissues, cell types and/or conditions** [4].

Entrez GEO DataSets

Human: 5,188 experiments on Affymetrix HG-U133 Plus 2.0
Mouse: 2,310 experiments on Affymetrix MG-430 2.0

To allow the selection of tissue- and/or condition-specific subsets, each experiment was classified (manual curation) according to: **anatomical annotation** (hierarchical MeSH ontology), **condition** (normal, tumor, other disease/condition), **stage** (adult, embryo), **sample type** (tissue, cell line)

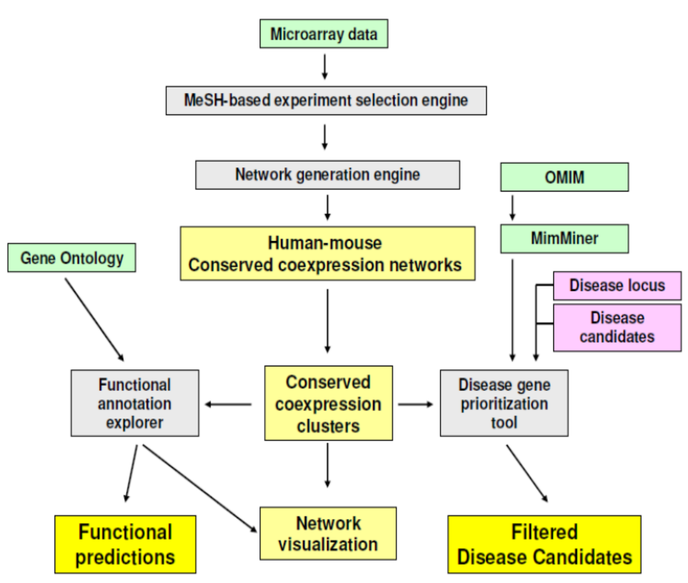
Multi-tissue and tissue-specific CONSERVED COEXPRESSION NETWORKS (CCNs):



- Single Species Networks (SSNs): significant coexpression if two genes are among the 1% most coexpressed genes of each other.
- CCNs: genes coexpressed in both species; intersection of human SSN and mouse SSN.
- Tissue-/condition-specific SSNs/CCNs: network construction from a selected subset of the expression data.
- Conserved coexpression clusters:
  - gene plus all its next neighbors
  - one cluster for each gene in the network

TS-COEXP WEBTOOL:

- Selection of 20 prebuilt CCNs both multi-tissue and tissue-specific
- CCN generation engine for custom selection of tissues, conditions, ...
- GO enrichment analysis for putative functional annotation in single or multiple networks
- Graphical network browser for visualization of gene clusters in single or multiple networks
- Disease gene prediction based on conserved coexpression with known disease genes



COMPLEMENTARITY OF INFORMATION CONTENT:

Both multi-tissue and tissue-specific CCNs are **enriched** for known functional relationships (Gene Ontology keywords) and protein-protein interactions (HPRD).
Most important: the information content is **complementary**!

Report for gene: SALL4 (GeneID 57167)

GO ID	P-value	E-value	Description	Network(s)
no functional annotation				

SALL4 in the CCN derived from "all samples"

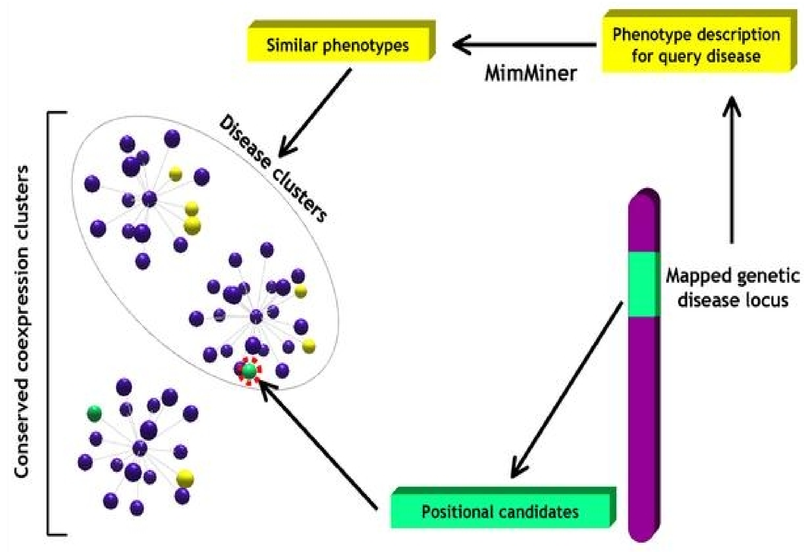
SALL4 in the "stem cells" CCN and the "embryonic tissues" CCN

Report for gene: SALL4 (GeneID 57167) — stem cells — embryo

GO ID	P-value	E-value	Description	Network(s)
GO:0030154	2.55242e-07	0.00263920228	cell differentiation	
GO:0048869	5.09597e-07	0.00526923298	cellular developmental process	
GO:0001829	5.24509e-06	0.0542342306	trophectodermal cell differentiation	
GO:0001825	9.04939e-06	0.0935706926	blastocyst formation	
GO:0009790	1.80276e-05	0.186405384	embryonic development	
GO:0040029	2.20393e-05	0.227886	regulation of gene expression, epigenetic	
GO:0030528	3.76275e-05	0.389068	transcription regulator activity	
GO:0045595	4.56878e-05	0.472411852	regulation of cell differentiation	
GO:0009653	5.6349e-05	0.58264866	anatomical structure morphogenesis	
GO:0001824	7.07775e-05	0.73183935	blastocyst development	

DISEASE GENE CANDIDATE SELECTION:

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



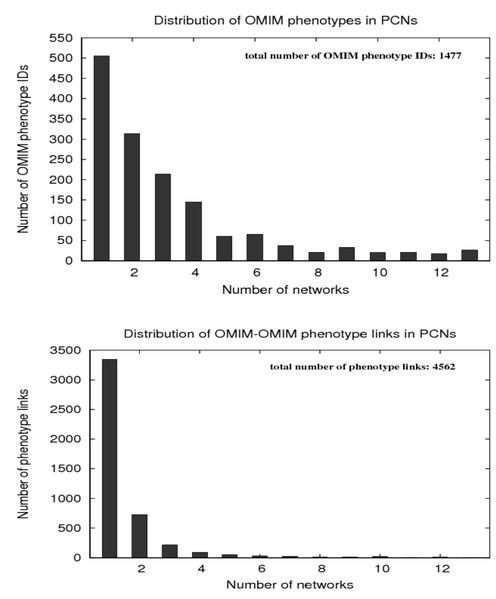
- Importance: 1,074 mapped "orphan loci" for 1,028 OMIM phenotypes with unknown molecular basis (June 2009)
- Contemporary use of gene gene clusters from **both multi-tissue and tissue-specific CCNs**
- **Application: our predictions for 238 loci associated to OMIM phenotypes with unknown molecular basis can be found in [4].**

PHENOTYPE SIMILARITY: MIMMINER

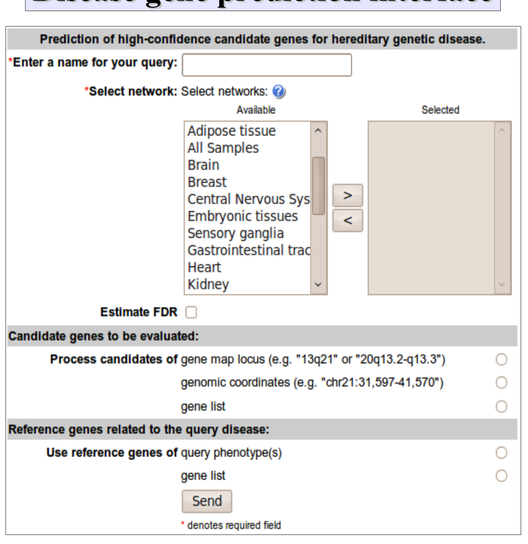
We use **MimMiner** as developed by van Driel et al. [5] to determine the similarity between two OMIM phenotype entries. MimMiner provides normalized scores for phenotype similarity (from 0 to 1); a threshold of 0.4 is used to denote similarity [5].
Important: the **integration of phenomics** allows to determine a set of disease-related genes also for OMIM phenotypes with so far **unknown molecular basis** (OMIM category: %).

PHENOTYPE COEXPRESSION NETWORKS (PCNs):

Translating gene-gene links from CCNs into **OMIM-OMIM links** between similar phenotypes shows that CCNs contains both **significant and complementary information** regarding relationships between disorders and their associated genes.



Disease gene prediction interface



CASE STUDY: X-LINKED MENTAL RETARDATION

Goal: **prioritization of mutational candidates** obtained from large-scale exome sequencing projects.
We compared our predictions with results obtained in a recent resequencing study of the X coding exons of 208 families affected by XLMR [6]. Tarpey et al. found 3 novel XLMR genes (SYP, ZNF711, CASK) and confirmed several known XLMR genes, but could find a reasonable genetic explanation only for 25% of the cases. Of 30 genes with truncating mutations only 8 could be associated to XLMR.

TEST: "RE-DISCOVERY" OF XLMR GENES

We took the genes resequenced by Tarpey et al. [6] as "candidate genes" and used only genes involved in similar phenotypes (MimMiner ≥ 0.5) as "reference genes", **pretending XLMR to be a phenotype with unknown molecular basis**. At a 10% FDR we obtained 102 high-confidence candidates (most from tissue-specific CCNs, mainly brain and CNS), 32% are known XLMR genes (p=2.5x10^{-6}). We predicted only 7 of the 30 genes with truncating mutations as high-confidence candidates, 6 of which are XLMR genes (including the novel XLMR genes SYP and ZNF711; p=3.1x10^{-4}). Notably, our 7th candidate is **USP9X** who's implication could neither be confirmed nor rejected by Tarpey et al. [6].

PREDICTION: NOVEL XLMR CANDIDATES

Taking **non-XLMR genes** re-sequenced by Tarpey et al. [6] as "candidate genes" and known XLMR genes as "reference genes": 49 novel high-confidence candidates (including **USP9X**), one of which has been confirmed recently (**RAB39B** [7]).

RAB39B as candidate

Legend.
★ = genes associated to related disorders
Brain — Central Nervous System

cluster size:
46 in brain; 65 in CNS

Report for gene: ITFG1 (GeneID 81533)

Gene ID	Gene Symbol	Description		Network(s)
51566	ARMCX3	armadillo repeat containing, X-linked 3	*candidate*	
3423	IDS	iduronate 2-sulfatase	XLMR	
5160	PDHA1	pyruvate dehydrogenase (lipoamide) alpha 1	XLMR	
5631	PRPS1	phosphoribosyl pyrophosphate synthetase 1	XLMR	
116442	RAB39B	RAB39B, member RAS oncogene family	*candidate*	
10479	SLC9A6	solute carrier family 9 (sodium/hydrogen exchanger), member 6	XLMR	
10971	YWHAQ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, theta polypeptide		

REFERENCES:
1. 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.
2. Stuart JM, Segal E, Koller D, Kim SK (2003) A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* 302: 249–255.
3. 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.
4. Piro RM, Ala U, Molineris I, Grassi E, Bracco C, *et al.* (submitted) An Atlas of Tissue-Specific Conserved Coexpression for Functional Annotation and Disease Gene Prediction.
5. 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.
6. Tarpey PS, *et al.* (2009) A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet* 41: 535-543.
7. Giannandrea M, et al. (2010) Mutations in the small GTPase gene RAB39B are responsible for X-linked mental retardation associated with autism, epilepsy, and macrocephaly. *Am J Hum Genet* 86:195-195.