HAL
open science

An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction

Ferdinando Di Cunto, Rosario Michael Piro, Ugo Ala, Ivan Molineris, Elena Grassi, Chiara Bracco, Gianpaolo Perego, Paolo Provero

To cite this version:

Ferdinando Di Cunto, Rosario Michael Piro, Ugo Ala, Ivan Molineris, Elena Grassi, et al.. An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction. European Journal of Human Genetics, 2011, 10.1038/ejhg.2011.96 . hal-00649447

HAL Id: hal-00649447

https://hal.science/hal-00649447v1

Submitted on 8 Dec 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction

Rosario Michael Piro[1*§], Ugo Ala[1*], Ivan Molineris[1*], Elena Grassi[1], Chiara Bracco[2], Gian Paolo Perego[2], Paolo Provero[1] and Ferdinando Di Cunto[1#]

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

2 Aethia Srl, Via Ribes 5, Colleretto Giacosa, Torino, Italy

* These authors contributed equally to the work

§ Current address: Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany

Corresponding author

Running title: Tissue-specific conserved coexpression and disease genes

Keywords: Disease-gene prediction, Functional annotation, Transcriptome, Phenome

Abstract

Gene coexpression relationships that are phylogenetically conserved between human and mouse have been shown to provide important clues about gene function that can be efficiently used to identify promising candidate genes for human hereditary disorders. In the past, such approaches have considered mostly generic gene expression profiles that cover multiple tissues and organs. The individual genes of multicellular organisms, however, can participate in different transcriptional programs, operating at scales as different as single cell types, tissues, organs, body regions or the entire organism. Therefore, the systematic analysis of tissue-specific coexpression could be, in principle, a very powerful strategy to dissect those functional relationships among genes that emerge only in particular tissues or organs. In this report, we show that, in fact, conserved coexpression as determined from tissue-specific and condition-specific datasets can predict many functional relationships that are not detected by analyzing heterogeneous microarray datasets. More importantly, we find that, when combined with disease networks, the simultaneous use of both generic (multi-tissue) and tissue-specific conserved coexpression allows a more efficient prediction of human disease genes than the use of generic conserved coexpression alone. Using this strategy, we were able to identify high-probability candidates for 238 orphan disease loci. We provide the proof of concept that this combined use of generic and tissue-specific conserved coexpression can be very useful to prioritize the mutational candidates obtained from deep-sequencing projects, even in the case of genetic disorders as heterogeneous as X-linked mental retardation.

Introduction

Despite the recent progress in mapping and sequencing technologies, the identification of genes involved in human diseases remains a very demanding task. Indeed, genome-wide techniques such as linkage analysis, SNP profiling or even deep sequencing of genetically heterogeneous disorders may select hundreds of candidates, whose experimental verification is time and resource-consuming [1]. A deep knowledge of the modular organization of biological functions may significantly increase the efficiency of this identification process. Indeed, biological phenomena are emergent properties of complex interaction networks, composed of proteins, DNA, small molecules and different classes of RNA, capable of self-organizing in discrete functional modules [2,3]. The dissection of the molecular basis of many diseases has evidenced that in most cases abnormal phenotypes are caused by the derangement of an entire module, due to single gene defects, to a combination of genetic abnormalities or to the interaction between gene variants and environmental factors [3,4]. Thus, when trying to identify the best candidates for a given phenotype or to establish the phenotypic significance of a particular gene variant, it would be very helpful to know whether the genes under study are involved in functional processes directly relevant to that phenotype. In theory, the use of functional gene annotations would represent the most straightforward support for this task. However, although this strategy has been used successfully in many cases [5,6], it is clearly limited by the lack of complete information about the function of most human genes.

Coexpression relationships derived from microarray data represent an extremely rich and less biased source of information, potentially relevant for functional annotation and disease gene prediction. Indeed, it has been extensively shown that functionally interacting genes tend to display very similar expression profiles, as a result of common regulatory mechanisms [7,8]. Moreover, the probability for two genes to be functionally correlated is remarkably higher when they are strongly coexpressed in more than one species

(conserved coexpression) [9,10]. Accordingly, the systematic integration of phenotype information with conserved coexpression may allow the efficient prediction of the best positional candidates in wide genomic regions associated to genetic diseases [11,12]. Despite these important results, most of the coexpression studies so far reported have been performed on heterogeneous datasets covering multiple tissues and organs, thus partly overlooking the complex control of gene expression within specific tissues or organs that is essential for higher eukaryotes. This is a very important shortcoming, because the transcriptional units that compose the human genome display an extremely flexible organization, allowing complex qualitative and quantitative control of gene expression in the different cell types [13].

In this report we show that the study of tissue-specific conserved coexpression allows an extended exploration of the transcriptional co-regulation of mammalian genes, with strong implications for their functional annotation. As an example, we identify a cluster of genes potentially implicated in the transcriptional programs of pluripotent stem cells. Moreover, we show that the simultaneous use of multi-tissue and tissue-specific conserved coexpression networks, combined with phenome analysis, allows efficient candidate gene prediction. In particular, we analyze the potential of combining our predictive strategy with high-throughput mutational screenings using X-linked mental retardation as a case study. Finally, we provide a user-friendly web resource allowing both the access to pre-computed predictions and the execution of custom analysis for functional annotation and disease gene identification.

Results

Generation of conserved coexpression networks and coexpression clusters.

To evaluate the potential of tissue-specific conserved coexpression, we studied a large microarray dataset downloaded from the Gene Expression Omnibus (GEO) [14], covering

many tissues, cell types and experimental conditions in both human (5188 experiments) and mouse (2310 experiments). Based on the description given in GEO, the samples were manually annotated to allow the selection of condition-specific subsets of the desired anatomical depth (see Methods), which we used to generate the corresponding single-species coexpression networks (SCNs) or human-mouse conserved coexpression networks (CCNs), by the procedures previously described [12]. Briefly, CCNs are obtained by intersecting a human SCN with a mouse SCN that reflect the same selection of tissues and/or conditions, keeping only the coexpression links that are observed in both species. We generated multi-tissue CCNs, covering for example all samples from normal (non-tumoral) tissues, and tissue-specific CCNs from normal samples from heart, kidney, breast, etc. In multi-tissue coexpression networks, it has been shown that phylogenetic conservation is a strong filter for functionally relevant links [10,12], because it is highly unlikely that the artifactual correlations generated in one species by possible sample outliers or by other factors are reproduced in a corresponding dataset of a second species. We verified that this principle holds true also in tissue-specific CCNs (see supplementary data for the details).

From each of the CCNs we then extracted a set of "coexpression clusters", consisting of a gene together with its nearest neighbours, i.e. together with those genes that show a conserved coexpression with it [12].

Complementarity and biological relevance of multi-tissue and tissue-specific conserved coexpression.

As expected, a direct pairwise comparison of all the different networks revealed in every case a highly significant overlap, consistent with the existence of transcriptional modules composed of genes highly coexpressed in most mammalian cell types. For instance, the least significant overlap was found between the embryonic tissues and the adipose tissue CCNs, where only 0.6% of the edges present in either network were found in both: such a

small overlap is however much larger than expected by chance (about 100 standard deviations above the mean overlap obtained from 100 nework randomizations). However, the overlap between the different tissue-specific CCNs or between the tissue-specific CCNs and multi-tissue CCNs was always much lower than the intersection between different multi-tissue CCNs (Figure 1A). Altogether, these data indicate that tissue-specific CCNs capture a large number of conserved coexpression relationships that cannot be detected using the promiscuous approach and vice versa.

To address the functional relevance of these correlations, we evaluated their capability to provide functionally characterized coexpression clusters, in comparison with those obtained from multi-tissue CCNs. Therefore, for every CCN, we analyzed the overrepresentation of GO keyword in all coexpression clusters extracted from the CCN. Both tissue-specific CCNs and multi-tissue CCNs showed a strongly increased number of functionally enriched clusters (Figure S2), when compared with 100 randomized versions of the same networks. Moreover, the distribution of the enriched GO terms showed that the different networks are strongly complementary, because a high percentage of them were specifically identified in one network and the majority was found in only a few CCNs (Figure 1B). Most importantly, we found that evolutionary conservation as a filter tends to preserve those coexpression edges that are of higher functional significance (see supplementary data for the details). Finally, tissue-specific CCNs displayed also a high prevalence of edges between proteins that are known to physically interact from the HPRD [15] (Figure S2).

As a case study that illustrates the added value of tissue-specific CCNs, we report the example of mammalian pluripotency genes. The elucidation of the molecular circuitry underlying the establishment and the maintenance of pluripotency is a crucial issue in biology [16], especially after the discovery that the expression of a small set of genes can reprogram differentiated cells to a pluripotent state [17]. Proteomic screenings have been so

far quite successful in mapping the interactome of some crucial pluripotency factors [18,19], but cannot identify functional correlations that do not require stable protein-protein interactions. We thus explored the possibility that this approach may identify pluripotency factors. In particular, we asked which genes are connected to the crucial pluripotency gene POU5F1 in the above-mentioned multi-tissue and tissue-specific CCNs. The network obtained from normal adult tissues contained no edges for this gene and very few edges were found in most tissue-specific networks (Table S3). In contrast, POU5F1 had a very high connectivity in the network obtained from embryonic tissues and was even more connected in a CCN obtained specifically from human and mouse stem cell datasets (Table S3). Strikingly, the neighborhood of POU5F1 in the latter networks contained most of the other genes that have been successfully used to reprogram differentiated cells to induced pluripotent stem (iPS) cells, such as NANOG, LIN28, SOX2 and NMYC and many other genes that have been functionally associated with the establishment and maintenance of the pluripotent state (Figure 2 and Table S3). This result indicates that in the stem cells CCN, most of the key pluripotency factors form a very tight cluster. Considering that the identification of this cluster by our method was completely independent from previous knowledge, it is very likely that it may comprise also a significant portion of the unknown core pluripotency machinery, including both other master control genes and at least some of the most direct targets of the core transcription factors.

In summary, these results indicate that the tissue-specific approach significantly extends the potential of human-mouse conserved coexpression to provide new functional hypotheses about mammalian genes. The observed complementarity to the multi-tissue approach suggests that it is best to use both approaches jointly, because both coexpression accross multiple tissues and coexpression within specific tissues can yield important clues to gene function.

Tissue-specific conserved coexpression and phenotype information

In our previous work we have demonstrated that the analysis of conserved coexpression in multi-tissue microarray datasets can be a powerful tool to evaluate candidate genes for human genetic diseases [12]. Therefore, we asked whether tissue-specific CCNs as well can provide important clues for disease gene prediction and whether the disease-related information they contain overlaps with or rather complements the information obtained from multi-tissue CCNs.

For this purpose we derived OMIM phenotype coexpression networks (PCNs) from 13 different tissue-specific CCNs and from one multi-tissue CCN, obtained from normal tissue data. In these PCNs, the nodes represent OMIM phenotypes and two nodes are joined by an edge if the phenotypes are similar [20] and there is at least one edge in the CCN between genes associated to the two phenotypes. In other words PCN edges indicate that, in the underlying CCN, genes involved in one phenotype are directly linked to genes involved in a similar phenotype and therefore that the CCN edges contain useful information to evaluate candidate genes for related disease phenotypes with so far unknown molecular basis.

Interestingly, 505 (34.2%) of 1477 OMIM phenotype IDs and 3348 (73.4%) of 4562 OMIM-OMIM edges were present in only one PCN. Although the multi-tissue PCN provides more unique OMIM IDs and edges than any of the tissue-specific PCNs, 68.1% of all unique phenotypes and 61.8% of all unique edges can be exclusively found in one of the tissue-specific networks (Figure 3A). These data suggest that tissue-specific CCNs can provide valuable information for disease gene prediction that is not contained in multi-tissue CCNs. A good example of this complementarity is given by the small subnetwork of OMIM phenotypes directly linked to Charcot-Marie Tooth disease type 4D (CMT4D; OMIM 601455) and their respective neighbors in both the central nervous system (CNS) and the normal tissues (NT) PCNs (Figure 3B). While many of the phenotypes and of the

relationships between them can be found in both PCNs, a significant part can be found in only one of the two PCNs.

Identification of high-confidence disease gene candidates by multi-tissue and tissue-specific conserved coexpression

As we have shown in our previous work, multi-tissue conserved coexpression can be efficiently combined with phenotype correlation data to provide high confidence candidates within genetic disease loci (see reference 12 and Methods). The results described above indicated not only that high-confidence candidates could be obtained from tissue-specific expression data, but also that such results would likely be complementary to those obtained from multi-tissue CCNs.

Thus, using the 14 CCNs from Figure 3A, we applied our procedure to 1074 mapped "orphan loci" associated to 1028 OMIM phenotype entries with unknown molecular basis. At a 1% false discovery rate (FDR), estimated separately for each CCN, we identified an average of 7.1 candidates for 238 of these "orphan loci", thus obtaining a total of 1692 high-confidence candidates (see the detailed table at http://87.253.99.109/ts-coexp/disease_gene_precomp_table.php).

This is a significantly reduced number, considering that the orphan loci contain on average 149 candidate genes. A leave-one-out test performed on artificial loci of the same size, centered on the known disease genes (see Methods), displayed a precision of 44%. Therefore, we expect that in approximately half of the cases the positional candidates selected by our procedure will contain the actual disease-causing gene.

Evaluation of XLMR candidates obtained by X-chromosome exome sequencing

Beside the proof of concept already provided by the leave-one-out test, we decided to specifically evaluate the usefulness of our method in the prioritization of the mutational candidates obtained by exome sequencing projects in genetically heterogeneous disorders. In particular, we concentrated on X-linked mental retardation (XLMR) as a case

study, because Tarpey and co-workers have recently resequenced a large fraction of X-chromosome coding exons of patients or carriers from 208 different XLMR families [21]. Interestingly enough, although this survey identified three novel genes (SYP, ZNF711 and CASK), as well as several other genes already implicated in XLMR, it also underscored the limits of such kind of mutational screenings [21]. For example, truncating mutations were identified in 30 genes, but after extensive validation only a minority of them could be proven to be actually disease-causing. Moreover, despite the identification of a very large number of missense mutations on conserved residues and of several potential splice site mutations, the statistical power of the study allowed to define a reasonable genetic explanation of the phenotypes only for 25% of the cases [21].

Therefore, to evaluate conserved coexpression as an additional criterion for scoring mutated genes, we asked how many known XLMR genes could be predicted as strong candidates using our approach. Importantly, to perform this test we considered as candidates all the genes that have been sequenced in the mentioned study and removed from the list of the reference genes all known XLMR genes, hence pretending XLMR to be a disease phenotype without any known molecular basis. Since the number of reference genes (disease genes involved in similar phenotypes) was still large at a MimMiner phenotype similarity threshold score (see Methods) of 0.4 (401 genes), we decided to use a more stringent MimMiner threshold of 0.5 that yielded a total of 64 reference genes. Afterwards, we used the previously mentioned CCNs to predict the candidates with a 10% FDR cutoff. We obtained a total of 222 predictions, corresponding to 102 candidate genes (Table S4). Although this number may appear very large, it must be considered that 90 genes on the X chromosome are known to be associated to some form of intellectual disability and that a similar number probably remains to be identified [22,23]. Among our candidates, only 25 were retrieved from the normal tissue CCN, while the majority was

derived from tissue-specific networks, with 40% coming from the central nervous system or from brain networks.

Our list of candidates was strongly enriched for known XLMR genes, since 32% of them were included (P= 2.5 E-06, Table S4). Moreover, the global prevalence of validated XLMR genes was 11.8% among the 288 genes in which the authors found missense substitutions of conserved residues, but almost three times as much among the 57 mutated genes that were also found as candidates by our approach (Tables S4).

Even more strikingly, the comparison of the 30 genes in which the authors found truncating mutations with our candidates showed only 7 common genes (AP1S2, BRWD3, CUL4B, SLC9A6, UPF3B, USP9X and ZNF711), 6 of which have been validated by the authors in the same study (P=2.2 E-08). The only non-validated gene of this kind is USP9X, whose role has remained uncertain because only a single truncating mutation was found in patients. Thus, our data further underscore USP9X as a strong XLMR candidate.

Considering the efficiency of our approach in 're-discovering' XLMR genes even when none of them was considered as a reference gene, we speculate that it may be even more effective in identifying new candidates when the validated genes are correctly considered as reference. We therefore provide in Table 1 a list of the candidates with an FDR of 10% or lower, obtained using only the known XLMR genes as reference. Also in this case, USP9X was highlighted as a promising candidate. The fact that it was predicted with two completely distinct sets of reference genes further supports the significance of its XLMR candidacy.

One of the most appealing features of our system is that, besides selecting promising candidate genes for a particular disease, it may give good indications about the underlying molecular mechanisms, through the functional analysis of the candidates' neighboring genes in the coexpression networks. Interestingly, the analysis of the genes that are connected with USP9X revealed a strong prevalence of enzymes involved in the ubiquitin

cycle (Figure 4). This result is fully consistent with USP9X being a deubiquitinating enzyme [24] and offers testable predictions about the specific proteins that may functionally cooperate with it and/or about the possible targets of its activity.

Discussion

Although deep sequencing techniques are setting new benchmarks for the analysis of gene expression, microarray data deposited in public databases still represent an incredibly rich source of information. Therefore, it is quite surprising to observe that, with notable exceptions [25-29], coexpression relationships based on DNA microarrays are not used extensively to support the identification of new disease-causing genes. The main reason for this discrepancy may well reside in the low specificity of the functional predictions obtained by coexpression analysis. In other words, it is well known that if two genes are involved in similar functions and disease phenotypes they tend to have strongly correlated expression profiles, but it is known as well that two or more genes may display strongly correlated expression profiles even though they are not functionally related. The studies conducted by different groups, including ourselves [10-12], have shown that very simple filters, such as phylogenetic conservation, can strongly increase the predictive power of coexpression analysis. Here we have demonstrated that another simple filter, i.e. the selection of tissue-specific subsets of a large data repository, can significantly extend the potential usefulness of the coexpression- based predictions. It is important to underscore that, if used in isolation, the tissue-specific approach would be even noisier than a multi-tissue approach, independently of the size of the starting microarray dataset (Table S1). However, when combined with conservation, the tissue-specific approach can provide a high percentage of accurate predictions that could not be obtained using heterogeneous datasets. In agreement with our previous studies, we showed that this information can be used effectively to identify high-confidence disease gene candidates.

Moreover, and perhaps most importantly, we provide a proof of concept that the combination of conserved coexpression and phenomic networks could be successfully used to support the identification of disease genes by massive sequencing strategies, which in the case of genetically heterogeneous disorders can identify a much higher number of mutated candidates than previously suspected [21].

As a side-product of our analysis, we obtained a list of high confidence candidates for XLMR. The relevance of this list is supported by two recent observations. Indeed, while the manuscript was in preparation, we realized that the RAB39B had been just discovered as an actual XLMR gene [30]. Moreover, during the revision process, the GSPT2 gene has been independently proposed as a strong XLMR candidate by a copy number variation study [31].

We think that the case of USP9X illustrates particularly well the potential usefulness of our approach for selecting the most promising candidates and making hypothesis about their functional interactions with the validated genes. The mutational screening by Tarpey and co-workers [21] found a single truncating mutation in this gene, which so far has not been validated by clear additional mutational evidence (Joseph Gecz, personal communication). USP9X is a de-ubiquitinating enzyme, encoded by a strongly conserved gene that escapes X inactivation [24,32]. Moreover, the Drosophila ortholog *faf* is known for its role in synaptic development and the expression pattern of the mouse ortholog is suggestive of a synaptic function [33,34].

Thus, the finding that the gene is consistently coexpressed with many other genes involved in intellectual disability (namely UBE3A, UBR1, OPA1 and CRBN) and with genes involved in the ubiquitin cycle is very suggestive. In particular, since UBE3A and UBR1, two of the most strongly connected genes, display both features, we propose the working hypothesis that the three proteins they encode may directly cooperate in regulating the

ubiquitination/deubiquitination cycle of common substrates involved in synaptic function, some of which could be possibly included in our list of potential partners.

In conclusion, we propose multi-tissue and tissue-specific conserved coexpression networks as highly valuable resources for predicting and exploring the functional properties of genes in relation to human diseases. To support the scientific community in this task, this manuscript is accompanied by the TS-CoExp browser (http://www.cbu.mbcunito.it/ts-coexp), an open user-friendly web interface described in more detail in the supplementary text.

Methods

Gene expression database annotation and normalization

The gene expression series that we used comprise 5188 human and 2310 mouse microarray experiments, performed on the Affymetrix platforms Human Genome U133 Plus 2.0 and Mouse Genome 430 2.0, respectively, obtained from the Gene Expression Omnibus (GEO). To allow the efficient selection of tissue- and condition-specific subsets of the whole database, the experiments were manually annotated according to the following scheme. First of all, the anatomical annotation of the samples was reported to the standard MeSH ontology, by associating every experiment to the most specific keywords that correctly described the sample. Secondly, we recorded for every sample whether the RNA was derived from whole tissues or from isolated cells. Third, were reported whether the sample corresponded to a normal, tumor-related or other diseased condition. Finally, we reported whether the sample was of adult or embryonic origin.

To avoid the spurious correlation links that may be introduced by the RMA procedure [35], we downloaded *CEL* files for each experiment and normalized them by using MAS5 algorithm. This procedure was used as implemented in software packages available from Bioconductor (http://www.bioconductor.org), using the default parameters.

Generation of conserved-coexpression networks

Generic and tissue-specific conserved coexpression networks were generated using the procedures described in [12]. Briefly, after selecting corresponding subsets of the human and mouse gene expression datasets, we first generated single species gene coexpression networks (SCN) and then integrated them on the basis of human-mouse orthology. SCNs were generated by calculating the Pearson correlation coefficients between all the probeset of the expression matrix. A directed edge was established from probeset p_1 to probeset p_2 if p_2 fell within the top 1% of all the correlation values calculated respect to p_1. The directed networks where then converted into undirected SCNs by mapping the rows to the corresponding Entrez Gene identifiers. Finally, CCNs were built from SCNs by mapping every Entrez Gene identifier to the corresponding homology cluster, defined as the union of the information contained in Homologene (built 63) and Ensembl homology (release 53) databases.

Measure of phenotype similarity

To measure the pairwise similarity of disorders (described by independent OMIM phenotype entries) we downloaded the OMIM database on June 17th, 2009 and processed it using MimMiner, essentially as described [20].

MimMiner scores are normalized and range from 0 (unrelated) to 1 (highly related or identical). Since van Driel et al. showed that similar phenotypes can be identified with reasonable accuracy considering a minimum score of 0.4, we used the same threshold for our work. Hence, when mentioning “similar phenotypes”, if not otherwise stated, we refer to pairs of OMIM phenotype entries that have a MimMiner similarity score of at least 0.4 in our updated database.

Generation of OMIM phenotype coexpression networks (PCNs):

Genes (nodes) in a given conserved coexpression network (CCN) were mapped to the disorders (OMIM phenotype IDs) in which they are known to be involved using Entrez

(mim2gene, downloaded on June 16th, 2009). Each edge in the CCN involving two genes both associated to OMIM phenotypes was translated into a corresponding OMIM-OMIM edge. For genes involved in multiple phenotypes several corresponding OMIM-OMIM edges could be obtained. For the final PCN, however, only OMIM-OMIM edges involving two similar phenotypes (MimMiner score ≥0.4; see above) were kept, while OMIM-OMIM edges that connect unrelated phenotypes were dropped. Therefore, PCNs can be considered to capture the valuable knowledge contained in CCNs for the purpose of disease gene prediction by means of identifying candidate genes that show a conserved coexpression with genes known to cause similar disorders.

Disease gene prediction

We applied our previously described method for disease gene prediction [12] to 1 multi-tissue and 13 tissue-specific CCNs. Here, we briefly summarize the method, for more details see reference 11.

We define gene clusters in the network, consisting of a given gene plus all its nearest neighbours (i.e. all genes that show conserved coexpression with the given gene). Hence each network composed of *N* genes will yield *N* gene clusters (one for each gene).

For these clusters we establish associations to OMIM phenotype entries in the following way: each cluster containing at least 2 genes known to be involved in a given phenotype *p* or similar phenotypes (see above) is consider as “potentially disease-relevant” with respect to *p*.

To identify promising candidate disease genes for phenotypes with currently unknown molecular basis (OMIM category “%”, as of June 2009) but already mapped gene map locus (e.g. via linkage analysis), we determine which (if any) of the candidate genes fall within the respective disease-relevant gene clusters taken from our networks. In other terms, we select the “best” candidates as those that show a conserved coexpression,

within a given context depending on the CCN (e.g. tissue-specific), with other genes involved in the given or similar phenotypes.

False discovery rate (FDR) for the disease gene prediction: In order to estimate the quality of our predictions we performed 20 randomizations of each CCN (keeping the network structure intact and instead randomizing only gene labels) and applied the same candidate selection method as described above. Please note that the FDR is computed for each CCN separately, hence there is some limited amount of multiple testing not yet considered by the computed FDRs. Therefore, with the exception of the evaluation of XLMR candidates, we present only predictions that satisfy a stringent FDR of 1%, instead of the 10% threshold used in our previous study [12].

Leave-one-out test procedure: To verify the effectiveness of the method in identifying disease genes, we performed a leave-one-out cross-validation on all genes known to be involved in phenotypes described in OMIM (Entrez mim2gene, downloaded on June 16th, 2009), with the exception of those genes that could not be mapped to their precise genome location via Ensembl 55 (http://www.ensembl.org). For each evaluated gene-phenotype pair we constructed an artificial locus, centred around the disease gene, containing 149 genes (Ensembl 55), corresponding to the average number of genes present in the orphan loci associated to phenotypes with currently unknown molecular basis (OMIM; June 17th, 2009). Additionally, we removed all associations regarding possible known disease genes from the given phenotype. We then applied the prediction procedure on the same CCNs, and considered as positive predictions those that satisfied a CCN-specific p-value threshold corresponding to the 1% FDR of the true predictions. We defined the precision as the fraction of cases in which the positive predictions contained the true disease gene.

Acknowledgements

We thank Jozef Jecz for critical reading of the manuscript. The financial support of FIRB-Italbionet program, of the Compagnia di San Paolo – Progetto Neuroscienze, of the Regione Piemonte Converging Technologies program and of the Italian Ministry of University and Research (MIUR)-PRIN program to FDC and of the "Associazione Italiana per la Ricerca sul Cancro" (AIRC) to PP, is gratefully acknowledged.

References

1 Yang JY, Yang MQ, Zhu MM, Arabnia HR, Deng Y: Promoting synergistic research and education in genomics and bioinformatics. *BMC Genomics* 2008; **9 Suppl 1**: I1.

2 Barabasi AL, Oltvai ZN: Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004; **5**: 101-113.

3 Oti M, Brunner HG: The modular nature of genetic diseases. *Clin Genet* 2007; **71**: 1-11.

4 Wang X, Dalkic E, Wu M, Chan C: Gene module level analysis: identification to networks and dynamics. *Curr Opin Biotechnol* 2008; **19**: 482-491.

5 Turner FS, Clutterbuck DR, Semple CA: POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* 2003; **4**: R75.

6 Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006; **78**: 1011-1025.

7 Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998; **95**: 14863-14868.

8 Zhou XJ, Kao MC, Huang H *et al*: Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol* 2005; **23**: 238-243.

9 Pellegrino M, Provero P, Silengo L, Di Cunto F: CLOE: identification of putative functional relationships among genes by comparison of expression profiles between two species. *BMC Bioinformatics* 2004; **5**: 179.

10 Stuart JM, Segal E, Koller D, Kim SK: A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* 2003; **302**: 249-255.

11 Oti M, van Reeuwijk J, Huynen MA, Brunner HG: Conserved co-expression for candidate disease gene prioritization. *BMC Bioinformatics* 2008; **9**: 208.

12 Ala U, Piro RM, Grassi E *et al*: Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput Biol* 2008; **4**: e1000043.

13 Carninci P, Kasukawa T, Katayama S *et al*: The transcriptional landscape of the mammalian genome. *Science* 2005; **309**: 1559-1563.

14 Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002; **30**: 207-210.

15 Peri S, Navarro JD, Amanchy R *et al*: Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003; **13**: 2363-2371.

16 Graf T, Enver T: Forcing cells to change lineages. *Nature* 2009; **462**: 587-594.

17 Takahashi K, Yamanaka S: Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006; **126**: 663-676.

18 van den Berg DL, Snoek T, Mullin NP *et al*: An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* 2010; **6**: 369-381.

19 Pardo M, Lang B, Yu L *et al*: An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell Stem Cell* 2010; **6**: 382-395.

20 van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA: A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006; **14**: 535-542.

21 Tarpey PS, Smith R, Pleasance E *et al*: A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet* 2009; **41**: 535-543.

22 Chiurazzi P, Schwartz CE, Gecz J, Neri G: XLMR genes: update 2007. *Eur J Hum Genet* 2008; **16**: 422-434.

23 Gecz J, Shoubridge C, Corbett M: The genetic landscape of intellectual disability arising from chromosome X. *Trends Genet* 2009; **25**: 308-316.

24 Wood SA, Pascoe WS, Ru K *et al*: Cloning and expression analysis of a novel mouse gene with sequence similarity to the Drosophila fat facets gene. *Mech Dev* 1997; **63**: 29-38.

25 Dobrin R, Zhu J, Molony C *et al*: Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol* 2009; **10**: R55.

26 Miller JA, Oldham MC, Geschwind DH: A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *J Neurosci* 2008; **28**: 1410-1420.

27 Mootha VK, Lepage P, Miller K *et al*: Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci U S A* 2003; **100**: 605-610.

28 Sibille E, Wang Y, Joeyen-Waldorf J *et al*: A molecular signature of depression in the amygdala. *Am J Psychiatry* 2009; **166**: 1011-1024.

29 Watanabe H, Darbar D, Kaiser DW *et al*: Mutations in sodium channel beta1- and beta2-subunits associated with atrial fibrillation. *Circ Arrhythm Electrophysiol* 2009; **2**: 268-275.

30 Giannandrea M, Bianchi V, Mignogna ML *et al*: Mutations in the small GTPase gene RAB39B are responsible for X-linked mental retardation associated with autism, epilepsy, and macrocephaly. *Am J Hum Genet* 2010; **86**: 185-195.

31 Whibley AC, Plagnol V, Tarpey PS *et al*: Fine-scale survey of X chromosome copy number variants and indels underlying intellectual disability. *Am J Hum Genet* 2010; **87**: 173-188.

32 Xu J, Burgoyne PS, Arnold AP: Sex differences in sex chromosome gene expression in mouse brain. *Hum Mol Genet* 2002; **11**: 1409-1419.

33 Huang Y, Baker RT, Fischer-Vize JA: Control of cell fate by a deubiquitinating enzyme encoded by the fat facets gene. *Science* 1995; **270**: 1828-1831.

34 Xu J, Taya S, Kaibuchi K, Arnold AP: Spatially and temporally specific expression in mouse hippocampus of Usp9x, a ubiquitin-specific protease involved in synaptic development. *J Neurosci Res* 2005; **80**: 47-55.

35 Lim WK, Wang K, Lefebvre C, Califano A: Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics* 2007; **23**: i282-288.

Figure legends

Figure 1: (A) Heatmap representing the fraction of common edges between different conserved-coexpression networks. The two random cases are representative examples of networks obtained from 10% of the experiments that compose the indicated dataset, chosen at random. The scale bar represents the percentage of common links (intersection/union). (B) Number of Gene Ontology keywords significantly enriched in the indicated number of conserved coexpression networks.

Figure 2: Neighbourhood of the POU5F1 gene in the conserved coexpression network obtained from stem cells microarray experiments. The size of the nodes and their distance

from the centre are a function of their connectivity. The genes displayed in orange have been successfully used to reprogram differentiated cells to iPS cells. The genes displayed in yellow have been experimentally linked to the pluripotent state or are considered as pluripotency markers.

Figure 3: (A) Number of OMIM-OMIM links unique to each of the phenotype coexpression networks (PCNs). (B) Charcot-Marie Tooth disease type 4D (CMT4D; OMIM 601455) and its first and second level neighbours in the CNS and NT PCNs. Grey nodes and black edges were found in both PCNs. Red and green nodes/edges were specifically found in the CNS or in the NT PCNs, respectively. Legend: CMT = Charcot-Marie-Tooth disease; MMZ = Myopathy, myofibrillar, ZASP-related; NEM = Nemaline myopathy; HMN = neuronopathy, distal hereditary motor; SNCV = slowed nerve conduction velocity; CHN = neuropathy, congenital hypomyelinating; HSAN = neuropathy, hereditary sensory and autonomic; RLHAD = Roussy-Levy hereditary areflexic dystasia; HNPP = neuropathy, hereditary, with liability to pressure palsies; HNDS = hypertrophic neuropathy of Dejerine-Sottas.

Figure 4: Representation of a subset of the neighbourhood of the USP9X gene in the three CCNs from which it was predicted as a candidate XLMR gene (normal tissues, CNS and skeletal muscle) via related phenotypes. It includes all the nodes that were connected to USP9X in at least one network. Thick edges represent links found in all three networks. The genes shown in orange are functionally involved in the Ubiquitin cycle, while those with a cyan border are involved in diseases related to XLMR (MimMiner score ≥ 0.4)

Table captions

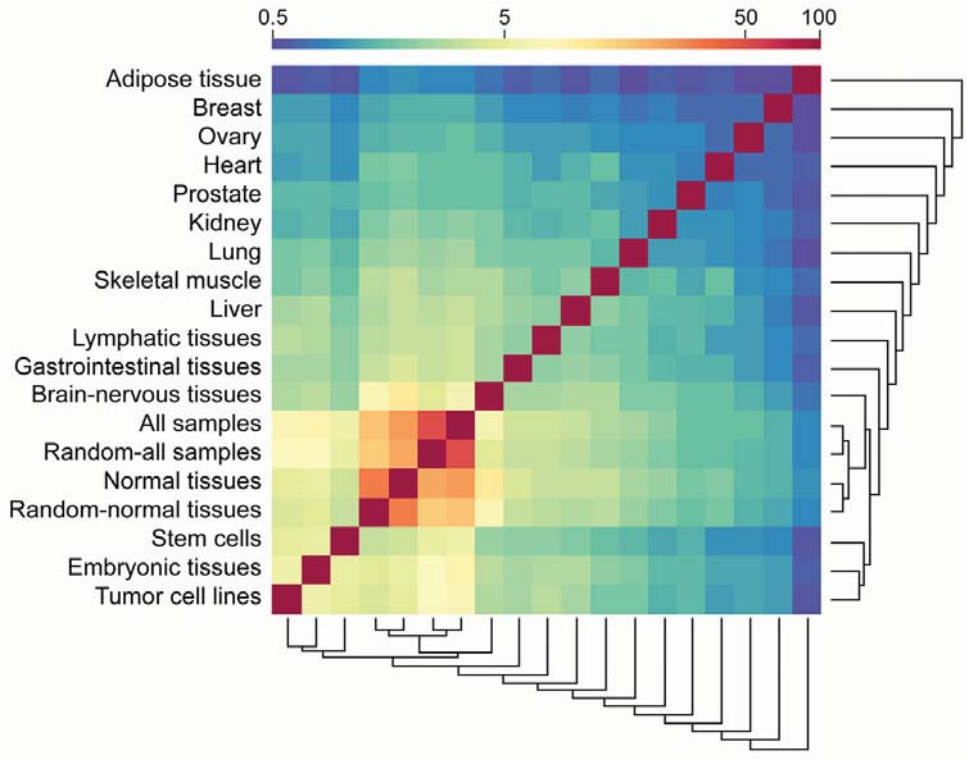
Table 1: List of the candidate genes for X-linked mental retardation generated from CCNs with a false discovery rate (FDR) of 0.1 or less, considering as reference genes all knownXLMR genes. The mutational score was obtained from reference 22, and “T” indicates that in the same study a truncating mutation was detected in the gene.

Table 1

Gene symbol	Number of networks	Mutational Score	Best p-value	Best FDR
GPRASP1	4	1	1.78E-07	0
GPM6B	1	NA	1.78E-07	0
GSPT2	2	11	3.49E-07	0
NAP1L3	2	4	3.49E-07	0
MAGEE1	2	2.04	3.49E-07	0
ARMCX1	2	NA	3.49E-07	0
NAP1L2	2	NA	3.49E-07	0
TCEAL1	2	NA	3.49E-07	0
USP9X	2	T	3.89E-07	0
ARMCX3	2	11.84	2.86E-06	0
RAB39B	2	NA	2.86E-06	0
CHIC1	1	NA	6.61E-06	0
DYNLT3	1	NA	6.61E-06	0
NKRF	2	NA	6.61E-06	0
TRO	2	NA	6.61E-06	0
RLIM	2	NA	1.00E-05	0.04
THOC2	1	10.528	1.71E-05	0
DIAPH2	1	3.87	1.71E-05	0
PLS3	1	NA	1.71E-05	0
REPS2	2	NA	1.71E-05	0
SH3BGRL	1	NA	1.71E-05	0
TMEM47	1	NA	1.71E-05	0
ZFX	1	NA	1.71E-05	0
ZXDA	1	NA	2.01E-05	0.08
ZXDB	1	NA	2.01E-05	0.1
CLCN4	2	21.359	2.47E-05	0.07
BEX2	2	1.37	2.47E-05	0.06
EIF1AX	1	NA	2.53E-05	0.03
TMEM164	1	NA	2.53E-05	0.05
YIPF6	2	NA	2.53E-05	0.04
ZMAT1	2	NA	2.53E-05	0.06
F9	1	13.068	3.14E-05	0.03
SMARCA1	1	24	3.30E-05	0
GPRASP2	3	4	3.30E-05	0
MAP7D2	1	3.172	3.30E-05	0
BEX1	3	NA	3.30E-05	0
BEX4	1	NA	3.30E-05	0
NUDT11	2	NA	3.30E-05	0
CHM	2	6.96	5.07E-05	0.08
CNKSR2	1	20.678	6.28E-05	0.07
PDZD4	1	NA	6.28E-05	0.08
PJA1	1	2.44	6.46E-05	0.07
LRCH2	1	16.2	6.51E-05	0.02
SYAP1	1	NA	7.67E-05	0.04
FGF13	1	NA	8.66E-05	0.08
USP11	1	NA	8.66E-05	0.08
MBNL3	1	6.354	9.92E-05	0.04
HMGN5	1	NA	9.92E-05	0.03
MAP3K7IP3	1	NA	9.92E-05	0.03

Figure 1

A


0.5
5
50
100
Adipose tissue
Breast
Ovary
Heart
Prostate
Kidney
Lung
Skeletal muscle
Liver
Lymphatic tissues
Gastrointestinal tissues
Brain-nervous tissues
All samples
Random-all samples
Normal tissues
Random-normal tissues
Stem cells
Embryonic tissues
Tumor cell lines

B

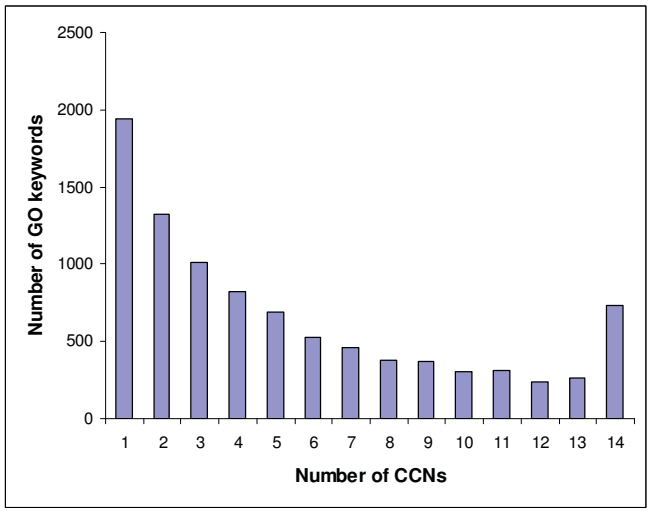

Number of GO keywords
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Number of CCNs

Figure 2

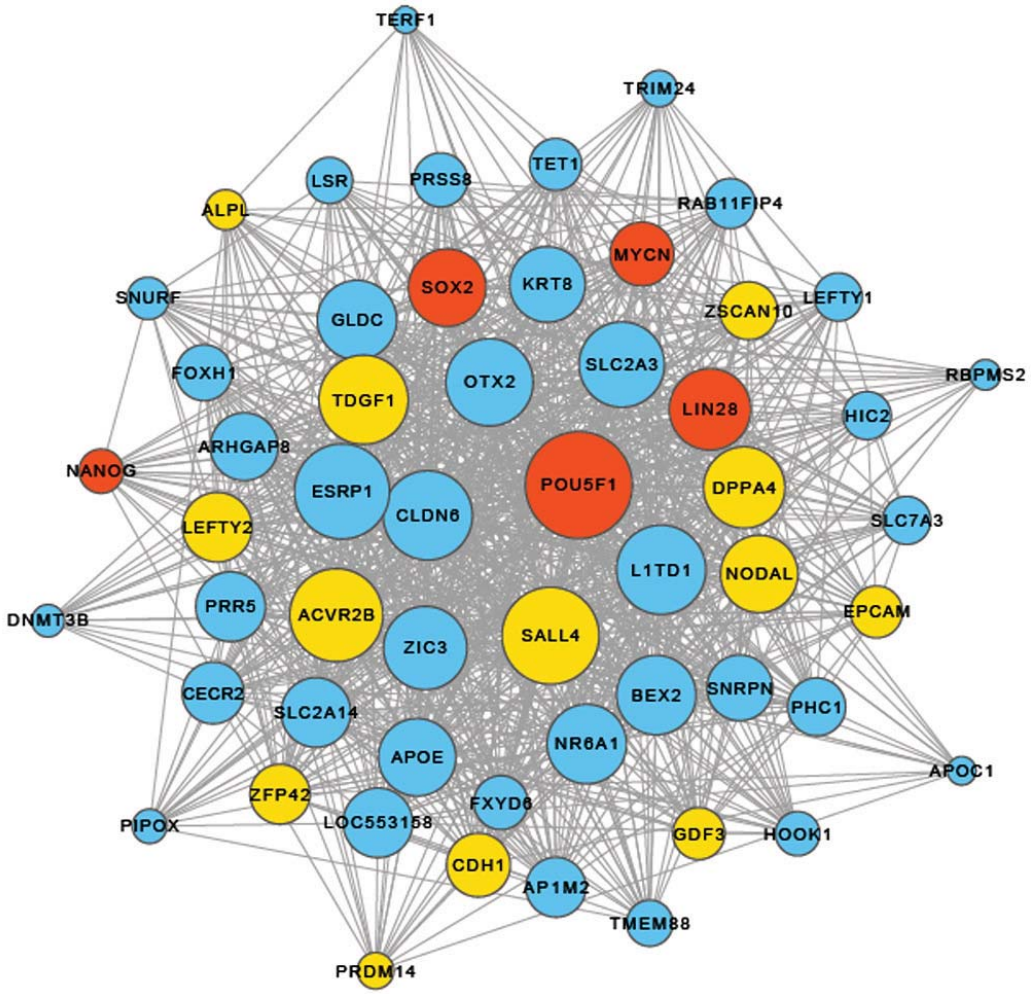
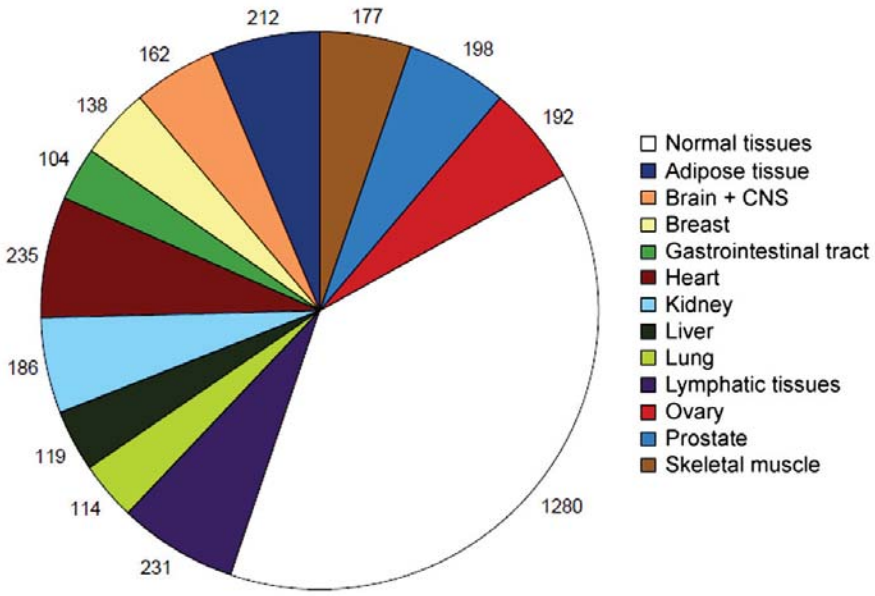

TERF1
TRIM24
TET1
LSR
PRSS8
ALPL
RAB11FIP4
MYCN
SOX2
KRT8
SNURF
ZSCAN10
LEFTY1
GLDC
SLC2A3
FOXH1
OTX2
RBPMS2
TDGF1
LIN28
HIC2
ARHGAP8
NANOG
ESRP1
POU5F1
DPPA4
CLDN6
SLC7A3
LEFTY2
L1TD1
NODAL
PRR5
ACVR2B
EPCAM
DNMT3B
ZIC3
SALL4
BEX2
SNRPN
CECR2
PHC1
SLC2A14
NR6A1
APOE
APOC1
ZFP42
FXYD6
PIPOX
LOC553158
GDF3
HOOK1
CDH1
AP1M2
TMEM88
PRDM14

Figure 3

A


212
177
162
198
138
192
104
235
186
119
114
231
1280
Normal tissues
Adipose tissue
Brain + CNS
Breast
Gastrointestinal tract
Heart
Kidney
Liver
Lung
Lymphatic tissues
Ovary
Prostate
Skeletal muscle

B

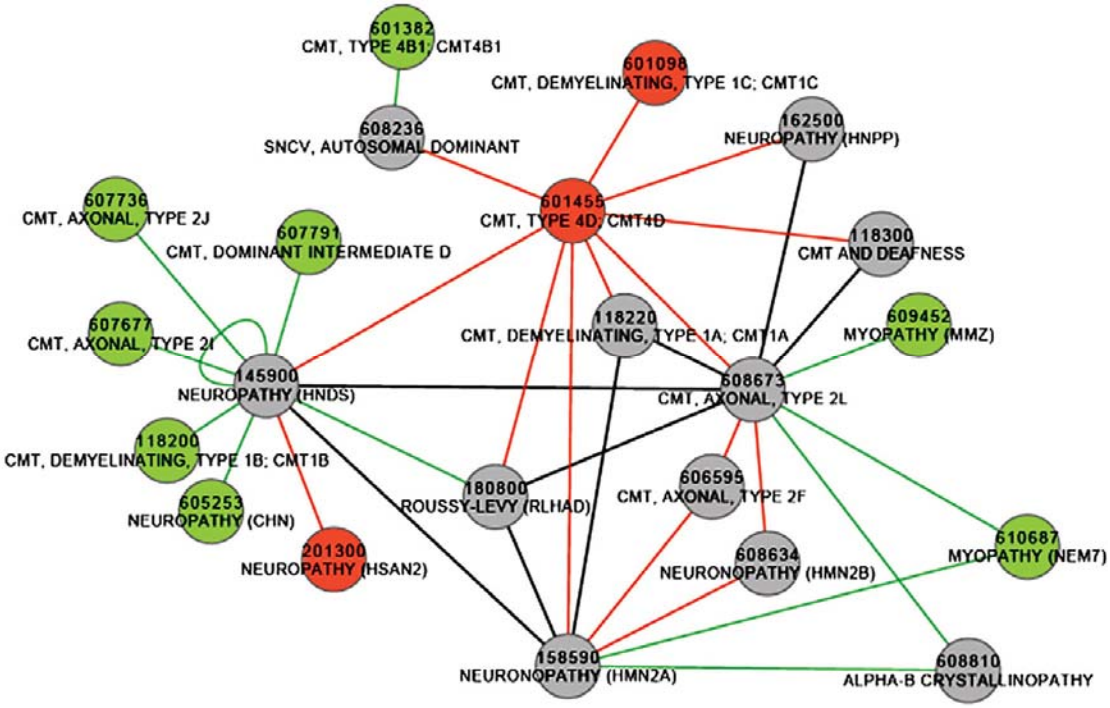
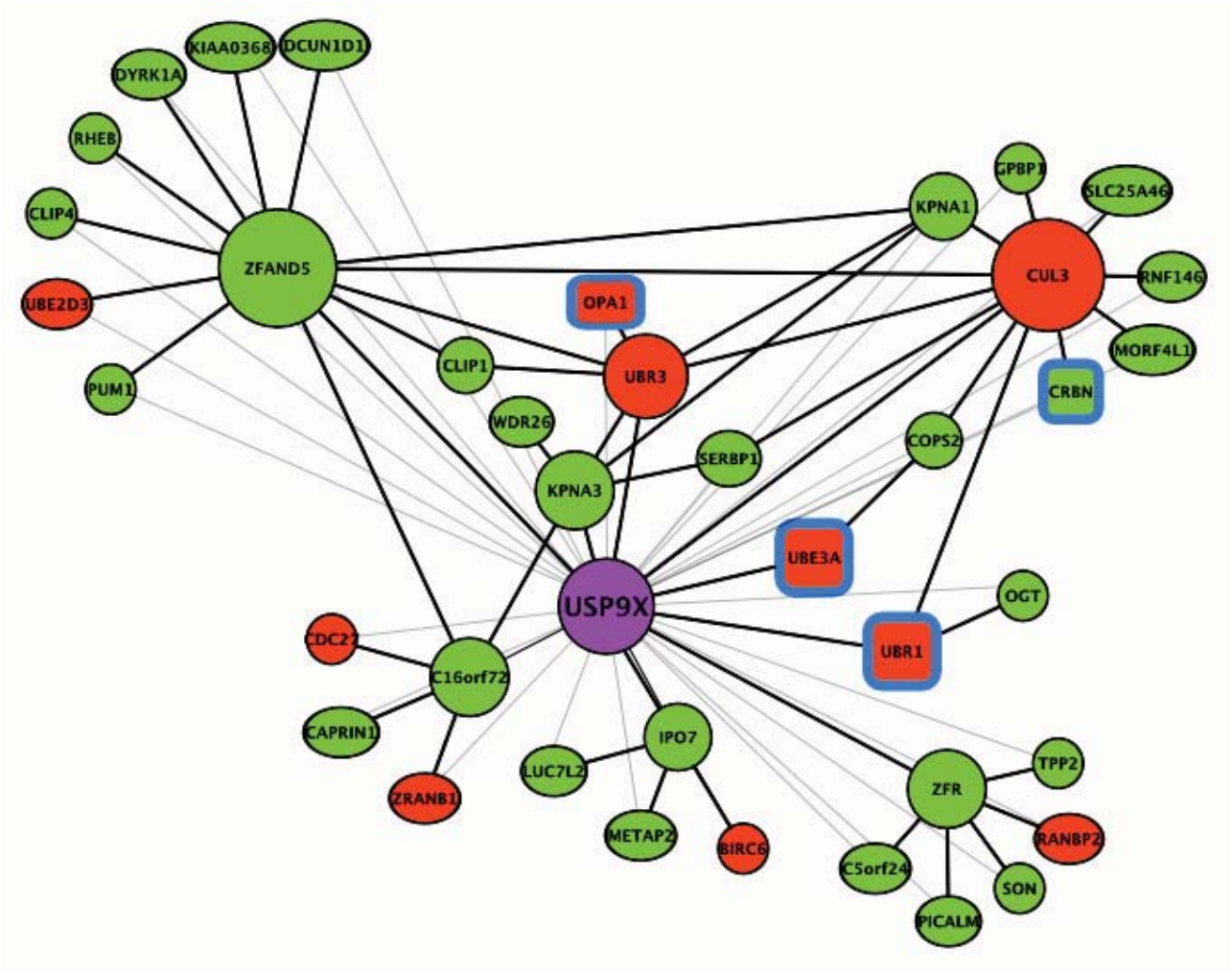

601382
CMT, TYPE 4B1; CMT4B1
601098
CMT, DEMYELINATING, TYPE 1C; CMT1C
608236
SNCV, AUTOSOMAL DOMINANT
162500
NEUROPATHY (HNPP)
607736
CMT, AXONAL, TYPE 2J
601455
CMT, TYPE 4D; CMT4D
607791
CMT, DOMINANT INTERMEDIATE D
118300
CMT AND DEAFNESS
607677
CMT, AXONAL, TYPE 2I
118220
CMT, DEMYELINATING, TYPE 1A; CMT1A
609452
MYOPATHY (MMZ)
145900
NEUROPATHY (HNDS)
608673
CMT, AXONAL, TYPE 2L
118200
CMT, DEMYELINATING, TYPE 1B; CMT1B
180800
ROUSSY-LEVY (RLHAD)
606595
CMT, AXONAL, TYPE 2F
605253
NEUROPATHY (CHN)
201300
NEUROPATHY (HSAN2)
608634
NEURONOPATHY (HMN2B)
610687
MYOPATHY (NEM7)
158590
NEURONOPATHY (HMN2A)
608810
ALPHA-B CRYSTALLINOPATHY

Figure 4


KIAA0368
DCUN1D1
DYRK1A
RHEB
CLIP4
ZFAND5
UBE2D3
PUM1
GPBP1
SLC25A46
KPNA1
CUL3
RNF146
MORF4L1
CRBN
OPA1
CLIP1
UBR3
WDR26
SERBP1
COPS2
KPNA3
UBE3A
USP9X
OGT
UBR1
CDC23
C16orf72
CAPRIN1
ZRANB1
IPO7
LUC7L2
METAP2
BIRC6
ZFR
TPP2
RANBP2
C5orf24
SON
PICALM