A COMMUNITY OVERLAP STRATEGY REVEALS CENTRAL GENES AND NETWORKS IN HEART FAILURE

AUTHORS

Pablo Cordero[1,2,*], Ayca Erbilgin[1,*], Ching Shang[1], Michael P Morley[3], Matthew Wheeler[1], Frederick Dewey[1], Kevin S Smith[4], Ray Hu[3], Jeffrey Brandimarto[3], Yichuan Liu[6,7], Mingyao Li[6], Hongzhe Li[6], Scott Ritter[3], Sihai D Zhao[8], Komal S Rathi[3], Liming Qu[3], Avinash Das[7], Stephen Montgomery[4], Sridhar Hannenhalli[9], Christine S Moravec[10], Wilson H Tang[10], Kenneth B Margulies[3], Thomas P Cappola[3], Euan A Ashley[1,2,11]

[1] Stanford Center for Inherited Cardiovascular Disease, [2] Program of Biomedical Informatics, [4] Department of Pathology, and [11] Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305
[3] Cardiovascular Institute and Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[6] Department of Biostatistics and Epidemiology, University of Pennsylvania
[7] Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, Philadelphia, PA 19104
[8] Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820
[9] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742
[10] Department of Cardiovascular Medicine, Cleveland Clinic, Cleveland, OH 44195

*Contributed equally

ABSTRACT

Heart failure is one of the leading causes of mortality worldwide, but its underlying molecular mechanisms are poorly understood. To obtain a systems view of the molecular networks that underlie heart failure, we harvested 1352 samples from 313 healthy and failing hearts directly from transplant operating rooms and obtained left-ventricular whole-genome gene expression and genotype measurements. From these data, we built directed regulatory gene networks and gene communities using an approach that combines network and community inference in one framework. Differences in co-expression and global and local centrality parameters pinpointed changes in the molecular interaction network associated with heart failure, as well as its network-wise genetic determinants. Connectivity of one gene, *PPP1R3A*, previously unassociated with heart failure, changed significantly between healthy and diseased states. Perturbation of *in vitro* and *in vivo* systems via time series transcriptome sequencing and murine cardiovascular phenotyping revealed that ablation of *PPP1R3A* alters disease progression.

Heart failure (HF) is a life-threatening syndrome characterized by an incapacity of the heart to meet the body's metabolic demands. It can result from ischemic, genetic, valvular, infectious, or toxic etiologies and costs the US medical system more than $34 billion a year to treat 6 million people (*1*), (*2*). Despite this, the underlying molecular mechanisms remain poorly understood with the few therapeutics available targeted towards secondary mechanisms (*3*, *4*). Studies have focused predominantly on determining gene expression signatures in small human groups or have used animal models to closely examine changes in known pathways (*5*–*11*) – the lack of expansive cardiac gene expression measurements in sufficient numbers of independent human tissues, especially for non-failing hearts, has precluded building detailed transcriptional maps and comparative analysis of the failing heart.

Here, we report the results of a comparative systems modeling of heart failure from genome-wide gene expression and genotyping measurements of 1352 different cardiac samples obtained in 313 human hearts, including 136 non-failure donor controls. These samples were collected directly from the explanted cardiac tissue in operating rooms of four transplant centers at the time of surgery using consistent and stringent tissue preservation procedures to mitigate degradation due to high oxidative and nitrosative stress (see Supporting Methods).

We devised a strategy, the Community LASSO with Intersecting Priors (CLIP, Figures S1 and S2 and Supporting Methods), to jointly infer gene co-expression network and community structure using a dense community overlap model that follows network/community structure relationships observed in gene interaction and pathway databases and ontologies (see Figure S3A and Supporting information). Using this method, we built undirected networks based on conditional independences for the failing and non-failing hearts, with confidence estimates obtained via jack-knifing (Figure 1, see Supporting Methods and Figure S4 for details on model selection). The non-failing network exhibited greater complexity, with an average of more than two times more edges (26833) than the failing network (10780) based on jackknife replicates and a 2-fold increase in the number of communities in the controls network when compared to the cases network (138 for cases, 339 for controls, Figure 1A). Pathway and annotation enrichment analysis revealed that the failing heart communities were significantly enriched for disease-specific pathways and terms, such as the hypertrophic and dilated cardiomyopathy-associated terms in OMIM and GO (Figure S5).

We leveraged genome wide genotypes to find network-wise genetic controllers for each community. We regressed the first principle component of each community's expression matrix (the so-called eigengene) against genotype using an additive model (*12*) with a conservative p-value cutoff of 10^{-10}. The failing hearts network had a larger number of loci correlating with the expression of multiple communities than the healthy network, despite having fewer edges and communities (see Figure 2A and B). SNPs correlating to communities of the failing heart network were more enriched with regulatory annotations according to hypergeometric enrichment tests using RegulomeDB (Figure S6 and Supporting Methods). These genetic controllers were further used to refine and partially-direct the network (see Supporting Methods).

We found that the rs16957898 SNP was linked to the expression of most communities in the failing heart network (max. p-value of significant associations to community eigengenes of 7.20×10^{-11} regulating 65 different communities, Figure 2D) – an association that held when permuting genotype while maintaining eigengene expression structure (*13*). The rs16957898 SNP lies in the first intron of the *CDH13* gene (encoding the cardio-protective cardiac T-cadherin receptor that binds adiponectin [31], [32]) and has evidence for transcription factor binding of more than 10 transcription factors by the ENCODE consortium). Communities linked to this SNP did not include *CDH13*, but 7 did include the neuronal T-cadherin gene, *CDH2* (see the union of the top five communities that were more significantly controlled by rs16957898 in Figure 2E), suggesting a *trans*-like regulation by this SNP.

Next, we used topological criteria to identify and prioritize genes that were differentially connected between the failing and non-failing heart networks: local connectivity (*lc*) that was dependent on a gene's number of edges and global connectivity (*gc*) that was calculated using the gene's community membership (see discussion of these metrics in the Supporting Information). Using *lc* and *gc* , we defined network roles for each gene (*14*). Non-hubs (N-hubs) were genes with low *gc* and *lc* that include genes with low co-expression with other genes. Local hubs (L-hubs) had a high *lc* but low gc, possibly indicating strong pathway specificity and importance to a set of neighboring genes, but that are uncoupled from the global network. Sparse hubs (S-hubs) were genes with a high *gc* but a low *lc*, belonging to a large number of communities and but without many direct connections. Finally, central hubs (C-hubs) had both a high *gc* and *lc* and belong to densely-connected communities and directly connected to a large number of genes.

We plotted gene-wise local and global connectivity metrics against each other in a scatter plot (see Figure 3A) to reveal genes in these network roles (N-hubs, lower-right quadrant; L-hubs, upper-left quadrant; S-hubs, lower-right quadrants; and C-hubs, upper-right quadrant). We prioritized genes involved in heart failure by focusing on those that gained significant global and/or local connectivity from non-failing to failing conditions (i.e. that changed roles from N-hub to L/S/C-hub or from L/S-hub to C-hub, see Figure 3C). This gave an informative list of genes that transitioned towards centrality in the disease state, are enriched in OMIM/KEGG cardiomyopathy terms and pathways (hypertrophic and dilated cardiomyopathy KEGG pathway and OMIM terms, test p-values < 0.001), and include the myosin heavy chain 7 *MYH7*, myosin binding protein C3 *MYBPC3*, cardiac troponin 2 *TNNT2*, and nebulette *NEBL* – genes that have previously been implicated in the Mendelian cardiac muscle diseases, hypertrophic cardiomyopathy and dilated cardiomyopathy (*15*, *16*). In contrast, prioritization by differential expression did not include many genes genetically associated with cardiovascular disease (1 of the top 20 were associated with cardiac pathologies compared to 7 out of the top 20 for the connectivity-derived list – a significant enrichment difference [Fisher exact test p-value < 0.001]). This global and local connectivity strategy was also better in pinpointing genes related to cardiovascular disease than other metrics that leveraged causal directionality and edge sign (see Supporting Results and Figure S7)

Among those genes whose network topology was maximally changed between failing and non-failing human heart tissue was protein phosphatase 1 regulatory subunit 3A (*PPP1R3A*; the top gene changing from L-hub in the non-failing network to a C-hub in the failing network), which encodes a muscle-specific regulatory subunit of protein phosphatase 1 (PP1) (*17*) and has not been previously associated with heart failure. Further inspection revealed that the union of the failing heart communities containing *PPP1R3A* included 11 out of the top 20 of the failing heart C-hubs (see Figure S8). Since *PPP1R3A* contains a glycogen-binding domain (*18*) and is thought to promote glycogen synthesis (*19*)(*20*), a central process in the heart's metabolism, we hypothesized that this gene would have an important transitional role in healthy and failing hearts.

To test whether *PPP1R3A* played a role in pathological cardiac remodeling, we determined the effect of perturbing *PPP1R3A* expression via RNA silencing (Figure 4A-C). We performed RNA sequencing in a standard cellular model of cardiac remodeling (stimulation of cell hypertrophy

through an alpha1-adrenergic receptor agonist, phenylephrine, in ventricular myocytes (*21*)). Two groups of cultured neonatal rat ventricular myocytes (NRVM) were phenylephrine-treated (48 hours after isolation) or untreated, respectively, and further split into four time point groups, measured in hours after isolation (Figure 4A): 36 (T0), 48 (T1), 72 (T2), and 96 (T3). Transcriptome sequencing was then performed in all groups (each group having at least 3 biological replicates each). The expression of genes in communities that contained *PPP1R3A* were significantly altered and exhibited a trend towards decreasing expression (Figure 4B), evidence of a "network knockdown". Under normal cell culture conditions, knockdown of this gene induced significant down-regulation of critical fatty acid metabolism genes, such as the pyruvate dehydrogenase *PDK2*, and the carnitine palmitoyltransferase *CPT1B* (Figure 4C). Under phenylephrine-treated conditions, *PPP1R3A* knockdown resulted in the alteration of genes critical to glucose metabolism: glucose transporters *GLUT1* and *GLUT4* (up-regulation and down-regulation, respectively, see Figure 4C). These measurements also allowed us to establish the robustness of the gene communities inferred (see Supporting Information and Figure S9) and to highlight portions of the failing and non-failing human networks that were consistent with the *in vitro* temporal expression patterns (see Figure S10).

To test whether *PPP1R3A* influenced failing cardiac phenotypes, we interrogated the effects of *PPP1R3A* knockout (*20*) in an *in vivo* mouse model of heart failure (Figure 4D-G). Wild-type and *PPP1R3A* null mice (all male, age 12 weeks) were exposed to a neurohormonal model of heart failure including 45 mg/kg/day isoproterenol and 1.6 mg/kg/day angiotensin II via continuous infusion for 10 weeks (8 mice per group). The metabolic gene program that was observed to be perturbed by knockdown *in vitro* which included the *GLUT1*, *CPT1B*, and *GYS1* genes was disrupted by knockout in the context of early heart failure *in vivo* (expression of *GLUT1, GLUT4, GYS1* is upregulated in the *PPP1R3A*-/- Figure 4B consistent with Figure 4F), following a consistent path of short and long term metabolic switch similar to the one observed with *GLUT1* overexpression in previous studies. We assessed both cardiac function through ultrasound scanning of the mouse hearts *in vivo* and fibrosis through trichrome staining of post mortem short axis sections of the formalin-fixed hearts *in situ*. We found that the loss of this key gene led to a statistically significant worsening of disease markers, namely an increase in fibrosis and decrease in fractional shortening (Figures 4D-E). Similarly, key members of the *PPP1R3A* disease network, *ATP5B, ATP5H, APOBEC*, and *HRC* that followed similar gene expression patterns in the *in vitro* and in human (see Supporting Results) were found to be up-regulated in the post

mortem tissue of knockout mice with heart failure (Figures 4F and 4G). These changes, along with the *in vitro* data, suggest a cardio-protective role for *PPP1R3A*.

In summary, a systematic multi-center approach to the collection of human heart tissue allowed genome wide genotypic and expression measurements in failing and non-failing human hearts. A novel approach to network inference incorporating measures of local and global centrality identified known and novel mediators of disease. Time series perturbation analysis following knockdown of one of these novel genes revealed its central nature in the context of heart failure, something further detailed in a gene null mouse model of the disease. Together, these data provide a comprehensive transcriptome map of human heart failure and identify *PPP1R3A* as a novel target for disease abrogation.

ACKNOWLEDGEMENTS

We thank members of the Ashley lab and the Stanford Biomedical Informatics training program for comments on the manuscript. The MAGNet heart failure study has been supported through grants for the US National Institutes of Health (HL105993, HL088577, HL098283, HL113933). This work was further supported by a pre-doctoral CONACyT fellowship (to PC) and a T32 post-doctoral fellowship (to AE).

REFERENCES

1. G. Felker, K. A. Jr, The problem of decompensated heart failure: nomenclature, classification, and risk stratification. *Am. Heart J.* **145**, S18–S25 (2003).

2. J. McMurray, S. Stewart, The burden of heart failure. *Eur. Hear. Journal-Supplements*. **4**, D50 (2002).

3. D. L. Mann, A. Deswal, B. Bozkurt, G. Torre-Amione, New therapeutics for chronic heart failure. *Annu. Rev. Med.* **53**, 59–74 (2002).

4. M. D. Murray *et al.*, Factors associated with exacerbation of heart failure include treatment adherence and health literacy skills. *Clin. Pharmacol. Ther.* **85**, 651–8 (2009).

5. M. Chen, E. Ashley, D. Deng, A. Tsalenko, Novel role for the potent endogenous inotrope apelin in human cardiac dysfunction. *Circulation*. **108**, 1432–1439 (2003).

6. M. Kittleson, K. Minhas, Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure. *Physiol. Genomics*. **21**, 299–307 (2005).

7. E. Ashley, R. Ferrara, J. King, A. Vailaya, Network analysis of human in-stent restenosis. *Circulation*. **114**, 2644–2654 (2006).

8. S. V Naga Prasad *et al.*, Unique microRNA profile in end-stage heart failure indicates alterations in specific cardiovascular signaling networks. *J. Biol. Chem.* **284**, 27487–99 (2009).

9. A. Tijsen, E. Creemers, MiR423-5p as a circulating biomarker for heart failure. *Circ. Res.* **106**, 1035–1039 (2010).

10. S. Greco, A. Perfetti, Heart failure modulates long noncoding RNA expression in human left ventricles. *Eur. Heart J.* **34**, P3250 (2013).

11. P. Han *et al.*, A long noncoding RNA protects the heart from pathological hypertrophy. *Nature*. **514**, 102–106 (2014).

12. A. Shabalin, Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. **28**, 1353–1358 (2012).

13. R. Breitling *et al.*, Genetical Genomics: Spotlight on QTL Hotspots. *PLoS Genet.* **4**, e1000232 (2008).

14. J. Scripps, P. Tan, A. Esfahanian, Node roles and community structure in networks. *Proc. 9th WebKDD 1st SNA-KDD 2007 Work. Web Min. Soc. Netw. Anal.* (2007).

15. L. Alday, E. Moreyra, *Hypertrophic Cardiomyopathy in Infants and Children*.

16. M. Maron, J. Finley, J. Bos, Prevalence, clinical significance, and natural history of left ventricular apical aneurysms in hypertrophic cardiomyopathy. *Circulation* (2008).

17. M. J. Hubbard, P. Cohen, Regulation of protein phosphatase-1G from rabbit/skeletal muscle. 1. Phosphorylation by cAMP-dependent protein kinase at site 2 releases catalytic subunit from the glycogen-bound holoenzyme. *Eur. J. Biochem.* **186**, 701–709 (1989).

18. C. B. Newgard, M. J. Brady, R. M. O'Doherty, A. R. Saltiel, Organizing glucose disposal: Emerging roles of the glycogen targeting subunits of protein phosphatase-1. *Diabetes*. **49** (2000), pp. 1967–1977.

19. M. P. Egloff *et al.*, Structural basis for the recognition of regulatory subunits by the catalytic subunit of protein phosphatase 1. *EMBO J.* **16**, 1876–1887 (1997).

20. Y. Suzuki *et al.*, Insulin control of glycogen metabolism in knockout mice lacking the muscle-specific protein phosphatase PP1G/RGL. *Mol. Cell. Biol.* **21** (2001), doi:10.1128/MCB.21.8.2683-2694.2001.

21. D. Eble, M. Qi, Contractile activity is required for sarcomeric assembly in phenylephrine-induced cardiac myocyte hypertrophy. *Am. J. …* (1998).

FIGURES

Figure 1: Global view of healthy and failing heart transcriptional landscapes (A) Node degree and community statistics for failing (red colors) and non-failing heart networks (gray colors). Degree (left) and community membership (right) distribution of the failing (red) and non-failing (grey) networks. Error bars are jackknifing standard deviations. (B) Total number of edges (top) and number of communities (bottom) of both networks. (C) Global view of the heart failure diseased (left, red) and healthy (right, gray) inferred networks with communities collapsed into

subnetworks (represented as nodes here) and labeled with their most common annotations. Arrows indicate connections between each part of the network, with thickness proportional to the number of edges between each subnetwork. Subnetworks of some of these parts of the network are presented in more detail as insets.

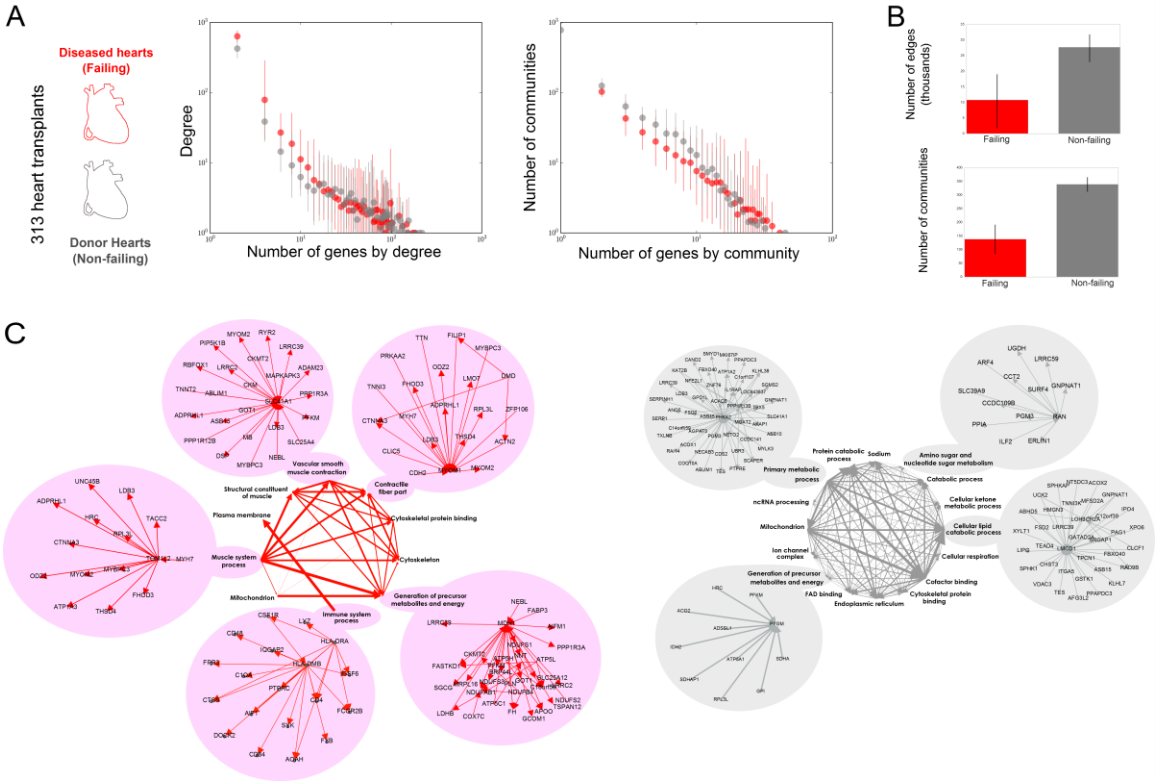


Figure 2: Controlling loci hotspots of gene community expression in healthy and failing hearts. (A) The genomic controllers for the 20 largest healthy modules. Red balls are communities with size proportional to the community's number of genes. Its position on the Y-axis is the genomic location of its most connected gene. Gray and black balls represent the controllers of each community, plotted in their genome coordinate along the X axis, size proportional to the association strength in $\log_{10}$ scale of their regression p-value. A yellow arc is drawn between communities if it they have at least one gene in common. (B) Genomic controllers for the 20 largest communities in the failing heart network (A). (C) SNPs that were found to be or are in linkage disequilibrium with statistically significant controllers of several modules in the non-failing network (X-axis: number of modules that the SNP healthy, Y-axis: regulatory impact as assessed by RegulomeDB; color map: effect size (r^2)). A SNP associated with 5 communities and with high evidence for regulatory impact and effect size is located in the *SOS1*. (D) Genetic markers correlating with community expression in the failing heart networks. A SNP in an intronic region of *CDH13* regulates 65 communities that include several genes involved in cardiac remodeling, cell-to-cell adhesion, and glycogen metabolism. (E) Subnetwork resulting from the top 5 communities that are most significantly controlled by rs16957898, which include the *CDH2* neuronal T-cadherin gene.

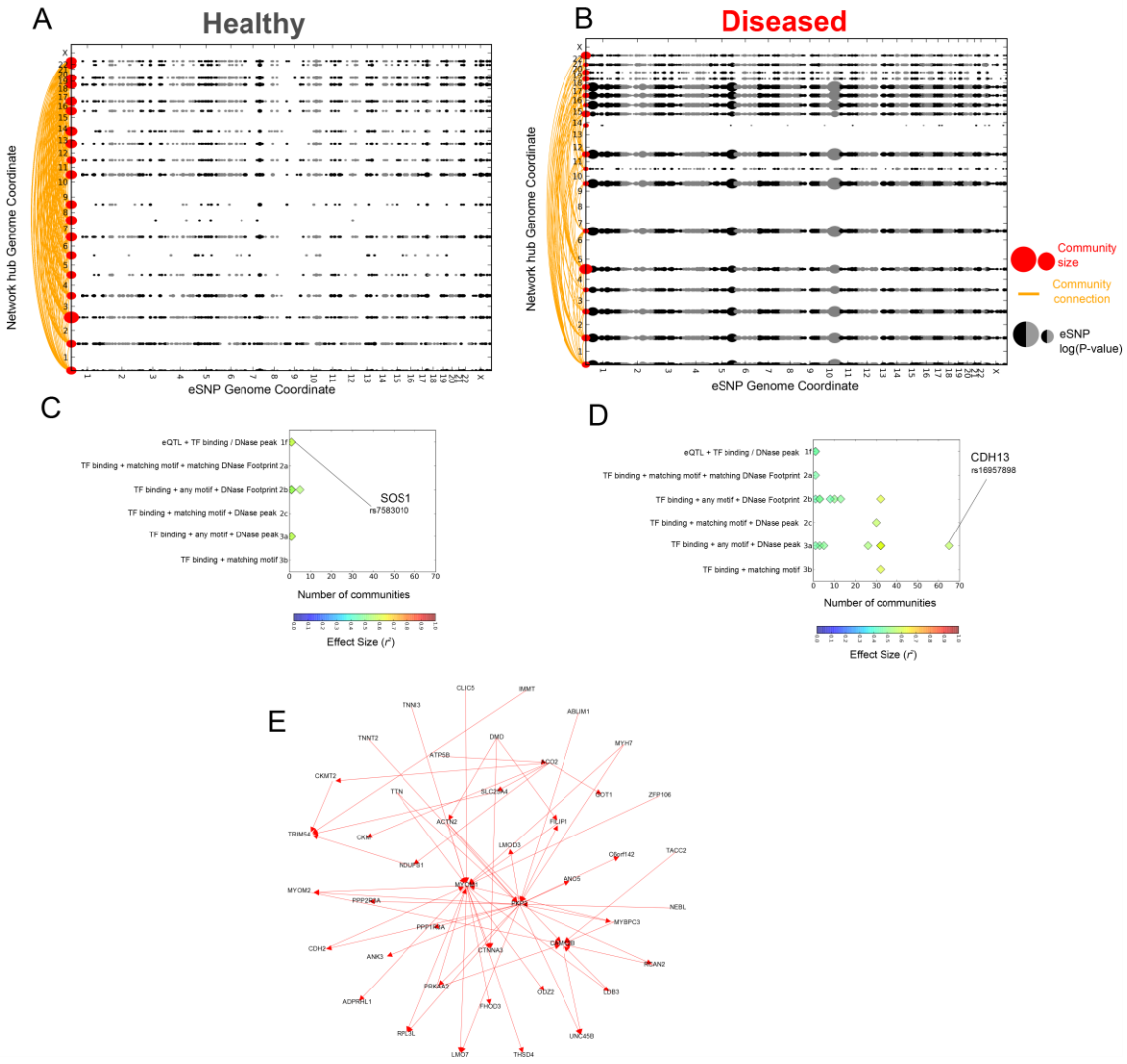

A
Healthy
B
Diseased
Network hub Genome Coordinate
eSNP Genome Coordinate
Community size
Community connection
eSNP log(P-value)
C
D
eQTL + TF binding / DNase peak 1f
TF binding + matching motif + matching DNase Footprint 2a
TF binding + any motif + DNase Footprint 2b
TF binding + matching motif + DNase peak 2c
TF binding + any motif + DNase peak 3a
TF binding + matching motif 3b
SOS1
rs7583010
CDH13
rs16957898
Number of communities
Effect Size (r^2)
E

Figure 3: Local and global centrality metrics define changes in gene role between healthy and failing states. (A) Depiction of gene roles, global connectivity (*gc,* X-axis) and local connectivity (*lc,* Y-axis). (B) Gene roles as defined by *gc* versus *lc* plots in the non-failing heart network (gray) and in the failing heart network. (C) The top 10 genes for each gene role change gaining connectivity from the non-failing to failing heart network. Values for *lc* and *gc* of these genes in the non-failing (light green and light blue) and in the failing heart network (dark green and dark blue). This prioritization strategy yields a gene list that is significantly enriched for genes with genetic associations with cardiomyopathies in contrast with a gene list derived from the top differentially expressed genes between conditions (p-val < 0.001, Fisher exact test).

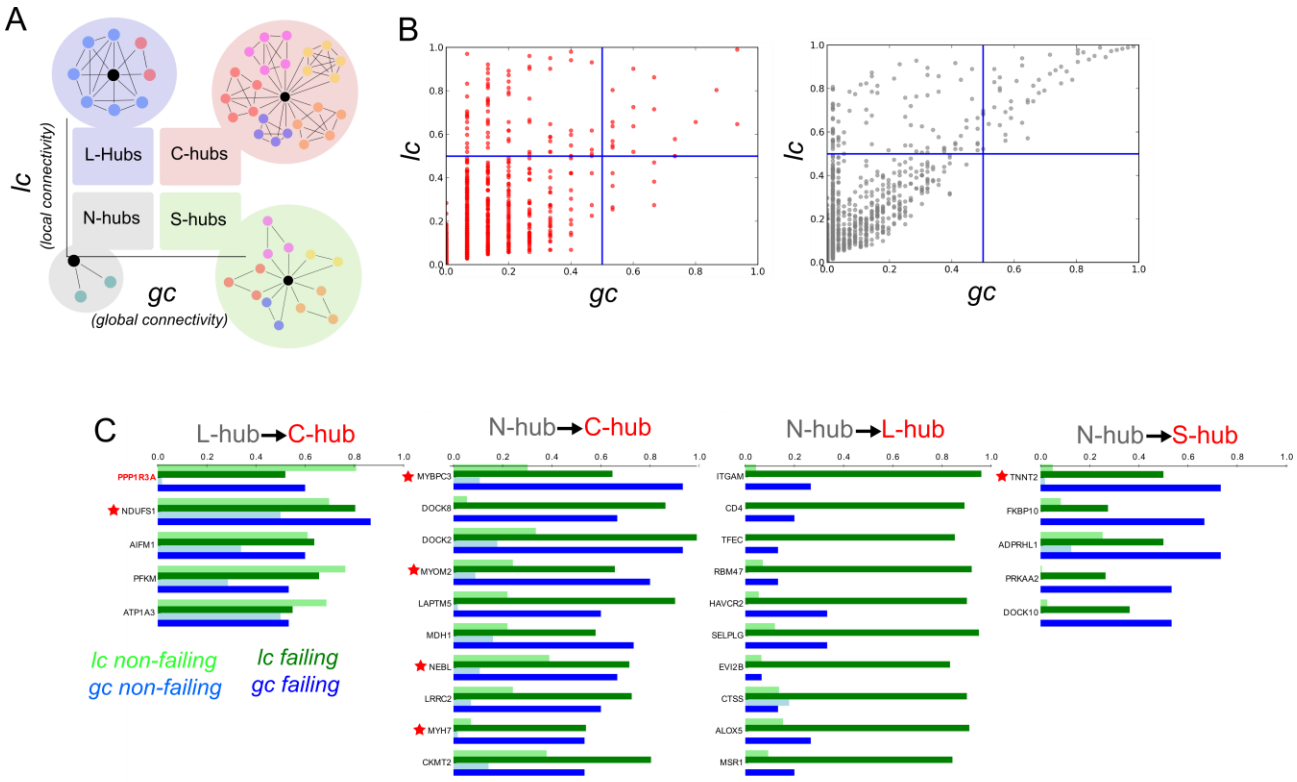
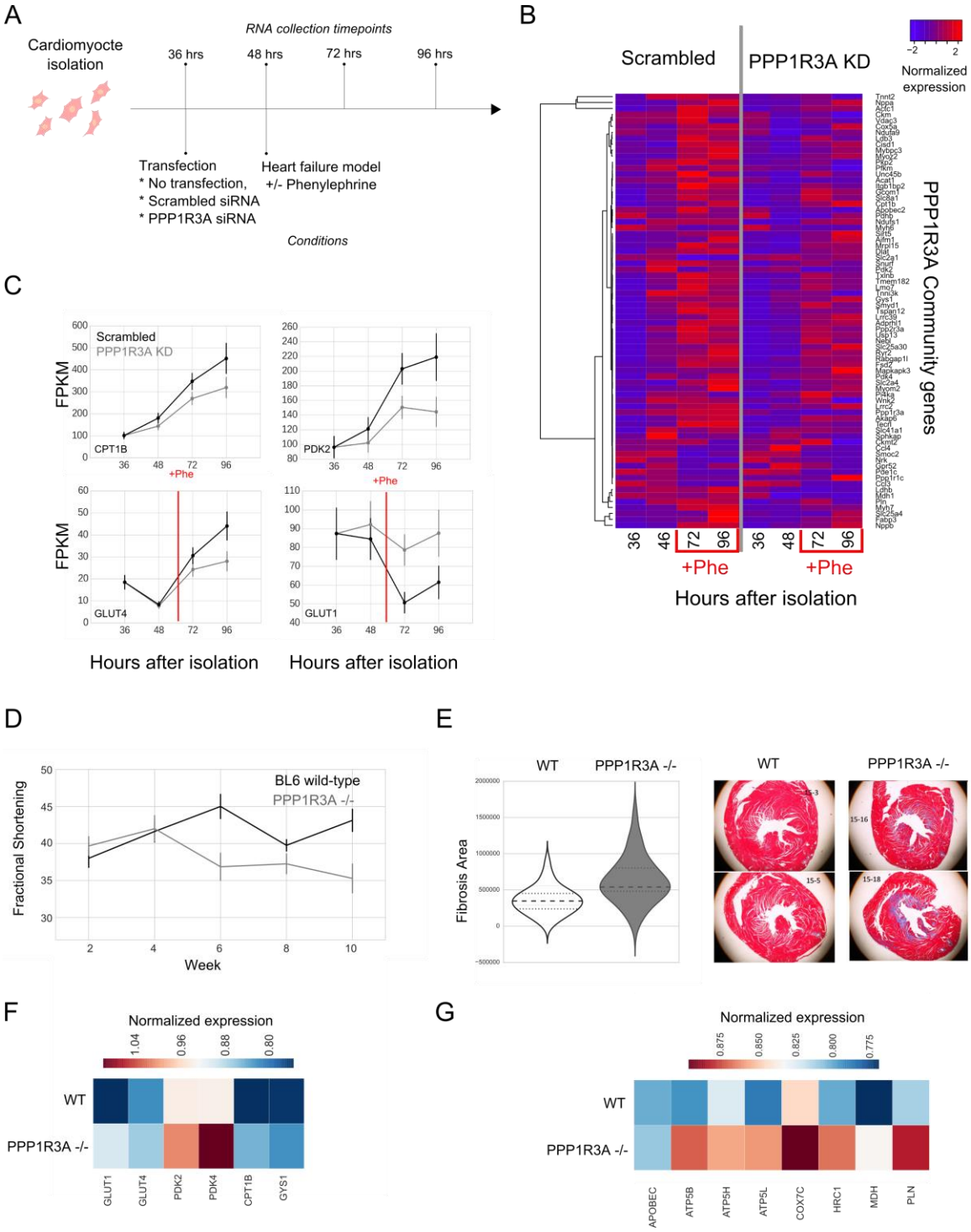


Figure 4: Cardiac and heart failure effects of *PPP1R3A* ablation *in vitro* and *in vivo*

Experimental design and analysis pipeline for whole-transcriptome, time-course measurements of an *in vitro* model of heart failure to test network-derived hypotheses. (A) Neonatal rat ventricular myocytes (NRVMs) are isolated and split into several groups: phenylephrine treatment induces a heart failure phenotype while *PPP1R3A* siRNA transfection perturbs a highly connected gene in the human cardiac networks. RNA gene expression is measured at 36, 48, 72, and 96 hours after isolation. (B) Clustered heatmap of normalized gene expression values (blue, lowly-expressed to red, highly-expressed) of a set of genes that include top predicted C-hubs from the failing hearts network, natriuretic peptide genes, and genes in the *PPP1R3A* communities at several time points after isolation under hypertrophic stimulation and scrambled/*PPP1R3A* knockdown conditions. (D) Expression level changes (X-axis) of the heart metabolic gene program as a function of time (Y-axis) and *PPP1R3A* knockdown conditions (scrambled siRNA conditions in black, *PPP1R3A* siRNA condition in gray, error bars represent standard deviations). Significant changes occur over time between *PPP1R3A* knockdown and scrambled healthy in the *PDK2* and *CPT1B* genes under normal conditions and in glucose transporters *GLUT1* and *GLUT4* under failure (phenylephrine) conditions. (C) Fractional shortening as measured through echocardiograms in each of the 10 weeks of continuous treatment of isoproterenol and angiotensin II treatment in WT mice (gray) and *PPP1R3A*$^{-/-}$ mice (black) – *PPP1R3A*$^{-/-}$ mice have significantly worse fractional shortening progression throughout the end of the time-course. (B) Fibrosis at the end of the 10 week treatment is significantly increased in *PPP1R3A*$^{-/-}$ mice compared to WT. Together with the deteriorated fractional shortening readings, this suggests that *PPP1R3A* plays a cardioprotective role in end-stage heart failure. (D) *PPP1R3A*$^{-/-}$ mice have the glucose metabolism program up-regulated when compared to WT, as assessed by *GLUT1, GLUT4, PDK2, PDK4, CPT1B*, and *GYS1* expression. (E) Genes from the *in vitro* time-course highlighted transcriptional network (Figure S10) are also up-regulated in KO mice with respect to WT.


A
Cardiomyocte isolation
RNA collection timepoints
36 hrs
48 hrs
72 hrs
96 hrs
Transfection
* No transfection,
* Scrambled siRNA
* PPP1R3A siRNA
Heart failure model
+/- Phenylephrine
Conditions
B
Scrambled
PPP1R3A KD
Normalized expression
PPP1R3A Community genes
+Phe
+Phe
Hours after isolation
C
Scrambled
PPP1R3A KD
FPKM
CPT1B
PDK2
GLUT4
GLUT1
+Phe
Hours after isolation
Hours after isolation
D
BL6 wild-type
PPP1R3A -/-
Fractional Shortening
Week
E
WT
PPP1R3A -/-
Fibrosis Area
F
Normalized expression
WT
PPP1R3A -/-
GLUT1
GLUT4
PDK2
PDK4
CPT1B
GYS1
G
Normalized expression
WT
PPP1R3A -/-
APOBEC
ATP5B
ATP5H
ATP5L
COX7C
HRC1
MDH
PLN

SUPPORTING INFORMATION FOR "A COMMUNITY OVERLAP STRATEGY REVEALS CENTRAL GENES AND NETWORKS IN HEART FAILURE"

AUTHORS

Pablo Cordero[1,2,*], Ayca Erbilgin[1,*], Ching Shang[1], Michael P Morley[3], Matthew Wheeler[1], Frederick Dewey[1], Kevin S Smith[4], Ray Hu[3], Jeffrey Brandimarto[3], Yichuan Liu[6,7], Mingyao Li[6], Hongzhe Li[6], Scott Ritter[3], Sihai D Zhao[8], Komal S Rathi[3], Liming Qu[3], Avinash Das[7], Stephen Montgomery[4], Sridhar Hannenhalli[9], Christine S Moravec[10], Wilson H Tang[10], Kenneth B Margulies[3], Thomas P Cappola[3], Euan A Ashley[1,2,11]

[1] Stanford Center for Inherited Cardiovascular Disease, [2] Program of Biomedical Informatics, [4] Department of Pathology, and [11] Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305
[3] Cardiovascular Institute and Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[6] Department of Biostatistics and Epidemiology, University of Pennsylvania
[7] Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, Philadelphia, PA 19104
[8] Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820
[9] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742
[10] Department of Cardiovascular Medicine, Cleveland Clinic, Cleveland, OH 44195

*Contributed equally

This document contains:

- Supporting Methods
- Supporting Results
- Data and Source Code Availability
- Supporting References
- Supporting Figures
- Supporting Tables

SUPPORTING METHODS

Quality control, data pre-processing, covariate correction, and differential expression

The MAGNet consortium coordinates a multi-institute collaboration of laboratories and clinics that seek to provide a comprehensive view of the genetic determinants of heart failure. The consortium put in place a series of best practices for procurement of explanted cardiac transplant tissue including harvesting explanted cardiac tissue at the time of cardiac surgery from subjects with heart failure undergoing transplantation and from unused donor hearts. Hearts are perfused with cold cardioplegia solution prior to cardiectomy to arrest contraction and prevent ischemic damage, and explanted cardiac tissue specimens are frozen in liquid nitrogen. DNA and RNA was extracted from the collected samples using the Gentra Puregene Tissue Kit (Qiagen) and the miRNeasy Kit (Qiagen), respectively. DNA material was then assessed for quality and quantified using the PicoGreen assay (Qiagen), whereas RNA quality was assessed with a NanoVue Plus™ spectrophotometer (GE Healthcare) and the Agilent 2100 RNA NanoChip. Sample DNA was then genotyped using an Affymetrix Genome Wide SNP Array 6.0 and gene expression was quantified via RNA hybridization with an Affymetrix Genechip ST1.1 arrays using manufacturer instructions.

Several technical and sample covariates can bias gene expression values inferred from our microarray data, such as array batch effects and individual ethnicity, gender. These covariates can greatly confound downstream analyses, resulting in false positive and negative associations and reducing the power of statistical analyses. We corrected for these biases in three ways: by normalizing the raw data using robust multiarray analysis (RMA), by applying batch normalization using ComBat (*1*), and by correcting for observed covariates (gender, age, and collection site) using robust linear regression (*2*). The Z-score normalized residuals after these corrections were then used as the gene expression values for downstream analyses. Differential expression analysis was performed using the Significance Analysis of Microarrays (SAM) (*3*). Genes shown in Figure S1B had the top 30 highest gene expression up and down regulation and were deemed significantly up or down regulated with a false discovery rate cutoff of 5%.

Network inference brief preliminaries: the gaussian graphical model (GGM) and the graphical lasso for gene network structure

Let $D \in \mathbb{R}^{p \times n}$ be a matrix of expression values with *p* genes (or features) and *n* samples. One way of modeling the data is assuming that the expression of these genes comes from *p*-variate Gaussian distribution, defined by its mean $\mu \in \mathbb{R}^p$ and symmetric positive-definite covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. The entries of the precision matrix, $\Theta = \Sigma^{-1}$, implicitly define an undirected network structure for each of the *p* genes, wherein two genes are connected if they have a non-zero precision. This type of model is known as a Gaussian Graphical Model (GGM) and its most important parameter, the precision matrix Θ, has been extensively used, both explicitly and implicitly, as a tool to analyze gene expression structure. To estimate Θ, the following log-likelihood function is maximized:

$$\log(|\Theta|) - tr(S\Theta)$$

Here, *S* is the sample covariance matrix and $tr(.)$ is the trace operator. One of the challenges that commonly arises when inferring Θ from gene expression data is accounting for the orders of magnitude less data than features (*p* >> *n*), in which case the estimation of the precision matrix is ill-posed. To address this, a regularization term that imposes sparsity on the precision matrix, reducing the number of non-zero components, is commonly added. This results in the likelihood function for the graphical LASSO (*4*):

$$\log(|\Theta|) - tr(S\Theta) - \lambda\|\Theta\|_1$$

Here, the tuning parameter λ controls the amount of model sparsity (the higher λ, the more sparse the resulting precision matrix) and is commonly obtained by cross-validation or through an information criteria (*5*). In the case where we want to compare gene expression measurements from several different conditions that may partially share an underlying gene coordination structure (e.g. our failing and non-failing cohorts, both derived from the same tissue), a joint inference of the covariance matrices for each condition may be performed by regularizing for precision differences between conditions (*5*). Formally, let $t = 1, \dots, T$ be the

different conditions and $\Theta^{(t)}$ be the precision matrix for condition t. The estimators of each precision matrix are those that maximize the likelihood function:

$$\sum_{t} \log\left(\left|\Theta^{(t)}\right|\right) - tr\left(S^{(t)}\Theta^{(t)}\right) - \lambda_1 \left\|\Theta^{(t)}\right\|_1 - \lambda_2 \sum_{t<t'} \sum_{i,j} \left|\Theta_{ij}^{(t)} - \Theta_{ij}^{(t')}\right|$$

This is known as the fused graphical LASSO because of the similarity of its group regularization function to the generalized fused LASSO. Here again, λ_2 is a tuning parameter that controls similarity across the precision matrix, resulting in similar precision matrices when λ_2 is small.

The community-based lasso with intersection priors (CLIP) model for gene co-expression

While the graphical LASSO has many desirable properties for establishing co-expression relationships, using the same sparsity prior for all edges may not match prior expectations of the structure of gene-gene interactions. Furthermore, it does not infer gene communities – obtaining these groupings has to be done *post facto*. We therefore decided to refine the network inference prior by using general relationships between currently known gene-gene interactions and gene groups. We posited that the observed gene expression values of biological systems are in most part the result of gene-gene interactions that are driven by the self-organization of genes into communities (e.g. pathways) with cohesive functions. As described in the Results section of the main text, we found that gene-gene interactions follow a general trend of dense community overlaps, regardless of whether the gene communities are defined ontologically or by pathways. We reasoned that this dense community overlap structure would serve as a good prior for network construction. Unfortunately, the communities underlying the gene expression measurements are not known *a priori*. Indeed, finding these gene communities is a central task in the analysis of gene expression that is necessary for many types of analyses such as constructing gene lists, reducing multiple testing burden, and assigning functionality to co-expression relationships. Therefore, in order to apply our dense community overlap prior, we were challenged with jointly inferring the gene community and co-expression structures.

As a solution to the joint inference of network and community structure, we developed the Community-based LASSO with Intersection Priors (CLIP), a probabilistic model of gene co-expression that decomposes the sparsity parameter of the graphical LASSO into several priors

that are set by the size of the intersection between hidden gene communities. Instead of modeling the expression values directly, CLIP models the observed precision between the expression of each pair of genes using the amount of communities that each gene pair shares, a model that is motivated by the dense overlap phenomenon observed in genes and pathways in gene-gene interaction networks. We reasoned that a gene-gene interaction will most likely result in a non-zero precision between their expression values. Therefore, in measured gene expression values with an underlying community structure, the more communities two genes share, the more likely their gene expression precision will be non-zero. We can summarize this observation in the following generative model. Assume that we know the number of communities, denoted by *K*, beforehand. Let *X* and *Y* be two genes indexed by the *i* and *j* variables respectively. We define $F_i, F_j \in \mathbb{R}^K$ as the genes' community affiliation vectors, where $F_{ic} = 1$ if and only if *i* belongs to community *c* and $F_{ic} = 0$ otherwise. We set a Laplace prior on each entry of Θ:

$$\Theta_{ij} \sim Laplace\left(0, F_i^T F_j\right)$$

$$F_i \in \{0,1\}, \forall i$$

Therefore, we can then write the probability of Θ_{ij} given the community affiliation vectors F as:

$$P\left(\Theta_{ij}|F\right) = \frac{1}{2F_i^T F_j} \exp\left(-\frac{|\Theta_{ij}|}{F_i^T F_j}\right) \qquad (1)$$

This is analogous to the model used in the BIGLAM algorithm for detecting overlapping communities given a network structure (*6*). This generative model allows the inference of gene community structure given known precisions between gene's expression variables. Modeling the data with a Gaussian graphical model (GGM) and given the community structure *C*, and letting Λ be a matrix such that $\Lambda_{ij} = 1/F_i^T F_j$, we can write the log-likelihood of the precision matrix as:

$$\log(|\Theta|) - tr(S\Theta) - \lambda_1 \sum_{i,j} \Lambda_{ij} \left|\Theta_{\mathrm{ij}}\right|_1 \qquad (2)$$

This functional form is essentially the same as the graphical LASSO, with the additional complexity of using different tuning parameters for each entry in the precision matrix, which depend on the community structure *C* (in fact, we are performing a non-negative factorization on the tuning parameters).

An advantage of modeling Θ in this manner is that we can apply other priors that further describe our expectation of the precision structure, such as minimizing the precision differences between two sets of samples (see, e.g. the joint and fused versions of the graphical lasso). For the failing and non-failing, we used a fused graphical lasso strategy to model the underlying similarities between failing and non-failing, resulting in the likelihood function:

$$\sum_{t\in\{failing,non-failing\}} \begin{array}{c} \log\left(\left|\Theta^{(t)}\right|\right) - tr\left(S^{(t)}\Theta^{(t)}\right) - \lambda_2 \sum_{i,j} \left|\Theta_{ij}^{(failing)} - \Theta_{ij}^{(non-failing)}\right| \\ -\lambda_1 \sum_{i,j} \Lambda_{ij}^{(t)} \left|\Theta_{ij}^{(t)}\right| - \sum_{i,j} \log(2\Lambda_{ij}^{(t)}) \end{array} \quad (3)$$

Here, the tuning matrices $\Lambda^{(cases)}, \Lambda^{(controls)}$ are calculated separately by optimizing (1) and the group tuning parameter λ_2 is selected by minimizing the Akaike Information Criterion (see Figure S4).

We note that CLIP's community scheme allows one gene to be part of multiple communities, a significant advantage over most popular methods for community inference, which simplify the task by allowing only one gene per community. This simplification can be problematic for model interpretation, since a gene is most of the time part of multiple pathways and the products that it codes for may perform multiple functions.

Estimation of the CLIP model parameters

Joint inference of the community and co-expression structure using the likelihood function in (3) requires estimating all the precision matrices $\Theta^{(t)}$ and the community membership vectors $F^{(t).}$ One way to achieve this inference is through hard expectation maximization (EM): by iterating between optimizing either the community vectors $F^{(t).}$ or the precision matrices $\Theta^{(t)}$ while fixing the other. Given the community membership vectors $F^{(t).}$ we maximized (3) for $\Theta^{(t)}$ using the same algorithm for the joint graphical LASSO, found in the JGL R package – the algorithm implementation was modified slightly to allow for using the multiple penalty matrices, $\Lambda^{(t)}$, in (3). When optimizing (3) for $F^{(t).}$, the function to maximize, for *(t)* being either (failing) or (non-failing), becomes:

$$-\sum_{i,j} \Lambda_{ij}^{(t)} \left|\Theta_{ij}^{(t)}\right| - \sum_{i,j} \log\left(2\Lambda_{ij}^{(t)}\right) = -\sum_{i,j} F_i^{(t)T} F_j^{(t)} \left|\Theta_{ij}^{(t)}\right| - \sum_{i,j} \log\left(2F_i^{(t)T} F_j^{(t)}\right) \quad (4)$$

$$F_i^{(t)} \in \{0,1\}$$

Since (4) is a binary factorization problem, obtaining even a local optimum for *F(t)* is not computationally feasible. Therefore, we relax the binary constraint to $F_i^{(t)} \in [0,1]$. However, even with this approximation, optimizing (4) is still potentially prohibitive since evaluating (4) and its gradient is of quadratic order on the number of genes: roughly 4×10^8 operations per iteration are required if we are performing the calculations across all known genes in the genome. We can instead maximize a lower bound on (4):

$$-\sum_{i,j} F_i^{(t)T} F_j^{(t)} \left|\Theta_{ij}^{(t)}\right| - \sum_{i,j} \log\left(2F_i^{(t)T} F_j^{(t)}\right) \geq -\sum_{i,j,\left|\Theta_{ij}^{(t)}\right| \neq 0} F_i^{(t)T} F_j^{(t)} \left|\Theta_{ij}^{(t)}\right| - \sum_i \log\left(2F_i^{(t)T}\mathbf{1}\right) \quad (5)$$

$$F_i^{(t)} \in [0,1]$$

Here, **1** is the vector with all of its entries set to one. Notice that (5), and its gradient, can be computed efficiently since $\sum_i \log\left(2F_i^{(t)T}\mathbf{1}\right)$ can be cached in advanced; also notice that this approximation requires that $F_i^{(t)} \in [0,1]$. To maximize (5) we used an adaptive stochastic gradient ascent algorithm.

In order to obtain confidence estimates on all the statistics obtained from the inferred network, we employed a jackknifing resampling strategy. We sampled without replacement 100 observations from the failing and non-failing cohorts and performed CLIP inference. We repeated this procedure 100 times to obtain a distribution of models that allowed reporting of confidence estimates in, e.g., degree and community distributions in Figure 1.

Eigengene expression QTL association and regulatory variant enrichment analysis

To obtain community-wise expression QTLs, we performed PCA on each gene community's expression matrix and obtained its eigengene (the first principle component). All gene community eigengenes explained more than 90% of the variance of the gene expression values in each community. We then regressed each eigengene against the samples' genotype using the additive model in the MatrixeQTL package (*9*) with a p-value cutoff of $10^{-10.}$ We tested for

regulatory variant enrichment in the loci that was found statistically significant by MatrixQTL by performing hypergeometric enrichment tests against the set of variants in categories 1 to 3 of RegulomeDB (*10*). We expanded each set of SNPs significantly associated with a community's eigengene by adding SNPs in likage disequilibrium with them (r > 0.8), counted the SNPs that fell into each RegulomeDB category and performed a Fisher exact test to evaluate the enrichment of RegulomeDB-annotated SNPs in loci that were associated with eigengene expression of each community. Histograms of the p-values for these tests for all community-associated SNPs are given in Figure S5. SNPs controlling the failing were significantly enriched with RegulomeDB-curated loci.

Inferring high-confidence local network directionality

Genotype measurements can be used to refine and help choose causal models from undirected networks. Because the GGM framework outlined above does not model edge directionality, we used local partial correlation testing that incorporates the genotype marker measurements to choose local causal models post-facto. For every co-expression relationship between two genes g1 and g2, given a set of genotype markers M, we considered the following models:

$$M \rightarrow g_1 \rightarrow g_2 \qquad \text{(i)}$$

$$M \rightarrow g_2 \rightarrow g_1 \qquad \text{(ii)}$$

$$g_1 \leftarrow M \rightarrow g_2 \qquad \text{(iii)}$$

Each of these models can be tested using the partial correlations that best describe the causal interactions. To choose between models, the next-best edge orientation (NEO) directionality likelihood ratio score is calculated (*11*). Letting L_i, L_{ii}, L_{iii} be the likelihood of models (i), (ii), and (iii) respectively. The directionality ratio for model (i), and analogously for models (ii) and (iii), is defined as

$$dLR_i(g_1, g_2, M) = \frac{L_i(g_1, g_2, M)}{\max\big(L_{ii}(g_1, g_2, M), L_{iii}(g_1, g_2, M)\big)}$$

The model with the highest directionality likelihood ratio is then chosen as the most likely local causal structure.

Many of the million genotype markers will not be associated with a particular set of co-expressed genes. To reduce the number of markers to include in the directionality tests of two genes g_1 and g_2, we only take markers that were significantly associated with the eigengene expression (see section above) of at least one of the communities that g_1 or g_2 form part of.

Partial correlations and other linear regression methods for testing causal models are usually performed in confirmatory rather than exploratory analyses, that is, in cases where there are specific hypothesized causal models that wish to be tested. Care must be taken when using these methods in an exploratory manner: there may be situations where the best scoring causal model is not very well supported by the data, e.g. when several casual models have very similar scores or when all causal models induce covariance matrices that are not good predictors of the observed covariance matrix. Therefore, we only declare an edge "directed" when: (1) the highest directionality likelihood ratio is greater or equal than 2 and (2) the likelihood of the best-scoring causal model is greater than 0.5.

This procedure resulted in partially-directed networks with high-confidence directionality. We observed that the percentage of edges that were successfully directed was markedly higher in the diseased HF network (71%) compared to the healthy network (58%) indicating higher confidence in the causal structures obtained.

Global and local topological parameters for gene ranking

After inferring the gene co-expression networks and gene communities for both cohorts, we calculated topological properties for each gene in each network in order to get a sense of a gene's role in the networks .To this end, we employed three metrics that measure different aspects of the inferred topological changes: the differences in global and local centrality parameters *gc* and *lc*. The global and local centrality parameters were calculated as:

$$gc(g) = |C(g)|$$

$$lc(g) = deg(g)$$

Where *C(g)* is the set of communities that *g* belongs to with size greater than 4 and $deg(g)$ is the degree of *g* (sum of their edge weights) in the network. These values are then quantile

normalized to the uniform distribution in the [0,1] interval, yielding values between 0 and 1 that can be readily compared across networks.

To define a gene's role using local and global connectivities, we partitioned the two dimensional space defined by global metrics into four quadrants, using the mean values of local and global connectivities as partitions: S-hubs, L-hubs, C-hubs, and N-hubs.

We note that while *lc* is directly proportional to node degree, *gc* is an additional metric that can be effectively used to assess a gene's importance to the network. Several metrics of network centrality, such as in-betweenness, have been employed to uncover critical genes and proteins in co-expression and protein-protein interactions (*12*). However, these metrics usually measure only local connectivity properties and are very well correlated with node degree (*13*), providing little additional information over simply counting the number of links of a node. In contrast, our global connectivity metric, *gc*, need not be correlated to local connectivity, as is shown in Figure 3B for the cases network.

Validation of gene communities through gene expression cohesion

To validate a predicted gene community using the RNA-seq timecourse data, we designed a permutation test for the time-dependent differences between gene expression values of the gene in the community. Intuitively, genes belonging to a community should have similar expression values, or present significant correlation. Since we did not have a big enough sample size to robustly calculate correlation coefficients, we used gene expression difference, summarized in the following "cohesion" statistic, *Coh*, defined for a community *C*:

$$Coh(C) = \frac{1}{|C|^2} \sum_{g_1, g_2 \in C} \sum_{t \in \{T0, T1, T2, T3\}} \left(ZE(g_1, t) - ZE(g_2, t) \right)^2$$

Here, $ZE(g, t)$ is the Z-score-normalized expression of gene *g* at timepoint *t*. If validating a community inferred from the failing hearts samples, we compared to the phenylephrine-treated cells, otherwise we compare to the un-treated group. We expected communities inferred by our CLIP algorithm to have low *Coh* values, indicating small distances between gene expression profiles across time. To assess the significance of *Coh* values for a community *C*, we sampled 10000 sets of genes with the size of *C* and calculated their *Coh* statistics. P-values (see Figure S9) were calculated as fractions of sampled sets of genes that had lower *Coh* than *C*.

Verification of network edges through search of causal paths of large effect size

To highlight causal paths stemming from the perturbed gene PPP1R3A, we looked at the size of differential gene expression trend across timepoints between PPP1R3A and scrambled transfection conditions. Starting from PPP1R3A, we fitted linear models to the timecourse expressions, in scrambled and PPP1R3A conditions separately, of each neighbor gene in the network and tested whether the slopes between transfection conditions varied more than two fold. We repeated this verification scheme for each neighbor of PPP1R3A that was verified, and so on successively following the directed edges of the network. Only genes inside a gene community that contained PPP1R3A were considered. Here again, to verify causal paths of the failing hearts network and communities, we compared to the phenylephrine-treated cells, and used the un-treated group to verify the non-failing hearts network and communities. Results of this analysis are given in the Supporting Results below and Figure S10.

Isolation, culture, perturbation, and visualization of cardiac myocytes

Cardiac myocytes were isolated from neonatal rats using standard collagenase protocols as described previously (*14*) and cultured in serum-free, glucose-free DMEM media. In order to attenuate the effects of fibroblast contamination a final concentration of 20 µM of the fibroblast inhibitor Ara-C (Sigma-Aldritch) was incorporated. At least 1 million cells were plated in a 12-well plate, corresponding to at least 70% confluency. For phenylephrine-treated cells, 50-µM of phenylephrine was added 48 hours after isolation. For the knockdown experiments, cells were transfected either with a siRNA targeted to *PPP1R3A* (Stealth siRNA, Invitrogen) or a scrambled siRNA using the RNAiMAX system (Invitrogen) according to manufacturer instructions; transfections were performed 24 hours after isolation. RNA extraction was performed using the Qiagen RNeasy kit according to manufacturer instructions and were DNAse-treated using the DNA-free RNA kit from Zymo research. CDNA was synthesized with the High-capacity cDNA reverse transcription kit from ABI and qRT-PCR assays were performed using KAPA SYBR FAST on a ViiA 7 ABI system. Primer sequences are given in Table S3.

RNA-seq measurements

After RNA extraction, RNA integrity was checked using a 2100 BioAnalyzer (Agilent); all RNA samples had an RIN of 7.0 or higher. Samples were screened for *PPP1R3A* knockdown efficiency and phenylephrine treatment using qRT-PCR prior to library construction. RNA-seq libraries were prepared using the TrueSeq Stranded mRNA kit (Illumina), according to manufacturer instruction. Libraries were barcoded, quality-checked using a 2100 BioAnalyzer and run in rapid run flow cells in a HiSeq 2500 (Illumina), producing at least 30 million paired-end reads.

Analysis of RNA-seq data

Sequencing reads were aligned to the Rattus Norvegicus rn5 UCSC reference genome using the STAR aligner (*15*). Quantification and differential expression analysis of RNA-seq data was performed using the Cufflinks package (*16*): full transcriptome assembly was performed with Cufflinks, quantified with Cuffquant, and analyzed for differential expression using Cuffdiff. All genes deemed to be significantly up or down-regulated in the main text were called as differentially expressed by Cuffdiff.

PPP1R3A knockout mice

The *PPP1R3A* -/- mice were maintained on a C57BL/6J background. 12-week old, male mice were treated with 45 mg/kg/day Isoprotenerol and 1.6 mg/kg/day Angiotensin II using an Alzet osmotic pump inserted subcutaneously for 10 weeks (each group had a total of 8 mice). Echocardiography was performed as previously described (*17*). Specifically, mice were anaesthetized using isoflurane inhalation (0.75-1.5% in oxygen) while maintaining the heart rates at 450-550 beats per minute. The fractional shortening (FS) of the left ventricle was defined as 100% x (1-LVIDs/LVIDd) where LVIDs and LVIDd were left ventricular internal dimensions at systole and diastole, respectively. After 10 weeks on treatment, mice were euthanized with isofluorane followed by cervical dislocation and hearts were dissected and frozen in liquid nitrogen. RNA was isolated from mouse hearts using the Qiagen RNeasy kit, cDNA was synthesized with the high-capacity cDNA reverse transcription kit from ABI, and qRT-PCR assays were performed using KAPA SYBR FAST on a ViiA 7 ABI system. Primer sequences are listed in the Table S3.

SUPPORTING RESULTS

CLIP's performance in an *in silico* network inference benchmark

To evaluate CLIP's ability to infer underlying gene networks of gene expression measurements, we generated an *in silico* benchmark of 10 networks. Each network was generated by sampling a set of 100 genes from the yeast gene interaction network using GeneNetWeaver (*7*) and simulating using the program's in-built differential equation expression simulator. Using these measurements, we challenged CLIP, the graphical LASSO (implemented in the JGL package), the shrinkage estimator for partial correlations (the R package pcor (*8*)), and a naive correlation inference (thresholded at correlation coefficients r=0.6) to infer the underlying networks. For the CLIP and the graphical LASSO, we optimized the sparsity parameter by choosing the one that minimized the AIC information criterion.

We evaluated the inferred networks using the area under the precision recall curve, as has been employed in the DREAM *in silico* challenges for network inference. CLIP's average performance was comparable to the joint graphical LASSO and the pcor package (see Figure S3B).

Community eigengene regression uncovers network genetic controllers

In contrast with the failing networks, the non-failing networks had a small number of statistically significant genetic controllers (Figure 3A, 3B, see also Table S1). This included, for example, rs7583010 (max. p-value of significant associations to community eigengenes of 9.8×10^{-14}, see Table 1), located in the 14th intron of the *SOS1* gene which had ENCODE-supported evidence of DNase hypersensitivity and previous evidence of transcription factor binding (RegulomeDB category 1f). *SOS1* is a critical regulator of the RAS pathway, and gain-of-function missense mutations in this gene have been found to cause Noonan syndrome (which includes cardiomyopathy) due to increased RAS and ERK activation (*18*).

Other differential topology metrics for gene prioritization

We also considered other metrics for gene prioritization using different aspects of the network. For example, in an attempt to identify genes that gain or lose a causal source role between

normal and disease, we calculated a causal flow metric: the normalized difference between gene-wise in-degree and out-degree in the failing and non-failing networks. We define the causal flow of gene *g*, *cf(g)* as:

$$cf(g) = \frac{indeg(g) - outdeg(g)}{deg(g)}$$

Here, *indeg(g)* and *outdeg(g)* are in-degree and out-degree of gene *g*, respectively and is calculated from the failing and non-failing networks separately yielding one causal flow value for each network. Genes with a large change in causal flow between positive (causal sources) to negative (causal sinks) between failing and non-failing pinpoint genes whose role is largely changed from effector to effected between conditions. For example, genes such as cardiac troponin T *TNNT2* and several ATPases such as *ATP5H* changed their causal flow from positive in the non-failing to negative in the failing (Figure S7A).

As a separate strategy for tracking gene topological changes between networks, the edge-wise l_1 difference between failing and non-failing was also calculated. For the gene *g,* this score was calculated as follows:

$$l_1(g) = \sum_{g'} |\Theta_{gg'}^{failing} - \Theta_{gg'}^{non-failing}|$$

Genes with highest $l_1(g)$ will be the ones with the most connections that change precision sign from failing to non-failing, i.e. that change roles from positive to negative regulators or vice-versa. The genes with the highest *l1*-diff metric are given in Supporting Figure S7B, include multiple transcription factors, and are enriched for nucleic acid binding GO terms as well as for genes involved in the NF- κ B response (Fisher exact test, p-value < 0.01), highlighting the regulatory genes that undergo the most dramatic change between the two networks.

The change in role highlighted by l_1 contrasts with the difference between local connectivities in failing and non-failing, which is maximized by genes that dramatically change their number of edges between networks, i.e. that go from being absent in one network to being a hub in another.

While the causal flow and *l1*-diff metrics gave interesting gene lists, they were not as enriched for genes previously associated with cardiovascular disease as the list yielded by the difference in local and global connectivity (the causal flow metric included only 1 such gene in the top 50 list, while the l1-diff metric did not include any cardiovascular disease genes).

Highlighting causal-paths consistent with time course data highlight transcriptional effects of *PPP1R3A* knockdown

To further investigate the effects of *PPP1R3A* perturbation on gene expression, we incorporated the knockdown data into our failing and non-failing human heart networks. We reasoned that we could use the large time-dependent gene expression effects to validate the relationships represented by the network. We therefore sought to highlight causal paths with high effect size that stemmed from *PPP1R3A* and therefore searched for highly-perturbed paths that originated in *PPP1R3A*. An edge between gene A and gene B was considered highly-perturbed if the slope of a linear model fitted to the gene expression values of A and B was 2-fold different between the control and *PPP1R3A* knockdown conditions. We used either the failing myocytes or the control group depending on whether the edge was part of the failing or non-failing heart network, respectively. Further, we constrained highly-perturbed edges to those with at least one highly-perturbed parent (see Figure S10A and Supporting Methods). This allowed us to highlight causal paths that were perturbed in the *in vitro* data, starting from *PPP1R3A* as a source.

The highlighted paths in the network contained *PPP1R3A* transcriptional partners that have been previously reported (Figure S10A and Figure S10B). These include phospholamban, a protein that regulates the calcium pump *SERCA2* and is de-phosphorylated by PP1 (*19*), and with which *PPP1R3A* has biophysical interaction (*20*). Other genes in this network include histidine rich calcium binding protein (*HRC*) that linked to cardiac arrhythmias (*21*, *22*)) and the malate dehydrogenase 1 (*MDH1)* gene, which is predominantly expressed in heart and skeletal muscle (*23*). *MDH1*, which has the strongest connection to *PPP1R3A* in this network, is a metabolism gene that is highly expressed during the fetal period followed by down-regulation in the adult heart (*24*), a pattern that is similar to the fetal gene pattern (*25*). The *MDH1* gene has also been shown to act as a transcriptional co-activator of p53 upon glucose deprivation (*26*), an intriguing role considering the importance of glucose metabolism in heart failure. Further down the network is the confirmed hypertrophy gene *TNNT2*. Mutations in this gene that disrupt its interaction with

myosin result in a hypertrophic phenotype, and as such this gene has been associated with dilated and hypertrophic cardiomyopathies (*27*). In addition, four of these disease network genes are differentially regulated under hypertrophy conditions in a knockout mouse model (Figure 4G).

DATA AND SOURCE CODE AVAILABILITY

Expression and genotype measurements for the human heart samples as well as clinical variables are being submitted to the Gene Expression Omnibus. Rat expression measurements are available via Amazon Web Services at http://s3.amazonaws.com/ashleylabrnaseq/timecourse_analysis

The CLIP algorithm as well as routines used to parse, correct, normalize, and analyze genotype and gene expression data are part of the genotype expression trait toolkit (gett) which is available at https://github.com/dimenwarper/gett.

REFERENCES

1. J. Schäfer, K. Strimmer, An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*. **21**, 754–64 (2005).

2. P. J. Huber, Robust Estimation of a Location Parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).

3. V. G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* **98**, 5116–21 (2001).

4. J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. **9**, 432–41 (2008).

5. P. Danaher, P. Wang, D. M. D. Witten, The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **76**, 373–397. (2014).

6. J. Yang, J. Leskovec, in *Proceedings of the sixth ACM international conference on Web search and data mining* (2013), pp. 587–596.

7. T. Schaffter, D. Marbach, D. Floreano, GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* (2011).

8. J. Schäfer, K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4**, 1175–1189 (2005).

9. A. Shabalin, Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. **28**, 1353–1358 (2012).

10. A. Boyle, E. Hong, M. Hariharan, Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).

11. J. Aten, T. Fuller, A. Lusis, S. Horvath, Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Syst. Biol.* **2**, 34 (2008).

12. M. Joy, A. Brock, D. Ingber, S. Huang, High-betweenness proteins in the yeast protein interaction network. *Biomed Res. Int.* **2005**, 96–103 (2005).

13. K. Faust, Centrality in affiliation networks. *Soc. Networks*. **19**, 157–191 (1997).

14. D. Eble, M. Qi, Contractile activity is required for sarcomeric assembly in phenylephrine-induced cardiac myocyte hypertrophy. *Am. J. Physiol. Physiol.* **274**, C1226–C1237 (1998).

15. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

16. C. Trapnell, A. Roberts, L. Goff, G. Pertea, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

17. P. Han *et al.*, A long noncoding RNA protects the heart from pathological hypertrophy. *Nature*. **514**, 102–106 (2014).

18. A. Roberts, T. Araki, K. Swanson, Germline gain-of-function mutations in SOS1 cause Noonan syndrome. *Nat.* … (2006).

19. P. Nicolaou, R. J. Hajjar, E. G. Kranias, Role of protein phosphatase-1 inhibitor-1 in cardiac physiology and pathophysiology. *J. Mol. Cell. Cardiol.* **47** (2009), pp. 365–371.

20. I. Berrebi-Bertrand *et al.*, Biophysical interaction between phospholamban and protein phosphatase 1 regulatory subunit GM. *FEBS Lett.* **439**, 224–230 (1998).

21. D. A. Arvanitis *et al.*, The Ser96Ala variant in histidine-rich calcium-binding protein is associated with life-threatening ventricular arrhythmias in idiopathic dilated cardiomyopathy. *Eur. Heart J.* **29**, 2514–2525 (2008).

22. V. P. Singh *et al.*, Abnormal calcium cycling and cardiac arrhythmias associated with the human Ser96Ala genetic variant of histidine-rich calcium-binding protein. *J. Am. Heart Assoc.* **2**, e000460 (2013).

23. T. Tanaka, J. Inazawa, Y. Nakamura, Molecular cloning and mapping of a human cDNA for cytosolic malate dehydrogenase (MDH1). *Genomics.* **32**, 128–130 (1996).

24. A. Lo *et al.*, Developmental regulation and cellular distribution of human cytosolic malate dehydrogenase (MDH1). *J. Cell. Biochem.* **94**, 763–773 (2005).

25. M. Rajabi, C. Kassiotis, P. Razeghi, H. Taegtmeyer, Return to the fetal gene program protects the stressed heart: A strong hypothesis. *Heart Fail. Rev.* **12**, 331–343 (2007).

26. S. M. Lee, J. H. Kim, E. J. Cho, H. D. Youn, A nucleocytoplasmic malate dehydrogenase regulates p53 transcriptional activity in response to metabolic stress. *Cell Death Differ.* **16**, 738–748 (2009).

27. N. Frey, M. Luedde, H. a Katus, Mechanisms of disease: hypertrophic cardiomyopathy. *Nat. Rev. Cardiol.* **9**, 91–100 (2012).

SUPPORTING FIGURES

Figure S1: Differentially-expressed genes between the failing and non-failing groups. (A) Gene expression distribution comparisons of HF expression markers *NPPA, NPPB, MYH7, MYH6, Serca2*, and *RYR2* between failing and non-failing. (B) Top 30 up-regulated and down-regulated genes between non-failing and failing.

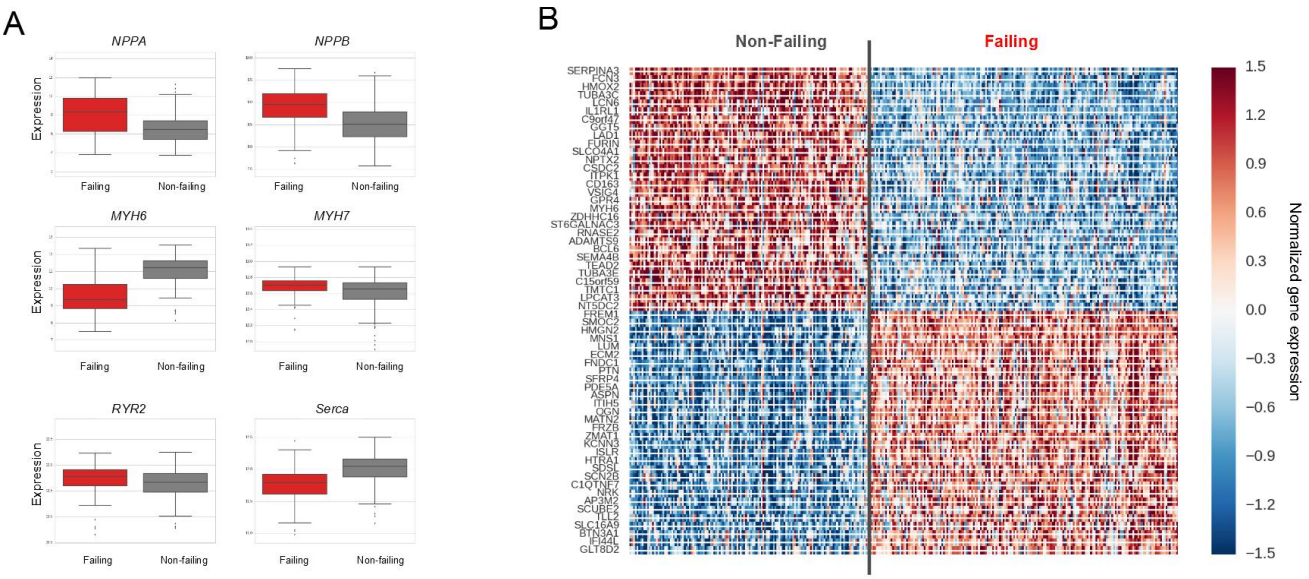


Figure S2: Network structure as a function of pathway-derived or ontology-derived communities. (A) Gene interaction network structure as a function of pathway overlaps – i.e. the more pathways two genes share the more likely they interact. The BIOGRID gene interaction network follows a dense pathway overlap structure according to the REACTOME, KEGG, PID, and BIOCARTA pathway databases. (B) Gene interaction network structure as a function of ontological term groupings; gene communities are defined by gene ontology (GO) terms whereby each term defines a community of genes that have are annotated with it. The BIOGRID interaction network also follows a dense community overlap when partitioned into GO communities. This relationship is maintained when restricting terms to cellular components and biological processes; terms in molecular function also follow the same trend, but with lower effect size.

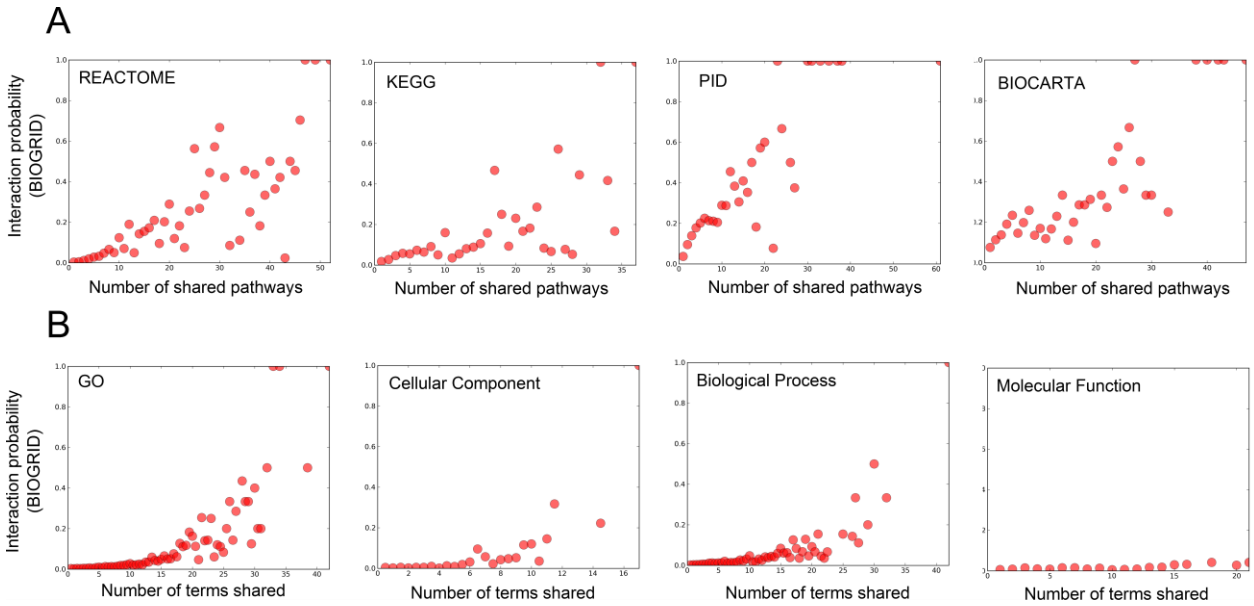


Figure S3: General framework for simultaneous calculation of community and network structure using an iterative method. We first calculate a sparse inverse covariance matrix and used the observed covariance strengths as input for inferring community structure. This community structure is then used to re-calculate the inverse covariance matrix using a LASSO type method. We iterate through this process until convergence of community and network structure. (B) Performance of CLIP against other methods in an *in silico* benchmark for network inference of 10 networks generated through the GeneNetWeaver program.

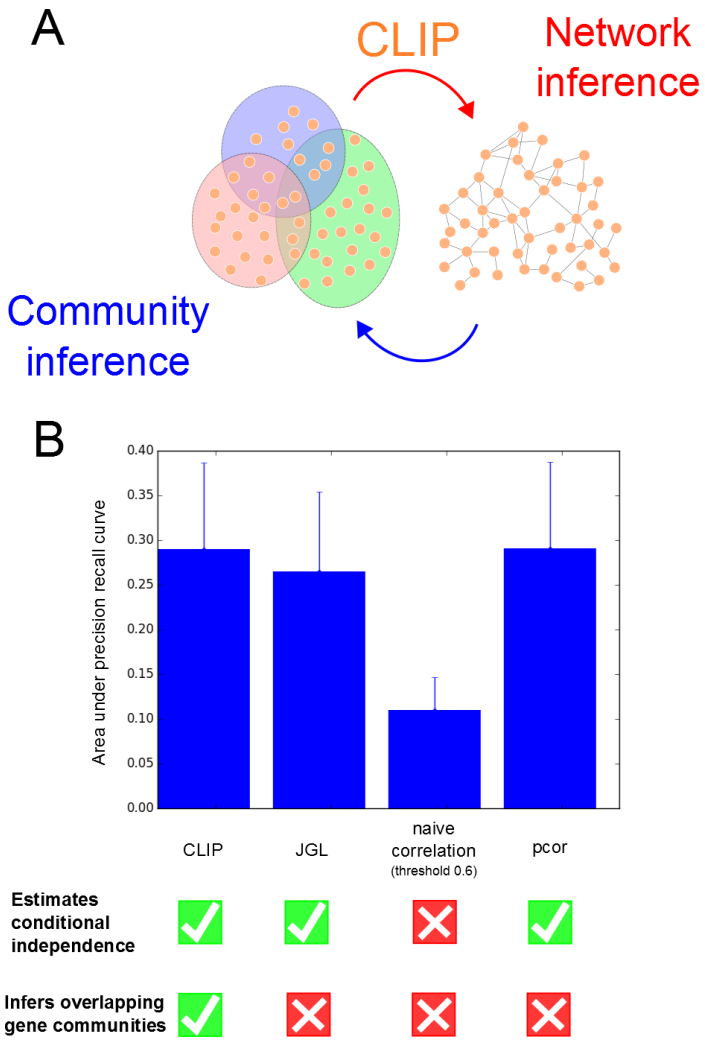


Figure S4: Choosing the tuning parameters of the model using the Akaike Information Criterion (AIC). (A) Scan of several values for λ_1 and λ_2 of the CLIP model – the model with the lowest AIC (AIC values shown as a heatmap) was chosen. (B) Number of non-zero edges between failing and non-failing as a function of the different tuning parameters.

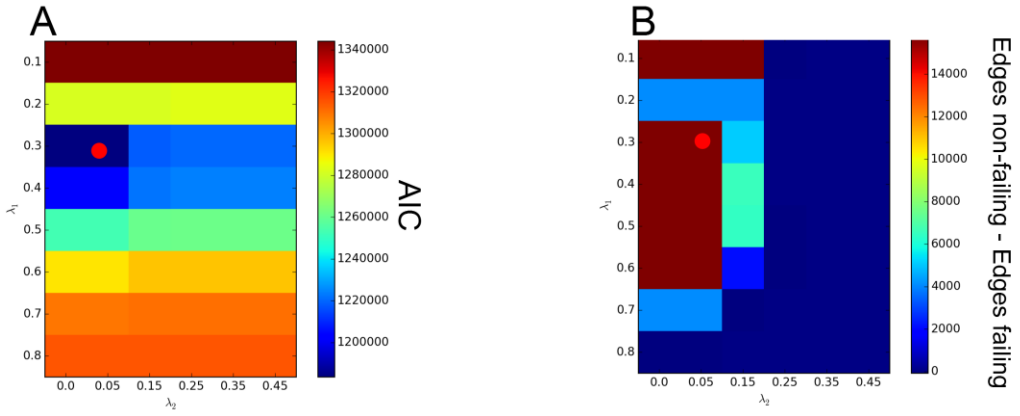


Figure S5: Ontology and pathway analysis of inferred communities in non-failing and failing groups. (A) GO terms for which most communities were enriched in failing (left; red) and non-failing (right; gray). (B) KEGG pathways for which most communities were enriched in failing and non-failing.

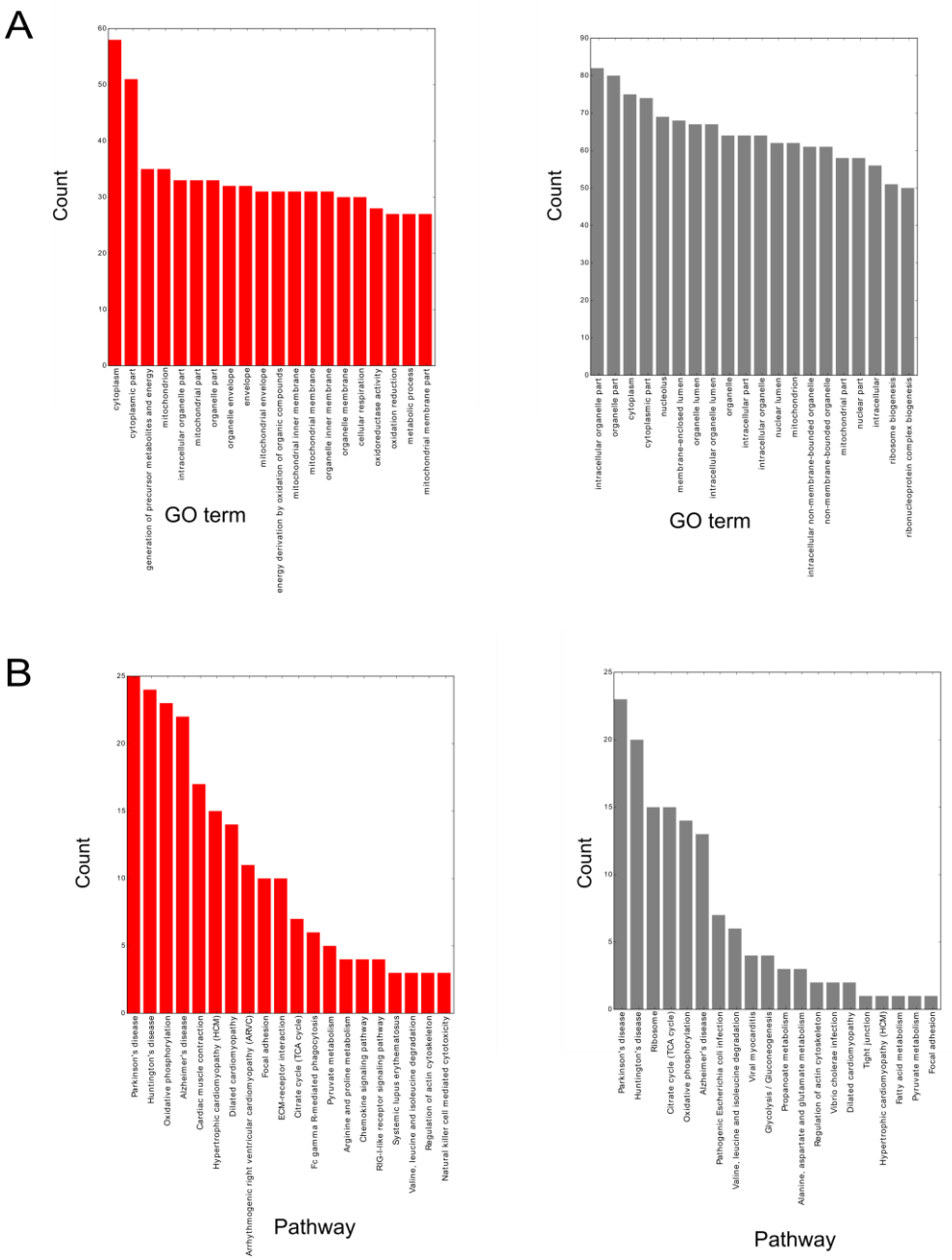


Figure S6: RegulomeDB enrichment p-value histograms, by RegulomeDB category, for genetic controllers of (A) the failing communities and (B) the donor non-failing communities. P-value distributions for the failing communities are skewed towards lower values than those for the non-failing communities.

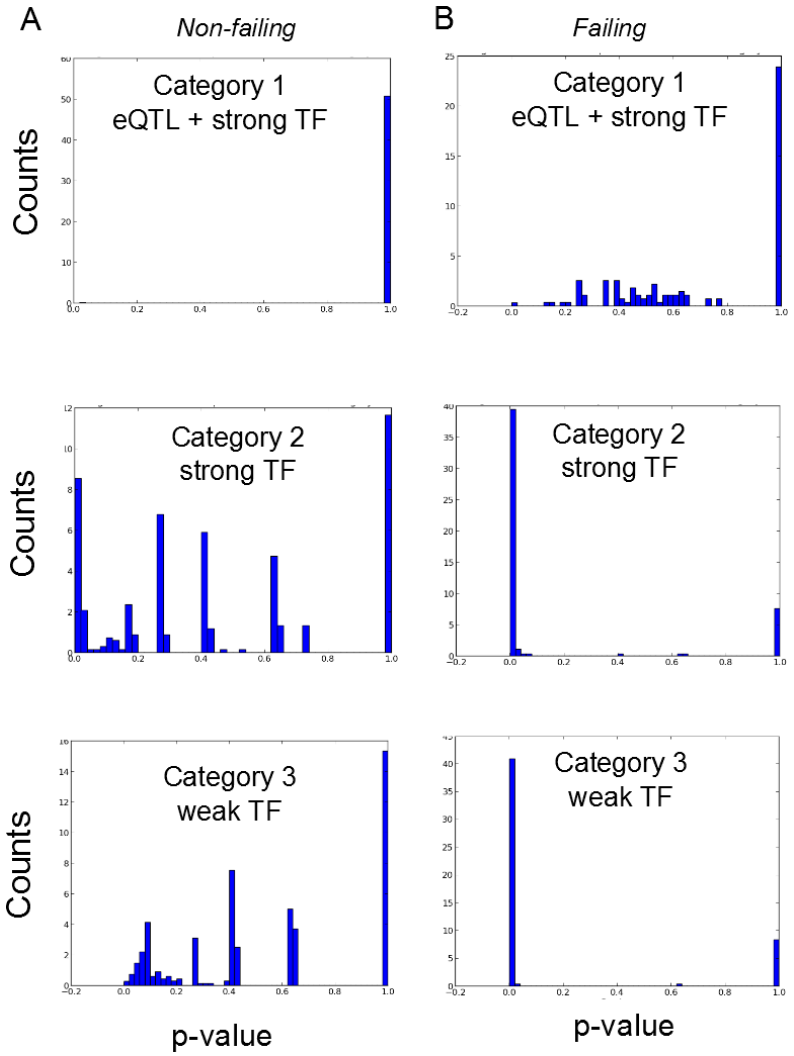
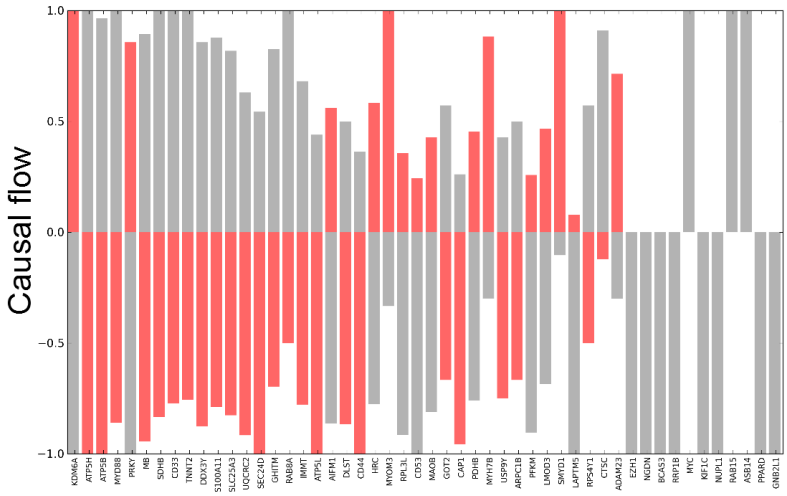


Figure S7: Other metrics for comparing gene topologies in non-failing and failing networks. (A) Genes with highest causal flow change between non-failing (gray) and failing (red) (B) Genes with highest l_1 difference between failing and non-failing networks.

A



B

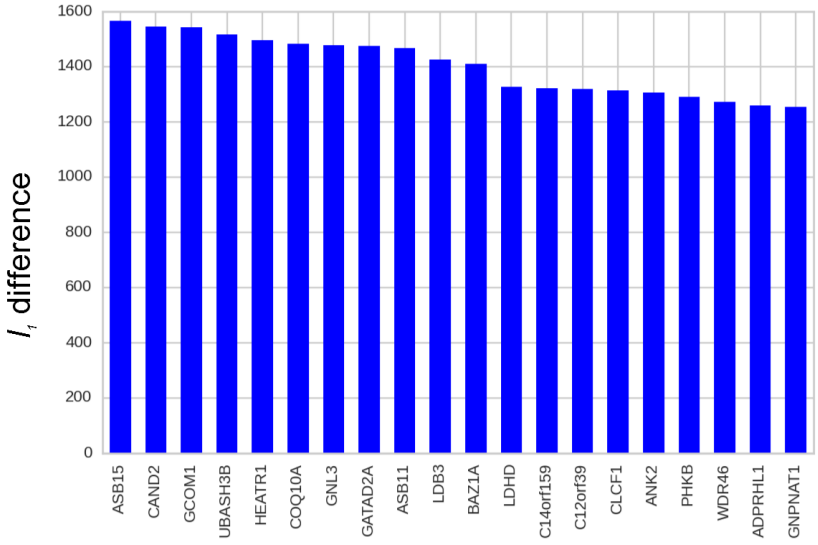


Figure S8: The PPP1R3A subnetwork obtained as the union of all communities in the failing network containing PPP1R3A, plotted with an approximate directed acyclic graph layout (nodes at the top are sources, nodes at the bottom are sinks). Node sizes are proportional to node degree and node color goes from low in-degree (red; all edges are outgoing) to high in-degree (green; all edges are incoming).

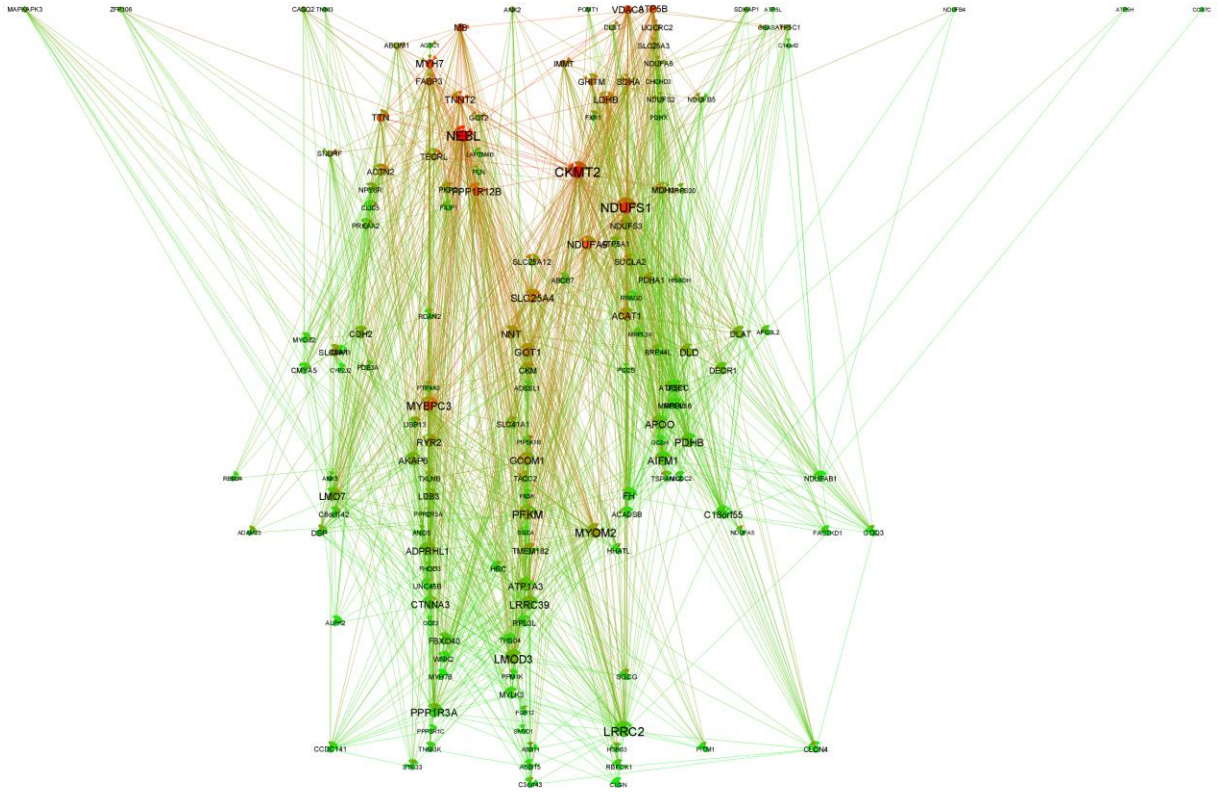


Figure S9: Validation of gene communities through gene expression cohesion across timepoints. Some of the previously inferred gene communities exhibit tight gene expression similarities across timepoints, providing validation of grouping these genes together. (A) Normalized gene expression profiles across timepoints in the phenylephrine-treated group of genes in a community inferred for failing hearts (top, red) and of a community inferred for non-failing counterparts in the un-treated group (bottom, gray). (B) Normalized gene expression of the randomly chosen genes in heart failure (top, red) and non-failing (bottom, gray) cells. These sets of genes do not have high expression cohesion. (C) P-values of gene expression cohesion permutation tests for all gene communities with more than 10 genes in the failing hearts (top, red), and non-failing counterparts (bottom, gray). Approximately half of the communities had significant p-values.

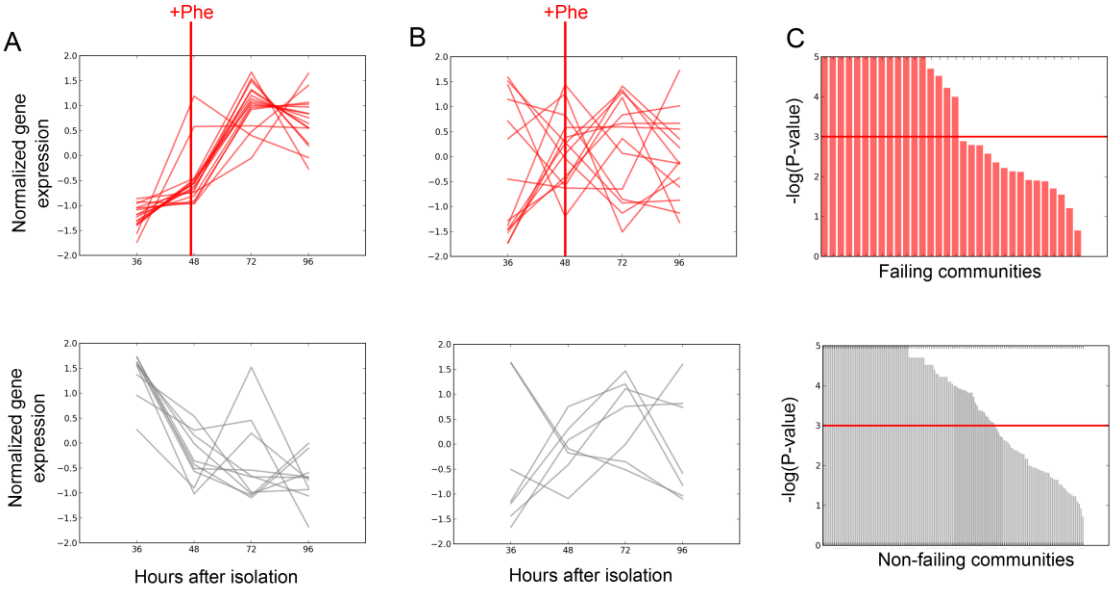
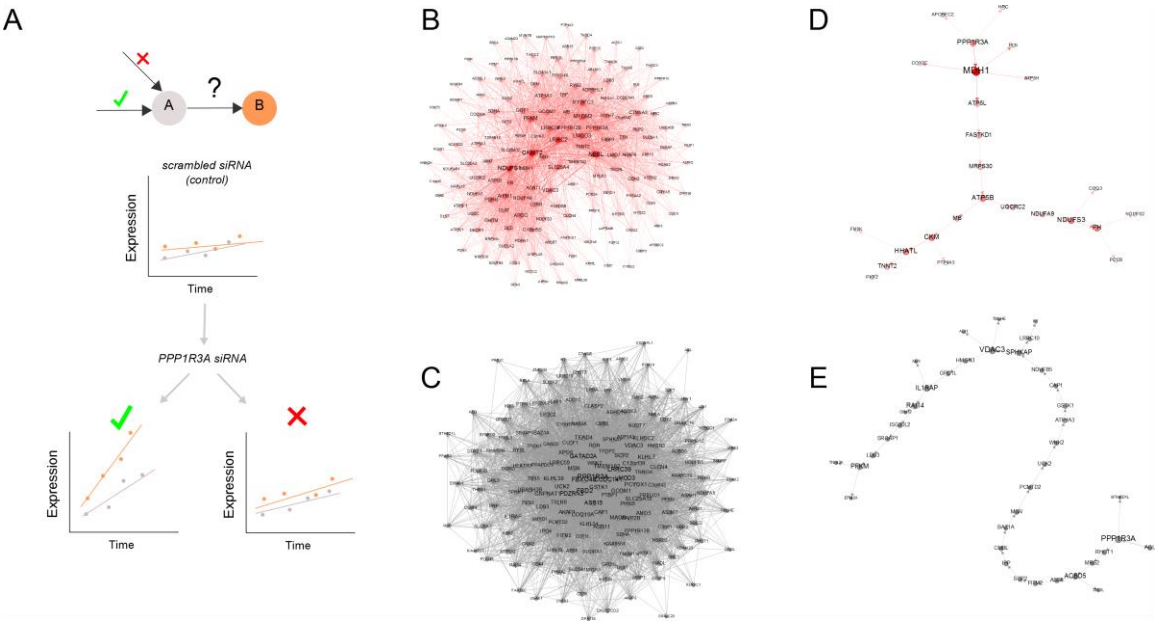


Figure S10: Network perturbation *in vitro* highlights causal paths stemming from *PPP1R3A*. (A) Scheme to verify directed co-expression with trends in the timed expression measurements. An edge between a gene A and a gene B is verified if both genes A and B are perturbed under *PPP1R3A* knockdown conditions as assessed by more than a 2-fold change in the slope of a linear fit to the timed gene expression values. Further, verification of at least one edge leading to A is a prerequisite for the verification of the edge from A to B. These highlighted paths start at the source of the perturbations, *PPP1R3A*. (B) Gene co-expression relationships of the union of gene communities containing *PPP1R3A* in the failing hearts network (top, red) and its non-failing counterpart (bottom, gray). Node colors and sizes represent the gene's degree. (C) Causal paths stemming from *PPP1R3A* that were highlighted in the failing hearts network (top, red), and non-failing counterparts (bottom, gray).



SUPPORTING TABLES

Table S1: SNPs with significant associations to community eigengenes. Max P-value and FDR are the maximum P-value and FDR, respectively, of all the significant association tests linking the SNP to all the communities. Transcription factors indicate Chip-Seq-supported transcription factors that fall in the SNP region.

Failing							
rsid	**No. controlled communities**	**Chrom.**	**Coord.**	**Ref Allele**	**Max P-value**	**Max FDR**	**Transcription Factors**
rs16957898	65	chr16	81220939	G	7.20E-11	5.20E-07	*TFAP2A, STAT3, STAT3, STAT3, TFAP2C, JUN, STAT3, BRCA1, JUN, CEBPB, EP300*
rs9467964	32	chr6	27096326	C	1.15E-11	9.73E-08	*POLR3A, POLR3A, STAT1, STAT1, GTF3C2, JUNB, GABPA*
rs7778483	32	chr7	78370332	T	1.15E-11	9.73E-08	*MEF2A, SPI1*
rs7769475	32	chr6	38649766	G	1.15E-11	9.73E-08	*JUND, FOS, JUN, MAFK, FOS, JUND*
rs7309559	32	chr12	1.27E+08	T	1.15E-11	9.73E-08	*CTCF*
rs16891488	32	chr6	26352130	G	1.15E-11	9.73E-08	*HNF4A, CDX2*
rs10763615	32	chr10	61078793	C	1.15E-11	9.73E-08	*MAFF, MAFK, MAFK, MAFK*
rs17010253	30	chr2	30854147	C	3.97E-11	3.08E-07	*AR*
rs9415787	26	chr10	52644567	A	2.50E-11	1.99E-07	*ZNF263*
rs7993297	13	chr13	41535439	C	8.09E-11	5.81E-07	*REST, REST, SIN3A*
rs16981555	10	chr20	55996422	C	8.24E-11	5.91E-07	*MYC, CTCF, CTCF*
rs6990187	8	chr8	40075801	T	8.26E-11	5.92E-07	*GATA2, EP300*
rs17063279	5	chr18	53062107	A	5.12E-11	3.88E-07	*TCF4, BHLHE40, ELF1, FOXA1, FOXA1, FOXA2, HDAC2, EP300, SP1, USF1, CEBPB*
rs4243559	3	chr14	48846937	G	6.27E-12	5.62E-08	*SP1*
rs3805691	3	chr5	1.41E+08	C	6.27E-12	5.62E-08	*HNF4A, REST, EP300, RAD21*
rs17119334	3	chr5	1.4E+08	G	6.27E-12	5.62E-08	*SRF, FOXA1*
rs3731815	1	chr2	86298029	T	7.68E-12	6.83E-08	*EGR1, JUN, MYC, MAFK, POLR2A*
rs16884834	1	chr5	55448597	C	4.48E-11	3.41E-07	*FOS, POLR2A*

rs12581753	1	chr12	1.04E+08	C	6.62E-12	5.92E-08	*CREBBP*
rs10494815	1	chr1	1.99E+08	C	3.70E-12	3.48E-08	*JUN, CEBPB, JUND, EP300, RFX5, STAT3*
rs10065426	1	chr5	99934293	A	4.21E-12	3.94E-08	*N/A*
Non-failing							
rs2279668	5	chr7	1.28E+08	C	4.75E-11	2.75E-05	*CEBPB*
rs923626	1	chr5	29041212	A	9.86E-14	9.18E-08	*CTCF, CTCF, CTCF, RAD21, CTCF, CTCF, CTCF*
rs7607098	1	chr2	1.5E+08	G	9.86E-14	9.18E-08	*GATA1, TRIM28, JUND, JUN, GATA2, GATA1, EP300*
rs7583010	1	chr2	39114755	T	9.86E-14	9.18E-08	*NA*
rs4328679	1	chr2	30418952	G	9.86E-14	9.18E-08	*NFKB1, RFX3*
rs1869720	1	chr17	36934322	A	9.86E-14	9.18E-08	*GATA3*
rs17075361	1	chr4	1.85E+08	T	2.12E-11	1.36E-05	*CEBPB*
rs16941165	1	chr16	84996559	C	2.12E-11	1.36E-05	*STAT1, FOS*
rs16829813	1	chr1	24734335	C	9.86E-14	9.18E-08	*EBF1, EBF1*
rs12304988	1	chr12	56519079	G	9.86E-14	9.18E-08	*GABPA, ELF1, SPI1, SPI1*
rs10400347	1	chr11	1.2E+08	T	9.86E-14	9.18E-08	*GATA2, GATA6*
rs10306183	1	chr9	1.24E+08	C	9.86E-14	9.18E-08	*SPI1*
rs10151341	1	chr14	60953240	T	2.35E-11	1.46E-05	*CREBBP, EBF1, EBF1*

Table S2: Local and global connectivity metrics of top 10 genes that most gained connectivity from non-failing to failing.

Gene	Role in failing	Role in non-failing	Gene neighbors in failing network	Community annotations in failing network
DOCK2	C-hub	N-hub	*DPYD, IL18, PPP1R13B, GSTM2, PXDN*	response to external stimulus, plasma membrane, guanyl nucleotide exchange factor activity, cytoskeleton organization, nucleoside triphosphatase regulator activity, natural killer cell mediated cytotoxicity, cytoskeleton, immune system process, iron ion binding
DOCK8	C-hub	N-hub	*RAB21, TMX1, IL18, ERBB3, DPYD*	nucleoside triphosphatase regulator activity, plasma membrane, natural killer cell mediated cytotoxicity, guanyl nucleotide exchange factor activity, immune system process
MYBPC3	C-hub	N-hub	*SCARNA10, CYC1, MYLK, RBP4, MT1F*	contractile fiber, dilated cardiomyopathy, muscle system process, muscle tissue morphogenesis,

				structural constituent of muscle, mitochondrial part, contractile fiber part, cytoskeleton, myosin filament
LAPTM5	C-hub	N-hub	*GSTM2, IL18, SERPINE2, RAB21, MARCH1*	cell activation, arachidonic acid metabolism, plasma membrane, guanyl nucleotide exchange factor activity, vacuole, cofactor binding, immune system process, innate immune response
MYOM2	C-hub	N-hub	*RBP4, SCARNA10, OLFML1, CYTIP, DOCK10*	non membrane bounded organelle, contractile fiber, contractile fiber part, vascular smooth muscle contraction, muscle system process, structural constituent of muscle, mitochondrial part, cytoskeleton, myosin filament
NEBL	C-hub	N-hub	*RBP4, DOCK10, CC2D2A, ANLN, FCGR3A*	insulin signaling pathway, cytoskeletal protein binding, generation of precursor metabolites and energy, vascular smooth muscle contraction, muscle system process, structural constituent of muscle, mitochondrial part
CD53	L-hub	N-hub	*IL18, GPR81, SERPINE2, RAB21, MDH2*	cell activation, guanyl nucleotide exchange factor activity, vacuole, cofactor binding, actin filament based process, immune system process, innate immune response
LRRC2	C-hub	N-hub	*LAIR1, MYLK, ANLN, DAPP1, RAB3A*	insulin signaling pathway, generation of precursor metabolites and energy, muscle system process, organelle part, structural constituent of muscle, mitochondrial part, protein c terminus binding, mitochondrion, ion transmembrane transporter activity
MDH1	C-hub	N-hub	*SNORA3, LAIR1, SLC2A3, IGHG1, KPNA4*	oxidative phosphorylation, cellular respiration, oxidative phosphorylation, oxidation reduction, generation of precursor metabolites and energy
CKMT2	C-hub	N-hub	*ANLN, THRSP, THBS3, CPA3, DAPP1*	generation of precursor metabolites and energy, muscle system process, organelle part, organelle inner membrane, mitochondrial part, mitochondrion
TNNT2	S-hub	N-hub	*FCGR3A, SCARNA10, KLF9, CYTIP, CCDC109B*	contractile fiber, dilated cardiomyopathy, muscle system process, structural constituent of muscle, muscle tissue morphogenesis, mitochondrial part
LCP1	L-hub	N-hub	*GSTM2, ERBB3, RAB21, GPR81, SERPINE2*	guanyl nucleotide exchange factor activity, plasma membrane, actin filament based process, immune system process, cell activation
NDUFS1	C-hub	L-hub	*RPL3, LAIR1, FGD2, CD163, MT1E*	generation of precursor metabolites and energy, organelle part, organelle inner membrane, mitochondrial part, mitochondrion, respiratory chain
NCKAP1L	L-hub	N-hub	*RAB21, IL18, PXDN, EEF1A2, DTX1*	nucleoside triphosphatase regulator activity, plasma membrane, fc gamma r mediated phagocytosis, immune system process, response to external stimulus
ADPRHL1	S-hub	N-hub	*CYTIP, CC2D2A, RBP4, MT1F, OLFML1*	cytoskeletal protein binding, non membrane bounded organelle, contractile fiber part, vascular smooth muscle contraction, muscle system process, muscle tissue morphogenesis, structural constituent of muscle, mitochondrial part, cytoskeleton, myosin filament

SELPLG	L-hub	N-hub	*DPYD, RAB21, CNTFR, DTX1, IL18*	response to external stimulus, plasma membrane, guanyl nucleotide exchange factor activity, immune system process
HAVCR2	L-hub	N-hub	*CASP4, DTX1, FAM7A1, DPYD, HADHB*	response to external stimulus, cytoskeleton, immune system process, cytoskeleton organization
PPP1R3A	C-hub	L-hub	*CCDC109B, CC2D2A, ACTG1, HSPB7, DOCK10*	insulin signaling pathway, cytoskeletal protein binding, generation of precursor metabolites and energy, vascular smooth muscle contraction, muscle system process, structural constituent of muscle, mitochondrial part
EPB41L3	L-hub	N-hub	*RAB21, NRAP, SRP9, PIK3CG, DPYD*	plasma membrane, guanyl nucleotide exchange factor activity, nucleoside triphosphatase regulator activity, guanyl nucleotide binding, actin filament based process, immune system process
MYH7	C-hub	N-hub	*CYTIP, FCGR3A, RBP4, KLF9, KY*	contractile fiber part, muscle system process, structural constituent of muscle, muscle tissue morphogenesis, mitochondrial part, cytoskeleton

Table S3: Local and global connectivity metrics of top 10 genes that most lost connectivity from non-failing to failing.

Gene	Role in non-failing	Role in non-failing	Gene neighbors in non-failing network	Community annotations in non-failing network
CCDC141	N-hub	C-hub	*SLC7A7, PRPF40A, MTIF3, LOC653075, C5orf41*	uropod, ribosome biogenesis, cytoplasm, primary metabolic process, cellular macromolecule metabolic process, cofactor binding, mitochondrion, protein catabolic process, sodium, intracellular, catabolic process, mitochondrial part, fad binding, membrane enclosed lumen, nuclear part, parkinson's disease, membrane bounded organelle, nucleolus, intracellular part, golgi apparatus, organelle, ion channel complex
FBXO40	N-hub	C-hub	*C5orf41, SMC5, UROS, LOC653075, COQ3*	modification dependent macromolecule catabolic process, uropod, protein import into nucleus, docking, cytoplasm, primary metabolic process, cellular macromolecule metabolic process, cellular lipid catabolic process, cofactor binding, mitochondrion, heart development, cellular ketone metabolic process, protein catabolic process, sodium, intracellular, catabolic process, mitochondrial part, fad binding, membrane enclosed lumen, nuclear part, ribosome biogenesis, nucleus, membrane bounded organelle, nucleolus, intracellular part, organelle

ASB15	N-hub	C-hub	*PRPF40A, C5orf41, COQ3, RPS24, GABRA4*	modification dependent macromolecule catabolic process, protein catabolic process, membrane enclosed lumen, nucleolus, intracellular, intracellular part, sodium, ribosome biogenesis, cytoplasm, primary metabolic process, uropod, organelle, catabolic process, cellular lipid catabolic process, mitochondrial part, nuclear part, mitochondrion, ion channel complex, cellular macromolecule metabolic process, cellular ketone metabolic process, pathogenic escherichia coli infection
VDAC3	L-hub	C-hub	*C5orf41, FADS3, TLR8, PRPF40A, HNRNPH3*	mitochondrial envelope, nuclear part, membrane enclosed lumen, nucleolus, intracellular part, cellular respiration, cytoplasm, parkinson's disease, envelope, organelle part, organelle inner membrane, anchoring junction, cellular lipid catabolic process, mitochondrial part, cellular ketone metabolic process, cytoplasmic part, nuclear envelope, mitochondrion, intracellular, huntington's disease, protein binding
KLHL38	N-hub	C-hub	*LOC653075, PRPF40A, ALG3, SMC5, CSRNP3*	modification dependent macromolecule catabolic process, protein catabolic process, membrane enclosed lumen, nucleolus, intracellular part, uropod, cytoplasm, primary metabolic process, intracellular, nuclear part, mitochondrial part, mitochondrion, cellular macromolecule metabolic process, cofactor binding, organelle
GCOM1	L-hub	C-hub	*MTIF3, C5orf41, FADS3, PRPF40A, COQ3*	protein catabolic process, membrane enclosed lumen, nucleolus, contractile fiber, intracellular part, cellular respiration, ribosome biogenesis, cytoplasm, valine, leucine and isoleucine degradation, parkinson's disease, nuclear part, cofactor binding, mitochondrial part, ion channel complex, cytoplasmic part, nuclear envelope, mitochondrion, intracellular, cellular ketone metabolic process, electron carrier activity, pathogenic escherichia coli infection
LRRC39	C-hub	C-hub	*PRPF40A, CD33, UROS, LOC653075, RRP9*	uropod, nuclear part, membrane enclosed lumen, nucleolus, mitochondrial envelope, intracellular part, sodium, ribosome biogenesis, cytoplasm, primary metabolic process, parkinson's disease, organelle, catabolic process, cellular lipid catabolic process, mitochondrial part, mitochondrion, intracellular, cellular ketone metabolic

				process, cofactor binding, anchoring junction
TMEM182	N-hub	C-hub	*COQ3, PRPF40A, DCAF6, UROS, C5orf41*	membrane enclosed lumen, nucleolus, intracellular part, ribosome biogenesis, cytoplasm, parkinson's disease, nuclear part, catabolic process, mitochondrial part, valine, leucine and isoleucine degradation, mitochondrion, intracellular, cofactor binding
CLIC1	N-hub	C-hub	*CSE1L, B4GALT3, DDHD2, SLC8A1, ARPC1B*	cell motion, membrane enclosed lumen, nucleolus, cytoskeletal protein binding, protein complex, intracellular part, localization, ribosome biogenesis, cytoplasm, envelope, ncrna metabolic process, cortical cytoskeleton, organelle, endoplasmic reticulum, mitochondrial part, cytoplasmic part, nuclear envelope, mitochondrion, pathogenic escherichia coli infection
TXLNB	N-hub	C-hub	*PRPF40A, BBS2, FAM35B2, NCF4, LOC653075*	organelle, membrane enclosed lumen, protein catabolic process, intracellular part, cytoplasm, primary metabolic process, intracellular, nuclear part, catabolic process, mitochondrial part, mitochondrion, nucleolus
KLHL24	N-hub	C-hub	*LOC653075, EMR2, RPS24, TMEM43, PRPF40A*	membrane enclosed lumen, nucleolus, intracellular part, cytoplasm, primary metabolic process, intracellular, organelle, endoplasmic reticulum, mitochondrial part, cytoplasmic part, mitochondrion, anchoring junction, pathogenic escherichia coli infection
COQ10A	N-hub	C-hub	*PRPF40A, SMC5, IRAK3, GSTM4, C5orf41*	modification dependent macromolecule catabolic process, membrane enclosed lumen, nucleolus, endoplasmic reticulum, protein import into nucleus, docking, cytoplasm, primary metabolic process, uropod, organelle, catabolic process, ribosome biogenesis, mitochondrial part, mitochondrion, heart development, cofactor binding
ANO5	N-hub	C-hub	*LOC653075, C5orf41, COQ3, UROS, DHTKD1*	membrane enclosed lumen, parkinson's disease, intracellular part, endoplasmic reticulum, cytoplasm, primary metabolic process, intracellular, organelle, catabolic process, mitochondrial part, ion channel complex, cytoplasmic part, membrane bounded organelle, mitochondrion, anchoring junction, pathogenic escherichia coli infection

LMOD3	C-hub	C-hub	*C5orf41, PRPF40A, MTIF3, DCAF6, COQ3*	protein catabolic process, membrane enclosed lumen, nucleolus, intracellular part, ribosome biogenesis, cytoplasm, parkinson's disease, nuclear part, endoplasmic reticulum, mitochondrial part, cytoplasmic part, valine, leucine and isoleucine degradation, mitochondrion, intracellular, fad binding, cellular ketone metabolic process, cofactor binding, organelle
MYH7B	N-hub	C-hub	*PRPF40A, KALRN, LOC653075, PSMD12, USP53*	endoplasmic reticulum part, protein catabolic process, membrane enclosed lumen, nucleolus, intracellular part, ribosome biogenesis, cytoplasm, organelle, endoplasmic reticulum, mitochondrial part, mitochondrion, heart development
MAOB	N-hub	C-hub	*DCAF6, EIF2C2, MTIF3, C5orf41, LOC728407*	membrane enclosed lumen, nucleolus, cytoskeletal protein binding, intracellular part, contractile fiber, sodium, ribosome biogenesis, cytoplasm, mitochondrial part, cytoplasmic part, mitochondrion, electron carrier activity
CAP1	L-hub	C-hub	*SLC8A1, GIN1, RAN, COX6A2, PLK1S1*	cell motion, protein catabolic process, nucleolus, cytoskeletal protein binding, protein complex, intracellular part, localization, cytoplasm, intracellular, nucleotide sugar metabolic process, cortical cytoskeleton, organelle, endoplasmic reticulum, mitochondrial part, pathogenic escherichia coli infection, cytoplasmic part, mitochondrion, cytoskeleton organization, cellular component morphogenesis
ANK2	N-hub	C-hub	*SLC3A2, MCAM, DCAF6, LOC653075, FADS3*	membrane enclosed lumen, protein catabolic process, intracellular part, cytoplasm, intracellular, envelope, mitochondrial part, ion channel complex, cytoplasmic part, mitochondrion, electron carrier activity, pathogenic escherichia coli infection
ASB11	N-hub	C-hub	*PRPF40A, C5orf41, UROS, MTIF3, ZC3H6*	membrane enclosed lumen, nucleolus, intracellular part, uropod, ribosome biogenesis, cytoplasm, catabolic process, mitochondrial part, mitochondrion, cofactor binding
C3orf43	N-hub	C-hub	*C5orf41, PRPF40A, RRP9, SMC5, UROS*	membrane enclosed lumen, intracellular part, ribosome biogenesis, cytoplasm, nuclear part, mitochondrial part, mitochondrion, cellular ketone metabolic process
NFE2L1	N-hub	C-hub	*LOC653075, GABRA4, PDHX, RPS24, PAK1IP1*	modification dependent macromolecule catabolic process, membrane enclosed lumen, protein catabolic process,

				intracellular part, cytoplasm, primary metabolic process, intracellular, organelle, endoplasmic reticulum, mitochondrial part, cytoplasmic part, mitochondrion, pathogenic escherichia coli infection, nucleolus

Table S3: Primers for qPCR measurements on several genes. Gene symbols starting in (r) are rat-specific genes, whereas those starting in (m) are mouse-specific.

Gene	Primer
rGLUT1	F; TCGTCGTTGGGATCCTTATTG
rGLUT1	R; GAAGATGACACTGAGCAGTAGAG
rGLUT4	F; GGCATGGGTTTCCAGTATGT
rGLUT4	R; CTCTGGTTTCAGGCACTCTTAG
rPDK2	F; TCACTCTCCCTCCCATCAAA
rPDK2	R; AGCCTCTCGATCTTCCTCAA
rPDK4	F; TGAACCAGCACATCCTCATATT
rPDK4	R; CTTCGACTACTGCTACCACATC
mPPP1R3A	F; GGAGTTGGAGAGAAGTTGGAAG
mPPP1R3A	R; CATCAGTCAGAGGGCTGAAAT
mCPT1B	F; TCCAAACGTCACTGCCTAAG
mCPT1B	R; CCAATGTCTCCATGCGGTAATA
mGYS1	F; ACGGGCTGAATGTGAAGAAG
mGYS1	R; GAAGTCCAGGTGCCCATAAA
mAPOBEC2	F; GAAGACCCTGAGAAGCTGAAAG
mAPOBEC2	R; TCCGGAACTGAAACTTGAAGAA
mMDH1	F; GATGACCTGCTCTACTCATTCC
mMDH1	R; TTTGCTGTCAGGTCCATCTT
mPLN	F; TATCAGGAGAGCCTCCACTATT
mPLN	R; CAGATCAGCAGCAGACATATCA
mHRC	F; CAGAGAAGGTTCCAGTGAAGAG
mHRC	R; CAAAGACCCTGGCTGGTAAT
mCOX7C	F; TCGCAGCCACTATGAGGA
mCOX7C	R; GGCAAACCCAGATCCAAAGTA
mATP5H	F; CAATTGGAAATGCCCTGAAGTC
mATP5H	R; CCTGTAGTAAGCCCAGTCAATC
mATP5L	F; GGCCAAGTTCATCCGTAACT
mATP5L	R; CAATCGAGGCTTCGAGTAAGT
mATP5B	F; CCAAGGTAGCGTTGGTATATGG
mATP5B	R; CCCTCCTGGTCTCTGAAGTAT