

1 Not by systems alone: replicability assessment of disease expression  
2 signals

3

4 Authors:

5 Sara Ballouz,<sup>1</sup>  
6 Max Dörfel,<sup>1,2+</sup>  
7 Megan Crow,<sup>1</sup>  
8 Jonathan Crain,<sup>1,3+</sup>  
9 Laurence Faivre,<sup>4,5</sup>  
10 Catherine E. Keegan,<sup>6</sup>  
11 Sophia Kitsiou-Tzeli,<sup>7</sup>  
12 Maria Tzetis,<sup>7</sup>  
13 Gholson J. Lyon, <sup>1,8+</sup>  
14 Jesse Gillis\*,<sup>1</sup>

15 Affiliations:

16 1 The Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY,  
17 11724, USA

18 2 Oxacell AG, Helene-Lange-Straße 12, 14469 Potsdam, Germany

19 3 Department of Chemistry, Stony Brook University, Stony Brook, NY, 11794, USA

20 4 GAD team, INSERM UMR 1231, Université Bourgogne Franche-Comté, Dijon, France

21 5 Centre de Référence Maladies Rares Anomalies du Développement et FHU TRANSLAD, CHU Dijon,  
22 Dijon, France;

23 6 Department of Pediatrics, Division of Genetics and Department of Human Genetics, University of  
24 Michigan, Ann Arbor, MI 48109, USA

25 7 Department of Medical Genetics, National Kapodistrian University of Athens, Athens, Greece;

26 8 Institute for Basic Research in Developmental Disabilities (*IBR*), Staten Island, New York, USA

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

29 + Current addresses

30 Contact Information:

31 JG: [jgillis@cshl.edu](mailto:jgillis@cshl.edu)

32 GJL: [gholson.j.lyon@opwdd.ny.gov](mailto:gholson.j.lyon@opwdd.ny.gov)

33 SB: [sballouz@cshl.edu](mailto:sballouz@cshl.edu)

34 MD: [doerfel@oxacell.com](mailto:doerfel@oxacell.com)

35 MC: [mcrow@cshl.edu](mailto:mcrow@cshl.edu)

1 JC: [jonathan.crain@stonybrook.edu](mailto:jonathan.crain@stonybrook.edu)

2 LF: [laurence.faivre@chu-dijon.fr](mailto:laurence.faivre@chu-dijon.fr)

3 CEK: [keeganc@med.umich.edu](mailto:keeganc@med.umich.edu)

4 MT: [mtzetis@med.uoa.gr](mailto:mtzetis@med.uoa.gr)

5 SKT: [ekanavak@cc.uoa.gr](mailto:ekanavak@cc.uoa.gr)

6

7 **Summary**

8 In characterizing a disease, it is common to search for dysfunctional genes by assaying the  
9 transcriptome. The resulting differentially expressed genes are typically assessed for shared features,  
10 such as functional annotation or co-expression. While useful, the reliability of these systems methods is  
11 hard to evaluate. To better understand shared disease signals, we assess their replicability by first  
12 looking at gene-level recurrence and then pathway-level recurrence along with co-expression signals  
13 across six pedigrees of a rare homogeneous X-linked disorder, *TAF1* syndrome. We find most  
14 differentially expressed genes are not recurrent between pedigrees, making functional enrichment  
15 largely distinct in each pedigree. However, we find two highly recurrent “functional outliers” (*CACNA1I*  
16 and *IGFBP3*), genes acting atypically with respect to co-expression and therefore absent from a systems-  
17 level assessment. We show this occurs in re-analysis of Huntington’s disease, Parkinson’s disease and  
18 schizophrenia. Our results suggest a significant role for genes easily missed in systems approaches.

19

20 **Keywords**

21 Systems biology analysis

22 Replicability

23 Gene set enrichment

24 Co-expression analysis

25 *TAF1* syndrome

26 Huntington’s disease

27 Parkinsons’s disease

28 Schizophrenia

29 Meta-analysis

30 Disease differential expression

1 **Introduction**

2 Gene dysfunction in disease is frequently studied through a systems biology framework (Kitano, 2002) in  
3 which genes are linked through their shared functional properties and pathways. To assay the networks  
4 underpinning a phenotype it is common to look to the transcriptome as a measure of the activity of the  
5 genes in the system (Cookson et al., 2009; Dermitzakis, 2008; Jirtle and Skinner, 2007; Lamb et al.,  
6 2006). Gene expression changes are detected through differential expression, and then systems-level  
7 signals are determined through gene set enrichment (Hosack et al., 2003; Subramanian et al., 2005),  
8 network connectivity assessments (Barabási et al., 2010; Greene et al., 2015; Lage et al., 2007), or gene  
9 property enrichment (Kircher et al., 2014; Lek et al., 2016). These methods all contextualize genes to  
10 known biological properties, functions and pathways. As this step aims to summarize the disease  
11 mechanism, it is typically the last step in an analysis, with little direct evaluation for efficacy (Nguyen et  
12 al., 2016; Pham et al., 2017; Yu et al., 2017). This reflects the challenge of developing a systematic  
13 framework for doing so, particularly where normal heterogeneity of phenotype and genotype may  
14 generate joint functional signals of their own. In this work, we seek to assess the replicability of disease  
15 and systems biology signals using a meta-analytic approach that exploits multiple pedigrees of a rare  
16 and homogeneous disorder, the *TAF1* syndrome.

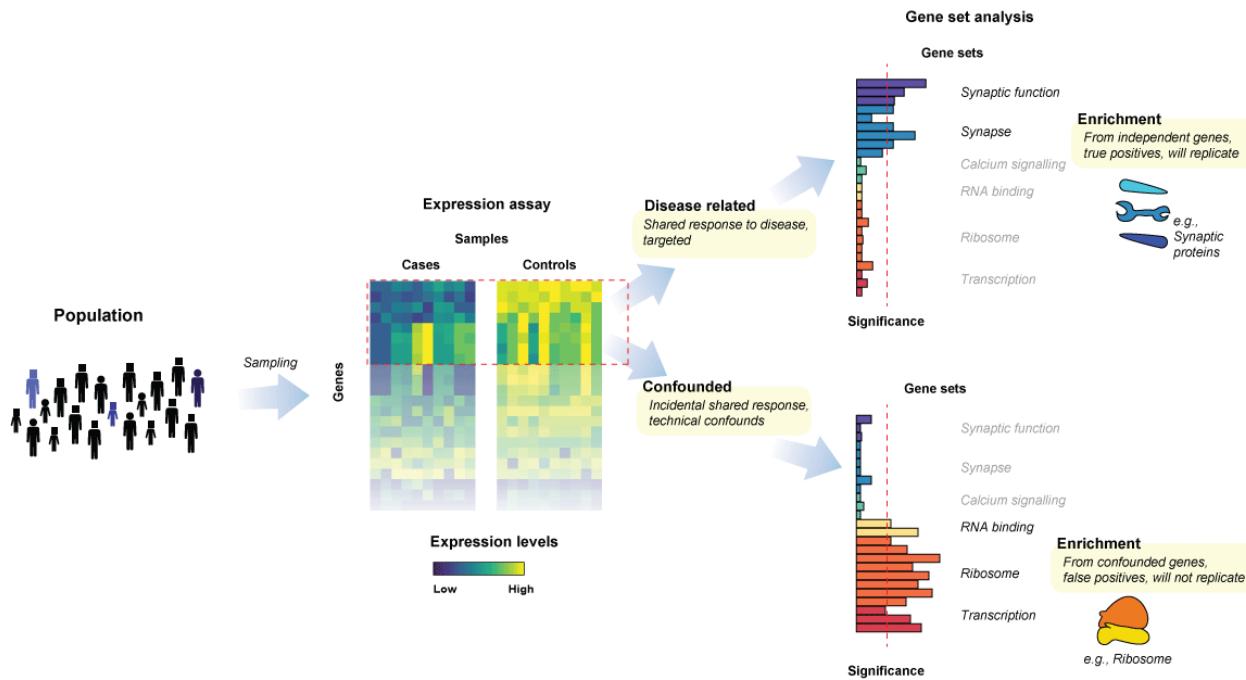
17 Genes may share signals for either biological or technical reasons. In gene or protein space, a “systems  
18 biology analysis” is used to define or assess these shared signals, with the assumption that the main  
19 driver behind the common features is biological (Draghici et al., 2007; Huang et al., 2008). A systems  
20 analysis can take the form of enrichment (e.g., Gene Ontology annotation overlaps) or more complex  
21 methods (e.g., *k*-nearest neighbors in co-expression networks), but ultimately outputs pathway-level  
22 summaries. Signals shared between genes due to technical properties such as sampling biases,  
23 confounded study designs, and batch effects can arise as false positives in a systems biology analysis and  
24 may be difficult to identify. False positive results can also easily arise due to unknown or uncontrolled  
25 biological variation, stemming from difficulties in phenotyping and the genetic heterogeneity of complex  
26 disorders (e.g., schizophrenia and autism (Purcell et al., 2014; Sanders et al., 2017)). There, variation in  
27 either phenotype or genotype may average away real disease effects and/or generate gene expression  
28 variation unrelated to disease (Hansen et al., 2011) (**Figure 1**). Replication is a central test of which of  
29 the two - either disease specific or untargeted variation - has driven the appearance of a characteristic  
30 signal across genes. In this work we assess the recurrence of candidate genes and pathways across  
31 separately analyzed data-sets. We characterize the degree to which replicate transcriptional signatures  
32 occur in groups of genes (e.g., co-expression) versus outliers (genes acting alone). A particular target of  
33 our analysis is the *TAF1* syndrome cohort, a rare and well-defined X-linked neurodevelopmental disorder  
34 with multiple pedigrees for assessment (see **Box 1** for further details).

35 We are able to look for replicability of disrupted gene expression signatures within the *TAF1* cohort for  
36 four main reasons: multiple pedigrees, phenotypic similarity, genetic homogeneity and a plausible  
37 mechanism for an impact on expression levels. We describe four classes of signals that can be extracted  
38 from disease analyses, reflecting whether signals are shared across families (recurrent/replicable) and  
39 across functional sets of genes (joint/disjoint signals). In *TAF1* syndrome, we find recurrence of gene  
40 expression change at the gene-level for two plausible candidates, *CACNA1I* and *IGFBP3*. At the systems  
41 biology level, we find little replicable enrichment, assessed via gene set enrichment of the Gene  
42 Ontology (GO) and via co-expression. Interestingly, the strongest replicable signal appears to arise from  
43 genes acting outside of the systems biology framework. We call genes meeting this property “functional

1 outliers". To see if our analysis is informative in other diseases and study designs, we also assess  
2 common heterogeneous neurodegenerative and neuropsychiatric disorders in over a thousand samples  
3 (~1.7k), including 5 Huntington's disease studies, 15 Parkinson's disease studies, and 10 schizophrenia  
4 studies. Once we extend the analysis to the more heterogeneous brain disorders, we are able to  
5 recapitulate known disease mechanisms, which appear as both recurrent joint signals (e.g., immune  
6 signaling pathways in schizophrenia), and recurrent functional outliers (e.g., SNCA in Parkinson's). The  
7 results from the four disorders studied here highlight the potentially important role of functional  
8 outliers, and suggest caution in applying gene-set based methods, such as enrichment or co-expression,  
9 in summarizing disease manifestation and mechanism.

10

11



12

13 **Figure 1 What does a systems biology approach tell us?** A systems biology assessment summarizes the properties  
14 of genes that are captured in an experiment, but can highlight both true and false positive results. In some cases,  
15 genes with a shared function are unlikely to arise in the experiment by chance. A gene set analysis will highlight  
16 their shared function (top panel), but not their independent value in identifying the enriched function. In other  
17 cases, a set of genes are so closely related both technically and biologically that if one arises, the others are almost  
18 certain to do so. A statistical analysis treating the genes as independent (bottom panel) will attach a misleading  
19 significance to the shared presence of the genes. These genes and gene sets will be unlikely to replicate in future  
20 studies.

21

22

23

24

25

1 **Box 1 *TAF1* syndrome cohort**

2 To probe in detail the functional gene signals that recur significantly due to disease, we focus a large part of our  
3 analysis on *TAF1* syndrome also known as “X-linked syndromic mental retardation-33” (MIM# 300966  
4 (O’Rawe et al., 2015)), an X-linked recessive neurodevelopmental disorder. *TAF1* syndrome is a rare, penetrant,  
5 and overall homogeneous disorder with no known disease mechanism. Genetically, it is defined by mutations in  
6 *TAF1* (TATA-Box Binding Protein Associated Factor 1), a key subunit of the general transcription factor TFIID  
7 (Louder et al., 2016; Müller et al., 2010). TFIID promotes transcriptional initiation by binding to the core promoters  
8 of genes, and recruits other transcription factor subunits that act as co-activators or co-repressors, encoding  
9 regulatory specificity (Pijnappel et al., 2013). Other subunits of TFIID are candidate genes in developmental and  
10 neurodegenerative diseases (Alazami et al., 2015; Bauer et al., 2004; El-Saafin et al., 2018; Hellman-Aharony et al.,  
11 2013), with reduced binding between the subunits playing a role in the pathogenesis. The characteristic  
12 phenotypic features of the *TAF1* disorder include global developmental delay, facial dysmorphology, generalized  
13 hypotonia, hearing impairments, microcephaly, and a characteristic gluteal crease with a sacral caudal remnant. All  
14 documented cases of the disorder affect males, and mostly have arisen *de novo*. In the few known inherited cases,  
15 female carriers do not show any features of the disease. This is generally a feature of X-linked disorders, with  
16 extreme X-skewed inactivation playing a role in phenotypic variation and protection in females (Migeon, 2007).

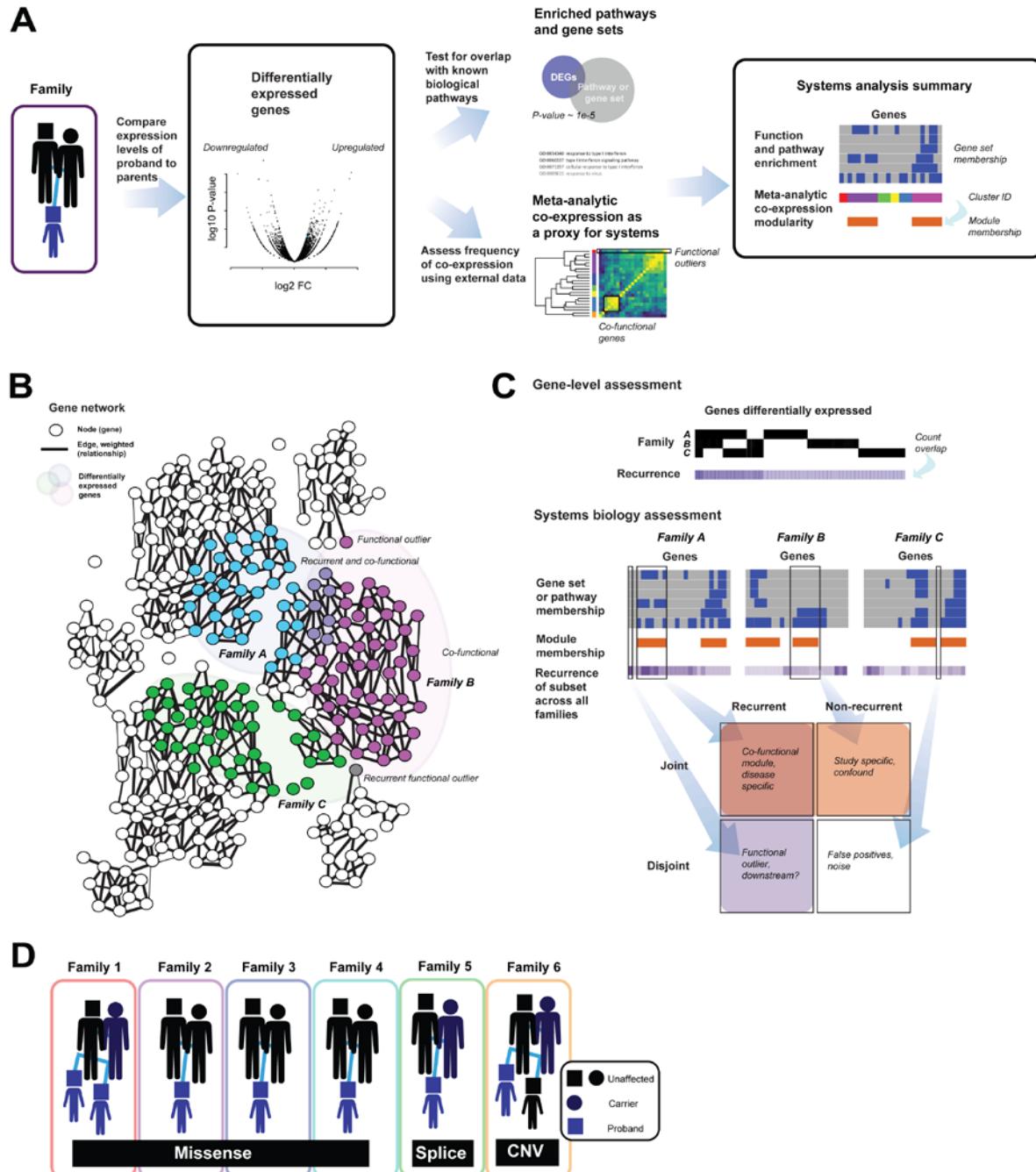
17 Despite being a relatively rare disorder, we have access to multiple pedigrees. In this study, six families were  
18 recruited from around the world, mainly of European descent and were between 5-21 years of age. All probands  
19 have point mutations in their *TAF1* transcription factor, except for a single CNV case. In three of the pedigrees, the  
20 mothers are carriers of the same mutation.

21 The four properties – global transcriptional impact, characteristic phenotype, genetic homogeneity, and multiple  
22 pedigrees - allow us to perform a disease replicability analysis using easily accessible blood transcriptional profiles.  
23 We can study each pedigree as a separate differential expression experiment: identifying differentially expressed  
24 genes and overrepresented pathways and then assessing these candidate genes and pathways for recurrence  
25 across pedigrees.

26 **Results**

27 **Replicability design overview**

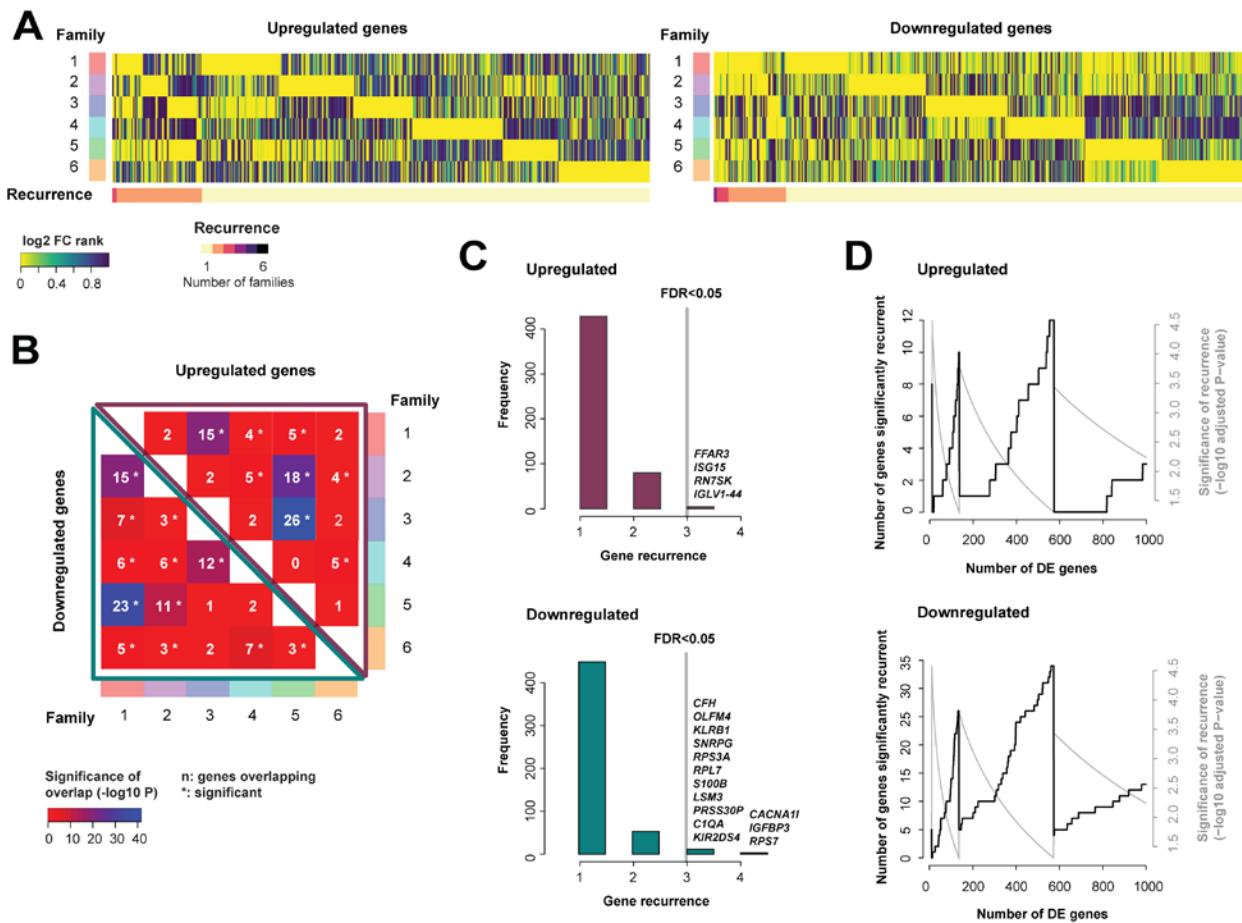
28 To understand shared disease signals, we assess replicability by testing the recurrence of candidate  
29 genes and pathways across separately analyzed differential expression experiments. We classify  
30 whether signals are replicated across families or datasets (and call this “recurrent”) and whether the  
31 signals involve sets of genes acting together (and call this “joint”). For the *TAF1* syndrome cohort  
32 analysis, we performed a family-based differential expression analysis, then tested for gene-level and  
33 pathway-level signals through gene set enrichment and a co-expression network modularity analysis  
34 (**Figure 2A**). We then looked across the pedigrees to evaluate where signals arise, summarized in **Figure**  
35 **2B**. There are four possibilities we can assess. 1) We may have a joint functional signal, where many  
36 genes are part of the same set/module, and it is this module which is replicated across families (i.e.,  
37 recurrent and co-functional). 2) We may have a recurrent disjoint signal: genes that replicate across  
38 families, but do not share common functions with other genes (i.e., recurrent functional outliers). 3)  
39 There is also the chance of a non-recurrent but joint signal: genes contribute to a shared signal, but  
40 uniquely within a family. 4) And finally, there is also an entirely disjoint signal, where we see one-off  
41 genes that are most likely false positives. We test all possible outcomes by measuring recurrence of  
42 genes, gene set enrichment and co-expression modularity (**Figure 2C**). We go through the results in the  
43 following sections.



1 Transcriptional replicability across the *TAF1* cohort occurs at the gene level  
2 Since *TAF1* is a transcription factor, we first wished to see if there was a common disease signature at  
3 the expression level. We identified differentially expressed genes (DEGs) using a family-based  
4 differential expression (DE) analysis (see **STAR Methods**). We saw only moderate overlaps of the  
5 differentially expressed genes between each of the pedigrees tested (**Figure 3A**) for both upregulated  
6 and downregulated genes. There were at most 26 (out of 100) genes in common between Family 3 and  
7 Family 5 (**Figure 3B**,  $p \sim 1e-40$ ). Even though few genes overlapped when assessed pairwise, there were a  
8 number of differentially expressed genes that were recurrently DE across families (significant if in at  
9 least three families  $FDR < 0.05$ ), with a modest number significantly recurrent. We find four genes  
10 recurrently upregulated (*ISG15*, *RN7SK*, *FFAR3* and *IGLV1-44*), and 14 downregulated (*C1QA*, *CFH*, *RPS7*,  
11 *SNRPG*, *LSM3*, *RPS3A*, *IGFBP3*, *RPL7*, *KLRB1*, *OLFM4*, *PRSS30P*, *KIR2DS4*, *S100B* and *CACNA1I* **Figure 3C**,  
12 see **Table S4**). These results suggest that even though a large fraction of the differentially expressed  
13 genes are unique to each pedigree, a recurrent gene transcriptional disease signature is present within  
14 the data. To test the dependence of these results on the DE threshold (top 100 genes), we repeated the  
15 recurrence using different DE thresholds (**Figure 3D**). As the change in the number of the DEGs called  
16 had an influence on the significance of recurrence, we see peaks and troughs of gene recurrence  
17 corresponding to the change in adjusted p-values. We found similar numbers of significantly recurrent  
18 genes within these ranges, with cut-offs between 50 and 200 genes most informative. We use the top  
19 100 DEGs for the remainder of the analyses.

20

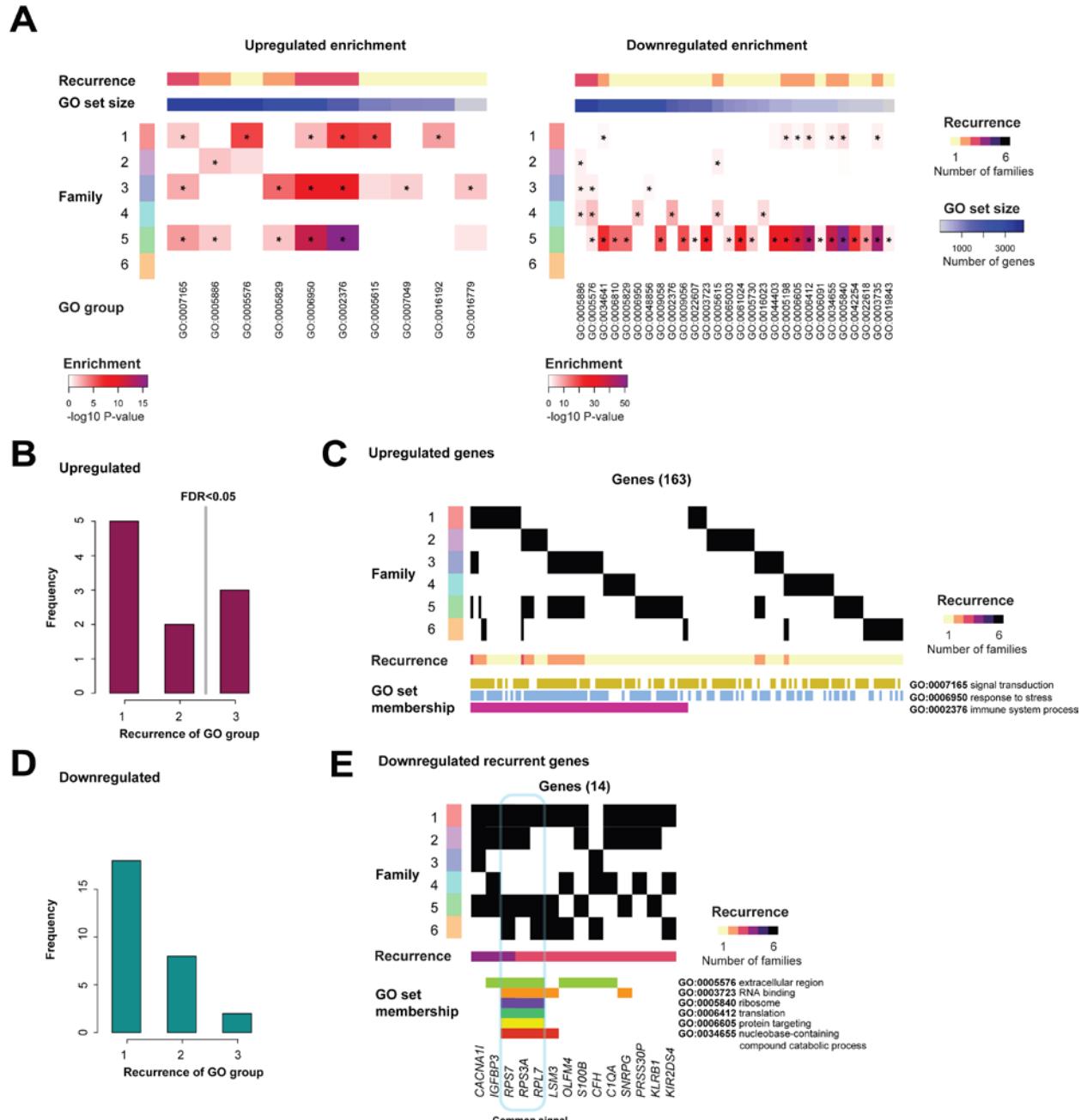
1



2

3 **Figure 3 Disease expression analysis with a family-based approach.** (A) The expression fold change for each gene  
4 is calculated within each family (top 100 up and down regulated genes are shown). (B) Overlaps in DE gene sets  
5 between the individual families (numbers in boxes), and the significance of this overlap (colored corresponding to -  
6 log10 P-value of the hypergeometric test). Overlaps are mostly small. (C) The replicable genes are those that are  
7 recurrent across families. The recurrence distributions for both up- and downregulated genes across the 6 families  
8 are shown. Using the binomial test, we find that genes recurring 3 or more times are significant (FDR<0.05). These  
9 genes are listed, with 4 up- and 14 downregulated genes significantly recurrent. (D) Robustness assessment of the  
10 DE threshold. The plot shows the number of recurrent genes as a function of the number of differentially  
11 expressed genes and the significance of the recurrence in grey.  
12

1 Gene set enrichment does not identify specific disease mechanisms  
2 While only a modest number of genes overlapped across the pedigrees, it is possible that non-recurrent  
3 genes still provide a shared disease signal, with variability in the exact genes identified due to technical  
4 limitations. To assess this possibility, we perform gene set enrichment in two ways. First, we test  
5 differentially expressed gene lists from each pedigree independently, and then measure the recurrence  
6 of the enriched pathways across the studies, as a parallel to the gene recurrence assessment. We then  
7 check for significantly recurrent pathways. Second, we test for enrichment of the recurrent genes  
8 themselves. We performed gene set enrichment using a subset of the Gene Ontology (GO slim) and  
9 found few pathways significantly enriched on a per family basis (FDR<0.05, **Figure 4A**), with the  
10 exception of Family 5, which had 22 downregulated pathways. Then, to assess the disease significance  
11 of the enrichment signal, we calculated the similarity and significance of recurrence of these pathways  
12 across the families. As in the case of gene recurrence, we expect the disease signal (here the enrichment  
13 signal) to replicate across the cohort if it is disease linked. Of the significantly enriched pathways, at  
14 most three were significantly upregulated and recurrent across the families (**Figure 4B** and **Table S5**).  
15 The pathways significantly recurrent were larger than average, implying broad properties (e.g.,  
16 GO:0002376 immune system process, GO:0006950 response to stress and GO:0007165 signal  
17 transduction, **Figure S3A**) with different genes contributing to the enrichment within each of the  
18 families (**Figure 4C**). While pathways were significantly recurrent across family-specific upregulated  
19 genes, the few recurrent upregulated genes do not show enrichment. In contrast, we found no  
20 downregulated pathways as significantly recurrent (**Figure 4D**) despite the gene recurrence and family-  
21 specific enrichment. However, the recurrent downregulated genes themselves (**Figure 4E**) are enriched  
22 for six GO groups related to ribosomal functions, including RNA binding, translation, and protein  
23 targeting. The discrepancy between the pathways enriched with the recurrent genes and the pathways  
24 significantly recurrent suggests that the weak enrichment signals are not specific to the disorder.

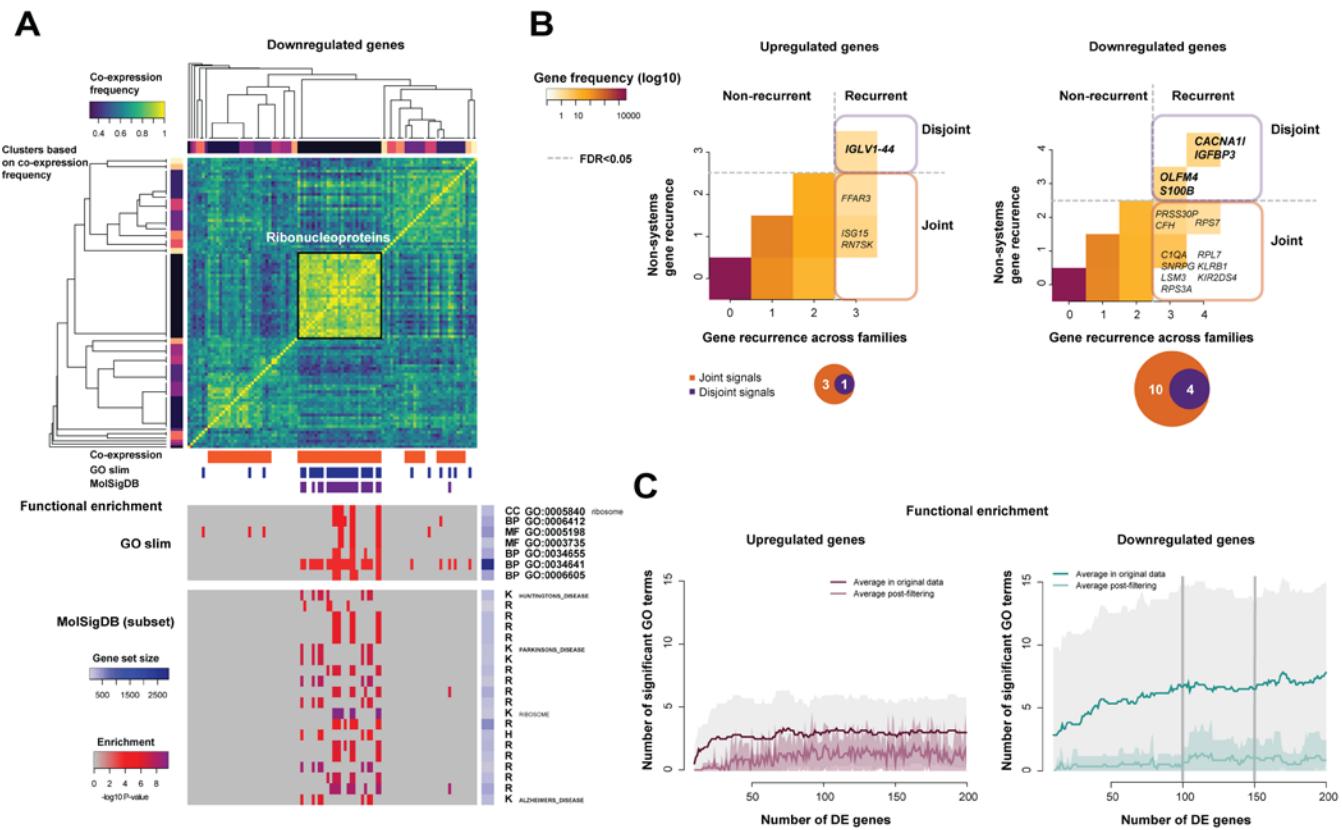


1  
2 **Figure 4 Gene set enrichment assessment of TAF1 cohort.** (A) Top GO enrichment results for each family for up-  
3 and downregulated genes. Significant terms (FDR < 0.05) are highlighted with an \*. (B) The frequency and  
4 significance of recurrence of each GO term is plotted for the upregulated genes. (C) Gene-GO membership matrix  
5 for the upregulated genes. Each column is a gene, and each row a family. The colored bars below highlight the GO  
6 terms that these genes belong to. The signal associated with the recurrent GO terms is distributed across different  
7 genes, shown by low overlap across the families. (D) There are no significantly recurrent pathways with the  
8 downregulated genes. (E) However, the recurrent genes themselves are enriched for ribosomal pathways (p-  
9 adjusted<0.05), as shown in the gene-GO membership matrix. The three RP\* genes seem to drive almost all the  
10 signal.

1 Top recurrent genes do not appear to act within a systems biology framework  
2 As we are limited by gene set annotations, the weak enrichment may be due to our lack of complete  
3 pathway knowledge (Thomas, 2017). Therefore, to test for joint signal with a different but  
4 comprehensive data modality, we looked to co-expression. Genes that are co-expressed are known to  
5 share functions, are co-regulated, or are parts of known pathways (Gaiteri et al., 2014). Unlike most  
6 curated or inferred gene annotations, co-expression can be assessed genome-wide. We used the gene  
7 pair co-expression frequency from a wide corpus of data (external to this study) as our co-functionality  
8 measure (see **STAR Methods**). For each family's set of differentially expressed genes, we found gene co-  
9 expression blocks comprising more than two thirds of these genes (**Figure S4A**). We show an example of  
10 this for the top 100 down-regulated genes from Family 1 in **Figure 5A**.  
11 We then asked where the recurrent genes sit in relation to the co-expression modules. Interestingly, we  
12 found the outlier or disjoint genes (those not in the large modules) were frequently among the most  
13 recurrent genes across the families (**Figure 5B**). Almost a third of the disease signature was not within  
14 common co-expression or functions, but rather appeared within very small modules or as outliers. For  
15 functional characterization to usefully summarize the DE list, candidate genes should be enriched within  
16 pathways. This was not true of *TAF1* disorder candidate genes. In particular, we find that the top  
17 downregulated genes *CACNA1I* and *IGFBP3* are not within modules. Once we perform a gene set  
18 enrichment analysis on the genes excluding those in modules, nearly all of the enrichment signal is lost,  
19 highlighting that the most recurrent candidates act outside of a common joint signal. If this is similarly  
20 true of the disease mechanism, then a systems-style analysis will fail to discover the disease signal.

21

1



2

3 **Figure 5 Co-expression of differentially expressed genes generates enrichment.** (A) As an example from family 1, 4 we show the co-expression frequency sub-network as a heatmap, where genes showing decreased expression 5 show co-expression. Co-expression blocks define modules as determined by the clustering (see rows). The modules 6 are enriched for particular genes, mainly ribonucleoproteins. Performing a gene set enrichment analysis on these 7 genes (Fisher's exact test on GO groups), genes (rows) that generate the enrichment (columns are enriched GO 8 terms) almost exclusively overlap with the co-expression blocks. The prominent pathways are ribosome related. 9 (B) The significantly recurrent genes can be divided into those present within co-expression modules (joint) and 10 those not (disjoint). The genes in bold are the functional outliers and the venn diagrams summarizes the number of 11 genes in each category. (C) If we look at the enrichment of these DE gene sets (pre-filtering dark line +/- SD 12 shadow), we see that filtering off the modules removes all but a few significant terms (lighter line, +/- SD shadow). 13

1 Recurrence of genes and pathways in other disorders

2 The *TAF1* cohort is interesting in part for being a rare disorder with a penetrant and distinct phenotype.  
3 In order to assess the role of functional outliers more broadly, we looked to three other disorders with  
4 substantial transcriptomic data and varying degrees of genetic heterogeneity. We focused on  
5 Huntington's disease (HD), Parkinson's disease (PD), and schizophrenia (SCZ). All studies used are listed  
6 in **Table S3**.

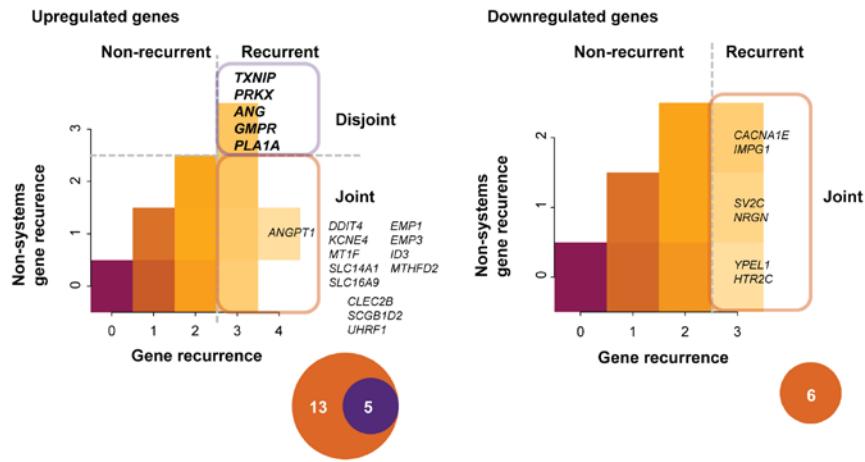
7 Huntington's disease is an inherited neurodegenerative disorder, characterized by the progressive  
8 degeneration of cells in the brain (primarily the striatum) and is associated with impaired movements,  
9 decline in cognitive abilities and depression (Bates et al., 2015). Similar to *TAF1* syndrome, HD is a  
10 monogenic disorder. The disease is caused by an expansion of the CAG repeats in the Huntington gene  
11 (*HTT*), which is believed to be toxic to other proteins once mutated. The exact functions of *HTT* are still  
12 unclear, along with the mechanisms (Bates, 2005). To test the possibility of a joint functional signal, we  
13 assessed five expression studies of HD, following the evaluative approach we took with the *TAF1* cohort.  
14 Interestingly, we found both joint and disjoint signals (**Figure 6A**). A total of 18 upregulated genes were  
15 significantly recurrent, with *ANGPT1* (Angiopoietin 1) recurrent in four of the five studies. This gene  
16 plays an important role in vascular development and angiogenesis. Along with other recurrent genes,  
17 such as *ANG* (angiogenin), *KCNE4* and *SLC14A1*, there seems to be an associated cardiovascular  
18 phenotype. Heart disease is comorbid with Huntington's, and these gene candidates suggest a link to  
19 cardiac development. Additionally, we found six genes recurrently downregulated in at least three of the  
20 five studies; these include synaptic genes (*SV2C*, *NRGN*, and *HTR2C*) and a calcium channel related to  
21 modulation of firing in neurons (*CACNA1E*). We found that all the recurrent downregulated genes were  
22 part of modules, thus showing a strong joint functional signal. And in the upregulated genes, *ANG* is a  
23 functional outlier, while all of the other recurrent genes were in shared functions. These results suggest  
24 a stronger joint functional signal than in the *TAF1* disorder. All recurrent genes are listed in **Table S6**.

25 Parkinson's disease is a progressive neurodegenerative disorder, characterized by the loss of  
26 dopaminergic neurons leading to decreased motor function. Unlike the *TAF1* cohort, Parkinson's has  
27 multiple genes implicated, each with different onset stages (e.g., *LRRK*, *SNCA*, *PRKN*, *FBXO7*, *PARK7*, and  
28 *PINK1*) (Poewe et al., 2017), increasing the genetic heterogeneity of the disorder and data. We collected  
29 15 differential expression gene lists and repeated our analysis. We find a subset of significantly recurrent  
30 downregulated genes but no upregulated genes as significant (**Figure 6B**). The most recurrently  
31 downregulated gene is *SNCA* (alpha-synuclein), a well-known Parkinson's disease gene (Siddiqui et al.,  
32 2016) which recurs six times (FDR < 0.05). We did not purposely select for studies with variants in this  
33 gene when selecting the studies for the meta-analysis, but could confirm it was the genetic cause in a  
34 few of the studies. Another top candidate was *SYT1* (synaptotagmin 1) which recurred in five of the 15  
35 studies (FDR < 0.05). Synaptotagmins are known to be involved in neurodegeneration (Glavan et al.,  
36 2009), and *SYT11* (Sesar et al., 2016) has been associated with the disorder through its interactions with  
37 *PARK7*. Other genes that are significantly recurrent include *SLC18A2* (linked to PD (Lohr and Miller,  
38 2014) ), *UCHL1* (mutations may be associated with PD (Healy et al., 2006; Maraganore et al.,  
39 2004)), *DLK1* and *SLC10A4*. A majority of the recurrent genes were present in co-expressed modules,  
40 but the joint signal was less predominate than in Huntington's disease. All recurrent genes are listed in  
41 **Table S7** and enriched pathways in **Table S8**.

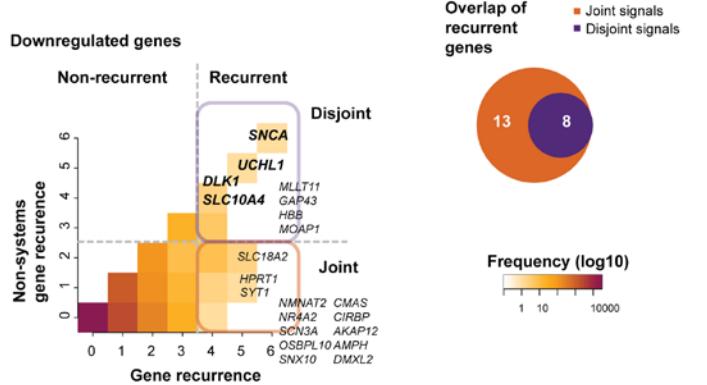
42 Our final use case was on schizophrenia, a neuropsychiatric disorder characterized by abnormal social  
43 behavior and psychosis, along with other cognitive impairments (Kahn et al., 2015). The disorder has  
44 strong environmental and genetic components (Gejman et al., 2010; Rees et al., 2015), with many genes  
45 increasing risk. The risk alleles are also shared amongst many other neuropsychiatric phenotypes,

1 making this a hard disorder to classify both genetically and phenotypically, and thus difficult to  
2 characterize molecularly. We assessed 10 disease expression studies and found both up- and  
3 downregulated transcriptional signatures (**Figure 6C**), but these recurred in three or four studies at  
4 most. Consistent with the genetic and phenotypic heterogeneity of the disorder, we find that recurrent  
5 genes have both joint and disjoint signals. Of the genes upregulated, we find recurrence of the *FCN3*  
6 gene (Ficolin 3), a recognition molecule in the lectin pathway of the complement system (Garred et al.,  
7 2009; Mayilyan, 2012). This is of interest in schizophrenia as it potentially interacts with C4, a known risk  
8 allele (Sekar et al., 2016). The remaining recurrent genes were enriched for an inflammatory signature,  
9 also recapitulating known schizophrenia etiology. Among recurrent downregulated genes is the protein  
10 phosphatase inhibitor *PPP1R17* which is primarily expressed in Purkinje cells in the cortex of the  
11 cerebellum, a brain region which may play a role in the disorder (Maloku et al., 2010). Interestingly, both  
12 *FCN3* and *PPP1R17* were functional outliers, with the other recurrent genes showing joint functional  
13 signals. All recurrent genes are listed in **Table S9** and enriched pathways in **Table S10**. Overall, across all  
14 the disorders, the joint signals via co-expression were much stronger than in the *TAF1* syndrome cohort,  
15 but there were important functional outlier candidates within each disease.

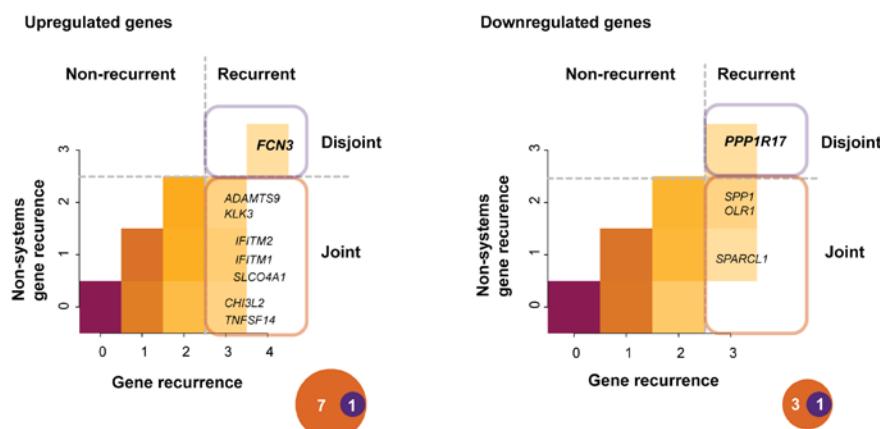
### A Huntington's disease



### B Parkinson's disease



### C Schizophrenia



1  
2 **Figure 6 Differential expression meta-analysis in three other disorders.** (A) Recurrence of genes in Huntington's  
3 disease (HD), (B) Parkinson's disease (PD) and (C) schizophrenia (SCZ), and whether they occur in groups (joint) or  
4 not (disjoint). The Venn diagrams summarize the number of recurrent genes and their joint or disjoint designation.

1      **Discussion**

2      The main contribution of this work is a rigorous analysis of replicability of functional signals implicated in  
3      disease through expression analysis. We describe four classes of signals that can be extracted from  
4      disease analyses, reflecting whether signals are shared across families or studies (recurrent) and/or  
5      across functional sets of genes (joint). These are: 1) recurrent joint signals, 2) recurrent but disjoint, 3)  
6      joint but not recurrent 4) not recurrent and disjoint. Our evaluation of the rare *TAF1* syndrome and  
7      three common disorders highlighted an important feature of disease that has been overlooked in the  
8      systems paradigm: the role of functional outliers, classified as recurrent but with no joint or shared  
9      function.

10     In the *TAF1* cohort, we believe that the recurrent disjoint genes are the disease signal for reasons within  
11     and outside the present analysis. Of the candidates we found, *CACNA1I* and *IGFBP3* had the strongest  
12     recurrent signal. *CACNA1I* is a calcium channel subunit, and mutations in calcium channels are known to  
13     have similar phenotypes to the cohort here, including intellectual disability, autism and dystonia (Fukai  
14     et al., 2016; Lu et al., 2012). In addition, there is a *TAF1* binding site upstream of the gene (Wang et al.,  
15     2012). Most convincing is that this gene has been implicated recurrently in other brain disorders; it is  
16     the only recurrent missense *de novo* in schizophrenia studies (Gulsuner et al., 2013), one of the few  
17     overlapping candidates between schizophrenia and autism. The other recurrent candidate was the  
18     insulin-like growth factor-binding protein 3 (*IGFBP3*). Downregulation of this gene has been recently  
19     implicated in a developmental disorder with a behavioral and cognitive phenotype (Perez et al., 2018).  
20     Despite these fairly convincing properties, it remains to be seen whether the candidate genes are  
21     specific to the *TAF1* cohort, or rather might reflect a more general signature of developmental disorders.

22     Consistent with the view that functional outliers and aberrantly expressed genes may play a particularly  
23     strong role in diseases with rare genotypic variation, we do not find them as the strongest signal in the  
24     other disorders. In our meta-analysis, we focused on disorders with varying genetic architecture and  
25     those also well powered for assessment. In each case, the meta-analysis was powered to call individual  
26     genes as recurrent, identifying both known and novel candidate disease genes. This was not dramatically  
27     more than those found assessing across individual families within the *TAF1* cohort, most likely due to  
28     the heterogeneity of the other disorders or their study designs. Within each disorder, we observed clear  
29     joint functional signals present through the co-expression analysis, but to varying degrees (e.g., strong  
30     immune signals in schizophrenia). Interestingly, despite the greater role of joint functional signals within  
31     these disorders, plausible functional outliers exist for each, most notably in the case of Parkinson's  
32     disease where known disease genes, such as *SNCA*, appear to be acting outside of their typical behavior  
33     as evaluated from co-expression.

34     Characterizing whether or not genes exhibit expected shared behavior bears strongly on the subject of  
35     disease mechanisms. In the transcriptomic analysis of rare disorders, a joint disruption is almost always  
36     assumed for disease. Yet, there is potential for unbuffered and uncharacteristic expression changes in a  
37     few single genes, particularly when the assumed pathogenic variant is regulatory or the disorder is  
38     monogenic (Cummings et al., 2017; Fresard et al., 2018; Kremer et al., 2017). In this case, genes that are  
39     no longer under regulatory control, or those that have gained regulation, will act out of place (Zeng et  
40     al., 2015; Zhao et al., 2016). These genes could be far downstream in a pathway or cascade, and thus not  
41     impact other pathway members directly or immediately. In general, enrichment and other systems  
42     biology analyses will miss these single genes that serve as unique bottlenecks. Our results suggest that

1 filtering based on enrichment will systematically remove interesting candidates. This does not mean that  
2 systems-style analyses should not be conducted, rather we suggest that they be treated as a tool to  
3 classify which genes are operating within a group and which are not. With time, strong candidates in  
4 either category may be identifiable and provide valuable and distinct information about disease  
5 mechanisms.

6

7 **Acknowledgments**

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

15

16 **Author Contributions**

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

21 All authors read and approved the final manuscript

22

23 **Declarations of Interests**

24 The authors declare no competing interests.

25

1      **References**

2      Alazami, Anas M., Patel, N., Shamseldin, Hanan E., Anazi, S., Al-Dosari, Mohammed S., Alzahrani, F.,  
3      Hijazi, H., Alshammari, M., Aldahmesh, Mohammed A., Salih, Mustafa A., *et al.* (2015). Accelerating  
4      Novel Candidate Gene Discovery in Neurogenetic Disorders via Whole-Exome Sequencing of  
5      Prescreened Multiplex Consanguineous Families. *Cell Reports* **10**, 148-161.  
6      Ballouz, S., Verleyen, W., and Gillis, J. (2015). Guidance for RNA-seq co-expression network construction  
7      and analysis: safety in numbers. *Bioinformatics* **31**, 2123-2130.  
8      Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2010). Network medicine: a network-based approach to  
9      human disease. *Nature Reviews Genetics* **12**, 56.  
10     Bates, G.P. (2005). The molecular genetics of Huntington disease — a history. *Nature Reviews Genetics*  
11     **6**, 766.  
12     Bates, G.P., Dorsey, R., Gusella, J.F., Hayden, M.R., Kay, C., Leavitt, B.R., Nance, M., Ross, C.A., Scahill,  
13     R.I., Wetzel, R., *et al.* (2015). Huntington disease. *Nature Reviews Disease Primers* **1**, 15005.  
14     Bauer, P., Laccone, F., Rolfs, A., Wüllner, U., Bösch, S., Peters, H., Liebscher, S., Scheible, M., Epplen, J.T.,  
15     Weber, B.H.F., *et al.* (2004). Trinucleotide repeat expansion in SCA17/TBP in white patients with  
16     Huntington's disease-like phenotype. *Journal of medical genetics* **41**, 230-232.  
17     Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease  
18     traits with global gene expression. *Nature reviews Genetics* **10**, 184-194.  
19     Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell,  
20     L.B., Sandaradura, S.A., O'Grady, G.L., *et al.* (2017). Improving genetic diagnosis in Mendelian disease  
21     with transcriptome sequencing. *Science translational medicine* **9**, eaal5209.  
22     Dermitzakis, E.T. (2008). From gene expression to disease risk. *Nature genetics* **40**, 492-493.  
23     Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and  
24     Gingeras, T.R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*.  
25     Draghici, S., Khatri, P., Tarca, A.L., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R. (2007).  
26     A systems biology approach for pathway level analysis. *Genome Res* **17**, 000.  
27     El-Saafin, F., Curry, C., Ye, T., Garnier, J.-M., Kolb-Cheynel, I., Stierle, M., Downer, N.L., Dixon, M.P.,  
28     Negroni, L., Berger, I., *et al.* (2018). Homozygous TAF8 mutation in a patient with intellectual disability  
29     results in undetectable TAF8 protein, but preserved RNA polymerase II transcription. *Human Molecular  
30     Genetics* **27**, 2171-2186.  
31     Fresard, L., Smail, C., Smith, K.S., Ferraro, N.M., Teran, N.A., Kernohan, K.D., Bonner, D., Li, X., Marwaha,  
32     S., Zappala, Z., *et al.* (2018). Identification of rare-disease genes in diverse undiagnosed cases using  
33     whole blood transcriptome sequencing and large control cohorts. *bioRxiv*.  
34     Fukai, R., Saitsu, H., Okamoto, N., Sakai, Y., Fattal-Valevski, A., Masaaki, S., Kitai, Y., Torio, M., Kojima-  
35     Ishii, K., Ihara, K., *et al.* (2016). De novo missense mutations in NALCN cause developmental and  
36     intellectual impairment with hypotonia. *J Hum Genet* **61**, 451-455.  
37     Gaiteri, C., Ding, Y., French, B., Tseng, G.C., and Sibille, E. (2014). Beyond modules and hubs: the  
38     potential of gene coexpression networks for investigating molecular mechanisms of complex brain  
39     disorders. *Genes, brain, and behavior* **13**, 13-24.  
40     Garred, P., Honoré, C., Ma, Y.J., Munthe-Fog, L., and Hummelshøj, T. (2009). MBL2, FCN1, FCN2 and  
41     FCN3—The genes behind the initiation of the lectin pathway of complement. *Molecular Immunology* **46**,  
42     2737-2744.  
43     Gejman, P.V., Sanders, A.R., and Duan, J. (2010). The Role of Genetics in the Etiology of Schizophrenia.  
44     *The Psychiatric clinics of North America* **33**, 35-66.  
45     Genevie, B., William, H., Soraya, S., Helena, K., and Soraya, B. (2018). A review of genome-wide  
46     transcriptomics studies in Parkinson's disease. *European Journal of Neuroscience* **47**, 1-16.

1 Glavan, G., Schliebs, R., and Živin, M. (2009). Synaptotagmins in Neurodegeneration. *The Anatomical Record* 292, 1849-1862.

2 Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., *et al.* (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* 47, 569.

3 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.

4 Hansen, K.D., Wu, Z., Irizarry, R.A., and Leek, J.T. (2011). Sequencing technology does not eliminate biological variability. *Nature Biotechnology* 29, 572.

5 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.

6 Healy, D.G., Abou-Sleiman, P.M., Casas, J.P., Ahmadi, K.R., Lynch, T., Gandhi, S., Muqit, M.M.K., Foltynie, T., Barker, R., Bhatia, K.P., *et al.* (2006). UCHL-1 is not a Parkinson's disease susceptibility gene. *Annals of Neurology* 59, 627-633.

7 Hellman-Aharony, S., Smirin-Yosef, P., Halevy, A., Pasmanik-Chor, M., Yeheskel, A., Har-Zahav, A., Maya, I., Straussberg, R., Dahary, D., Haviv, A., *et al.* (2013). Microcephaly Thin Corpus Callosum Intellectual Disability Syndrome Caused by Mutated TAF2. *Pediatric Neurology* 49, 411-416.e411.

8 Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. *Genome biology* 4, R70-R70.

9 Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* 4, 44.

10 Jirtle, R.L., and Skinner, M.K. (2007). Environmental epigenomics and disease susceptibility. *Nature reviews Genetics* 8, 253-262.

11 Kahn, R.S., Sommer, I.E., Murray, R.M., Meyer-Lindenberg, A., Weinberger, D.R., Cannon, T.D., O'Donovan, M., Correll, C.U., Kane, J.M., van Os, J., *et al.* (2015). Schizophrenia. *Nature Reviews Disease Primers* 1, 15067.

12 Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* 46, 310.

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

14 Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., *et al.* (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature Communications* 8, 15824.

15 Lage, K., Karlberg, E.O., Størling, Z.M., Ólason, P.Í., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., *et al.* (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology* 25, 309.

16 Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., *et al.* (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929-1935.

17 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.

18 Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285-291.

19 Li, X., and Teng, S. (2015). RNA Sequencing in Schizophrenia. *Bioinformatics and Biology Insights* 9, 53-60.

1 Lohr, K.M., and Miller, G.W. (2014). VMAT2 and Parkinson's disease: harnessing the dopamine vesicle.  
2 Expert review of neurotherapeutics 14, 1115-1117.

3 Louder, R.K., He, Y., López-Blanco, J.R., Fang, J., Chacón, P., and Nogales, E. (2016). Structure of  
4 promoter-bound TFIID and model of human pre-initiation complex assembly. Nature 531, 604.

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

7 Maloku, E., Covelo, I.R., Hanbauer, I., Guidotti, A., Kadriu, B., Hu, Q., Davis, J.M., and Costa, E. (2010).  
8 Lower number of cerebellar Purkinje neurons in psychosis is associated with reduced reelin expression.  
9 Proceedings of the National Academy of Sciences of the United States of America 107, 4407-4411.

10 Maraganore, D.M., Lesnick, T.G., Elbaz, A., Chartier-Harlin, M.-C., Gasser, T., Krüger, R., Hattori, N.,  
11 Mellick, G.D., Quattrone, A., Satoh, J.-I., et al. (2004). UCHL1 is a Parkinson's disease susceptibility gene.  
12 Annals of Neurology 55, 512-521.

13 Mayilyan, K.R. (2012). Complement genetics, deficiencies, and disease associations. Protein & Cell 3,  
14 487-496.

15 Migeon, B.R. (2007). Why females are mosaics, x-chromosome inactivation, and sex differences in  
16 disease. Gender Medicine 4, 97-105.

17 Müller, F., Zaucker, A., and Tora, L. (2010). Developmental regulation of transcription initiation: more  
18 than just changing the actors. Current Opinion in Genetics & Development 20, 533-540.

19 Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and  
20 Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nature  
21 methods 12, 453-457.

22 Nguyen, T., Diaz, D., and Draghici, S. (2016). TOMAS: A novel TOpology-aware Meta-Analysis approach  
23 applied to System biology. Paper presented at: Proceedings of the 7th ACM International Conference on  
24 Bioinformatics, Computational Biology, and Health Informatics (ACM).

25 O'Rawe, Jason A., Wu, Y., Dörfel, Max J., Rope, Alan F., Au, PY B., Parboosingh, Jillian S., Moon, S., Kousi,  
26 M., Kosma, K., Smith, Christopher S., et al. (2015). TAF1 Variants Are Associated with Dysmorphic  
27 Features, Intellectual Disability, and Neurological Manifestations. American Journal of Human Genetics  
28 97, 922-932.

29 Perez, Y., Menascu, S., Cohen, I., Kadir, R., Basha, O., Shorer, Z., Romi, H., Meiri, G., Rabinski, T., Ofir, R.,  
30 et al. (2018). RSRC1 mutation affects intellect and behaviour through aberrant splicing and transcription,  
31 downregulating IGFBP3. Brain 141, 961-970.

32 Pham, N.C., Haibe-Kains, B., Bellot, P., Bontempi, G., and Meyer, P.E. (2017). Study of Meta-analysis  
33 strategies for network inference using information-theoretic approaches. BioData Mining 10, 15.

34 Pijnappel, W.W.M.P., Esch, D., Baltissen, M.P.A., Wu, G., Mischerikow, N., Bergsma, A.J., van der Wal, E.,  
35 Han, D.W., Bruch, H.v., Moritz, S., et al. (2013). A central role for TFIID in the pluripotent transcription  
36 circuitry. Nature 495, 516.

37 Poewe, W., Seppi, K., Tanner, C.M., Halliday, G.M., Brundin, P., Volkmann, J., Schrag, A.-E., and Lang, A.E.  
38 (2017). Parkinson disease. Nature Reviews Disease Primers 3, 17013.

39 Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert,  
40 K., Bergen, S.E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in  
41 schizophrenia. Nature 506, 185.

42 Raser, J.M., and O'Shea, E.K. (2005). Noise in Gene Expression: Origins, Consequences, and Control.  
43 Science 309, 2010-2013.

44 Rees, E., O'Donovan, M.C., and Owen, M.J. (2015). Genetics of schizophrenia. Current Opinion in  
45 Behavioral Sciences 2, 8-14.

46 Sanders, A.R., Drigalenko, E.I., Duan, J., Moy, W., Freda, J., Göring, H.H.H., and Gejman, P.V. (2017).  
47 Transcriptome sequencing study implicates immune-related genes differentially expressed in  
48 schizophrenia: new data and a meta-analysis. Translational Psychiatry 7, e1093.

1 Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J.,  
2 Baum, M., Van Doren, V., *et al.* (2016). Schizophrenia risk from complex variation of complement  
3 component 4. *Nature* *530*, 177.  
4 Sesar, A., Cacheiro, P., López-López, M., Camiña-Tato, M., Quintáns, B., Monroy-Jaramillo, N., Alonso-  
5 Vilatela, M.-E., Cebrián, E., Yescas-Gómez, P., Ares, B., *et al.* (2016). Synaptotagmin XI in Parkinson's  
6 disease: New evidence from an association study in Spain and Mexico. *Journal of the Neurological  
7 Sciences* *362*, 321-325.  
8 Siddiqui, I.J., Pervaiz, N., and Abbasi, A.A. (2016). The Parkinson Disease gene SNCA: Evolutionary and  
9 structural insights with pathological implication. *Scientific Reports* *6*, 24475.  
10 Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A.,  
11 Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* (2005). Gene set enrichment analysis: A knowledge-based  
12 approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of  
13 Sciences* *102*, 15545-15550.  
14 Thomas, P.D. (2017). The Gene Ontology and the Meaning of Biological Function. In *The Gene Ontology  
Handbook*, C. Dessimoz, and N. Škunca, eds. (New York, NY: Springer New York), pp. 15-24.  
15 Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A.,  
16 Cheng, Y., *et al.* (2012). Sequence features and chromatin structure around the genomic regions bound  
17 by 119 human transcription factors. *Genome Res* *22*, 1798-1812.  
18 Yu, C., Woo, H.J., Yu, X., Oyama, T., Wallqvist, A., and Reifman, J. (2017). A strategy for evaluating  
19 pathway analysis methods. *BMC Bioinformatics* *18*, 453.  
20 Zeng, Y., Wang, G., Yang, E., Ji, G., Brinkmeyer-Langford, C.L., and Cai, J.J. (2015). Aberrant Gene  
21 Expression in Humans. *PLOS Genetics* *11*, e1004942.  
22 Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T.J., Lee, Ciaran M., Banskota, S., Marigorta, Urko M., Bao, G.,  
23 and Gibson, G. (2016). A Burden of Rare Variants Associated with Extremes of Gene Expression in  
24 Human Peripheral Blood. *The American Journal of Human Genetics* *98*, 299-309.  
25 Zoubarev, A., Hamer, K.M., Keshav, K.D., McCarthy, E.L., Santos, J.R.C., Van Rossum, T., McDonald, C.,  
26 Hall, A., Wan, X., and Lim, R. (2012). Gemma: a resource for the reuse, sharing and meta-analysis of  
27 expression profiling data. *Bioinformatics* *28*, 2272-2273.  
28  
29  
30  
31

1 **Figure Legends**

2 Figure 1 What does a systems biology approach tell us? A systems biology assessment summarizes the properties  
3 of genes that are captured in an experiment, but can highlight both true and false positive results. In some cases,  
4 genes with a shared function are unlikely to arise in the experiment by chance. A gene set analysis will highlight  
5 their shared function (top panel), but not their independent value in identifying the enriched function. In other  
6 cases, a set of genes are so closely related both technically and biologically that if one arises, the others are almost  
7 certain to do so. A statistical analysis treating the genes as independent (bottom panel) will attach a misleading  
8 significance to the shared presence of the genes. These genes and gene sets will be unlikely to replicate in future  
9 studies.

10 Figure 2 Disease expression analysis schematic. (A) We calculate expression fold change between probands and  
11 parents and pick out the top 100 up- and down-regulated genes (increased and decreased expression,  
12 respectively). We test for joint functional properties through gene set enrichment and co-expression modularity.  
13 (B) Given multiple pedigrees/families of the disorder, we can piece together whether disease signals are recurrent  
14 across families or non-recurrent, and due to multiple genes (joint) or independent genes (disjoint). (C) This can be  
15 done by assessing recurrence at the gene and pathway levels. (D) The *TAF1* syndrome cohort pedigrees used in this  
16 analysis. Four cases have missense mutations, one case a splice site mutation, and the last case a CNV duplication.  
17 Three of the mothers are carriers with no distinguishing characteristics.

18 Figure 3 Disease expression analysis with a family-based approach. (A) The expression fold change for each gene is  
19 calculated within each family (top 100 up and down regulated genes are shown). (B) Overlaps in DE gene sets  
20 between the individual families (numbers in boxes), and the significance of this overlap (colored corresponding to -  
21 log10 P-value of the hypergeometric test). Overlaps are mostly small. (C) The replicable genes are those that are  
22 recurrent across families. The recurrence distributions for both up- and downregulated genes across the 6 families  
23 are shown. Using the binomial test, we find that genes recurring 3 or more times are significant (FDR<0.05). These  
24 genes are listed, with 4 up- and 14 downregulated genes significantly recurrent. (D) Robustness assessment of the  
25 DE threshold. The plot shows the number of recurrent genes as a function of the number of differentially  
26 expressed genes and the significance of the recurrence in grey.

27 Figure 4 Gene set enrichment assessment of *TAF1* cohort. (A) Top GO enrichment results for each family for up-,  
28 and downregulated genes. Significant terms (FDR < 0.05) are highlighted with an \*. (B) The frequency and  
29 significance of recurrence of each GO term is plotted for the upregulated genes. (C) Gene-GO membership matrix  
30 for the upregulated genes. Each column is a gene, and each row a family. The colored bars below highlight the GO  
31 terms that these genes belong to. The signal associated with the recurrent GO terms is distributed across different  
32 genes, shown by low overlap across the families. (D) There are no significantly recurrent pathways with the  
33 downregulated genes. (E) However, the recurrent genes themselves are enriched for ribosomal pathways (p-  
34 adjusted<0.05), as shown in the gene-GO membership matrix. The three RP\* genes seem to drive almost all the  
35 signal.

36 Figure 5 Co-expression of differentially expressed genes generates enrichment. (A) As an example from family 1,  
37 we show the co-expression frequency sub-network as a heatmap, where genes showing decreased expression  
38 show co-expression. Co-expression blocks define modules as determined by the clustering (see rows). The modules  
39 are enriched for particular genes, mainly ribonucleoproteins. Performing a gene set enrichment analysis on these  
40 genes (Fisher's exact test on GO groups), genes (rows) that generate the enrichment (columns are enriched GO  
41 terms) almost exclusively overlap with the co-expression blocks. The prominent pathways are ribosome related.  
42 (B) The significantly recurrent genes can be divided into those present within co-expression modules (joint) and  
43 those not (disjoint). The genes in bold are the functional outliers and the venn diagrams summarizes the number of  
44 genes in each category. (C) If we look at the enrichment of these DE gene sets (pre-filtering dark line +/-SD  
45 shadow), we see that filtering off the modules removes all but a few significant terms (lighter line, +/-SD shadow).

1 Figure 6 Differential expression meta-analysis in three other disorders. (A) Recurrence of genes in Huntington's  
2 disease (HD), (B) Parkinson's disease (PD) and (C) schizophrenia (SCZ), and whether they occur in groups (joint) or  
3 not (disjoint). The venn diagrams summarize the number of recurrent genes and their joint or disjoint designation.

4

1 **STAR Methods**

2 **CONTACT FOR REAGENT AND RESOURCE SHARING**

3 Further information and requests for resources should be directed to and will be fulfilled by the Lead  
4 Contact, Jesse Gillis ([tgillis@cshl.edu](mailto:tgillis@cshl.edu)).

5 **EXPERIMENTAL MODEL AND SUBJECT DETAILS**

6 *The TAF1 syndrome cohort*

7 We assessed 6 pedigrees of a genetically and phenotypically homogeneous X-linked *TAF1* syndrome  
8 cohorts. The original cohort was assembled by O’Rawe et al. (O’Rawe et al., 2015), which included 11  
9 pedigrees from around the world. The probands are male, between 5-21 years of age, have intellectual  
10 disability, distinct facial dysmorphology, general hypotonia, hearing impairments, and a characteristic  
11 intergluteal crease. Of the 6 pedigrees we studied, all probands had a point mutation in their *TAF1*  
12 transcription factor, except for a single CNV case with a duplication of a ~0.42 Mb region at Xq13.1 that  
13 includes *TAF1* and other genes.

14 **METHOD DETAILS**

15 *RNA-sequencing and processing*

16 Blood was collected in PAXgene Blood RNA tubes and the RNA was isolated with the PAXgene Blood  
17 RNA kit (QIAGEN) according to the manufacturer’s recommendations. The RNA was quantified using  
18 NanoDrop. To increase downstream sensitivity, globin mRNA was depleted from the samples using the  
19 GLOBINclear Kit (Life Technologies). Briefly, RNA was precipitated with ammonium acetate, washed and  
20 resuspended in 14 µl TE (10 mM Tris-HCl pH 8, 1 mM EDTA). Subsequently, for each sample 1.1 µg RNA  
21 were hybridized with the provided streptavidin beads and purified. To control for variation in RNA  
22 expression data, 1 µl of a 1:100 dilution of ERCC RNA Spike-In control (Thermo Fisher) was added to 1 µg  
23 RNA and libraries generated according to the TruSeq Stranded mRNA Library Kit-v2 (Illumina) with the  
24 index primers as indicated in

25 **Table S1.** Quality control of the generated libraries was performed on a Bioanalyzer High Sensitivity DNA  
26 chip (Agilent) and the concentration was measured using Qubit dsDNA HS Assay (Life Technologies). To  
27 eliminate primer dimers in the libraries, additional purifications were performed using the Agencourt  
28 AMPure XP system (Beckman Coulter). The libraries were pooled to 2-10 nM total concentration and  
29 sequenced on an Illumina NextSeq 500, PE100, mid output. Libraries were generated independently for  
30 each family and family-pools multiplexed and sequenced on separate lanes. ERCC spike-ins included in  
31 the preparation were not used for normalization, but rather as a measure of quality control. Families 2,  
32 3 and 4 showed the lowest variation in the ERCCs between family members, while family 5 and 6 had  
33 higher technical noise (**Figure S2**). Reads were filtered for QC and artifacts using the fastX toolbox, and  
34 then the reads were paired up using an adapted python script  
35 (<https://github.com/enormandeau/Scripts/blob/master/fastqCombinePairedEnd.py>). The reads were  
36 aligned to the genome (GRCh38, GENCODE v22 (Harrow et al., 2012)) using STAR (2.4.2a)(Dobin et al.,  
37 2012).

38 *Differential expression analysis*

39 We calculated fold change between parents and probands for the differential expression analysis. We  
40 first calculated the CPM (counts per million) for each individual, and then took the average CPM for the  
41 parents and compared it the CPM of the proband. Fold change was defined as the log2 of the ratio of

1 these values after adding a pseudocount of 1. We exploited within-family variance to detect noisy  
2 genes, removing genes that showed strong differential expression between the parents (i.e., top 100 up-  
3 regulated and top 100 down-regulated genes). After removing these highly variable genes, top up- and  
4 down-regulated genes were defined based on ranked fold change. We assessed each family in a  
5 separate batch (library preparation and sequencing run), holding technical variation constant in each  
6 family and independent across families, so that gene-level recurrence is not expected to differ from the  
7 null. By way of analogy, our experimental design resembles the analysis of *de novo* variants in DNA  
8 analyses, in which as many factors as possible are held constant in the control group for the proband.  
9 The use of unaffected family members as controls provides the closest possible genetic and  
10 environmental match for the probands, constraining variability of known importance for expression  
11 analysis (Raser and O'Shea, 2005). Although each family-specific analysis is confounded with age and sex,  
12 we anticipate that genes detected as differentially expressed that are due to these overlaps can be  
13 assessed and identified directly, as these are well-powered properties in many previous studies.

14 *Common co-expression frequency network*

15 Human RNA-seq expression data was downloaded from Gemma (Zoubarev et al., 2012). From the total  
16 collection of approximately 300 human experiments, we selected 75 expression experiments (3,653  
17 samples) that we could ascertain derived from tissues and not cell lines (listed in **Table S2**). For each  
18 experiment, we consolidated our list of genes/transcripts to the ~30K genes with Entrez gene identifiers,  
19 and did not limit either expression level or occurrence of expression. For each experiment with at least  
20 10 samples, we generated a co-expression network by calculating Spearman's correlation coefficients  
21 between every gene pair (Ballouz et al., 2015) and calculated the frequency that a pair of genes was  
22 positively co-expressed (Spearman's correlation coefficient  $r_s > 0$ ). We used this tally network as a  
23 measure of the frequency of common co-expression of the gene pairs. The more observations with a  
24 positive correlation, the more commonly co-expressed the pairs are (see **Figure S1**).

25 *Gene set enrichment*

26 To calculate gene set enrichment of the differentially expressed genes, we used an in-house gene set  
27 enrichment R script based on the hypergeometric test. For each gene set, we calculated the significance  
28 of the overlap of the differentially expressed genes with that set, correcting for multiple tests with  
29 Benjamini-Hochberg (FDR, *p.adjust* in R). We used an in-house parsed version of GO (downloaded July  
30 2015). We report GO slim (filtered to 132 GO groups) results primarily in the text but also assess a  
31 subset of GO based on gene set size to remove redundancy (10-100 genes per group, 4605 GO terms). In  
32 addition, we assess enrichment using four gene set lists from MSigDB (v6, 1127 gene sets, HALLMARK,  
33 KEGG, REACTOME and BIOCARTA).

34 *Co-expression module and outlier detection*

35 In a set of differentially expressed genes, we defined co-expression modules as highly co-expressed  
36 genes, seen as blocks in a co-expression network when clustered and represented as a matrix. Genes  
37 that did not cluster or clustered weakly were considered outliers. These genes potentially have stronger  
38 co-expression links with other genes outside of the gene set, but are not as well-linked within the  
39 subset. To identify these two classes in our list of differentially expressed genes (DEGs), we first  
40 extracted the sub-network of the differentially expressed genes from the co-expression frequency  
41 network. Then, thresholding on the median co-expression value of the co-expression network, we used  
42 this binary network as distance matrix, and performed hierarchical clustering of the genes. This  
43 clustering returned a dendrogram of genes that are closer in distance, and we used this dendrogram to

1 define modules within the data. We used the R dynamicTreeCut (Langfelder et al., 2008) package to  
2 select modules within the data with a cut height 0.995. We used these clusters to define our co-  
3 expression modules, where clusters with more than five genes were labelled as modules and those  
4 smaller as co-expression outliers.

5 *Recurrence analysis*

6 To test for replicability of the disease signal, we measured differentially expressed gene recurrences and  
7 the significance of recurrence as the probability of observing the differentially expressed genes across all  
8 the pedigrees. We first calculated the significance of the pairwise overlap between families using  
9 Fisher's exact test (phyper in R). We calculated the significance of recurrence of the differentially  
10 expressed genes using the binomial test (pbinom in R), and then corrected for multiple tests using  
11 Benjamini-Hochberg (FDR, p.adjust in R). After we filter on differentially expressed genes within  
12 modules, we recalculate the significance of recurrence, this time with a permutation test to obtain an  
13 FDR. Similarly, we use a permutation test to calculate significance of recurrence of the pathways from  
14 the gene set enrichment assessment.

15 *Meta-analysis datasets of neurodegenerative and neuropsychiatric disorders*

16 We sought to repeat the systems biology evaluations in other disorders. We collected the reported  
17 differentially expressed genes in Parkinson's disease, Huntington's disease and schizophrenia studies.  
18 The majority of the studies were collected from recent review articles (Genevie et al., 2018; Li and Teng,  
19 2015), and from a search within the Gemma (Zoubarev et al., 2012) database for "parkinson's disease",  
20 "huntington's disease" and "schizophrenia", respectively (listed in Table S3). We downloaded fold  
21 changes, p-values and adjusted p-values. We removed studies from our analysis where we either could  
22 not assess the direction of the differential expression and where no genes passed significance based on  
23 log2 fold changes ( $\log_2 \text{FC} \geq 1$ ) and adjusted P-values ( $q < 0.05$ ).

24 **QUANTIFICATION AND STATISTICAL ANALYSIS**

25 All statistical analyses were done in R. Significance was defined as an FDR of 0.05 for all statistical tests.

26 **DATA AND SOFTWARE AVAILABILITY**

27 All R code, scripts and network data is available for download from our github repository  
28 (<https://github.com/sarbal/redBlocks>). The RNA-seq data has been deposited in GEO/SRA under  
29 accession number **GSE84891**. All other software used in this analysis is freely available and has been  
30 listed in the key resources table.

31

1 KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Raw and analyzed data	This paper	GEO: GSE84891
Software and Algorithms		
STAR v2.4	(Dobin et al., 2012)	<a href="https://github.com/alexdobin/STAR/">https://github.com/alexdobin/STAR/</a>
Combine paired end reads	Online	<a href="https://github.com/enormandeau/Scripts/blob/master/fastq/CombinePairedEnd.py">https://github.com/enormandeau/Scripts/blob/master/fastq/CombinePairedEnd.py</a>
FASTX - Toolkit	Online	<a href="http://hannonlab.cshl.edu/fastx_toolkit/">http://hannonlab.cshl.edu/fastx_toolkit/</a>
R	Online	<a href="https://cran.r-project.org/">https://cran.r-project.org/</a>
dynamicTreeCut R package	(Langfelder et al., 2008)	<a href="https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/">https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/</a> <a href="https://cran.r-project.org/web/packages/dynamicTreeCut/index.html">https://cran.r-project.org/web/packages/dynamicTreeCut/index.html</a>
Outlier redBlocks algorithm	This paper	<a href="https://github.com/sarbal/redblocks">https://github.com/sarbal/redblocks</a>
CIBERSORT	(Newman et al., 2015)	<a href="https://cibersort.stanford.edu/">https://cibersort.stanford.edu/</a>
Other		
GRCh38 genome files: GRCh38.p2.genome.fa.noPatches.gz	(Harrow et al., 2012)	<a href="https://www.gencodegenes.org/human/release_22.html">https://www.gencodegenes.org/human/release_22.html</a>
GENCODE v22 annotation files: gencode.v22.annotation.gtf	(Harrow et al., 2012)	<a href="https://www.gencodegenes.org/human/release_22.html">https://www.gencodegenes.org/human/release_22.html</a>
Gemma	(Zoubarev et al., 2012)	<a href="https://gemma.msl.ubc.ca/home.html">https://gemma.msl.ubc.ca/home.html</a>

2

3

4

1    **Supplementary**

2    **Figure S1** Meta-analytic co-expression frequency network generation. Related to Figure 2.

3    **Figure S2** ERCC spike-ins QC. Related to Figure 3.

4    **Figure S3** Pathway recurrence and GO group size. Related to Figure 4.

5    **Figure S4** Robustness assessment of DE threshold. Related to Figure 3.

6

7    **Table S1** Library numbers and adapter sequences used in this study. Related to Figure 3.

8    **Table S2** RNA-seq experiments used. Related to Figures 3-6.

9    **Table S3** Studies used in the meta-analysis. Related to Figure 6.

10    **Table S4** *TAF1* syndrome recurrent genes. Related to Figure 5.

11    **Table S5** *TAF1* syndrome enrichment results. Related to Figure 4.

12    **Table S6** Huntington's disease recurrent genes. Related to Figure 6.

13    **Table S7** Parkinson's disease recurrent genes. Related to Figure 6.

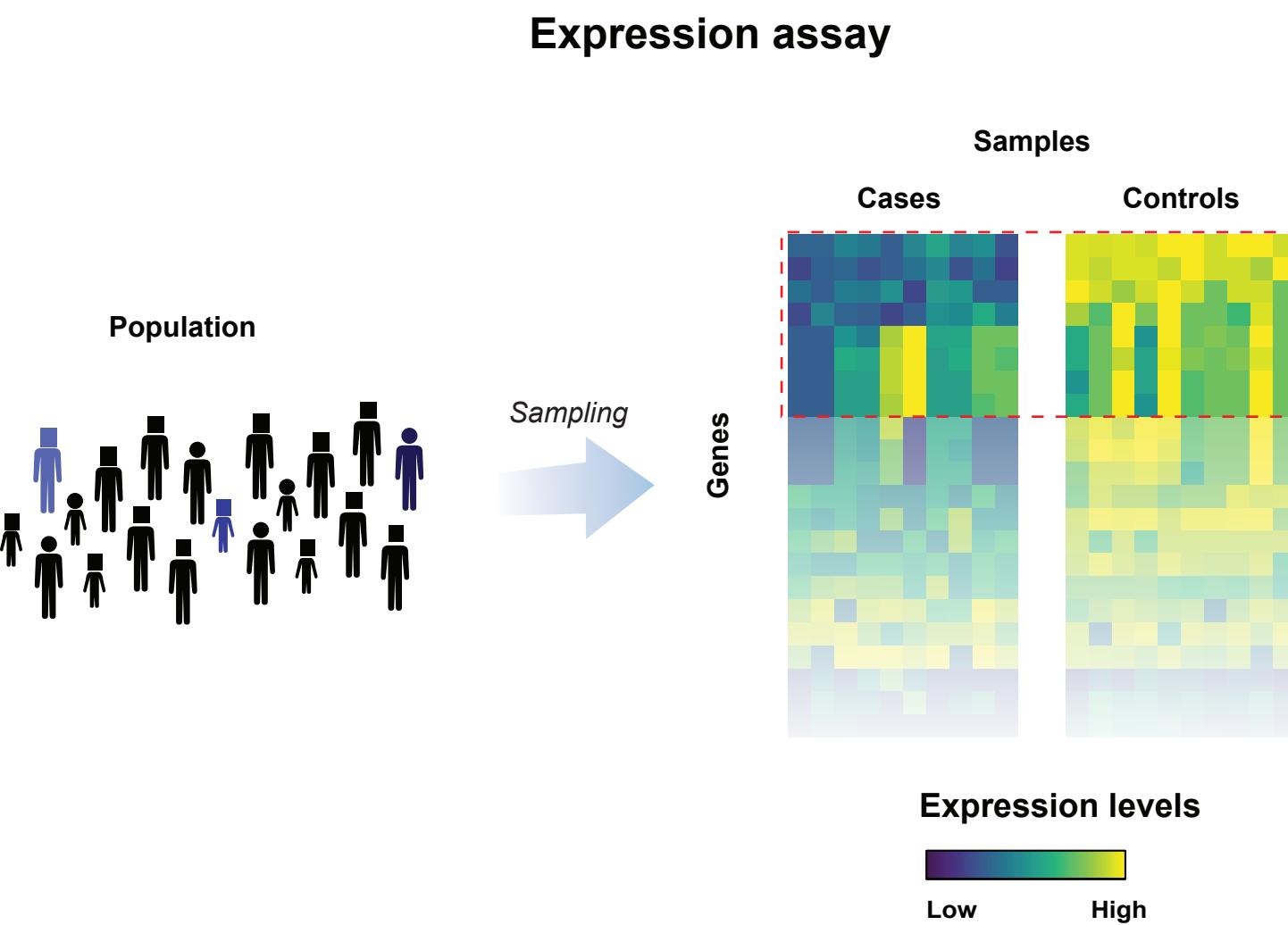
14    **Table S8** Parkinson's disease enrichment results. Related to Figure 6.

15    **Table S9** Schizophrenia recurrent genes. Related to Figure 6.

16    **Table S10** Schizophrenia enrichment results. Related to Figure 6.

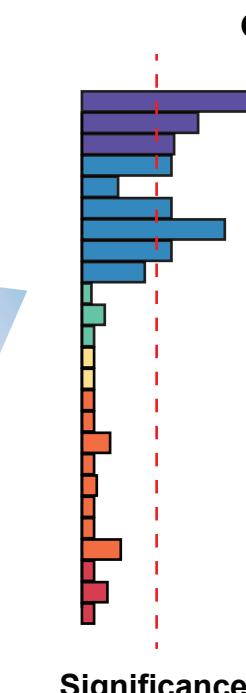
17

# Gene set analysis



**Disease related**  
Shared response to disease, targeted

**Confounded**  
Incidental shared response, technical confounds



## Gene sets

*Synaptic function*

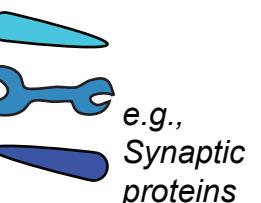
*Synapse*

*Calcium signalling*

*RNA binding*

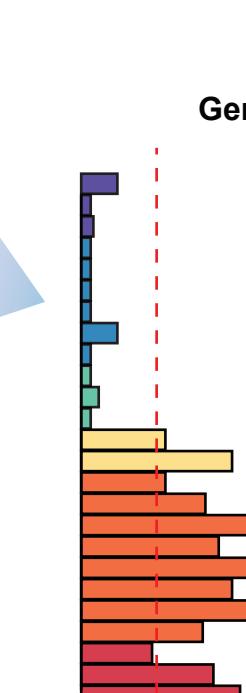
*Ribosome*

*Transcription*



## Enrichment

From independent genes, true positives, will replicate



## Gene sets

*Synaptic function*

*Synapse*

*Calcium signalling*

*RNA binding*

*Ribosome*

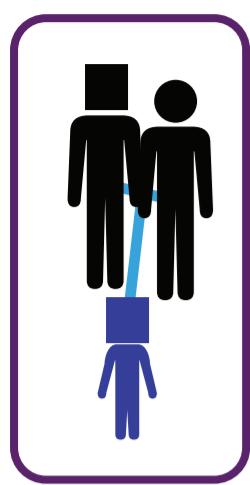
*Transcription*

## Enrichment

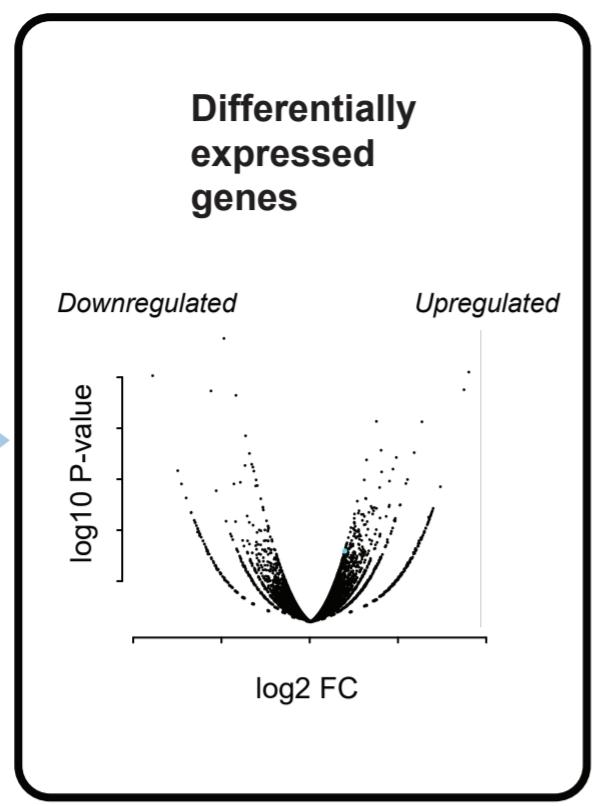
From confounded genes, false positives, will not replicate



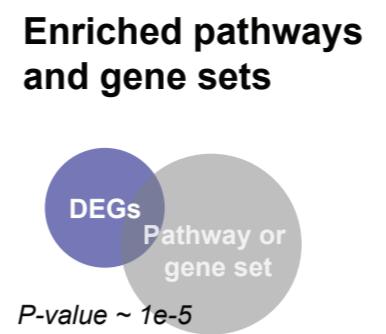
## Significance

**A****Family**

Compare expression levels of proband to parents



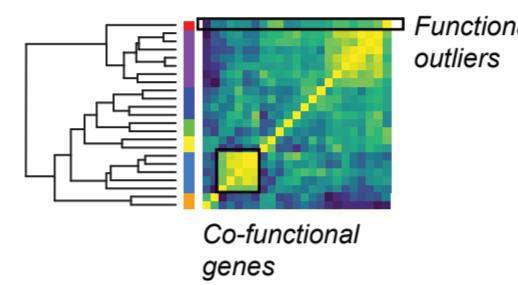
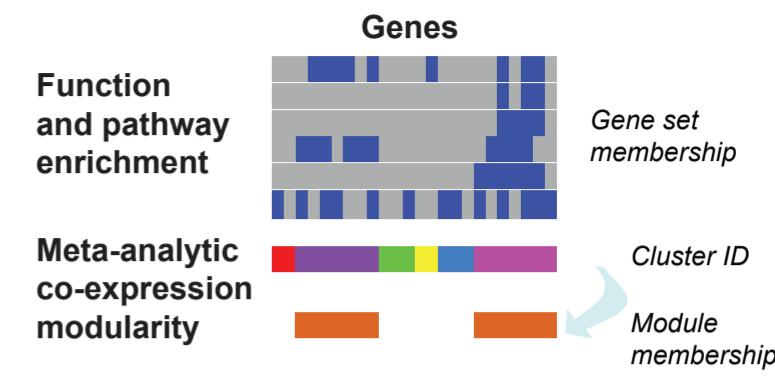
Test for overlap with known biological pathways



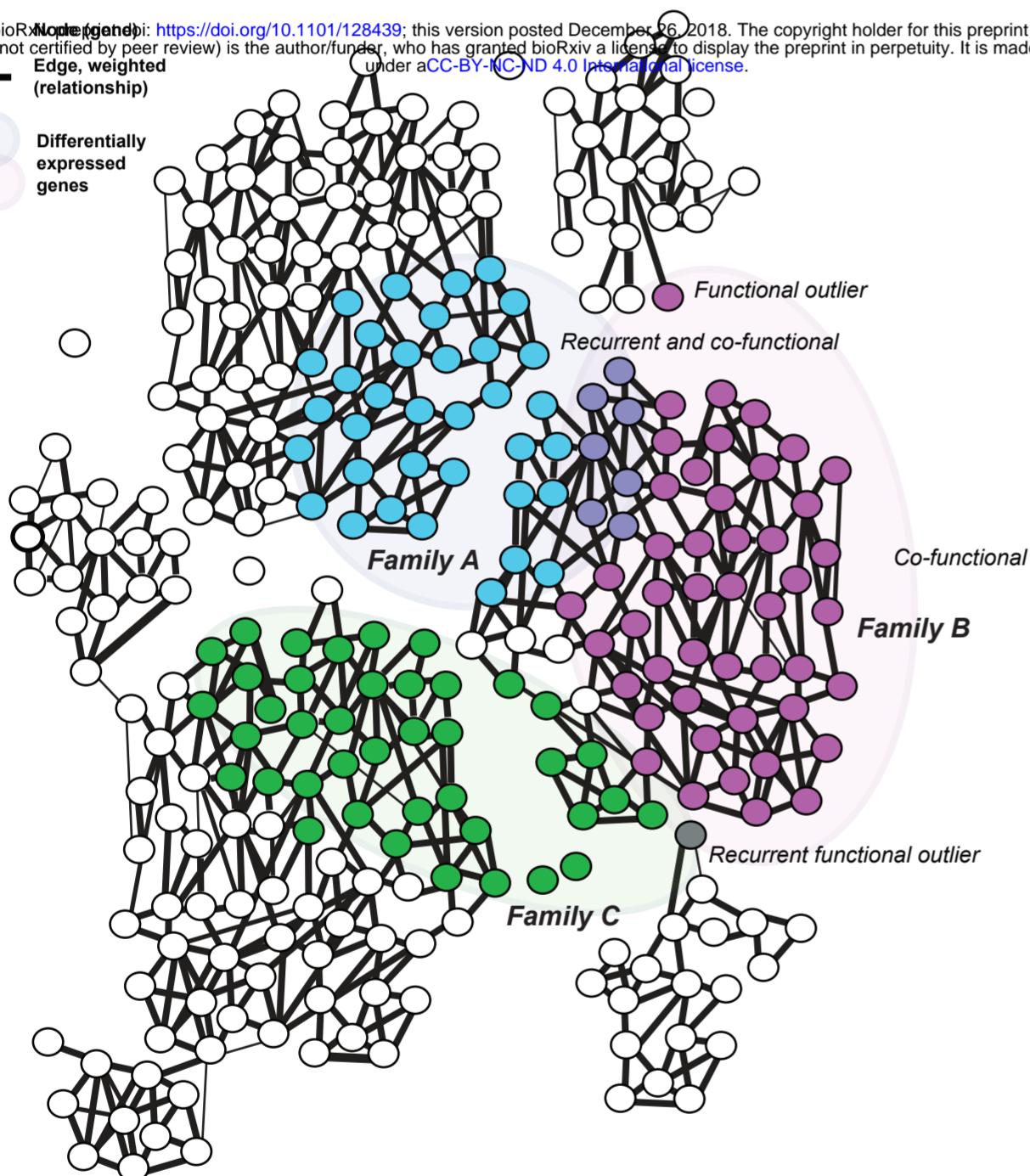
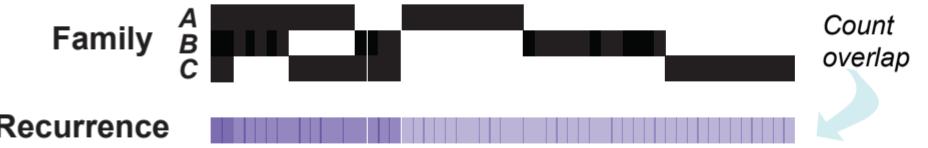
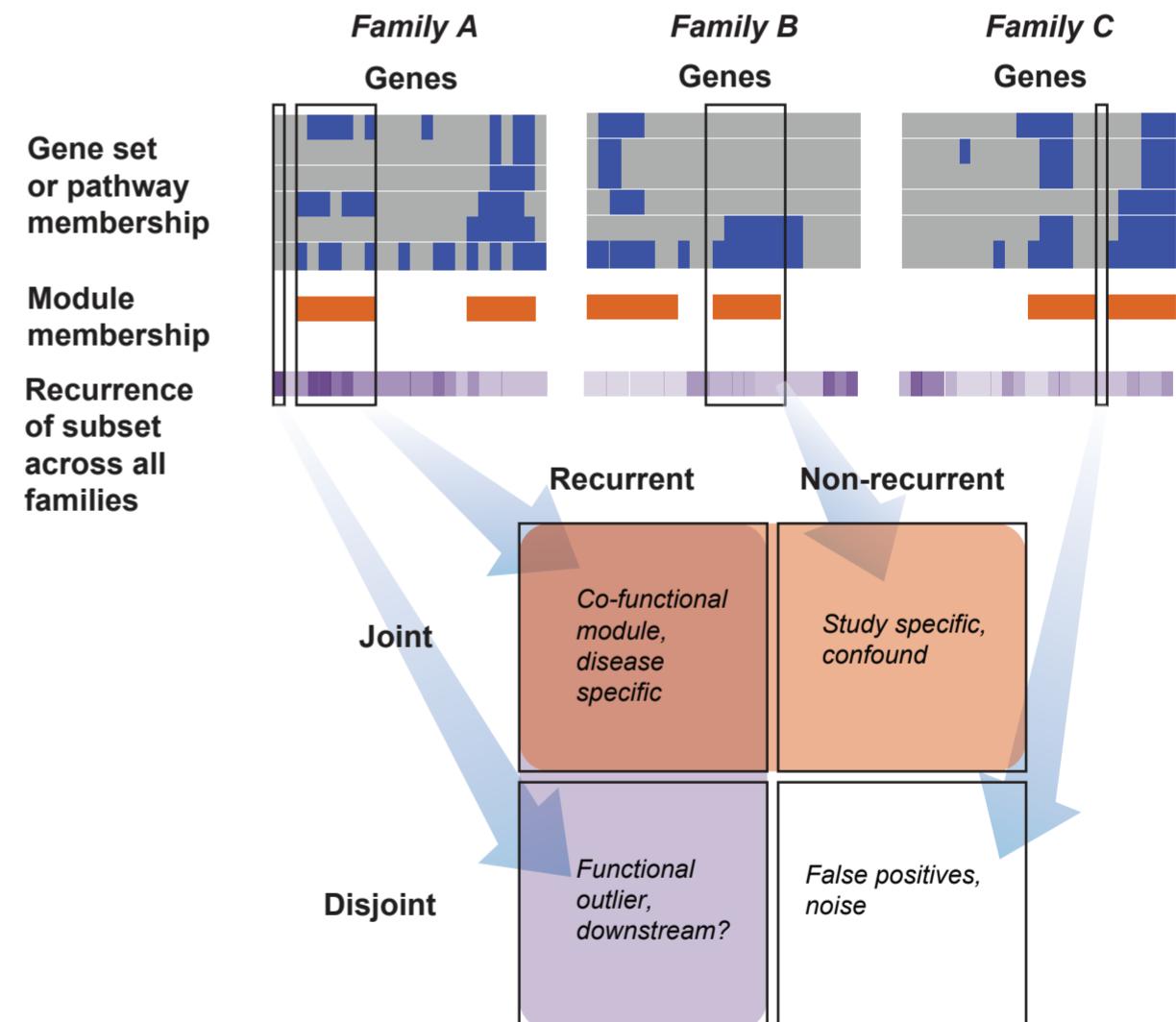
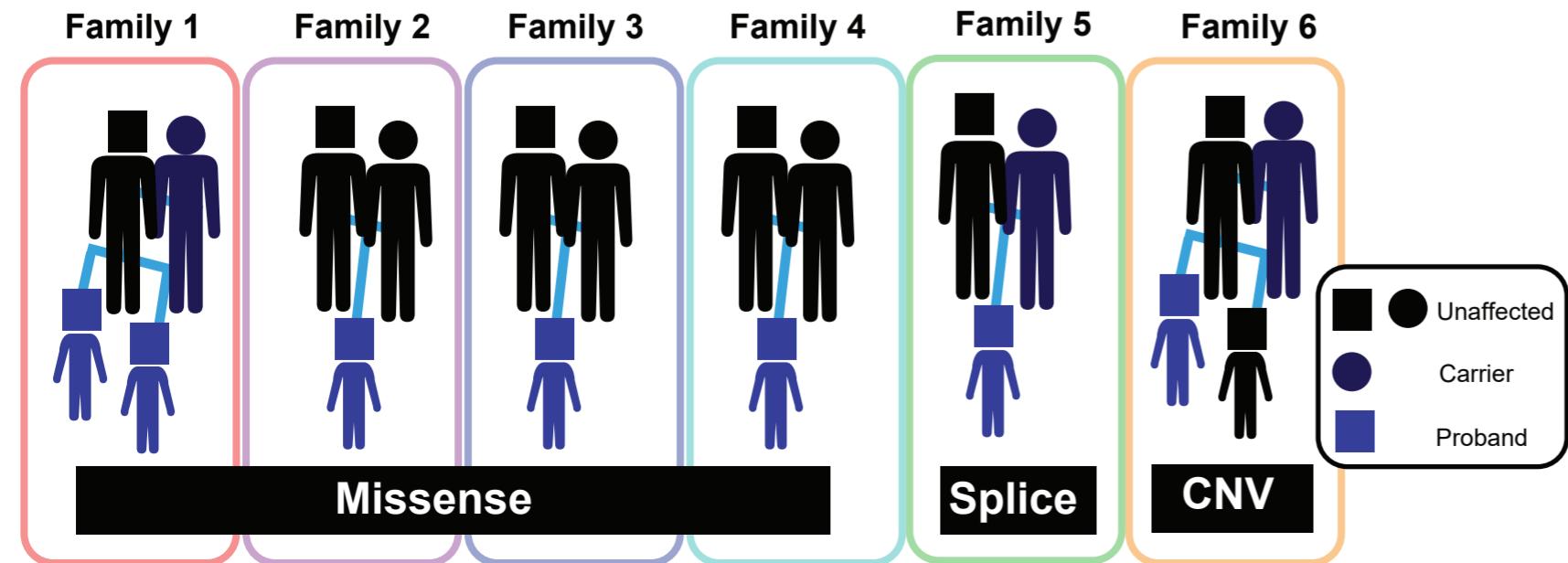
Assess frequency of co-expression using external data

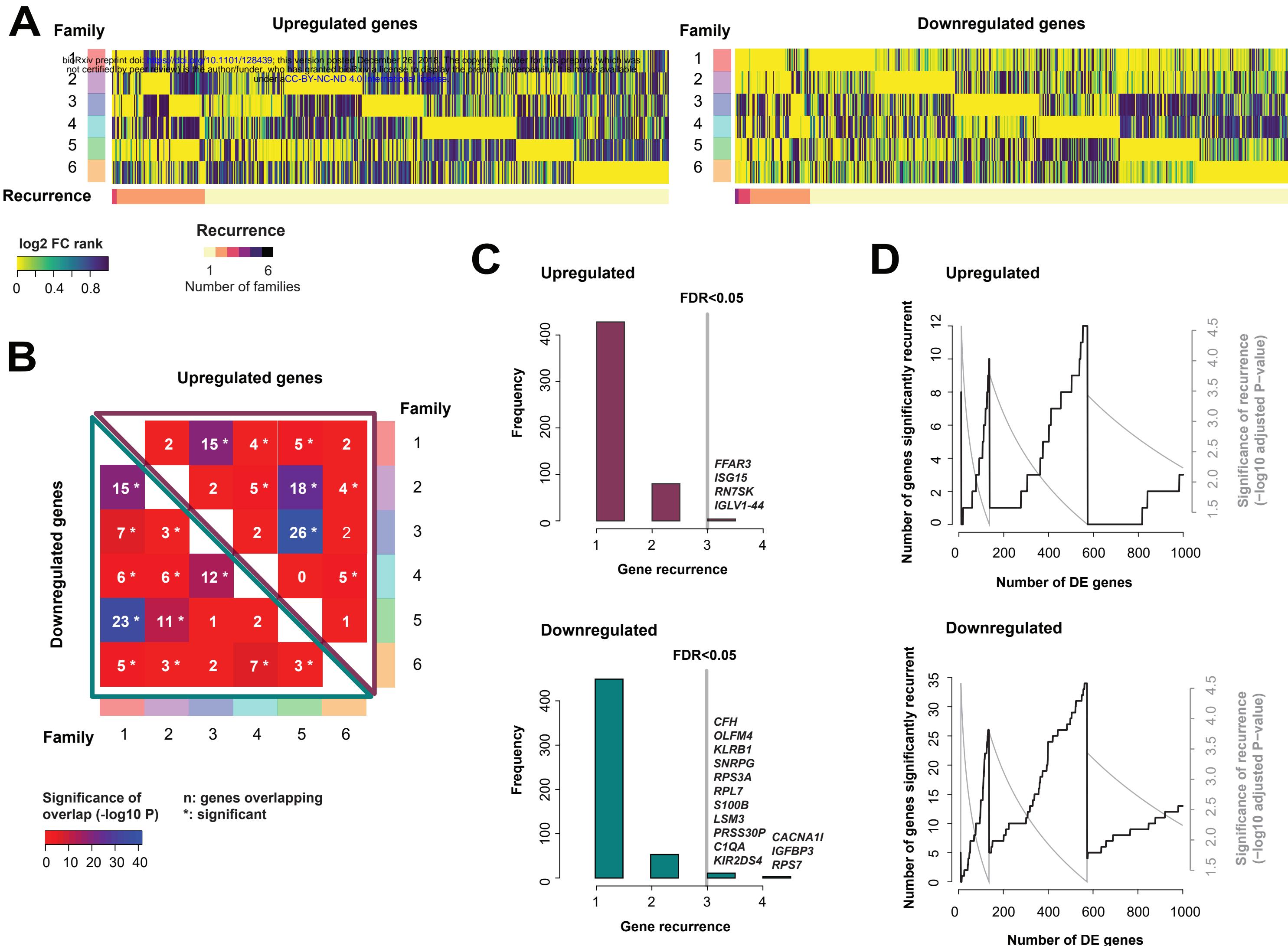
GO:0034340 response to type I interferon  
GO:0060337 type I interferon signaling pathway  
GO:0071357 cellular response to type I interferon  
GO:009615 response to virus

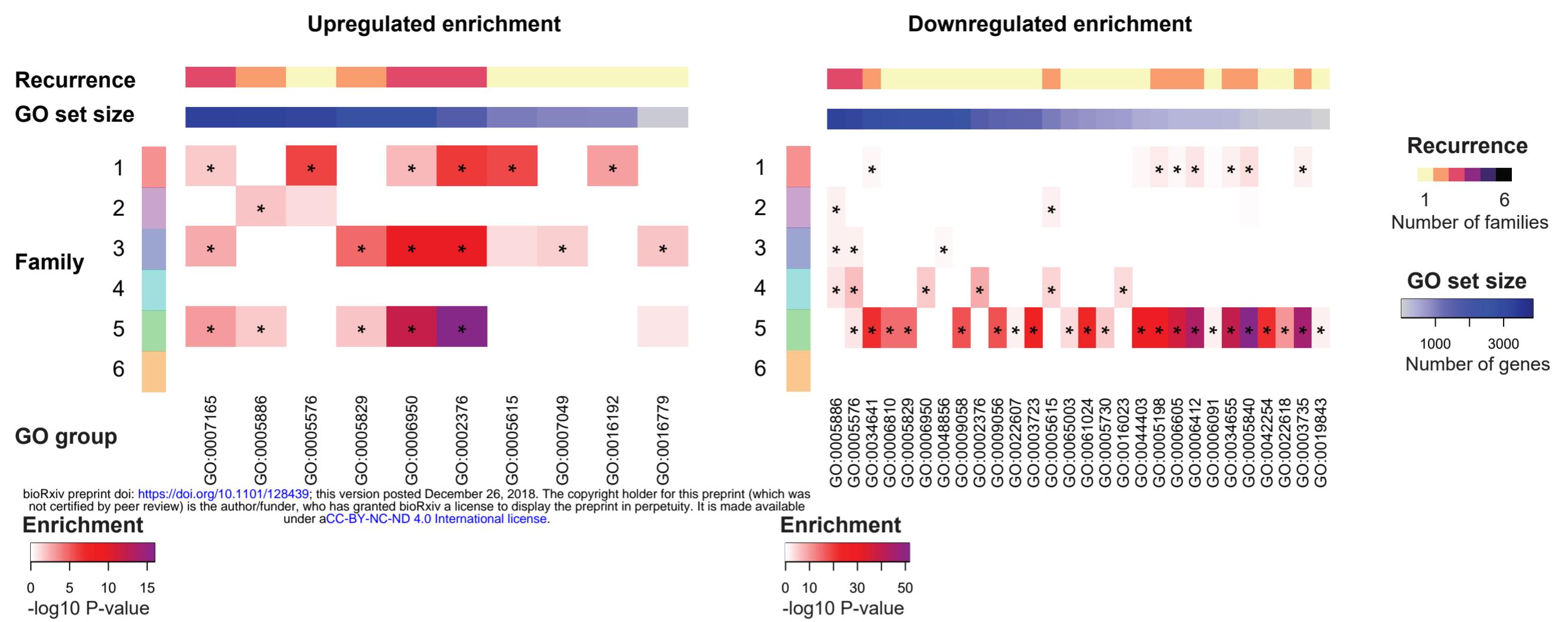
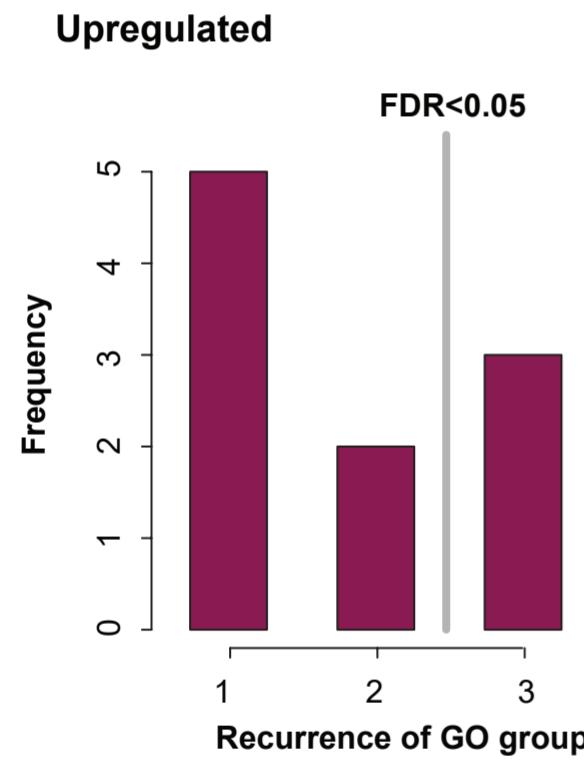
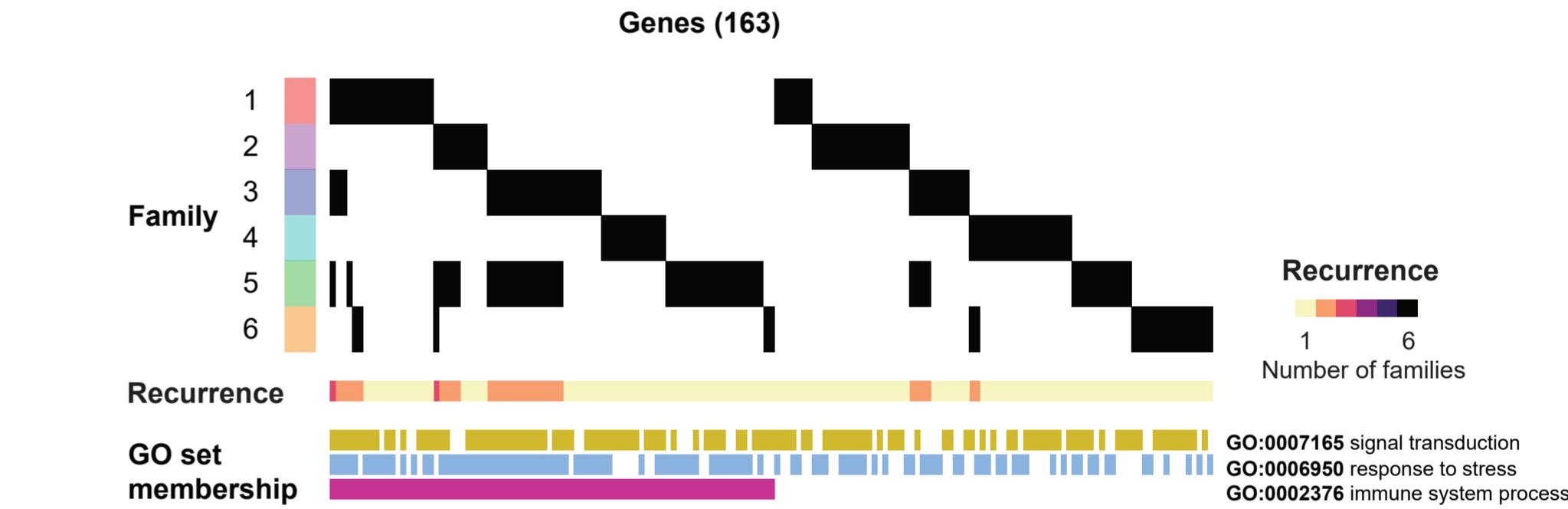
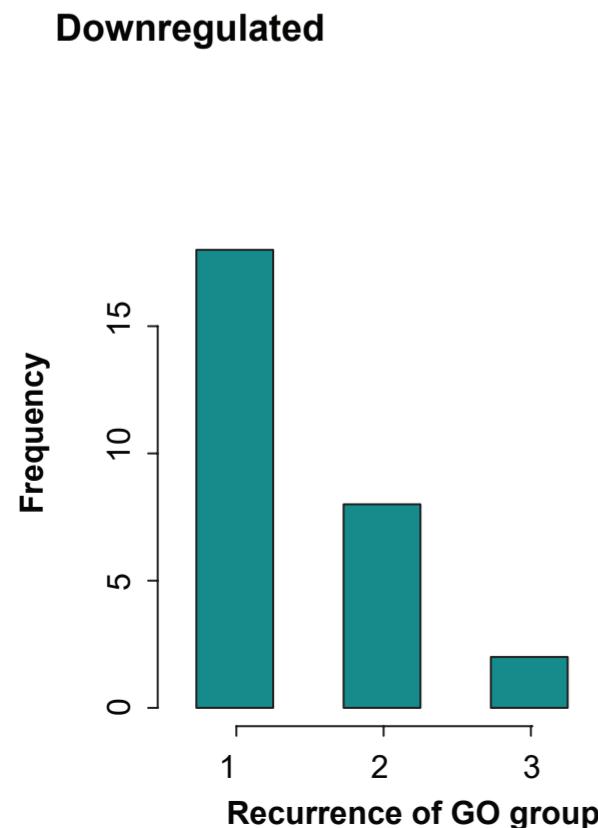
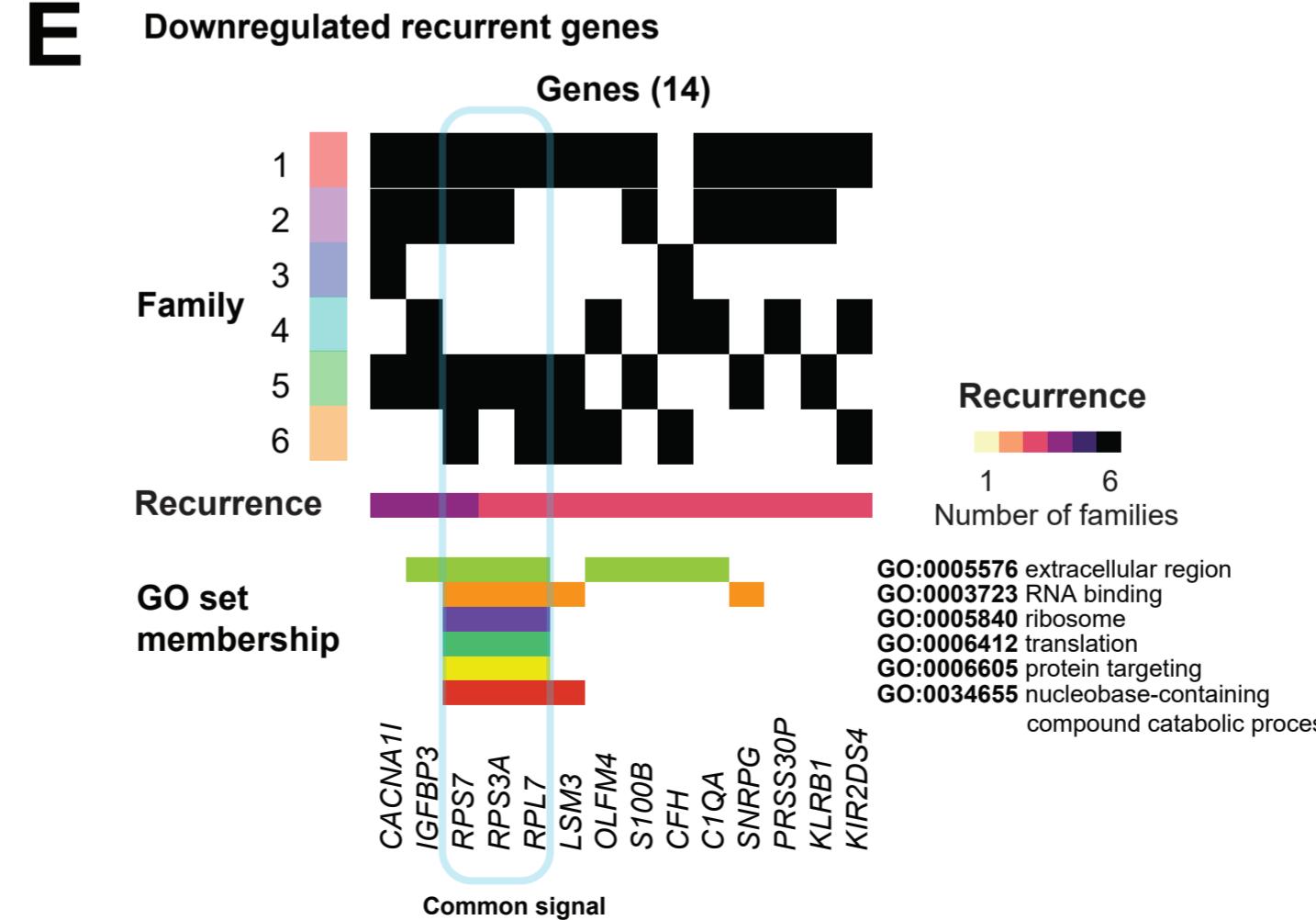
**Meta-analytic co-expression as a proxy for systems**

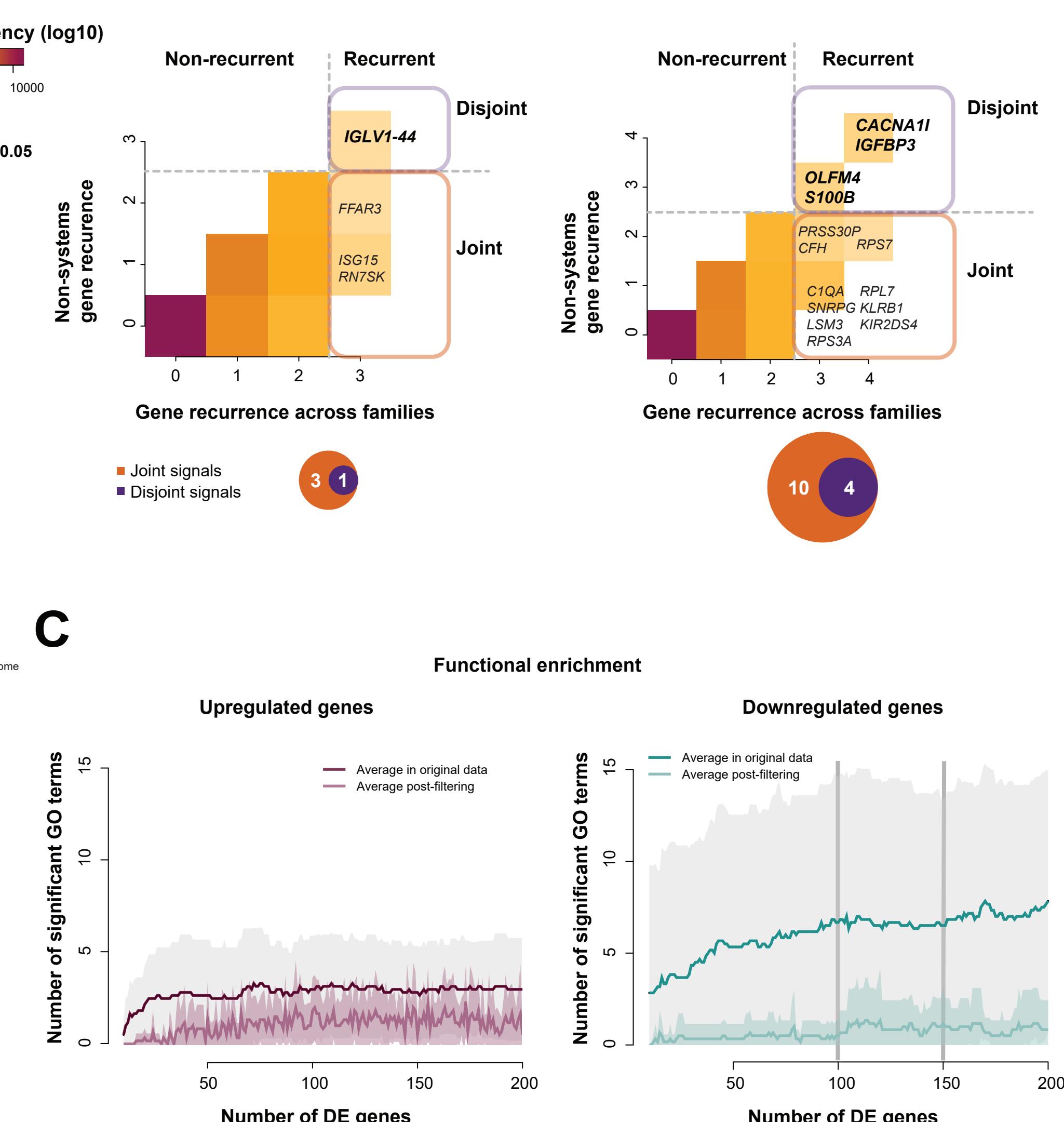
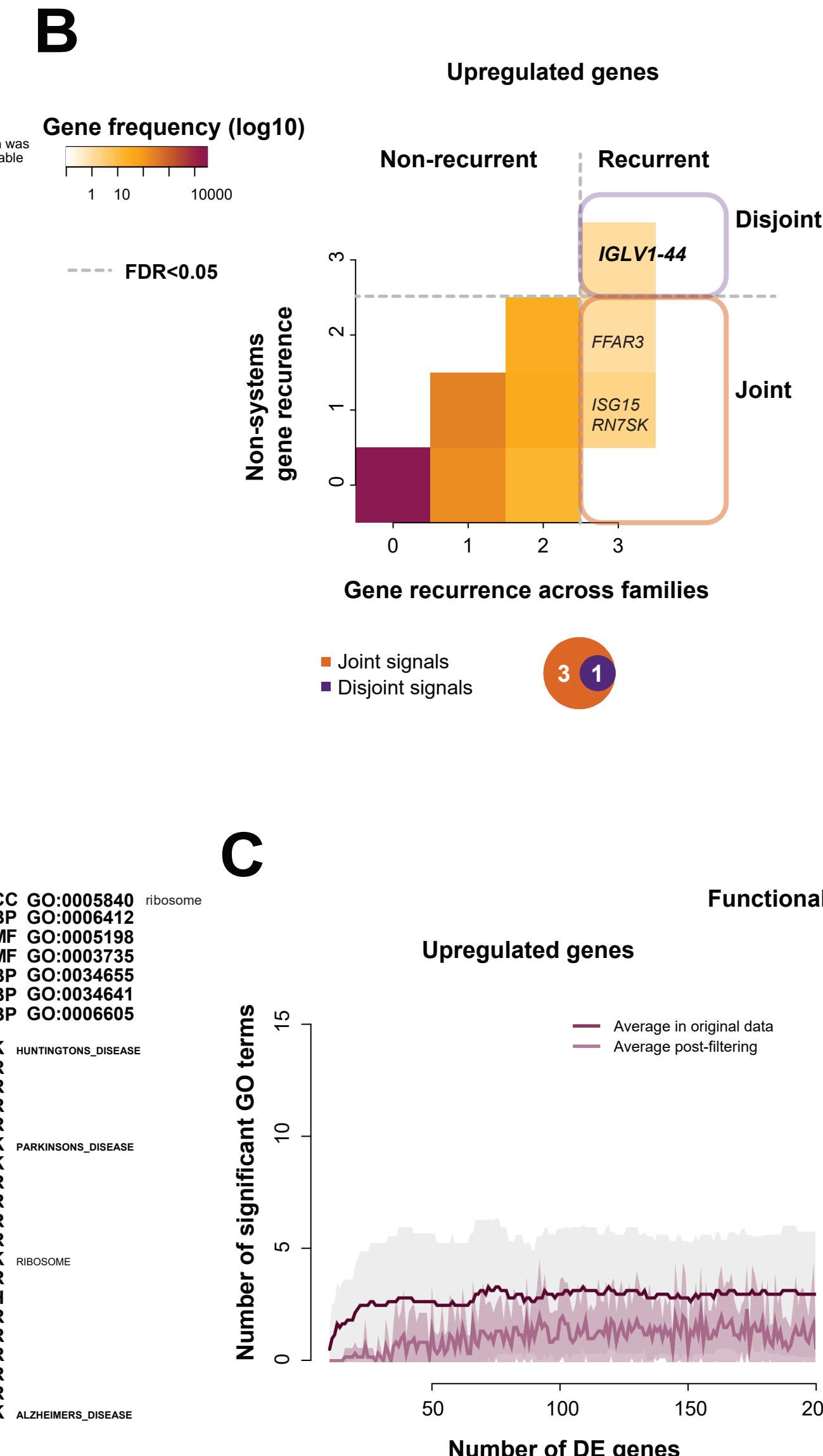
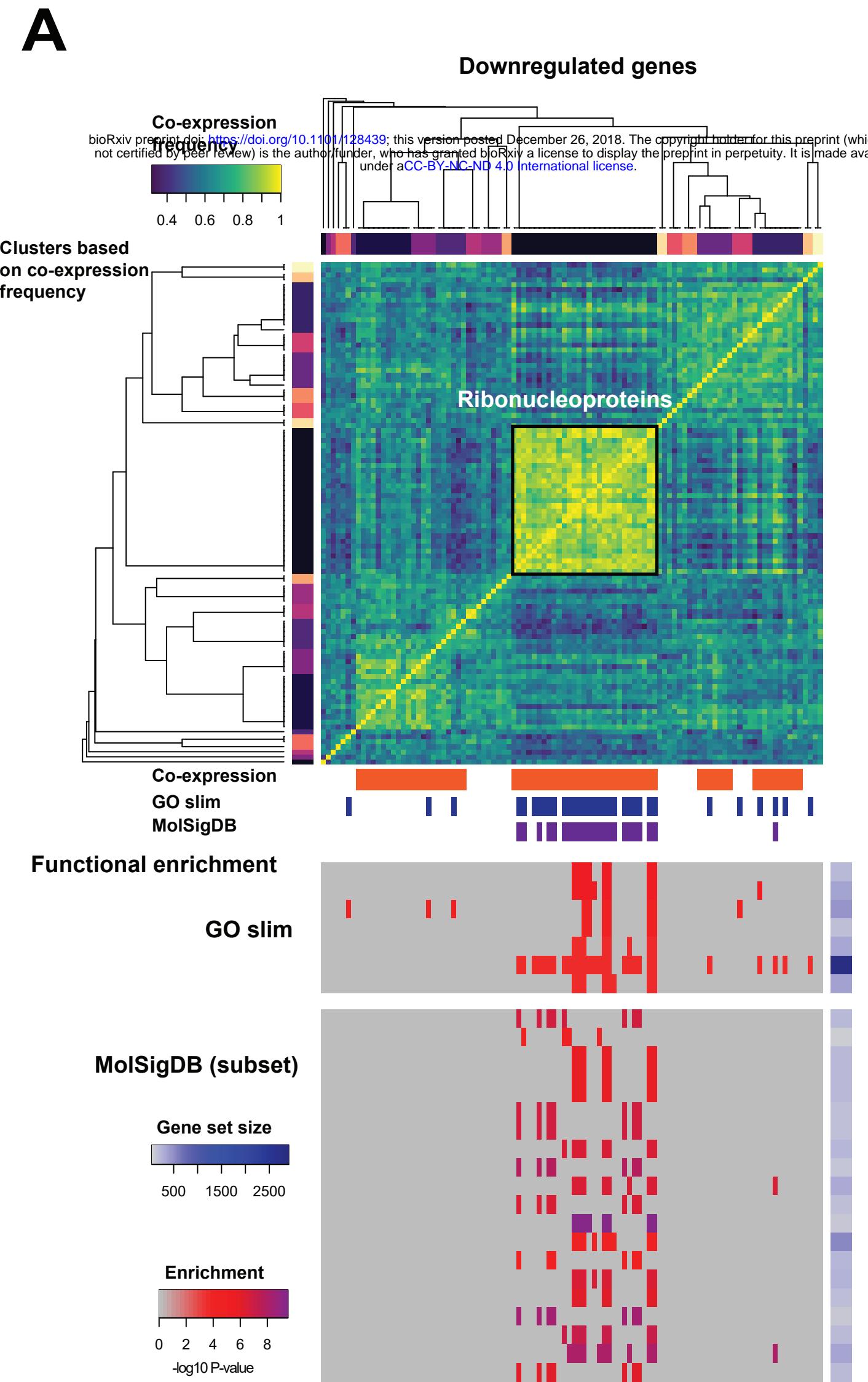
**Systems analysis summary****B****Gene network**

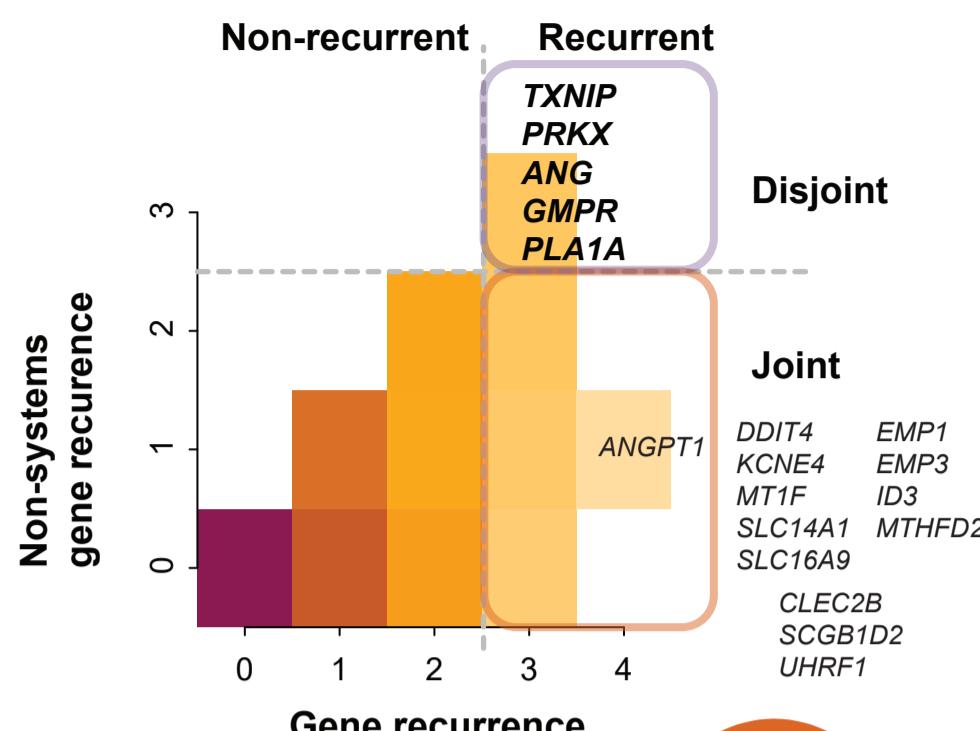
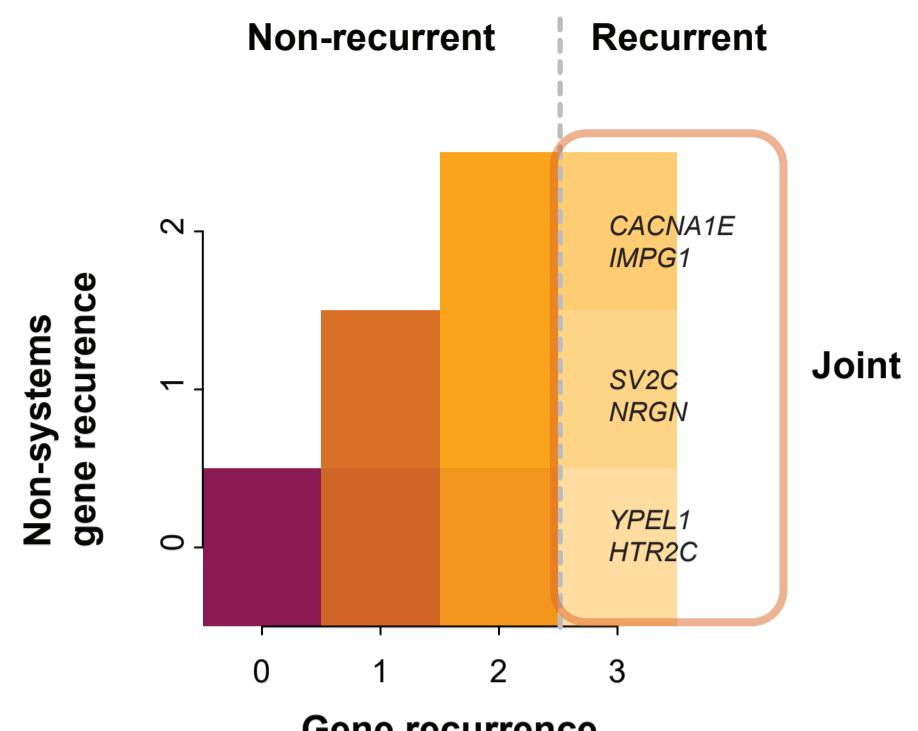
bioRxiv preprint doi: <https://doi.org/10.1101/128439>; this version posted December 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**C Gene-level assessment****Genes differentially expressed****Systems biology assessment****D**



**A****B****C Upregulated genes****D****E**



**A****Huntington's disease****Upregulated genes****Downregulated genes****B****Parkinson's disease**

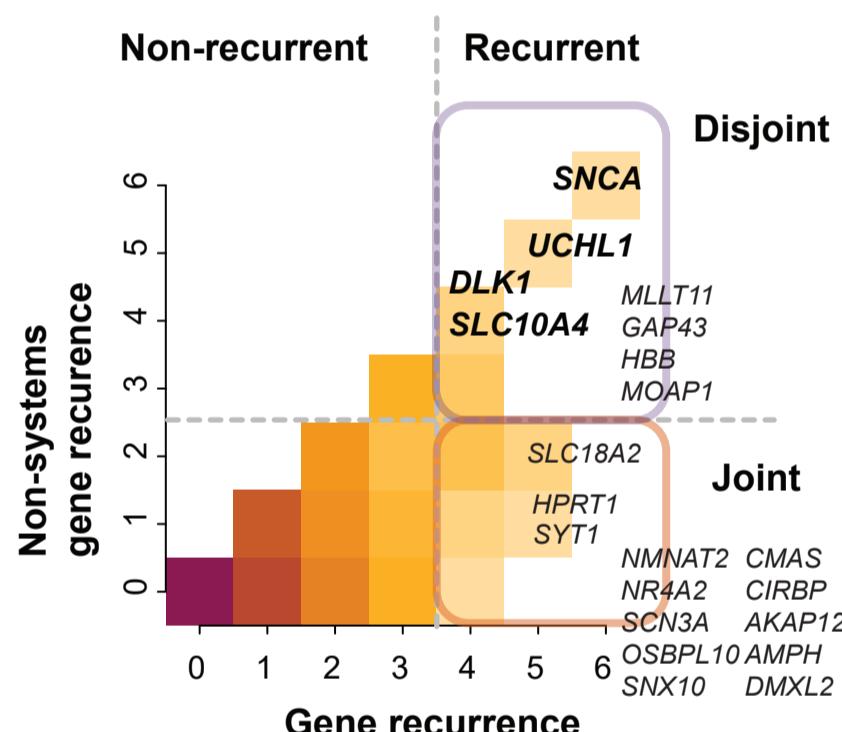
bioRxiv preprint doi: <https://doi.org/10.1101/128439>; this version posted December 26, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Downregulated genes**

**Overlap of recurrent genes**

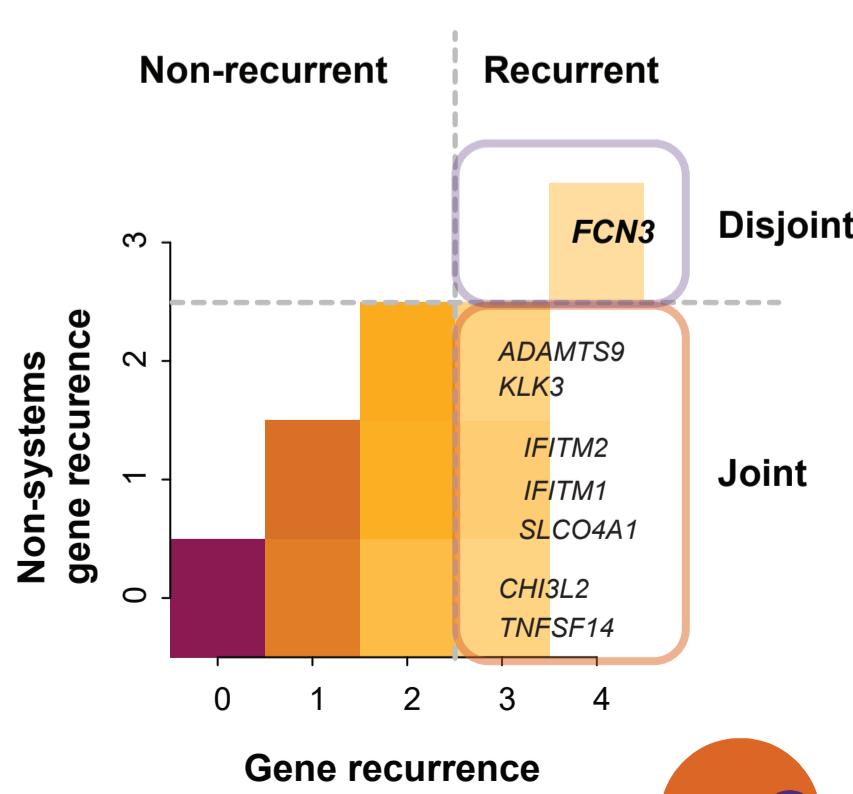
Joint signals (orange)

Disjoint signals (purple)



Frequency (log10)

1 10 10000

**C****Schizophrenia****Upregulated genes****Downregulated genes**