Not by systems alone: identifying functional outliers in rare disease pedigrees

Authors:

Sara Ballouz,[1]
Max Dörfel,[1]
Jonathan Crain,[1]
Megan Crow,[1]
Laurence Faivre,[2,3]
Catherine E. Keegan,[4]
Sophia Kitsiou-Tzeli,[5]
Maria Tzetis,[5]
Gholson J. Lyon, [1,6,7]
Jesse Gillis*,[1]

Affiliations:

1 The Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA;

2 GAD team, INSERM UMR 1231, Université Bourgogne Franche-Comté, Dijon, France;

3 Centre de Référence Maladies Rares Anomalies du Développement et FHU TRANSLAD, CHU Dijon, Dijon, France;

4 Department of Pediatrics, Division of Genetics and Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA;

5 Department of Medical Genetics, National Kapodistrian University of Athens, Athens, Greece;

6 Graduate Genetics Program, Stony Brook University, Stony Brook, NY, 11794, USA;

7 Utah Foundation for Biomedical Research, Salt Lake City, UT, 84107, USA;

* Corresponding author: Dr Jesse Gillis, The Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA

Contact Information:

JG: jgillis@cshl.edu

GJL: glyon@cshl.edu

SB: sballouz@cshl.edu

MD: mdoerfel@cshl.edu

JC: jcrain@cshl.edu

MC: mcrow@cshl.edu

LF: laurence.faivre@chu-dijon.fr

CEK: keeganc@med.umich.edu

MT: mtzetis@med.uoa.gr

SKT: ekanavak@cc.uoa.gr

Summary

In disease expression analysis, looking for shared functional signals from a set of genes which exhibit differential expression is commonplace. We examine the complement as a possibility, that disease genes display "outlier" or unexpected expression relative to broader patterns of functional expression variation. Using six families from the rare *TAF1* syndrome disease cohort, we performed family-specific differential expression analyses and find that functional characterization of top candidates enriches for common pathways unlikely to be specifically linked to disease. However, by filtering away common expression changes using known co-expression, we lose all functional enrichment and are left with a small number of outliers characteristic of each proband. Two of these outlier genes are highly recurrent across pedigrees (FDR <2.63e-05) and are the primary commonality among the cohort as a whole. This suggests that systems analysis may be relevant to rare diseases principally as a means of filtering out biological signals unrelated to disease.

Keywords

Rare disorders; TAF1, differential expression; co-expression; family-based differential expression; outlier expression analysis; rare expression; recurrent transcriptomic dysregulation

Introduction

Systems biology takes as its fundamental premise that biological systems can be best understood through the relationships between their parts, rather than detailed examination of those parts in isolation (Kitano, 2002). Both systems approaches and purely reductionist alternatives will be useful in different contexts, but systems analyses are particularly relevant to functional properties, where genes or their downstream products generally must be selected to work in coordination. The groupings defined by this coordinated activity may be assessed transcriptionally (Segal et al., 2003), proteomically (Huynen et al., 2003), by sequence relationship (Marcotte et al., 1999), chemical activity (Keiser et al., 2009), or a host of other factors (Guan et al., 2014; Zhou et al., 2014), in each case yielding gene networks to define relationships at a systems level. This framework not only captures important elements of the underlying biology, but also places genes within a simple joint feature space for computational analyses. Methods to discover commonality between genes range in methodology, at the most basic level using enrichment to find overlap with known processes and pathways (Irizarry et al., 2009) to more sophisticated machine learning algorithms that reveal hidden patterns in the data (Libbrecht and Noble, 2015) . Biologists have come to rely on these methods to study complex biology and disease (Gaiteri et al., 2014). But, what does a systems approach exclude?

Systems methods determine signals defined by genes jointly within the data. Schematically, this is shown in **Figure 1**. Simply, genes can be linked based on overall shared functionality whether defined explicitly (e.g., KEGG pathways (Kanehisa et al., 2008)) or implicitly (e.g., PPI data (Chatr-aryamontri et al., 2017)). A set of genes of interest (e.g., differentially expressed) are placed within that overall network (**Figure 1A**). Many of the genes will, as expected by the systems framework, be involved in the same process, and it is these genes, along with their associated shared function, that systems approaches will identify as salient to the experiment. In this work, we will be interested in the genes specifically not part of the system that is most prominent within the data. Even if those remaining are

involved in many functional processes, they are not characteristic of the dominant systems signal within the data. Instead, these genes appear as outliers (**Figure 1B**). One way to think of these genes is to consider disease, where systems break down or respond unusually to a perturbation. It might be that genes acting uncharacteristically are of relevance to the dysfunction. We call these rogue actor genes "functional outliers", and it is their possible disease significance which we hope to identify within this work, focusing on rare disease.

Rare diseases pose both a statistical and functional problem in genomics; statistical, because they are hard to power, and functional, because the biology may not naturally generalize from other systems. The challenges notwithstanding, rare disorders are important to study both for their role in population disease burden, which may be substantial in aggregate, and for the unique window they offer into molecular processes underlying human biology (Forrest et al., 2011). Rare diseases are close to an ideal hunting ground in the search for functional outliers because they are usually caused by a few rare and disruptive variants (Boycott et al., 2013). Since functional outliers can arise in interest only by recurrence across individuals (rather than shared activity in a common process), it is preferable to have a cohort whose genetic architecture is similar so that gene-level recurrence may be more likely. The disease cohort which we focus on is a recently reported set of families with rare mutations in the *TAF1* transcription factor (TATA-Box Binding Protein Associated Factor 1) contributing to a well-defined phenotypic alteration (O'Rawe et al., 2015). Genetically defined rare cohorts like the *TAF1* syndrome are extremely unusual, given the current state of genomic knowledge where complex diseases such as schizophrenia and autism are (devastatingly) genetically heterogeneous. However, rare diseases defined at the level of DNA variation may account for subtypes of many more common diseases (see, for example this n= 15 cohort in autism (Bernier et al., 2014)). *TAF1* syndrome has a number of features that may enrich for the presence of functional outliers within the expression data including a phenotype not solely localized to the brain and a molecular basis where unusual or unbuffered expression

alterations may be likely (Lee and Young, 2013; O'Rawe et al., 2015). In general, our viewpoint is that the less the disease looks like a convergent regulatory response, the greater the chance we will enrich for functional outliers. However, this is not a supposition upon which our analysis depends, merely a factor which will have implications for the scope of applicability of our approach and findings.

In this work, we develop a means to characterize functional outliers, genes which are exhibiting anomalous differential expression specifically in the context of the expression of the other genes in the data. We first treat each of our pedigrees as a separate differential expression experiment, assess the candidate gene list for systems convergence and determine the remainder not showing systems convergence. We then exploit the known overlaps between our pedigrees to use recurrence of signals across them – and only that – as our measure of significance for both classes of result. Within our experimental paradigm, optimized to detect functional outliers, we find that most differential expression exhibits expected co-variation between genes, but there are exceptions which can be robustly characterized. It is these exceptions which are validated by the analysis of the cohort as a whole. The most prominent functional outlier within our data is a highly plausible candidate to play a role in *TAF1* syndrome. We close our results by assessing the implications from our targeted family-based analysis to applying functional outlier detection in case-control experiments. Finally, we discuss whether the unusually clear role functional outliers appear to play in *TAF1* syndrome can be expected to generalize to other disorders.

Results

Using co-expression as a filter to define unexpected differential expression

In this work, we rely on largescale co-expression meta-analysis to define functional relationships between genes, not just as a standard method of capturing functional relationships (Lee et al., 2004), but also because it captures the expectation of co-incidence within a hit list derived from a differential expression (DE) experiment in a purely technical sense (i.e., co-expression is co-variation of expression). Many of the elements of our approach are relatively conventional (Robles et al., 2012); however, two atypical elements are worth highlighting. First, our analysis is highly conservative; we aggregate across individual networks thresholded as having any positive correlation at all (and then take the complement). This would be overly permissive in identifying functional links and is therefore conservative at identifying outliers, likely at a cost to performance. Second, we consider only positive relationships. We will consider our DE lists directionally; i.e., treating positive and negative differential expression as separate hit lists. In this context, only genes jointly positively differentially expressed can be said to be expected to be part of the same list. A summary of our approach is detailed in **Figure 2** (see experimental procedures for more details).

To generate a common co-expression frequency network, we tally up all poorly co-expressed pairs (Spearman's correlation coefficient $r_s \leq 0$) from 75 co-expression networks across 3,653 samples (see experimental procedures and **Figure 2A,** experiments listed in **Table S1**). One potential problem with filtering by co-expression for outliers is the possibility that outliers are simply unusual due to noise, in the form of low expression. For example, the absence of joint expression might be observed because the level of expression of these genes is weak or noisy. However, we see little relationship between the average expression levels and the presence within our co-expression frequency network (Spearman's correlation coefficient $r_s=0.03$). Another possibility is that the genes are not measured or detected,

which would not affect the mean expression, but affect co-expression. Once again, we see close to no relationship with the number of detected genes and the node degree of our co-expression frequency network (Spearman's correlation coefficient r_s=-0.03). As another check of the information within the network, we measured its performance in terms of being able to recapitulate gene membership across Gene Ontology (GO (Ashburner et al., 2000)) terms through a simple machine learning method (Ballouz et al., 2016) (see experimental procedures). As expected, the performance metric of 0.26 (AUROC) across all GO terms is far lower than random (0.5), indicating the tally of non-relationships strongly captures the absence of shared function.

Family-based differential expression is dominated by expected functional co-variation

We perform family-based differential expression in order to benefit from the shared genetics of the trio (see experimental procedures), but potentially introducing other sources of common co-variation, which our approach then looks to filter. For each pedigree, we sequence the RNA of the parents and proband, and then compare the transcriptional profile of the proband to the parents (**Figure 2B**). We take the top differential genes in both directions: those showing increased (up-regulated) or decreased (down-regulated) expression, and quantify and visualize their occurrence co-expression pattern in the co-expression frequency network (**Figure 2C**). Genes that are never seen as co-expressed will be closer to 1 (yellow/white), while those often seen are close to 0 (red). For each family, we observed large modules (red blocks) in the data, indicating that those genes coming up as differentially expressed are also co-expressed and are also broadly reflective of functions expected to be represented in blood. We show a representative plot of this co-expression matrix as a heatmap for family 3 (**Figure 3A**, **Table S2** for DE gene lists for each family). We detect these co-expression modules or "co-expression blocks" through hierarchical clustering of the genes to generate a dendrogram, and then a dynamic tree cutting algorithm to identify the modules (see experimental procedures).

Our first observation is that the genes in these modules are clearly functionally related, or belong to gene families. Since our co-expression frequency network is in essence an inverted co-expression network, and co-expressed genes are believed to be functionally related, this was not surprising. The genes within these co-expression blocks also dominate enrichment analysis within the DE hits. For example, we see a large block of immunoglobulins, one of interferon-related proteins and a third related to mitosis and the cell cycle in the top 100 up-regulated genes in family 3. Performing gene set enrichment on this set of genes, we obtain 63 significant (after multiple hypothesis test correction) GO terms enriched in our example shown in **Figure 3B**. Here, we show the overlap of the genes (rows) and the enriched GO terms (columns). The genes are clustered as they are in part **Figure 3A**; co-expression blocks remain in one piece. We see that the significant GO terms contain genes almost exclusively in those blocks, as indicated by the colored segments in the heatmap.

The terms enriched for are immune related (e.g., GO:0071357 cellular response to type I interferon, p~3.27e-21) and cell-cycle related (e.g., GO:0000278 mitotic cell cycle p~0.008). Removing the co-expression modules weakens or removes most enrichment. That the genes showing expected functional behavior (co-expression blocks) appear to be enriched for functions that seem likely to be confounds provides some support for the possibility that their opposite (outliers) may represent dysfunction within the data. Top DE enrichment results are listed in **Table S3**.

Common functional variation can be robustly filtered

For each of our 6 families, we filter away these co-expression modules in order to retain the genes with rare co-expression and unexpected (in the context of the other hits) differential expression. As an example (**Figure 4A**), we show the genes left for assessment once we remove the co-expression blocks from the top 100 down-regulated genes. In this case, we are left with 28 genes, and on average 42 genes differentially expressed in each family.

We wished to test the dependence of our results on our parameter choices (i.e., the cutoffs for selecting a gene to be termed DE and the size of the module to filter). We varied the number of DE genes we consider by selecting thresholds from 10 to 1000 (**Figure 4B**). As we increase the number of genes, we filter roughly between a third to two-thirds of the hits. We used the GO enrichment of the sets as a check for the promiscuity of expected function since we observed that most enrichment is from these co-expression blocks, as might be expected from the use of co-expression to predict function. We see an increase in the number of significant terms returned as enriched as a function of the number of DE genes considered (**Figure 4C,** black line is the average across all families), and almost a complete loss of enrichment once we filter the co-expression blocks (red line in **Figure 4C** with remaining families in **Figures S1-S2**). A few GO terms did remain enriched across some of these DE gene sets. However, they were mostly still generic or blood related (e.g., GO:0030218 erythrocyte differentiation, GO:0005576 extracellular region), although a few were potentially interesting (e.g., GO:0048812 neuron projection morphogenesis), especially given the tiny number now present.

Identification of functional outliers

Statistical significance does not arise from analysis within each family (n=1), but by recurrence of high fold-change genes across families. To detect and characterize functional outliers, we begin by measuring the recurrent overlap of the top up and down regulated genes (**Figure 5B**). Relatively few genes overlapped between the families, with at most 26/100 overlapping between any pair of families. To calculate the probability of recurrence of differentially expressed genes (**Figure 5C**), we used the binomial distribution. A modest number of genes are significant across the families at this level, but mostly due to common functional variation (red highlighted genes in **Figure 5C**). We see that once we filter common co-expression, we lost almost all recurrence (**Figure 5D**), with a very small number of functional outliers remaining. Note that this is, if anything, a mildly positive result since “easy” or promiscuous significance is precisely what functional outliers should select against.

We once again performed a threshold analysis, to observe the trends and effects on our results when we alter the stringency of our parameters. As in the previous analysis, we varied the number of genes to consider as a hit. The number of recurrent genes increases the more genes we take as differentially expressed (dashed lines in **Figure 5E**). Of course, this is balanced by the threshold required for a gene to be considered as significant (grey lines in **Figure 5E**). There is generally an increase in the number of significantly recurrent genes with per family fold-change "hit" thresholds into the low 100s. After thresholding a hit to the top 200 genes, no genes are significantly recurrent for the up-regulated genes, and the same drop-off occurs at close to 400 genes for the down-regulated genes (**Figure 5E**), when the threshold required to pass as recurrent jumps. These trends repeat, as we see more significantly recurrent genes farther down in the fold-change threshold. However, these genes are not the same as those found as recurrent in the earlier part of the analysis and given the fold changes by this stage (500 down in the list for some families) are hard to discern from potentially overlapping technical noise (i.e., SEQC recommendations of < log2 FC 1 and low expressing (Consortium, 2014)). The analysis generally suggests a threshold of top 100 is reasonable, which was chosen based on our sense of what is typically regarded as biologically plausible. In the reverse, to the extent the statistics support what is usually an ad hoc decision motivated by biological intuitions, it suggests our analytical framework is well-calibrated to the underlying biology.

Filtering for co-expression blocks leaves three genes as significantly recurrent (**Figure 5D**). One of two highest ranking candidates is the calcium channel subunit *CACNA1I* (FDR<2.63e-05), and a potentially interesting candidate for follow-up work. Mutations in calcium channels are known to have similar phenotypes to the cohort here, including intellectual disability, autism and dystonia (Fukai et al., 2016; Lu et al., 2012), and a *TAF1* binding site is located upstream of the gene (Wang et al., 2012). Most convincing however is that this gene has been implicated recurrently in neurological diseases; it is the only recurrent missense *de novo* in schizophrenia studies (Gulsuner et al., 2013), one of the few

overlapping candidates between schizophrenia and autism (Iossifov et al., 2015; Lu et al., 2012) , and also one of the hits in the largest schizophrenia analysis (Schizophrenia Working Group of the Psychiatric Genomics, 2014). It is almost a uniquely strong candidate for missense variation to play a role in a complex-phenotype neuropsychiatric disorder, and is a "hit" in 4-5 (depending on threshold) of our 6 families. Given this, it is interesting that only in family 4 is this gene not differentially expressed at all. Family 4 is the only family in which the mutation in the proband is a CNV duplication, and this proband is not characterized as possessing one of the key phenotypic features making this disorder distinctive with respect to other neuropsychiatric diseases (e.g., the unusual intergluteal crease).

The other strong outlier candidate, also seen in 4-5 of the 6 families, was the insulin-like growth factor-binding protein 3 *IGFBP3* (FDR <2.63e-05). This gene was also not found in family 4, and also not at our default settings in family 3. In this case, it showed modest differential expression in both these families, but had high variability between the parents and so was filtered away in family 3 (see experimental procedures), and in family 4 was below the default threshold. In general, and in contrast to *CACNA1I*, *IGFBP3* showed at least some co-expression with other DE genes.

Exploiting functional outliers in conventional case-control analysis

Our data can be repurposed as a conventional case-control analysis with 6 batches blocked between case and control but not controlled for age and sex. We repeated the differential expression analysis, but this time using all the probands and all the parents as if a case-control while also correcting for batch effects. Interestingly, *CACNA1I* and *IGFBP3* are the top ranked candidates (**Figure 6A**). The similarity of our outlier-based DE and regular case-control DE result might seem to negate the necessity of detecting outliers, but we suggest the opposite is true. It is startling that we remove the majority of hits within each family as being definitely not outliers and still obtain the same results as case-control analyses from the entire cohort. The strong implication is that we are, in essence, predicting which hits

within a family are unlikely to be replicated across the entire cohort (those that filter away). To better quantify this possibility, we downsampled the families assessed for both case-control and recurrence (of outliers) analysis and use the top hits within the whole case-control cohort as positives to assess the performance of the downsampled data. We use the rankings of these two genes as our true positive set, for all family-proband combinations (see experimental procedures), including family 4. Although recurrence is not significant when using only two families, we still see on average much better ranks when using the recurrence analysis as compared to the case-control analysis (**Figure 6C** and **D**), typically varying by orders of magnitude in precision. As we increase the number of cases, the difference in average rank shrinks, although only because the recurrence reaches near perfect performance after only a small number of probands. As suggested, the identified systems signal associated with all co-expression and enrichment is precisely what is averaged away as the cohort is expanded across all families; however, the same signal can be identified within families by filtering away the systems signal directly.

Having characterized where DE arises within this data when assessed by one pipeline, we wanted to determine the generalizability to DE pipelines other than the comparatively simple one we employed. To this end, we re-analyzed our data using DESeq2, a popular mainstream choice (Love et al., 2014). The hit lists identified are highly similar as is the presence of co-expression blocks within each analysis (**Figure S4**). This seems plausible to us because those signals are real and biological, just less likely to be associated with a rare disease (which DESeq2 cannot assume). Moreover, those co-expression blocks typically rank above the positive hits validated by the cohort as a whole, with a mean rank (of both *CACNA1I* and *IGFBP3*) improving from ~58th to ~9th after filtering, similar to our earlier results (**Figure S5**), although recurrence drops once we filter too stringently. While this suggests that identification of functional outliers is largely robust to DE pipeline, numbers of genes to include (e.g., top 100), exact cut depths, and other factors which were robust within our data may vary in unassessed pipelines. There is

also potential for variation on a per family basis as each family has its unique (and potentially artifactual) properties.

More broadly, our observation is that the benefit of assessing the individual families separately is the ability to filter off pathways that are generic, but do characterize that sample-set idiosyncratically (i.e., are strong biological artifacts of that data unrelated to diseases). While co-expression blocks are still visible in the top 100 candidate genes in the full case-control version of our analysis, the pattern has become blurred (**Figure 6B**) and filtering parameters may need to be carefully considered in high 'n' studies.

Discussion

The two main contributions of this work are the description of a means of determining functional outliers in expression data and providing a proof-of-principle of their importance in the analysis of the *TAF1* cohort. There are many considerations which motivated our interest into whether functional outliers exist. Medically, functional outliers seem like unusually strong candidates to have a retained signal across tissues, including those not principally affected by the disease. In an analysis where genes show some joint change in activity reflective of shared regulation within a pathway, it is likely that the pathway is itself variable across tissues, especially if the disease is. In contrast, functional outliers look like they have "missed" regulation to some degree. When other regulated processes change as expected across tissues, a gene that is not regulated in one tissue will be more likely to also be unregulated in other tissues. Hypothetically, this could be because the system cannot buffer the transcriptional variation: new interactions are generated or control is lost. Similarly, functional outliers may represent a critical point of causation for rare disorders if they provide a substantial portion of the signal; in essence, serving as a bottleneck through which all later joint dysregulation passes.

From the perspective of network biology, functional outliers are of interest because they are more likely to occur in topologically unusual locations in gene networks and particularly may overlap with "critical connections" encoding non-redundant functional information (Gillis and Pavlidis, 2012). Along related lines, our approach for identifying functional outliers subtly resembles permutation testing in enrichment software, since that holds constant correlations between genes; it is this very correlation we are estimating from other data. More broadly, as the significant remainder from many types of conventional analysis, the presence of functional outliers helps define "unknown unknowns" in interpreting expression profiles. For all these reasons, we expected the presence or absence of functional outliers to be worthy of consideration. What was unexpected, however, was the large degree to which selecting for them appeared to improve the interpretability of our expression data.

This is not to say that functional outliers dominate the underlying biology, even of rare disease. We do not think this is the case, but the picture our results paint is that common functional co-variation does dominate the space of false positives. Thus, even though all our samples are derived from blood, genes related to blood show differential expression and functional enrichment because the precise degree of signal is not held constant from sample to sample. From that perspective, identifying functional outliers is simply a form of "data clean-up", albeit a comparatively challenging one since it is untargeted biological signals, rather than technical ones, which are being cleaned up. After this clean-up, rather than finding lots of commonplace signal, we find a very small number of potential candidates.

The clearest candidate, *CACNA1I* appears highly plausible both from the disease-variation known to be linked to it, and also from phenotype overlaps in knock-out models. Mouse knock-outs show a poor contact righting reflex, change in number of caudal vertebrae, and an impaired auditory response (Koscielny et al., 2014). In comparison, the probands suffer from hypotonia, an unusual intergluteal crease, a prominent protruding coccyx, and impaired hearing (O'Rawe et al., 2015). While our analysis

does not incorporate the post-hoc observation that family 4 is distinct from the others with respect to this phenotype (and genotype), the fact this observation was brought to light specifically by the expression data seems promising since an expression-first characterization of disease would be highly desirable, where possible. Consequently, other known roles of *CACNA1I* may help define follow-up characterization of *TAF1* syndrome. *CACNA1I* encodes a pore-forming alpha subunit of a voltage-gated calcium channel (Cav3.3), also known as a low voltage activated T-type calcium channel. Channels of this type are involved in variety of calcium-dependent processes, including the modulation of neuronal firing patterns (Talley et al., 1999), and may play roles in pacemaking activity, hormone secretion, cell growth and differentiation (Carbone, 2009). The gene is highly expressed in the brain (GTEx (Lonsdale et al., 2013)), and the protein is seen expressed in bone marrow mesenchymal stem cells and the prefrontal cortex (proteomicsDB (Wilhelm et al., 2014)).

It is possible to imagine a number of factors which made functional outliers of particular importance to our study. To whatever extent this is true, it may make them of less value or importance in other analyses. For example, as our samples were derived from blood, it is possible that our RNA-seq analysis was in particular need of the type of clean-up filtering for commonplace co-expression which our approach provides. Or in a nearly opposite view, it is possible our data is unusually clean due to the clarity and homogeneity of the disorder we were studying and that in most cases, recurrence reflecting broader phenotypic overlaps can only arise downstream of any functional outliers. And, of course, many of the caveats that would apply to any expression analysis can apply to ours and, in particular, gene-specific variation in the degree to which we are powered to detect changes may be important; we followed the guidelines derived from the SEQC experiments, but genes can be outliers for novel reasons and care will need to be taken. For instance, it would be problematic if a gene passed co-expression blocking simply because it was more variable than its functional group in general. In our analysis, signals were clear enough to make results robust to these considerations, and while we have provided a

general pipeline, the methods mostly provided straightforward results that were visible by eye. Our thresholds and parameters are likely data dependent and potentially DE method dependent, and we recommend the approach to complement researchers' preexisting methods, not as an alternative. Experimentalists can directly use the co-expression block data we have provided to assess their own results in more detail. We suggest assessment for functional outliers is a useful characterization of expression data wherever rare expression variation potentially plays a role.

Conclusion

In our *TAF1* disease cohort, most of the differential expression reflects co-variation that is expected between genes, indicating differences between the salience of those signals unrelated to disease. Filtering off these signals reveals functional outlier genes, whose presence as differentially expressed is specifically not expected in the context of the other hits within an experiment. The most prominent functional outlier, *CACNA1I*, is a very plausible candidate to play a molecular role in *TAF1* syndrome. Characterization of functional outliers should be incorporated as a default into studies of rare disorders and possibly more broadly, where unique phenotypic convergence is present.

Experimental Procedures

RNA-sequencing and processing

We collected blood from 6 of the pedigrees, renaming them from the original work and listed in **Table 1**. For more information on the families, refer to (O'Rawe et al., 2015).For RNA sequencing, blood was collected in PAXgene Blood RNA tubes and the RNA was isolated with the PAXgene Blood RNA kit (QIAGEN) according to the manufacturer's recommendations. The RNA was quantified using NanoDrop. To increase downstream sensitivity, globin mRNA was depleted from the samples using the GLOBINclear Kit (Life Technologies). Briefly, the RNA was precipitated with ammonium acetate, washed and resuspended in 14 µl TE (10 mM Tris-HCl pH 8, 1 mM EDTA). Subsequently, for each sample 1.1 µg RNA

were hybridized with the provided streptavidin beads and purified. To control for variations in RNA expression data, 1 µl of a 1:100 dilution of ERCC RNA Spike-In control (Thermo Fisher) was added to 1 µg RNA and libraries generated according to the TruSeq Stranded mRNA Library Kit-v2 (Illumina) with the index primers as indicated in **Table S5.** Quality control of the generated libraries was performed on a Bioanalyzer High Sensitivity DNA chip (Agilent) and the concentration was measured using Qubit dsDNA HS Assay (Life Technologies). To eliminate primer dimers in the libraries, additional purifications were performed using the Agencourt AMPure XP system (Beckman Coulter). The libraries were pooled to 2-10 nM total concentration and sequenced on an Illumina NextSeq 500, PE100, mid output. Libraries were generated independently for each family and family-pools multiplexed and sequenced on separate lanes. ERCC spike-ins included in the preparation were not used for normalization, but rather as a measure of quality control. Families 2, 3 and 6 showed the lowest variation in the ERCCs between family members, while family 4 and 5 had higher technical noise (**Figure S3**). Reads were filtered for QC and artifacts using the fastX toolbox, and then the reads were paired up using an adapted python script (https://github.com/enormandeau/Scripts/blob/master/fastqCombinePairedEnd.py). The reads were aligned to the genome (GRCh38, GENCODE v22 (Harrow et al., 2012)) using STAR (2.4.2a)(Dobin et al., 2012). The data has been deposited in GEO/SRA under accession number GSE84891.

Data collection and co-expression analysis

RNA-seq expression data was collected from Gemma (Zoubarev et al., 2012) for human subjects. From the collection, we selected 75 expression experiments (3,653 samples) that we could ascertain were of tissues and not cell lines. These are listed in **Table S1**. For each experiment, we consolidated our list of genes/transcripts to the 30K with Entrez gene identifiers, and did not limit either expression level or occurrence of expression. For each experiment with at least 10 samples, we generated a co-expression network using Spearman's correlation coefficient (Ballouz et al., 2015) and calculated the frequency that a pair of genes was negatively co-expressed (Spearman's correlation coefficient $r_s \leq 0$).

Network analysis using neighbor-voting

To measure the information content of the network, we use the performance of the *n*-fold cross validation task of a neighbor voting algorithm. If we can hide known information about genes in a gene set and then "learn" this information from the network, then our network has, to a degree, information that is reflective of the known biology of that GO term. This is based on the "guilt-by-association" principle, which states that genes with shared functions should be connected preferentially in the network. The reported performance metric from this task is the averaged AUROC (area under the ROC curve) for each group across the n-folds. We used the Bioconductor package EGAD (Ballouz et al., 2016) and GO to perform this analysis on the frequency of co-expression network. AUROC performance of 0.5 is random, with 1 representing perfect identification of positives and 0 representing perfect failure to represent positives (not just random). AUROC values above 0.7 are generally regarded as good and the significance of those performances is symmetric (around 0.5). Thus, performances below 0.3 indicate the network is unusual in the degree to which it is not linking genes by function. The AUROC can be converted into a p-value by virtue of its overlap with the Mann-Whitney test-statistic (after a standardization), but because of our data size, even modest deviations from 0.5 are extremely significant.

Differential expression analysis

We calculate a fold change between the parents and the probands for the differential expression analysis. We first calculate the CPM (counts per million) for each individual, and then take the average CPM for the parents and compare it the CPM of the proband. We take the log2 of the CPM (adding 1) of the ratio of these values. To exploit within family variance to detect noisy genes, we remove genes that showed strong differential expression between the parents themselves (i.e., top 100 up-regulated and top 100 down-regulated genes). After removing these highly variable genes, we take the top 100 up- and down-regulated genes based on ranking the fold change. Note that each family-specific analysis is

perfectly confounded with age, and strongly confounded with sex. However, because these factors are frequently present in other data, genes jointly affected by these conditions should co-vary across previous data and thus not generate artefactual outliers. We assess each family in a separate batch (library preparation and sequencing run). This holds likely technical variation constant in each family and independent across families, so that gene-level recurrence is not expected to differ from the null. By way of analogy, our experimental design resembles the analysis of *de novo* variants in DNA analyses, in which as many factors as possible are held constant in the control group for the proband (e.g., siblings, parents). For the case-control version of the analysis, we take all the parents as controls, and all the probands as cases, and perform the standard DE analysis. First we run Combat on all the samples, making it aware of the model and the batches. We then average the parents and the probands CPM expression levels, and calculate the fold change, and use both the one-sided and two-sided Wilcoxon-rank-sum tests to calculate a p-value, adjusted for multiple tests using Benjamani-Hochberg correction (FDR, p.adjust in R). For the DESeq2 version of the analysis, we used the counts data filtering away low expressing genes, and ran the default steps. All DE methods output the log2 fold-change and a *p*-value for each gene. We rank genes based on fold-change in the within family analysis, and *p*-value for the case-control analyses and all DESeq2 analyses.

The downsampling experiment involved taking combinations of the families and performing the case-control DE, such that once again the probands were the cases and the parents the controls. This was done for all 56 combinations of 2,3,4 and 5 probands across our 6 families. For each combination, we used the Combat corrected data and ranked genes based on the adjusted p-value of the Wilcoxon test (wilcox.test, p.adjust in R).

GO enrichment

To calculate GO term enrichment of the top differentially expressed genes, we used a simple gene set enrichment based on the hypergeometric test. For each gene set in GO, we calculate the significance of the overlap of the differentially expressed genes and that GO group, correcting for multiple tests with Benjamani-Hochberg (FDR, p.adjust in R).

Module detection and outlier analysis

To measure the joint differential activity of a set of genes, we extract the sub network of these genes from the frequency of co-expression network. Then, taking threshold on the median value, we use this binary network as distance matrix, and perform a hierarchical clustering of the genes. This clustering returns genes that are closer in distance, and we use this dendrogram to define modules within the data. We used the R dynamicTreeCut (Langfelder et al., 2008) package to select modules within the data of a minimum cluster size 2. We used these clusters or modules to define our co-expression blocks, and filtered away blocks greater than size 5, to keep “functional-outliers”. We calculate the significance of overlap using Fisher’s exact test (phyper in R). We calculate the significance of recurrence of the differentially expressed genes using the binomial test (pbinom in R), and then correcting for multiple tests using either Bonferroni (FWER) or Benjamini-Hochberg (FDR, p.adjust in R). Code and the network is available for download from our github repository (https://github.com/sarbal/redBlocks). For the downsampling version of the recurrence analysis, we calculate the significance of recurrence as above, but this time for a combination of the probands. The recurrence or adjusted p-values were used to rank the genes.

List of abbreviations

ROC: Receiver Operating Characteristic

AUROC: Area Under the ROC

DE: Differential Expression

FDR: False Discovery Rate (Benjamini-Hochberg correction)

FWER: Family-Wise Error Rate (Bonferroni correction)

GO: Gene ontology

CPM: Counts Per Million (relative RNA expression based on read counts)

Author Contributions

JG and GJL conceived the project. JG and SB designed experiments. SB performed computational experiments. MD and JC performed wet-lab experiments. SB and JG wrote the manuscript. GJL arranged for blood donation from which RNA was isolated. LF, CEK, MT, and SKT provided blood samples. JG, SB, MD, GJL, MC interpreted data and edited text. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the families for participating in the study. The authors would like to thank Sanja Rogic and Paul Pavlidis for their assistance with the Gemma RNA-seq data. The authors would also like to thank the following groups for their samples: Micheil Innes from the Department of Medical Genetics and Alberta Children's Hospital Research Institute and Rosemarie Smith from the Department of Pediatrics at the Barbara Bush Children's Hospital. This work was supported by a gift from T. and V. Stanley and a grant from the Collaborative Center for X-linked Dystonia Parkinsonism (CCXDP).

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Ballouz, S., Verleyen, W., and Gillis, J. (2015). Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics *31*, 2123-2130.

Ballouz, S., Weber, M., Pavlidis, P., and Gillis, J. (2016). EGAD: Extending guilt by association by degree (R package).

Bernier, R., Golzio, C., Xiong, B., Stessman, H.A., Coe, B.P., Penn, O., Witherspoon, K., Gerdts, J., Baker, C., Vulto-van Silfhout, A.T., *et al.* (2014). Disruptive CHD8 mutations define a subtype of autism early in development. Cell *158*, 263-276.

Boycott, K.M., Vanstone, M.R., Bulman, D.E., and MacKenzie, A.E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet *14*, 681-691.

Carbone, E. (2009). Calcium Channels–An Overview. In Encyclopedia of Neuroscience (Springer), pp. 545-550.

Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., *et al.* (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Res *45*, D369-D379.

Consortium, S.M.-I. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol *32*, 903-914.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics.

Forrest, C.B., Bartek, R.J., Rubinstein, Y., and Groft, S.C. (2011). The case for a global rare-diseases registry. The Lancet *377*, 1057-1059.

Fukai, R., Saitsu, H., Okamoto, N., Sakai, Y., Fattal-Valevski, A., Masaaki, S., Kitai, Y., Torio, M., Kojima-Ishii, K., Ihara, K., *et al.* (2016). De novo missense mutations in NALCN cause developmental and intellectual impairment with hypotonia. J Hum Genet *61*, 451-455.

Gaiteri, C., Ding, Y., French, B., Tseng, G.C., and Sibille, E. (2014). Beyond Modules & Hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes, brain, and behavior *13*, 13-24.

Gillis, J., and Pavlidis, P. (2012). "Guilt by association" is the exception rather than the rule in gene networks. PLoS Comput Biol *8*, e1002444.

Guan, D., Shao, J., Deng, Y., Wang, P., Zhao, Z., Liang, Y., Wang, J., and Yan, B. (2014). CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data. Bioinformatics *30*, 1190-1192.

Gulsuner, S., Walsh, T., Watts, A.C., Lee, M.K., Thornton, A.M., Casadei, S., Rippey, C., Shahin, H., Nimgaonkar, V.L., Go, R.C.P., *et al.* (2013). Spatial and Temporal Mapping of De novo Mutations in Schizophrenia To a Fetal Prefrontal Cortical Network. Cell *154*, 518-529.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., *et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res *22*, 1760-1774.

Huynen, M.A., Snel, B., Mering, C.v., and Bork, P. (2003). Function prediction and protein networks. Curr Opin Cell Biol *15*, 191-198.

Iossifov, I., Levy, D., Allen, J., Ye, K., Ronemus, M., Lee, Y.-h., Yamrom, B., and Wigler, M. (2015). Low load for disruptive mutations in autism genes and their biased transmission. Proc Natl Acad Sci U S A *112*, E5600-E5607.

Irizarry, R.A., Wang, C., Zhou, Y., and Speed, T.P. (2009). Gene Set Enrichment Analysis Made Simple. Statistical methods in medical research *18*, 565-575.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T.*, et al.* (2008). KEGG for linking genomes to life and the environment. Nucleic Acids Res *36*, D480-D484.

Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B.*, et al.* (2009). Predicting new molecular targets for known drugs. Nature *462*, 175-181.

Kitano, H. (2002). Systems Biology: A Brief Overview. Science *295*, 1662-1664.

Koscielny, G., Yaikhom, G., Iyer, V., Meehan, T.F., Morgan, H., Atienza-Herrero, J., Blake, A., Chen, C.-K., Easty, R., Di Fenza, A.*, et al.* (2014). The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic Acids Res *42*, D802-D809.

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics *24*, 719-720.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression Analysis of Human Genes Across Many Microarray Data Sets. Genome Res *14*, 1085-1094.

Lee, Tong I., and Young, Richard A. (2013). Transcriptional Regulation and Its Misregulation in Disease. Cell *152*, 1237-1251.

Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nat Rev Genet *16*, 321-332.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N.*, et al.* (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Lu, A.T.-H., Dai, X., Martinez-Agosto, J.A., and Cantor, R.M. (2012). Support for calcium channel gene defects in autism spectrum disorders. Mol Autism *3*, 1-9.

Marcotte, E.M., Pellegrini, M., Ng, H.-L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999). Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. Science *285*, 751-753.

O'Rawe, Jason A., Wu, Y., Dörfel, Max J., Rope, Alan F., Au, PY B., Parboosingh, Jillian S., Moon, S., Kousi, M., Kosma, K., Smith, Christopher S.*, et al.* (2015). TAF1 Variants Are Associated with Dysmorphic Features, Intellectual Disability, and Neurological Manifestations. Am J Hum Genet *97*, 922-932.

Robles, J.A., Qureshi, S.E., Stephen, S.J., Wilson, S.R., Burden, C.J., and Taylor, J.M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics *13*, 484.

Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421-427.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet *34*, 166-176.

Talley, E.M., Cribbs, L.L., Lee, J.H., Daud, A., Perez-Reyes, E., and Bayliss, D.A. (1999). Differential distribution of three members of a gene family encoding low voltage-activated (T-type) calcium channels. The Journal of neuroscience : the official journal of the Society for Neuroscience *19*, 1895-1911.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y.*, et al.* (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res *22*, 1798-1812.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H.*, et al.* (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582-587.

Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human symptoms–disease network. Nat Commun *5*, 4212.

Zoubarev, A., Hamer, K.M., Keshav, K.D., McCarthy, E.L., Santos, J.R.C., Van Rossum, T., McDonald, C., Hall, A., Wan, X., and Lim, R. (2012). Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. Bioinformatics *28*, 2272-2273.

Figures

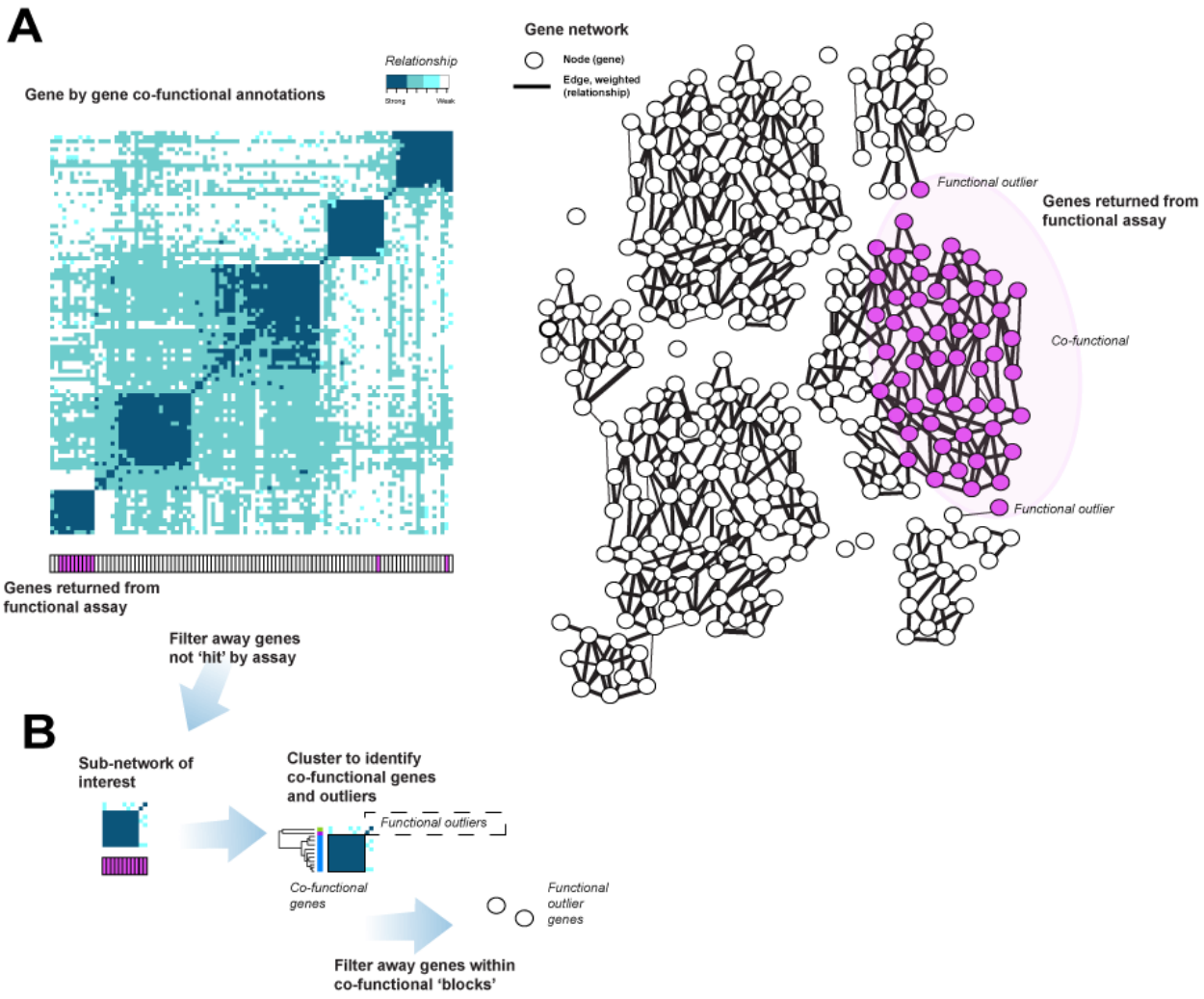

Figure 1 Schematic of functional outlier identification

(A) An adjacency matrix representing joint functionality defines a network, where genes are the nodes, and the strength of the relationship between genes defines the edges. This can be represented as a gene by gene matrix, where each entry is the edge weight, or visually as a graph. A functional assay, or a gene list, allows us to look at a subset of genes. In most cases, researchers look for functional enrichment within that set. (B) Here we take those genes and instead identify those grouped together within the network. This allows us to define genes that share common functions, and also pick out the genes which are functional outliers.

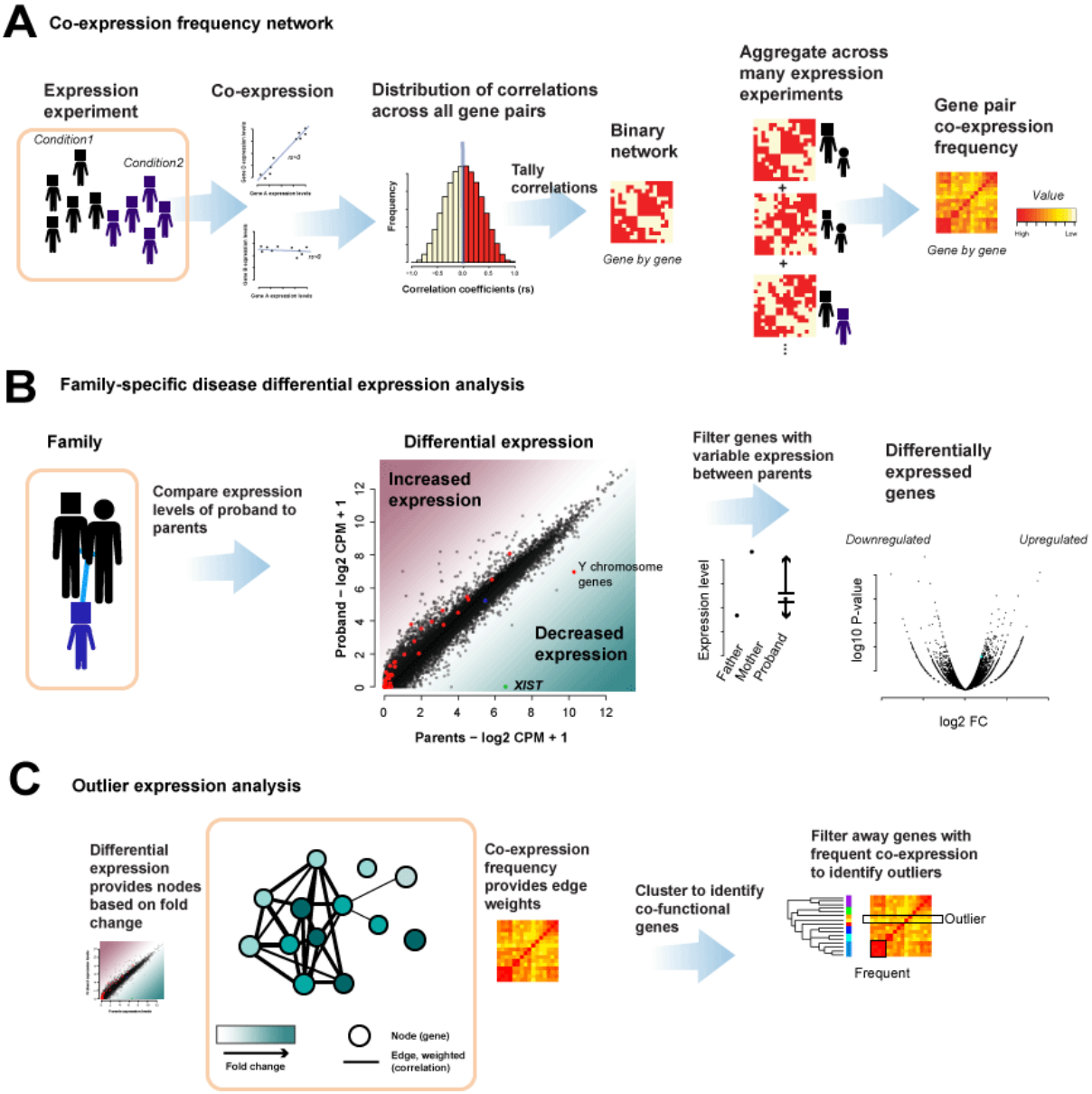

Figure 2 Differential expression and co-expression blocking within a family.

(A) Study design to find differential expression. We compare the average expression (counts per million - CPM) of the parents to the proband. Taking the genes with the highest fold changes, we look to see how often these genes are rarely seen as co-expressed. Genes that are often co-expressed show as “red blocks” in a co-expression module, while rare or outlier DE genes are not part of a module. A sample-sample plot showing the increased (up-regulated) and decreased (down-regulated) expression changes between the parents (x-axis) and the proband (y-axis). Genes with high differentials between the parents are filtered away (e.g., *XIST* highlighted in green) as to minimize the role of outliers present due to constitutively high variability. (B) Co-expression frequency analysis to calculate co-functional annotations. Taking each co-expression distribution, we threshold so that all Spearman’s correlation coefficient rs<0 are given a value of 1 (light yellow shading), and all others 0 (red shading). Repeating for all human RNA-seq experiments, we aggregate these individual networks into an occurrence network, such that for each gene pair we have the frequency these genes share any change in expression across samples. (C) Study design to determine functional (co-expression) outliers. This analysis uses the genes selected as differentially expressed and their co-expression frequency.

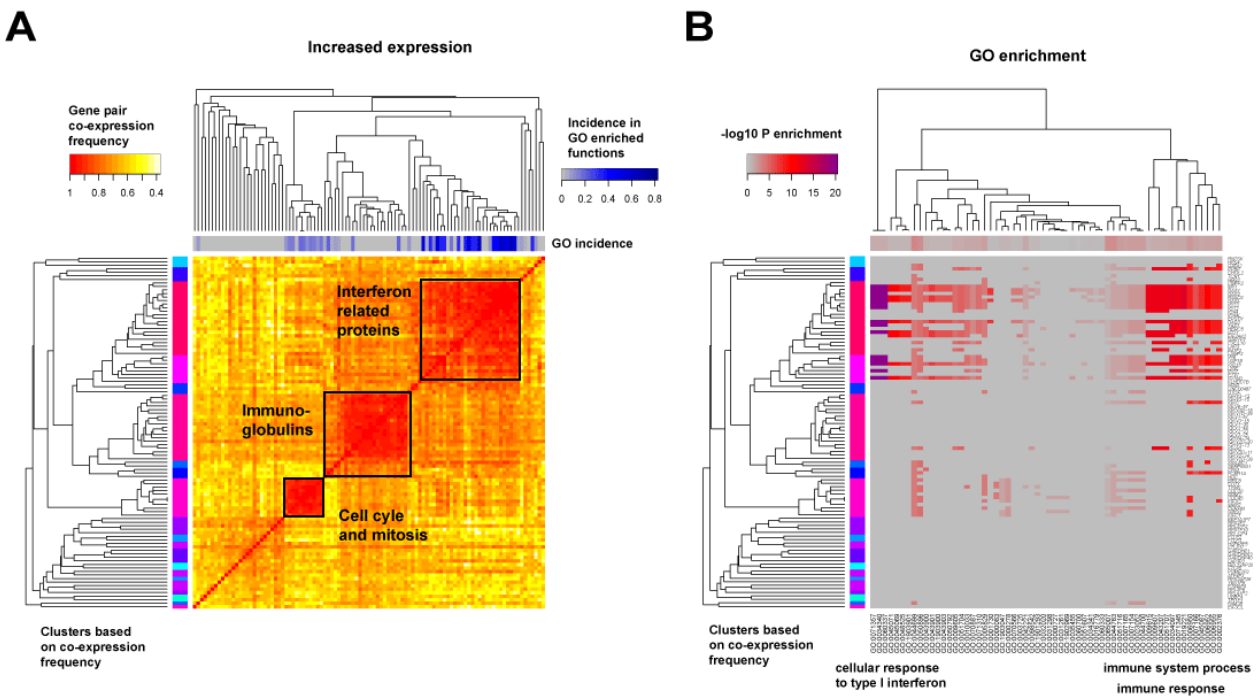

Figure 3 Co-expression blocks show gene set enrichment

(A) As an example, we show the co-expression frequency sub-network as a heatmap, where genes showing increased expression show co-expression. We see modules (co-expression blocks) as determined by the clustering (see rows). The modules are enriched for particular genes, such as interferon related proteins, immunoglobulins and cell-cycle genes. (B) Performing a gene set enrichment analysis on these genes (Fisher's exact test on GO groups), we see that the genes (rows) that generate the enrichment (columns are enriched GO terms) almost exclusively overlap with the co-expression blocks.

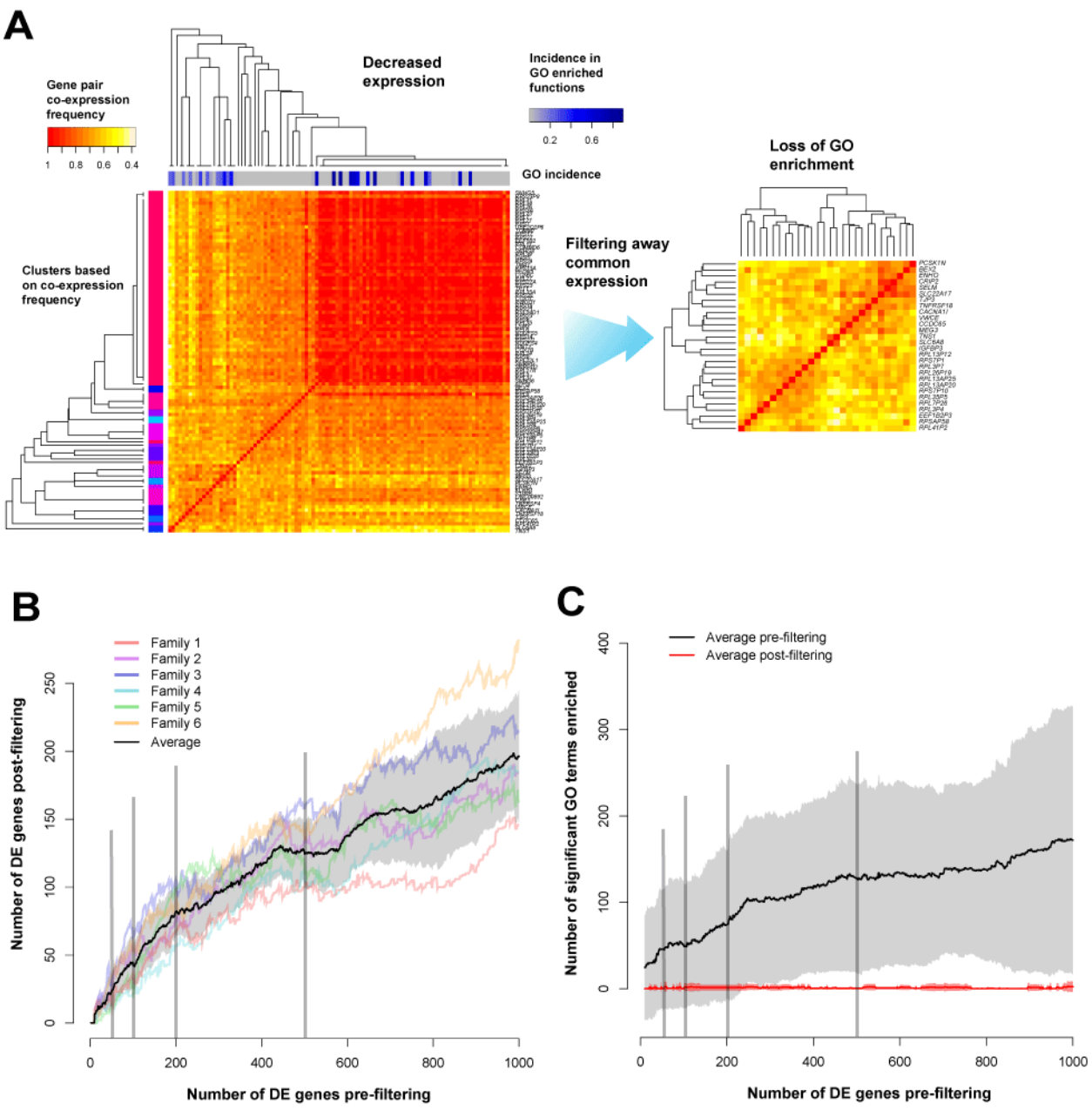

Figure 4 Filtering away common co-expression reveals rare differential expression.

(A) Given the set of DE genes for a family, removing "red blocks" leaves a smaller subset with less common co-expression. (B) Robustness analyses showing the number of genes pre and post-filtering for each family (colored lines) and averaged (black line, SD grey shadow). As we increased the number of genes considered for DE, the number of genes we filter also increases, averaging to approximately two-thirds removed. (C) If we look at the enrichment of these DE gene sets (pre-filtering in black +/-SD shadow), we see that filtering off the modules removes all but a few significant terms (red line, +/-SD shadow).

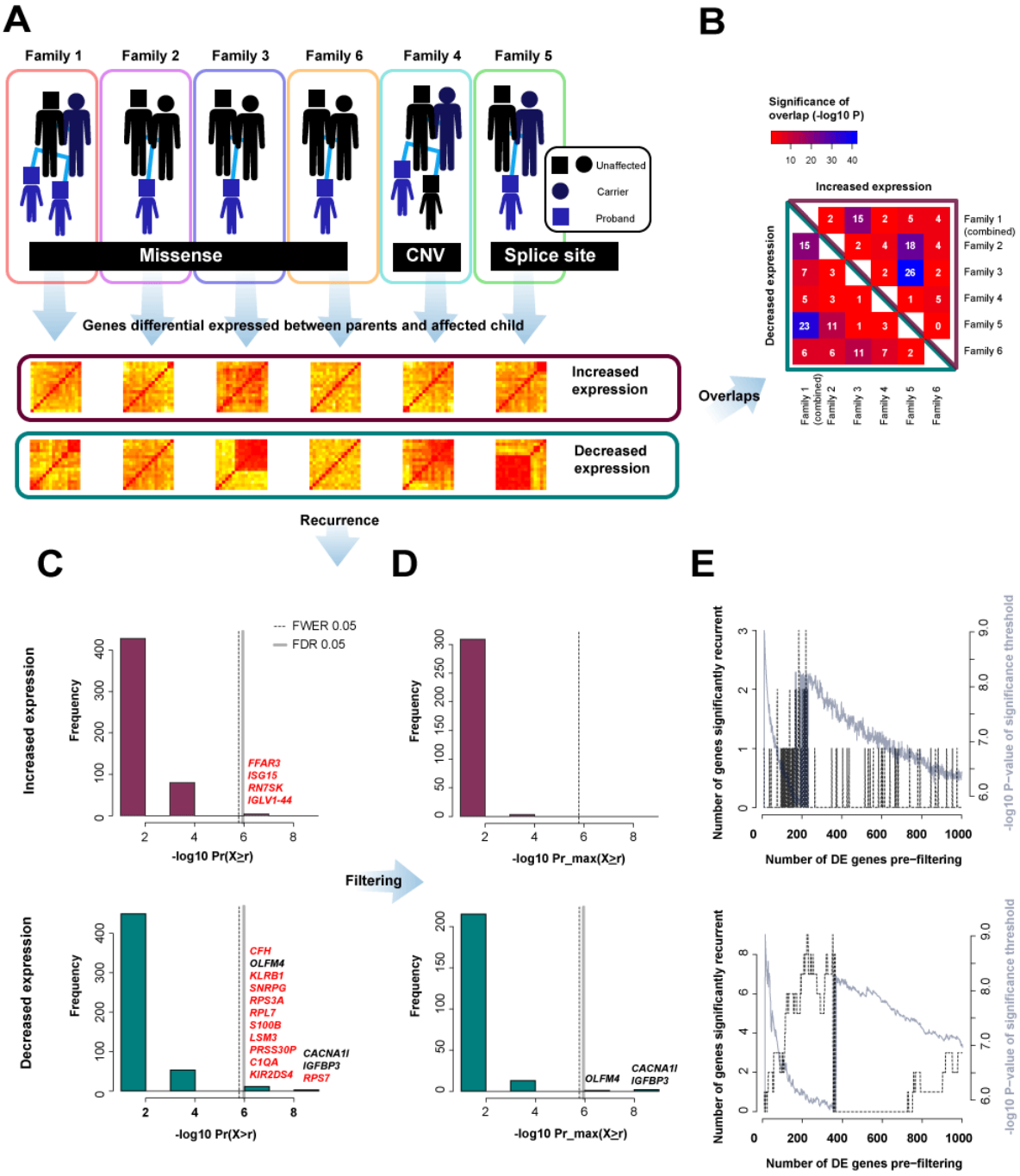

Figure 5 Recurrence of gene hits across families.

(A) Pedigrees of the families used in this study. For each, we calculate the expression fold change and pick out the top 100 up and down regulated genes (increased and decreased expression respectively).(B) First, we calculate the overlap between the individual families (numbers in boxes), and the significance of this overlap (colored corresponding to -log10 P-value of the hypergeometric test). Overlaps are mostly small and where significant, dominated by functions not clearly related to disease. (C) The more interesting genes are those that are recurrent across the families. The probability and significance of these genes recurring are shown in the plot (binomial test), pre-filtering by co-expression, for the 6 families, compared to the significance if we increase the number of families

tested. The FDR and FWER corrected P-values are similar. (D) For both increased and decreased expression, the number of significantly recurrent genes pre-filtering are few (listed). The genes in red are those that are found in the modules, and are "lost" as we filter. The genes remaining post filtering (outliers) are of specific interest. (E) Changing the threshold to characterize a gene as a "hit" modestly increases the number of significantly recurrent hits (black dotted line) while weakening the significance of those hits (grey, using the FWER). There are few genes positively differentially expressed in the proband (top), but multiple hits negatively differentially expressed (bottom).

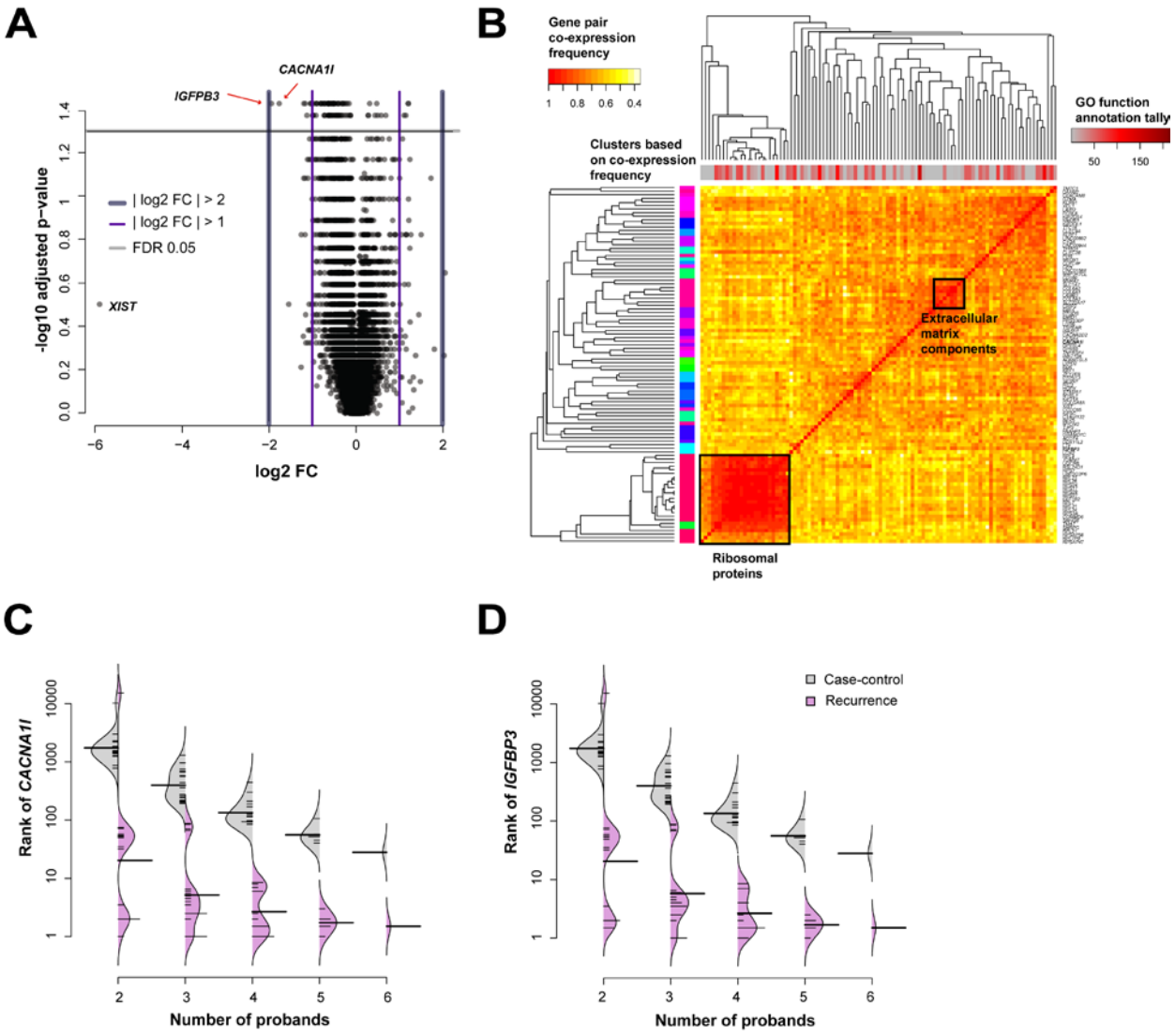

Figure 6 Case-control version of differential expression

(A) The log2 FC versus the adjusted P-value (-log10, FDR). (B) Visualizing the top ranked genes in the co-expression frequency network, still showing some blocks. (C) Rank of *CACNA1I* when performing the DE analysis using case-control (in grey) versus the recurrence analysis (in light purple) of the proband downsampled experiments . (D) Same as in (C) but for *IGFBP3*. The median value shows the recurrence analysis outperforms the case-control. Note log10 scale.

Tables

Table 1 Disease cohort used in the study

	Member	Sex	Mutation	Age (when sample collected)
Family 1	Proband 1	M	Missense	15
Family 1A	Proband 2	M	Missense	13
	Father	M		
	Mother	F	Missense[a]	
Family 2	Proband	M	Missense	21
Family 6A	Mother	F		
	Father	M		
Family 3	Proband	M	Missense	5
Family 2A	Father	M		
	Mother	F		
Family 4	Proband	M	CNV	19
Family 10A	Father	M		
	Mother	F	CNV[a]	
	Sibling	M		15
Family 5	Proband	M	Splice site	3
Family 5A	Father	M		
	Mother	F	Splice site[a]	
Family 6	Proband	M	Missense	9
Family 4A	Mother	F		
	Father	M		

[a] Carrier of mutation

Supplemental Information

Table S1. List of experiments used to generate co-expression networks, related to Figure 2.

Table S2. List of top 1000 differentially expressed genes for each family, related to Figures 3-5.

Table S3. GO enrichments for top 100 up and down regulated genes for each family, related to Figures 3-5.

Table S4. List of top 1000 differentially expressed genes for the case-control analysis related to Figure 6.

Table S5. Library indices for RNA-seq analysis of *TAF1* families, related to Table 1.

Figure S1. Decreased expression robustness analyses pre- and post-filtering, related to Figure 4.

Figure S2. Increased expression robustness analyses pre- and post-filtering, related to Figure 4.

Figure S3. ERCC spike in distributions, related to Table 1.

Figure S4. Visualizing family-specific DE of the top 100 down-regulated genes related to Figures 3-6.

Figure S5. Precision and recall for *CACNA1I* and *IGFBP3* with alternate methods and parameters related to Figure 6.