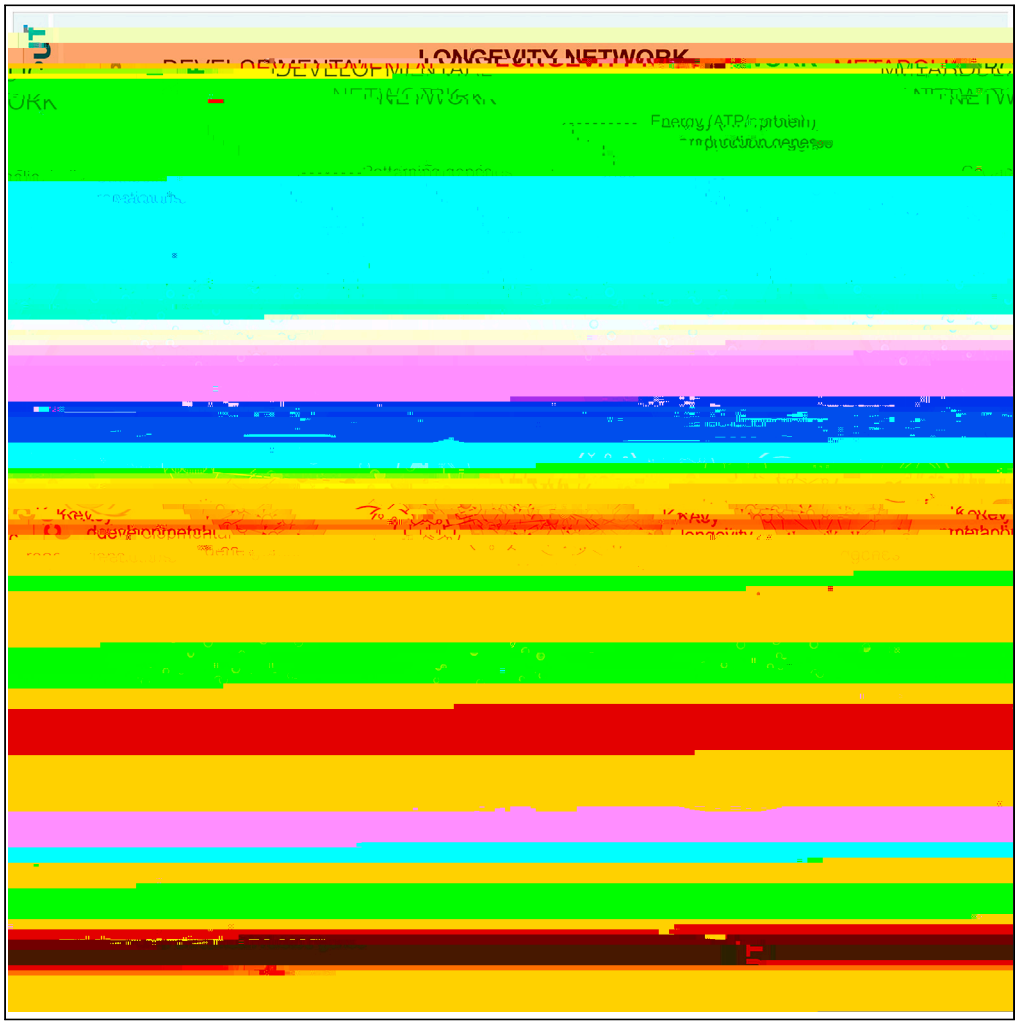
iScience

Article

# Gene regulatory network inference in long-lived reveals modular properties that are predictive of novel aging genes




Manusnan Suriyalaksh, Celia Raimondi, Abraham Mains, ..., Simon Andrews, Marta Sales-Pardo, Olivia Casanueva

marta.sales@urv.cat (M.S.-P.)
olivia.casanueva@babraham.ac.uk (O.C.)

## Highlights

Gene-regulatory inference provides global network of long-lived animals

The large-scale topology of the network has an hourglass structure

iScience

CellPress
OPEN ACCESS

Article

# Gene regulatory network inference in long-lived [illegible] reveals modular properties that are predictive of novel aging genes

Manusnan Suriyalaksh,[1,5] Celia Raimondi,[1,5] Abraham Mains,[1] [illegible]

reporters and a novel functional network covering 5,497
make mechanistic predictions. We used genetic epistasis
predictions, uncovering a novel transcriptional regulato
alongside DAF-16/FOXO. We present a framework with
can accelerate discovery in ***C. elegans*** and potentially hun

## INTRODUCTION

Reductionist single-gene perturbation approaches using [illegible]s as a model organism have led to fundamental discoveries in aging (Kappeler et al., 2008; Kenyon, 2011). Two important pathways include the highly conserved insulin-like signaling (ILS) pathway (Kappeler et al., 2008; Kenyon, 2011) and the signals from thC. elegans

g/10.1016/j.isci.

To assess the effect of the input information on output GRNs, we considered different combinations of input regulators (2,795 genes, which are either in GenAge database (Tacutu et al., 2018) (22.6%), known transcription factors in [illegible] (25.8%), or display high-variability in their expression (51.6%); STAR Methods), time series length, and sets of priors; and inferred a total of 50 GRNs with binary, directed edges from regulators to targets (Table S4). Despite the fact that NI approaches aim to minimize the number of regulatory interactions (for instance, by explicitly incorporating regularization terms in the regression), the inferred GRNs are very dense. On average each regulator has more than 653.6 targets, almost two-fold the maximum number of regulators per target reported in ModERN's ChIP-seq datasets of 350 (Kudron et al., 2018), suggesting that the number of interactions is overestimated. Each of the NI methods produces different scores for each predicted interaction and there is no standard objective criterion to filter edges that is not based on [illegible] knowledge (Arrieta-Ortiz et al., 2015). Therefore, to identify spurious interactions, we compared the edge scores of each network to the distribution of edge scores obtained from a randomized time-series for each input combination and retained interactions that fall below the 5% significance level, reducing to almost a third of the number of targets per regulator (224 on average).

We analyzed the accuracy of inferred networks against WT-GS and found that none of the GRNs performs systematically better than the others (Figures S1C and S1D). Therefore, as a final step in our pipeline, we considered all 50 GRNs to build consensus networks. First, we identified nine groups of networks based on edge overlap (Figure S1E; STAR Methods). Then, for each group we built a consensus network by keep-

inherent technical variability of the RNAi technique (Kamath and Ahringer, 2003) (Figures S2A and S2B) and calculated the Pearson correlation coefficient (PCC) between the changes in expression levels of regulator-target pairs from at least six replicates (Figure S2B, Table S7, STAR Methods, https://s-andrews.github.io/wormgrn/qpcr/)

To determine whether the knockdown (KD) of a regulator affects a target, we chose the PCC cut-off value at the inflection point of a curve obtained when plotting the average precision versus cut-off values (PCC = 0.75, Figure 2

regulators and targets, which is equal to 0.9% (STAR Methods, for similar analyses see Marbach et al., 2012; Siahpirani and Roy, 2016; Miraldi et al., 2019). We also find that as we increase the number of tested targets per regulator, the number of correctly predicted edges for each regulator fluctuates less and approaches the mean precision (Figure S2C). The apparent convergence toward the mean suggests that despite this being a very partial validation (we only tested targets for 10 out of 1396 regulators), the errors we observe are unbiased (Figure S2

complexes. In addition, we find genes encoding metabolic enzymes (carboxypeptidase, adenosyl-hydrolase, and ubiquinone oxidoreductases); a glutamate receptor, a gene encoding a protein folding chaperone and several novel genes of unknown function (Table 220(s5)-218/Cs9cs000scn3.372.0486D.010bl)83.24sa7a7-251t240tt2l2e888201d242.4.t2la7242.

reporters. First, we quantified in a semi-automatic manner (STAR Methods) the " " expression of a

Figure 5. Global characterization of the novel aging genes reveals genes sharing the same metabolic features and pathways as DAF-16/FOXO and ILS

(A) Comparison of the fluorescence measures of -3 :GFP versus -3 -3 in -1 animals at day 4 of adulthood. -3 -3 -3 is a translational reporter which localizes to the intestinal lipid droplets (LD). - 3 :GFP is a transcriptional reporter for the expression of superoxide dismutase 3 ( -3), a direct target of DAF-16. Colors correspond to lifespan phenotype as shown in the figure. L4440 is the control/empty vector (Table S15).

(B) Novel aging genes sharing known aging and metabolism targets with -1 and -1

g c c

multi-omics datasets. From the WormExp, modERN, Cis-BP, and GEO databases, we manually identified context-specific datasets, reconciled and curated 380,023 interactions in young adult . , freely accessible to the community. We obtained for the first time, genome-wide GRNs that are contextual to aging in . whose modular structure is biologically meaningful, providing a systems-view of the regulatory interactions underpinning the aging process. This study presents a novel approach that integrates NI with large-scale network analysis tools applied to networks containing many errors at the "local" level, typical of NI-derived networks. Our work provides a compelling example where a network that contains unbiased errors at a local level, can be predictive as long as the global structure is robust. The study led to the discovery of 50 novel regulators of *-1( )* longevity, augmenting the number of regulators of the pathway by 62.5% and the majority of which have an identifiable human orthologue. This pipeline presents a minimum of 4.8% hit rate, more than a two-fold increase compared to the blind genetic screening in the *-1( )* which reported a 2.1% success rate (Berman and Kenyon, 2006). Although the fold change may

the gene interaction network, where genes in this category share a similar gene expression profile with *-1* and *-1* (Figures 5B and S6E).

From our mechanistic studies we have placed the transcriptional activator *-3* alongside with *-1* regulation. *-3* has been previously identified as a DAF-16 target by chromatin precipitation analysis (Wook Oh et al., 2006) further strengthening our findings. It is interesting to also notice that the PCC network also shows that *-3* as a positive regulator of *-12*, *-4* and the PTEN homologue and ILS pathway component *-1*

- Two-step -1( ) lifespan screening
- -3( ) lifespan screen
- L4 -1( ) screen
- Lifespan epistasis experiments
- Microscopy
- Construction of a mechanistic TF-gene (physical) and gene-gene (functional) interaction database
- Definition of empirical modules

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Adjusted p-values for correlation calculation in RNAi knockdown experiments
  - Statistics used to analyse significance in the two-step -1( ) lifespan screening
  - Statistics used to analyse lifespan epistasis experiments
  - Statistics used to analyse microscopy data
  - ATAC-seq data reconciliation
  - Genome-wide . gene regulatory network inference
  - Significance calculation of enrichment of sets of genes by resampling
  - Performance metrics for validations against the gold standard
  - Precision for empirical validation experiments and random expectation
  - Bayesian model selection with stochastic block models (SBM) and definition of structural modules
  - Recovery quantification of empirical gene modules by structural gene modules
- ADDITIONAL RESOURCES
  - Gene interaction network interactive webpage

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103663.

## ACKNOWLEDGMENTS

We acknowledge Wolf Reik, Len Stephens, Rebeca Taylor, David Weinkove, Ben Lehner, Rob Jelier, Stefan Schoenfelder for comments on the manuscript. We acknowledge Nicolas Gambardella, Sven Bergmann and Janna Hastings for advice on network inference. MS acknowledges PhD scholarship from Cambridge Trust. MSP and RG acknowledge funding from project PID2019-106811GB-C31 from MCIN/AEI/10.13039/501100011033 and by the Government of Catalonia (2017SGR-896). OC acknowledges funding from ERC (award 638426) and BBSRC (award BBS/E/B00C0421).

## AUTHOR CONTRIBUTIONS

M.S. wrote code and performed computational experiments; analyzed data; designed experiments; discussed and interpreted results; wrote manuscript.

C.R. performed wet-lab experiments; analyzed data; designed experiments; discussed and interpreted results; wrote manuscript.

A.M., S.M., S.M., and R.A. performed experiments.

F.K. assisted on bioinformatics.

A.S-P. performed statistical analysis.

R.G. discussed and interpreted results.

S.A. analyzed data; discussed and interpreted results; created the visualization database.

M-S.P. analyzed data; designed experiments; discussed and interpreted results; wrote manuscript.

O.C. analyzed data; designed and performed wet-lab experiments; discussed and interpreted results; wrote manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 23, 2021
Revised: September 9, 2021
Accepted: December 15, 2021
Published: January 21, 2022

## REFERENCES

Aalto, A., Viitasaari, L., Ilmonen, P., Mombaerts, L., and Gonçalves, J. (2020). Gene regulatory network inference from sparsely sampled noisy

resistance and Mn-superoxide dismutase gene expression in Caenorhabditis elegans. FASEB J. *13*, 1385–1393. http://www.ncbi.nlm.nih.gov/pubmed/10428762.

Hou, L., Wang, D., Chen, D., Liu, Y., Zhang, Y., Cheng, H., Xu, C., Sun, N., McDermott, J., Mair, W.B., and Han, J.D.J. (2016). A systems approach to reverse engineer lifespan extension by dietary restriction. Cell Metab. *23*, 529–540. https://doi.org/10.1016/j.cmet.2016.02.002.

Hsin, H., and Kenyon, C. (1999). Signals from the reproductive system regulate the lifespan of C. elegans. Nature *3* , 362–366. https://doi.org/10.1038/20694.

Ighodaro, O.M., and Akinloye, O.A. (2018). First line defence antioxidants-superoxide dismutase (SOD), catalase (CAT) and glutathione peroxidase (GPX): their fundamental role in the entire antioxidant defence grid. Alexandria J. Med. *54*, 287–293. https://doi.org/10.1016/j.ajme.2017.09.001.

Kamath, R.S., and Ahringer, J. (2003). Genome-wide RNAi screening in Caenorhabditis elegans. Star Methods (San Diego, Calif.) *30*, 313–321. https://doi.org/10.1016/s1046-2023(03)00050-1.

Kamath, R.S., Martinez-Campos, M., Zipperlen, P., Fraser, A.G., and Ahringer, J. (2000). Effectiveness of specific RNA-mediated

## STAR★METHODS

### KEY RESOURCES TABLE

*Continued*

| REAGENT | RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|---|

*Continued*

| REAGENT RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| OL309 T21H3.1 F CCGTCAGTCCACAGGACAAG | | |
| OL310 T21H3.1 R GCAGCCCATGGAGTGTGAT | | |
| OL311 Y26D4A.21 F AAAATGTAAGAT TCATCCAACCGAC | | |
| OL312 Y26D4A.21 R CTCGTCCGATTGAATTGCCTG | | |
| OL313 Y51H7C.13 F GCTATTGGAACCAAGTGCTGC | | |
| OL314 Y51H7C.13 R TTTGGGATGGAAGGTTCGGG | All oligos were designed in house | Does not apply to any oligos in this list/ NA |
| OL315 Y54F10AM.5 F TCTTTGAAGCGTTCCTGCCT | | |
| OL316 Y54F10AM.5 R TGGTCGTCATGGCACTTTGA | | |
| OL317 Y51H1A.3 F GGAATTGGGGAGAGCTCGTG | | |
| OL318 Y51H1A.3 R CGATCGAGTTCCAGGTGGTG | | |
| OL319 Y71G12B.10 F GTTATGAGGTCTCGCTGGGC | | |
| OL320 Y71G12B.10 R GTTTTTCGGCTGGCACACTT | | |
| OL321 Y43F8B.1 F CGAGCAACTAGCCCAGAAGT | | |
| OL322 Y43F8B.1 R GAGCTCGGAATCAGCTACCC | | |
| OL323 lpl-1 F AGGTGTTAGGCATTTTGTGGAC | | |
| OL324 lpl-1 R ACCCATACACCTGTATTCGCA | | |
| OL325 atp-1 F AAACCGGAAAGACCGCCATT | | |
| OL326 atp-1 R GACAGCGACGTAGATGCAGA | | |
| OL327 mdh-2 F CTTCCAGCAAAGACCCTCGT | | |
| OL328 mdh-2 R AGAAGAGCGACCTTTGGAGC | | |
| OL329 paf-2 F AGTTGGTCATGTCATCCGCT | | |
| OL330 paf-2 R TTGCTTTTTGGAAGTCCGTTGT | | |
| OL331 str-7 F TTTCACATCAAACGGCAATTCG | | |
| OL332 str-7 R GGAGGAACGTGTGAAACAAGTAT | | |
| OL333 tag-120 F TATTTTCACTCTCTCGGCAGCA | | |
| OL334 tag-120 R TCCACTGCATACTGTGGTGAT | | |
| OL335 sucl-1 F GGGAGCTGCTCGTTTCTACA | | |
| OL336 sucl-1 R AGGTACCCTGCTTTCCTGTG | | |
| OL337 tald-1 F GAATTCGGGCTGCTAACACG | | |
| OL338 tald-1 R GGCGAGATTAGGGTGACTCC | | |
| OL339 tiam-1 F CCTTGTGATGAGCAGCCAGA | | |
| OL340 tiam-1 R ACGCGAAACATTCCAGCAAAT | | |
| OL341 rhy-1 F TGACACTTGTCGTCATCGGAA | All oligos were designed in house | Does not apply to any oligos in this list/ NA |
| OL342 rhy-1 R TTCGAGGATCT.9(ACG)-T | | |
| oliTCGGCAGCA | | |

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

Most assays performed in this study used sterile *-1( 2144)* or *-3(20)* . s that were maintained at 16 C on NGM with OP50 . . To induce sterility, L1 synchronised larvae were added to NGM plates containing HTT115 at 25 C.

### Worm maintenance and synchronization to obtain time-series RNA-seq datasets

Worms were maintained at 16 C on NGM with OP50 . . Synchronised experimental populations were

## METHOD DETAILS

### RNA-sequencing

Eggs were collected by bleaching and L1 larvae were hatched in the absence of food. Synchronisation was obtained by adding L1 larvae to plates with HTT115, strain containing the empty vector plasmid L4440 on standard NGM plates containing 50 μg/ml Carbenicillin and 1 mM IPTG until harvesting and grown at 25

well. Contaminated plates, starved plates, or plates with outwardly defective or arrested animals were discarded (104 conditions).

*-1( )* lifespan screen 2. 93 candidates were selected based on novelty as explained in the main text. The experiment was set-up as described above and the percentage of survival was assayed every 2 to 3 days (at days: 3, 5, 7, 10, 12, 14, 16, 17, 19, 21, 24 and 26) by scoring animals based on movement. Three biological replicates were conducted and each of them included two technical replicates. To ensure the identity of the genes that were knocked down, each one of the clones that significantly change the lifespan was sequenced using M13 forward primer. Day 19 normalised percentage of survival was calculated using the same procedure. For this screen, we used day 19 survival rate as we found that it is closer to the mean survival.

### ***Fem-3(ts)*** lifespan screen

To evaluate the effect of RNAi clones in normal lived animals we used *-3(20)* .The experiment was set-up at 25 C as described in the two-step *-1( )* lifespan screening, and the percentage of survival was assayed at day 13 of adulthood, which is the time point when animals grown in bacteria expressing empty vectors reached roughly 50% survival. *-3(20)* survival at day 13 was compared with *-1( )* survival at day 19, at which time both strains have reached 50% survival.

### L4 ***glp-1(ts***

motifs, using the window command with 1000 bp window size. For eY1H data which already are interactions between TF and genes, we only included an interaction if the TSS site of the target gene overlaps with an open ATAC-seq region by at least one base pair according to the output (version 2.29.0, Quinlan and Hall, 2010).

We used the following sources of TF-gene interactions: 1) 115 L4 or young-adult ChIP-seq datasets from ModERN (Kudron et al., 2018), 2) two young-adult ChIP datasets (GSE28350, GSE81521) (Hochbaum et al., 2011; Li et al., 2016), 3) 202 unique TF DNA recognition motifs using "direct evidence" option from CiS-BP motif database (Weirauch et al., 2014), obtained through RTFBSDB R package (Wang et al., 2016), and 4) 13,501 TF-gene interactions from eY1H assay (Fuxman Bass et al., 2016). Regulatory sequences were obtained using biomaRt R package (accessed on 31st Oct 2017) (Durinck et al., 2009). This study used WBcel235/ce11 version of the genome, and WormBaseWS260 genome annotations.

For gene-gene interactions, we based our curation on the WormExp v1.0 database (Yang et al., 2016) which has compiled nearly all published expression data over the past decade (last updated on 27/07/2017) (Yang et al., 2016). Out of the 361 studies, 298 studies were in 'Mutants', 'DAF/Insulin/food', 'Devel-

then keep the largest PCC (in absolute value). After the 1,000 iterations we obtain a distribution of extreme (largest in absolute value) PCCs conditioned on the expression vector of the KD and target genes and we

## Genome-wide *C. elegans* gene regulatory network inference

. To select genes whose gene-expression time series would be fed to network inference algorithms, we applied a threshold of a minimum of log2-difference between the highest and the lowest values across all time conditions. Out of 20,191 protein coding genes, 12,884 genes were above that threshold and thus, RNA-seq data for these genes was used as input for the inference algorithms.

. We used Inferelator (Arrieta-Ortiz et al., 2015), MERLIN-P (Siahpirani and Roy, 2016) and Time-lagged Ordered Lasso (Nguyen and Braun, 2018) (TOL) inference al-

amount of the potential number of interactions that can be inferred. Because of this, we resorted to performance metrics that are more suitable to measure whether the observed signal differs from random or not. Specifically, we use the precision fold enrichment introduced in (Roy et al., 2013) and the area under the precision fold enrichment curve.

- It measures the precision (that is the fraction of predicted positive interac-

nodes or not. Using the same procedure described above we calculated precision and accuracy for each randomisation. We then obtained from this ensemble the expected average precision and accuracy as well as their standard deviations to obtain Z-scores.

### Bayesian model selection with stochastic block models (SBM) and definition of structural modules

Stochastic block models are simple generative models that assume that there are underlying groups of nodes in the network and that the probability that there is an edge running between nodes (i,j) only depends on the group memberships of i and j. As generative models, SBMs are amenable to Bayesian inference, and therefore to model selection techniques that allow us to find the best division of the nodes into groups. Nodes in the same group have statistically similar connection patterns and are thus interpreted to play a similar role in the network. Note that there is no a priori selection neither of the number of groups nor of the interactions between the groups. SBMs have been shown in the literature to be appropriate models for real network topology being successful at both error prediction (Guimerà and Sales-Pardo, 2009) and community detection (Peixoto, 2014).

We use a minimum description length approach (MDL), which is equivalent to maximizing the posterior, to find the best division of nodes into groups (https://graph-tool.skewed.de/; Peixoto, 2014). Specifically, we use the MDL approach to identify the best SBM variant (non degree-corrected and degree-corrected with and without hierarchical priors for the groups). Because the minimisation process is heuristic, we ran the algorithm 1,000 times to identify the best model (with minimum description length, $\Sigma$) and therefore best division of nodes into structural modules. We find that the best model is a degree-corrected SBM with hierarchical priors. We therefore obtain a hierarchical tree of network divisions into structural modules. Note that there is no a priori selection neither of the number of groups nor of the interactions between the groups. The inference methodology finds the division into groups that best describes the observed topology.

- - . We represent structural modules, , at the second most coarse-grained level in the hierarchy– level 1 (Figure S2D). We select this level because it summarises the networks and it is the first level that is significantly correlated with empirical modules for the two smallest GRNs we select (max AUFE and max PFE). In Figures 3, S3, andS4, connections between structural modules have a weight equal to the number of connections between genes in the two modules. We only represent connections with a weight > 260 to represent the main structure of the network of structural modules. All selected consensus GRNs have an input-core-output structure. In networks with this kind of topology it is possible to define three layers with different topological properties. The layer has genes that are either connected to genes in other layers or with genes within the same layer. The layer has genes that are connected either to genes in the layer or to genes in the layer. Genes in the layer only connect to genes in that same layer. This type of networks thus has a clear direction of 'information' or regulatory flow from the input layer to the output layer.

To better characterize the input-core-output structure, the tables below show the aggregate connections between the input, core and output layers.