Supplementary Text for "Detection of transcriptional triggers in the dynamics of microbial growth: application to the respiratory-versatile bacterium *Shewanella oneidensis*"

Table of Contents

List of additional Supplementary Material available online

Supplementary Table S2: For each of the nonessential predicted trigger genes we identify significant phenotypic patterns upon oxygen, lactate, nitrogen or stress signals. In each column the resulting p-values from the t-test comparison are reported. In the last row the significance from permutation test reflects how likely it is to randomly draw, among the totality of 3,355 nonessential mutants, an equal or greater number of genes with significant phenotypic patterns for each of the tested conditions (MS excel file)

Supplementary Dataset 1: Excel spreadsheet of all genes with log expression profiles in LAC minimal and LB rich media, opportunely clustered (MS excel file)

Supplementary Dataset 2: Excel spreadsheet of enrichment analysis of 5 clusters resulting from LAC and LB growth conditions (MS excel file)

Supplementary Dataset 3: Excel spreadsheet of significantly correlated genes according to GDM (MS excel file)

Supplementary Dataset 4: Excel spreadsheet of significantly perturbed genes inferred by D2T2 with log expression profiles in LAC and LB media (MS excel file)

Supplementary Dataset 5: Source file for raw data of all figures in the manuscript (MS excel file)

Supplementary Dataset 6: Excel spreadsheet of genes that were found common in GDM and D2T2 analysis

1. K-means clustering

Standard K-means clustering was performed to generate Supplementary Dataset 1 for LAC and LB growth conditions, respectively. The K-means clustering algorithm, using Pearson Correlation as the distance metric between gene expression profiles, was used to assign each gene profile to a unique cluster. The number of clusters was computed using the rule of a significant mode as suggested in (1). In particular, a comparison of the relative variance of each component obtained through singular value decomposition of the expression matrices revealed 5 significant eigenvalues (i.e. above the threshold: 0.7/number of modes) in both growth conditions (LAC and LB). The gene expression profiles from both growth conditions were hence clustered into 5 clusters each by the K-mean clustering algorithm (Supplementary Figure S1). Gene Ontology (GO) analysis performed on each individual cluster revealed enriched functional categories (see Supplementary Dataset 2). The enrichment analysis was performed using Gene Ontology Enrichment Analysis Software Toolkit (2)

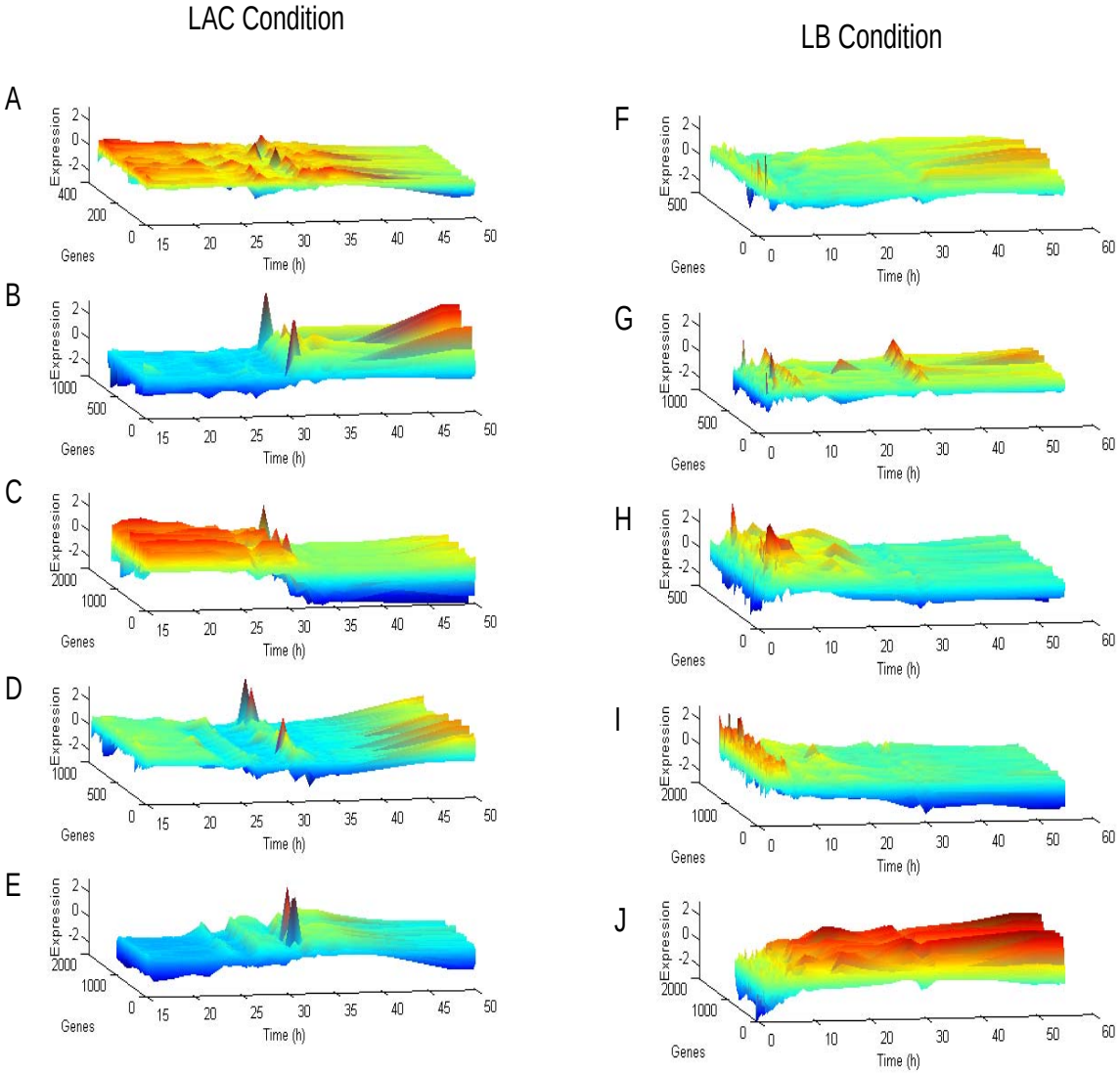


Figure S1: Gene expression profiles grouped in 5 clusters, by a K-means clustering algorithm for LAC (A-E) and LB (F-J) growth conditions. Details about number of genes in each cluster and their actual expression profile can be found in Supplementary Dataset 1. Functional enrichment analysis for each cluster is available in Supplementary Dataset 2.

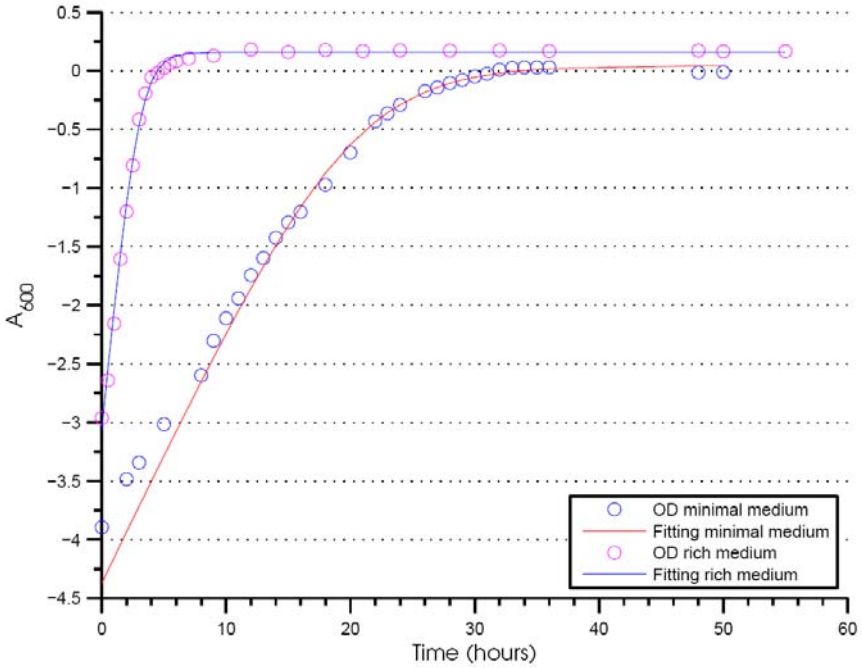
2. Growth derivative Mapping (GDM)

The Growth Derivative Mapping (GDM) analysis was implemented in order to compare gene expression profiles during batch growth under two radically different nutrient conditions. The goal was to put side by side the expression states of each gene across the two conditions, upon appropriately rescaling the corresponding time-scales. Specifically, we sought to compare expression levels at times that correspond to the same relative growth rate across the two conditions. The correct mapping was first inferred based on the OD measurements in the two conditions (i.e. by finding the parameters that would transform one curve into the other). The ensuing mapping (i.e. time-rescaling) was then applied to all genes, generating gene expression profiles comparable between the two conditions. This process basically maps gene expression profiles from a time-dependent to a growth-rate dependent domain. The goal, next, will be to identify genes likely to be dependent on growth-related processes, regardless of the different environmental conditions (i.e. nutrients availability).

In what follows, we start by describing in detail the different steps of the GDM process. The first step consists of modeling bacterial growth with a logistic model:

$$y = \frac{k}{1 + e^{a - bt}} \quad (2.1)$$

The parameters *a*, *b* and *k* are inferred by fitting this curve to the two reported growth data for the minimal (LAC) and rich (LB) media (**Supplementary Figure S2** and table below).



Supplementary Figure S2

Parameter	K	a	b
Minimal medium	1.0455	4.4033	0.2220
Rich medium	1.1713	3.1806	1.1081

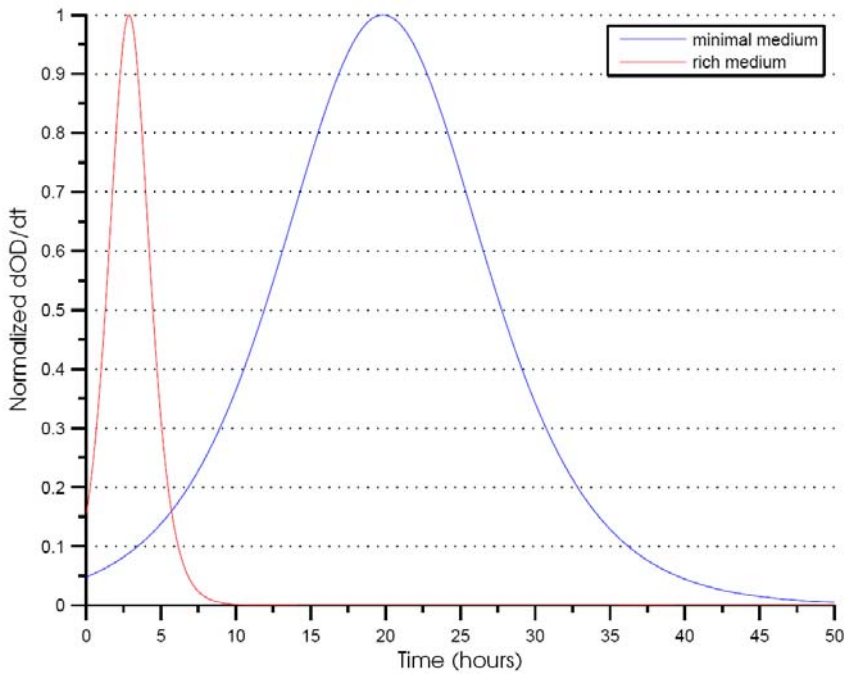
It is worth noting that we also tested the ability of another model (Gompertz model), commonly used in alternative to the logistic model, to describe our observed bacterial growth dynamics. In the table below we show slightly worse fitting results for the Gompertz model, in favor of the logistic one (Eq. 2.1):

$$y = ke^{e^{a-bt}} \qquad (2.2)$$

Residual sum of squares	LAC Minimal medium	LB Rich medium
Logistic	0.0149	0.0198
Gompertz	0.0519	0.0187

From the logistic model (Eq. 2.1) one can compute the analytical derivative of the two sigmoidal growth curves, as shown in Fig.S3, where instantaneous growth-rates have been normalized with respect to their highest values.

$$c = \frac{dy}{dt} = \frac{bke^{a-bt}}{\left[1+e^{a-bt}\right]^2} \qquad (2.3)$$



Supplementary Figure S3: Growth rate (c) as a function of time in minimal (LAC) and rich (LB medium)

For any given value of growth rate *c* (within [0,1]) there exist two time points at which such growth rate is reached ($z_{1,2}$), one before reaching the maximal relative growth-rate (i.e. 1), and one afterwards, when nutrient depletion slows down cell replication:

$$z_{1,2} = \frac{-(2\gamma - 1) \pm \sqrt{1 - 4\gamma}}{2\gamma} \qquad (2.4)$$

with $\gamma = \frac{c}{bk}$.

We then divide the time window of each growth experiments into two distinct growth phases: the first one, when growth rate is accelerating over time, and a second one where growth rate is decreasing. This distinction allows a univocal relationship between time points of the two different growth conditions. On the basis of this analytical framework the expression profiles of each transcript can be directly related to the instantaneous growth-rate. By matching time points corresponding to similar instantaneous growth-rates and growth-phases in the two conditions, we generate a non-linear mapping of the time course expression profiles during the exponential and stationary phases from the rich to the minimal medium profiles.

The final step is to asses the similarity between the two obtained rescaled expression profiles of the same gene. As a preliminary step we use a linear interpolation in order to generate a pseudo continuous description of the time dependent expression patterns. We then compute the similarity between the matched profiles using a Pearson Correlation index. To evaluate the significance of the above index we perform a permutation test that randomly shuffles the time order of the original dataset, recalculates the interpolated profile and the corresponding Pearson correlation value. From a comparison between the original similarity between rescaled expression profiles and the background distribution generated from random permutations we extrapolate empirical p-values, which are then corrected for multiple tests (3). Finally we select the genes with an associated q-value lower or equal to 0.01 (Supplementary Dataset 3).

3. Dynamic Detection of Transcriptional Triggers (D2T2)

In the following we describe in detail the D2T2 procedure, which is aimed at predicting systemic changes caused by external inputs on large scale gene-gene transcriptional dependency networks, using reverse engineering of time-dependent transcriptomic profiles (4, Reference 40 from main text). The underlying idea, first proposed in (Reference 42 from main text), is to employ a model where, for each gene, the rate of transcription is expressed as a function of all gene levels, as well as an external perturbative term (Eq. 3.1). A key feature of this approach is the assumption that this function is linear. With this simplification, the transcriptional rate (dx/dt) of each gene is considered to be a linear combination of the current mRNA transcripts level (x) of all genes in the genome, plus possibly an additive external stimulus/perturbation. Similarly to (Reference 42 from main text), the responses of the system to the various perturbations can be formalized as a system of linear differential equations (ODEs) in which the input vector $u = [u_1; \ldots; u_p]$ is modeled as a linear combination of the external inputs.

$$\frac{dx_i}{dt} = \sum_{j=0}^{n} a_{ij} x_j + \sum_{l=1}^{p} b_{il} u_l \quad , \quad x_0 = 1 \tag{3.1}$$

or in matrix form:

$$\dot{x} = Ax + Bu \tag{3.2}$$

where, a_{ij} represents the influence of gene j on gene i, while a_{i0} is the basal transcriptional level. The variable u can be taken as a vector ($p \times 1$) identical to 1. The n × p input matrix B collects the influences of the external perturbations u on each single gene. The k-th column of B indicates how effective/intense the k-th perturbation u_k is on the state vector (Reference 40 from main text). Typically, the inference of A and B constitutes an underdetermined problem, because the number of variables is much greater than the number of observations/microarray experiments.

In order to cope with this problem, we combine a time-course profile, usually restricted to a limited number of time samples, with a large compendium of steady-state measurements $X_{Training}$. Under steady state conditions the variation in time of the mRNA concentration is by definition equal to zero. Given this assumption, Eq. 3.2 can be reduced to a simpler system of algebraic equations:

$$Ax + Bu = 0. \tag{3.3}$$

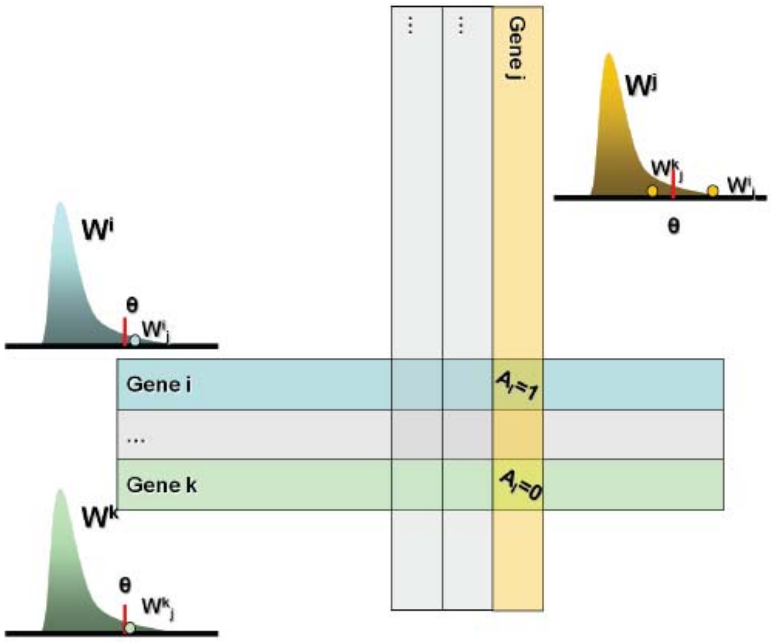
Hence, we will first seek to identify a nominal gene network model (matrix A), based on imposing a steady state assumption to the linear ODE description. This will lead to an initial network of significant transcriptional inter-dependencies. In a second step, this initial model will be used as the basis for integration of a time course experiment, leading to the identification of condition-specific gene targets (matrix B). In what follows, the two steps: "static network inference" and "dynamic enrichment" are described in details.

First step: Static Network Inference

The static gene-network inference procedure (largely similar to our previously proposed approach [see Reference 44 from main text] is further separated in three main sub-sections. The first sub-step is the reconstruction of an adjacency matrix (connectivity map) between gene pairs (A_I). This step will result in a gene network as an undirected graph, where nodes represent genes and edges are non-causal relationships between genes. The second stage (edges weighting) assigns weights to the edges of this graph, reflecting the strength of the associated interaction. In the third step (cut-off selection) an objective function is used to select the best inferred model from an ensemble of possible A_I matrices.

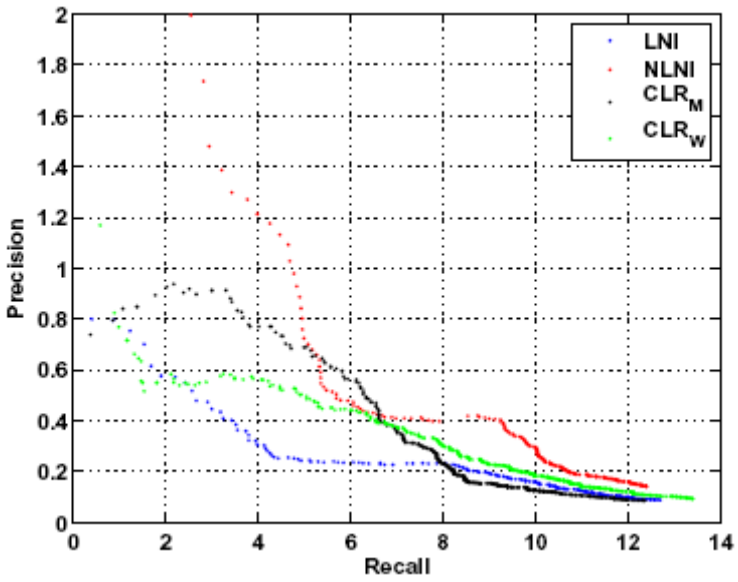
Connectivity map: In order to perform a truly system-level analysis of gene expression profiles, we employ a procedure belonging to the family of the relevance network algorithms. Relevance network algorithms have the remarkable advantage of being computationally feasible for genome-wide applications. Two different metrics are initially used: Pearson Correlation and Mutual Information. The former is a parametric statistic, which assumes normally distributed data and performs linear correlation tests. The latter is a dependency measure without any linearity assumptions. We call these two procedures Linear and Non-linear Network Inference (LNI and NLNI) respectively. A square matrix ($n \times n$) represents the pairwise similarities between gene expression patterns (D). Each row/column of D_{ij} can be considered as a distribution of pairwise similarity coefficients (D^i)
between a gene *i* and all other genes. This distribution can be transformed in the corresponding vector of rankings, from *1* to *n* (W^i). As a consequence of this transformation, the correlation between *i* and *j* is associated to two ”ranking values”. Denote $W^i(j)$ and $W^j(i)$ the rank of D_{ij} with respect to D^i and D^j respectively. Genes (i and j) are considered as putative interactors if and only if both their ranking indexes ($W^i(j)$ and $W^j(i)$) are above a certain common threshold θ. A graph of putative interactions A_I is obtained by the following relationship:

$$(i, j) \in A_I \Leftrightarrow W^i(j) \geq \theta \,\&\, W^j(i) \geq \theta \tag{3.4}$$



Supplementary Figure S4: Linear/Nonlinear Network inference: This scheme represents the selecting criterion implemented in Eq. 3.4. The similarity between gene i and j ($D(i,j)$) is above the same threshold (θ) in both the ranking distributions (W^i and W^j). Therefore, an edge connects the two genes ($A_I(i,j)=1$). On the contrary, $D(k,j)$ is satisfying the criterion only with respect to W^k and therefore the edge is pruned ($A_I(k,j)=0$).

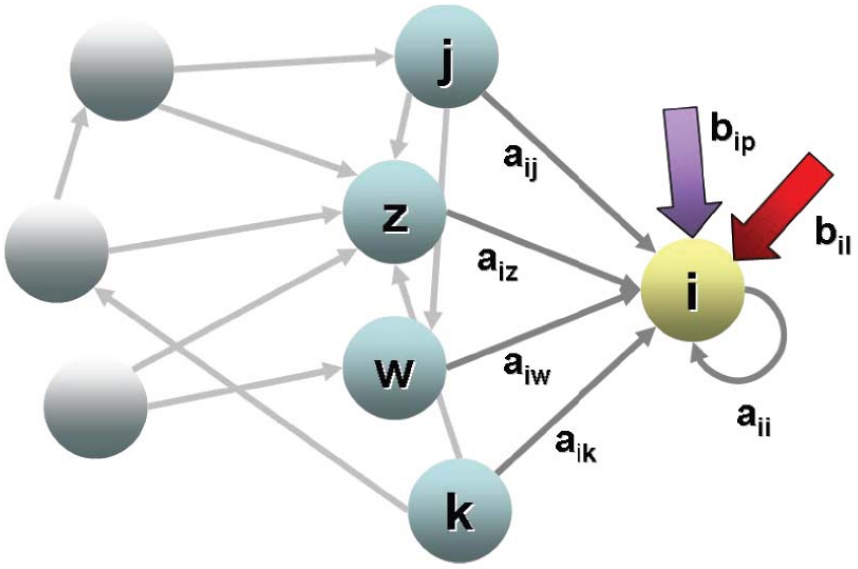
The above procedure prunes indirect gene-gene interactions differently from the strategy used in the Context Likelihood of Relatedness algorithm (CLR) (Reference 41 from main text), where a joint likelihood estimates based on mutual information is used. A point worth noting is that the selected threshold in Eq. 3.4 constrains the maximum number of possible interactions for each gene. This results in a total number of retained pairwise interactions different than the one obtained by just selecting the portion of interactions with a similarity greater than a threshold (corresponding to the fraction of highest indexes among all the possible pair wise combinations, i.e. $(n^2-n)/\theta$). In Supplementary Figure S5 we compare the ability of CLR (applied to the Pearson correlation / CLR_W and mutual information / CLR_M matrices), NLNI and LNI to recover the transcriptional regulatory network of *E. coli*, showing similar performances.



Supplementary Figure S5: CLR comparison: Precision *vs* Recall curves is reported for the transcriptional regulatory network of *E.coli* (regulonDB). The CLR criterion yields slightly better results if applied on Pearson Correlation in agreement with (Reference 41 from main text) and slightly worse if applied on mutual information.

In an ideal scenario, assuming that the underlying regulatory network is known, the model of Eq. 3.1 can be further simplified. The rate of synthesis of gene i is dependent on the expression of a sub-set of genes it interacts with, and on the external perturbation (u) (see Supplementary Figure S6 below), hence Eq. 3.3 can be read as:

$$0 = \sum_{(i,j)\in A_I} a_{ij} x_j + \sum_l b_{il} u_l \qquad (3.5)$$



Supplementary Figure S6: **Gene network model:** gene i interacts with a limited number of genes (i.e. $A_I = \{i, j, z, w, k\}$) and experiences the influence of external inputs (i.e. $u = [u_p, u_l]$): $\frac{dx_i}{dt} = f(x_i, x_j, x_z, x_w, x_k, u_p, u_l)$.

Solving Eq. 3.3 implies solving a linear system of n algebraic equations, where A_I now is a $n \times n$ matrix with a zero entry for non-interacting genes, and a non-zero value for each interacting gene pair.

In order to estimate the parameters of the model we employ a large compendium of gene expression profiles, containing many different external perturbations (Reference 39 from main text). The key assumption is that the matrix A is invariant over all the different experimental conditions, and that the columns of B are sparse (i.e. have few nonzero elements). This last assumption reflects the fact that the different perturbations collected in the dataset are likely to affect just a small portion of genes, if compared to the entire genome. This is justified for example in case of mutants, where a specific gene has been knocked out. Therefore, for each transcript, expression changes caused by the direct effect of an external stimuli are likely to occur just in few of the experimental conditions of the training dataset, and to be ”uniformly” distributed over the rows of B (i.e. different perturbations do not act always on the same transcripts).

Hence, when looking at a large compendium of gene expression profiles collected from diverse experiments it is possible to consider B to be sparser than matrix A, and hence negligible. A sufficiently accurate estimation of the non-zero terms of the A matrix (a_{ij}) can still be drawn from the following approximation

$$Ax \approx 0. \tag{3.6}$$

With the advantage of an appropriate sparse adjacency matrix A_I, and a suffiently large collection of gene expression profiles, we can solve Eq. 3.6 as a least square

problem (i.e. the solution corresponds to the least sum of squared residuals). More specifically, we used here a multiple regression framework to learn the network coefficients a_{ij} from the training dataset $X_{Training}$. A solution is then found by minimizing the L^2 norm between the predictions of the model and the experimental values.

In summary, the LNI method is initially adopted here to infer a subset of putative "linearly interacting" partners for each gene, followed by the actual estimate of the set of coefficients a. This step is equivalent to reducing the dimensionality of the least squares problem (Eq. 3.3). Another critical reason for using this selection process criterion (instead of eg. CLR) is the necessity to control for each gene the maximum number of possible interactions (to avoid under-determination).

Edges weighting: As mentioned above, once A_I is inferred, we can use a multiple regression framework to learn the network model coefficients. We will show now that this simplification is more effective and computationally convenient than the procedure used in (Reference 43 from main text). For each gene we search for the vector of coefficients (a) that minimizes Eq. 3.7 (Least square regression).

$$\bar{a} = \underset{a}{argmin} \left\| 0 - \sum_{(i,j)\in A_I} a_{ij} x_j \right\|^2 \tag{3.7}$$

In order to avoid the trivial solution of $\bar{a} = 0$ we assume that each gene potentially is under the effect of a self-feedback loop (a_{ii}), i.e. it can self-regulate itself. This assumption allows us to perform the following useful manipulation of Eq. 3.1: the term $a_{ii}x_i$ can be extracted from the summation and moved to the right hand side of the equation (Eqs. 3.8, 3.9):

$$\sum_{(i,j)\in A_I} a_{ij} x_j + a_{ii} x_i = 0 \tag{3.8}$$

$$x_i = - \sum_{(i,j)\in A_I} \frac{a_{ij}}{a_{ii}} x_j . \tag{3.9}$$

Upon dividing both terms by a_{ii} we can use the least square multiple regression scheme to find the set of coefficients $\bar{a}$. The result is a set of normalized coefficients ($\bar{a}_c$).

$$\bar{a}_{c_i} = \underset{a}{argmin} \left\| x_i - \sum_{(i,j)\in A_I} \frac{a_{ij}}{a_{ii}} x_j \right\|^2 \tag{3.10}$$

The solution of Eq. 3.10 is correct up to an undetermined scaling factor (the diagonal term a_{ii}) by which we rescale each row of A. The outcome of the regression problems is a weighted asymmetric matrix A_c, as well as a vector of residues (r_i), resulting from the

difference between observed and predicted values for each gene. The A_c matrix has the characteristic that all the elements on the diagonal are equal to one, while the non-zero elements represent the normalized influence of each transcript on the rate of synthesis of a specific gene:

$$A_c = C^{-1}A, \quad C = diag([a_{11}, a_{22}, ..., a_{nn}]) \tag{3.11}$$

where, C is a diagonal matrix.

Cut-off selection: A general issue in relevance network algorithms is the cut-off selection. The choice of the threshold in Eq. 3.4 can be chosen arbitrarily (e.g. as in (7)) and can strongly affect the fitting results of Eq. 3.10. Therefore, a rational choice of the cut-off (and therefore of A_I) would be preferable. In this last sub-step we define an objective function guiding the selection of the cut-off θ used in Eq 3.4 which will substitute the arbitrary selection implemented before. Towards this goal, we utilize a metric describing the goodness of the fit for Eq. 3.10 (i.e. how well the model predict the experimental data).

When comparing models with different numbers of free parameters, the R-square values are statistically irrelevant, since an increase in the number of free parameters in a fitting equation always reduces the absolute sum of residuals (r) and hence increases the R-square value. Therefore, we fitted multiple models with different choices of θ in Eq. 3.4 and estimated the corresponding fitness qualities via the Bayesian Information Criterion (BIC). This criterion attempts to find the best compromise between the model complexity, in terms of the number of parameters (i.e. non zero entries in the connectivity map A_I) and fitness quality. BIC (Eq. 3.12) is the most common criterion for the fitness of a mathematical model to observed data, leaning more towards lower dimensional models (i.e. sparse matrices)

$$BIC_i = N\left[ln\left(\frac{RSS_i}{N} \right) \right] + ln(N)K_i \tag{3.12}$$

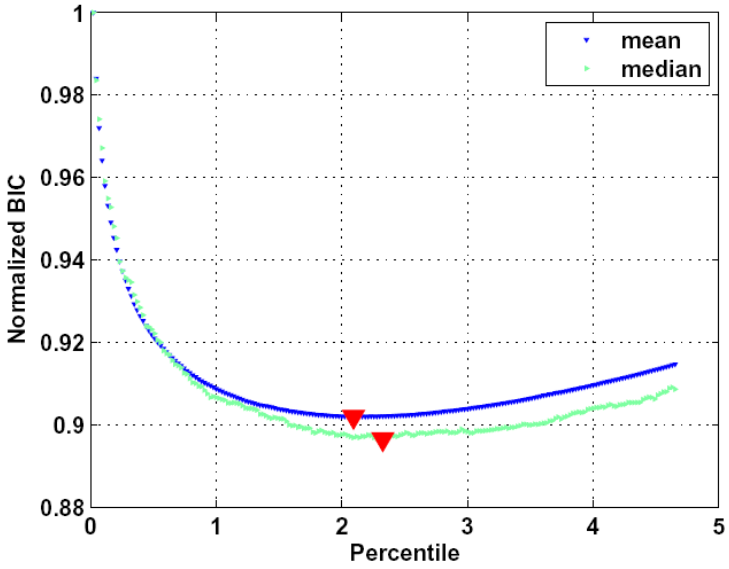
where K_i is the number of parameters in the model for gene i (i.e. number of edges of i), N the number of observed data points (i.e. number of chips) and RSS_i the residual sum of squares ($\sum_{z=1}^{N} r_{i_z}^2$). The mean of BIC indexes over all the genes leads to a scoring function.

The value of θ corresponding to the minimum of the mean of BIC_i indexes ($\tilde{\theta}$) is selected for the next step of the analysis (Eq. 3.4):

$$\tilde{\theta} = \underset{\vartheta}{argmin}\left(\sum_{i=1}^{n} BIC_i \right). \tag{3.13}$$

The identification of the graph A_I is the most under-determined step of this

reverse engineering approach. However, coherently with (Reference 43 from main text), we have found that small changes of θ (and hence A_I) do not drastically change the structure of A_c and its numerical entities (see Supplementary Figure S7). In summary, the described procedure, provides an automated method to reverse engineer the gene-gene dependencies from a large compendium of steady-state microarray experiments, where we do not have to fix any parameters in order to extrapolate a weighted un-directed gene network.



Supplementary Figure S7: BIC objective function. Different thresholds (i.e. θ values used in Eq. 0.4) are reported on the horizontal axis. The mean (blue triangle) and median (green triangle) BIC values, normalized by the highest score, are reported on the y-axis. The mean curve is smoother. Nevertheless, the thresholds θ associated to the two minima are close to each other.

Second step: Dynamic Enrichment

In the previous sections we showed how we infer the weighted matrix of gene-gene relatedness, A_c. We will now investigate the possibility of using this model to discriminate between significant and neglible variations around these nominal profiles during a time course experiment. The inferred gene-network (A_c) is now used to filter the information content of perturbed cells dynamically switching from their original steady state to a new one. In order to take advantage of the time content information embedded in time course gene expression profiles we use the dynamic model.

$$\dot{x}_i(t) = \sum_{(i,j)\in A_I} a_{ij} x_j(t) + \sum_l b_{il} u_l(t) \qquad (3.14)$$

or in matrix form:

$$\dot{x}(t) = Ax(t) + Bu(t). \tag{3.15}$$

Since the transcript levels of each gene are now measured at different time points, we estimate the transcription rate change by approximating the continuous derivative in time following Euler scheme.

$$\Delta x_i(t_{q+1}, t_q) = \Delta x_i = \frac{x_i(t_{q+1}) - x_i(t_q)}{\Delta t} \tag{3.16}$$

The higher the sampling frequency, the more accurate the approximation. As previously mentioned, we approximate the external function $u(t)$ as a constant ($u(t) = 1$). These approximations lead to the following formulation of the problem:

$$\Delta x = CA_c x + B. \tag{3.17}$$

The resulting system of linear differential equations can be solved by a second linear least square regression. Note that we now face a regression problem in which, according to Eq. 3.17, the unknown parameters (for each gene) are only two: the normalizing vector C and the perturbation intensities B. Therefore, for each gene the entire set of time-points can be now used to fit only these two parameters, avoiding the typical disadvantages of a highly under-determined problem in the anlysis of time-course datasets.

$$[c_i, b_i] = \underset{c,b}{argmin} \left\| \Delta x_i - c_i \sum_{(i,j)\in A_I} a_{c\,ij} x_j - b_i \right\|^2 \tag{3.18}$$

By solving Eq. 3.18 we find the set of normalizing factors (C) and external perturbations (B) that minimize the difference between the observed mRNA change in time and the model predictions. By doing this, we effectively perform an integration over time of the perturbation experienced by gene i. This second regression unveils a condition-specific gene-network, where a weighted directed graph describes the most relevant relationships during dynamic changes (CA_c). Most important, this regression yields also a vector of perturbations intensities (B) that can be further used to infer which are the genes mediating the response to the external perturbation. A major advantage of this procedure with respect to the Time Series Network Identification (TSNI) procedure is the absence of any tuning in parameter selection by the user (i.e. the pre-selection of the genes considered to be relevant in the TSNI procedure), leading to a truly genome-wide approach.

Statistical test

As already mentioned, those genes that are associated with a high perturbation value are most likely to be the target genes of an external agent. However, instead of just ranking the genes on the basis of the associated B values, like in (10) we asses in parallel a

statistical significance of the estimation (i.e. a q-value). In fact, one could ask whether a high perturbation value for gene i (b_i) can be due just to the inadequacy of the initial inferred nominal model (A_c) or from the rough approximation of the derivative of Eq. 3.16.

In order to assign a statistical relevance index to each gene perturbation we take advantage of the first regression step (Eq. 3.1). The hypothesis made in Eq. 3.6 is that each gene is observed under steady state condition with no external perturbation. Hence, we consider the residues (r) derived from the estimation of A_c (Eq. 3.10) as a background null distribution (corrected by the scaling factor C, see Eq. 3.21) against which to compare the estimates of the condition-specific perturbation intensities B of (Eq. 3.18).

$$A_c x \approx 0 \Rightarrow A_c x = r \tag{3.19}$$

$$CA_c x \approx 0 \Rightarrow CA_c x = r_c \tag{3.20}$$

$$r_c = Cr \tag{3.21}$$

A p-value is now assigned on the basis of this background distribution. Once the distribution of empirical p-values have been generated, the q-value is used to correct the measure of significance for multiple testings (3). Genes with a q-value smaller or equal than 0.01 are consider to be significantly perturbed.

Method comparison

In this last section we summarize differences, advantages and limitations of our method when compared to the state-of-the-art procedures which aim at inferring stimulus-induced transcripts. In (Reference 40 from main text) the authors proposed an approach (Mode-of-action by Network Inference - MNI) where a static model is used to analyze the expression changes in compound-treated cells. An alternative solution is explored in (10) (SSEM-Lasso algorithm). Both strategies aimed at inferring the perturbed genes by ranking them according to the estimated B values. While it has been shown that SSEM performs better than MNI in identifying genetic perturbations (i.e. knockout/overexpression) (10), results are inconsistent when inferring compound gene-targets. As a matter of fact these methodologies do not take into account any time-dependency. This gap has been filled by the TSNI method, which was applied to infer norfloxacin targets in *E.coli* from time-course experiments (Reference 43 from main text). In (Reference 40, 43 from main text) a singular value decomposition (SVD) is used to map the gene expression profiles in a smaller number of characteristic profiles. This procedure allows to project an n dimensional expression data into an arbitrary lower dimensional space, thus making Eq. 3.3 solvable. Briefly, the discrete form of Eq. 3.1 is first formulated as follow:

$$x(t_{k+1}) = A_d x(t_k) + B_d u(t_k) \tag{3.22}$$

which can be rewritten in a compact way:

$$X = H\,Y \tag{3.23}$$
$$H = [A_d \;\; B_d] \tag{3.24}$$
$$Y = [X^T \;\; U^T]^T. \tag{3.25}$$

It is worth noting that A_d and B_d are different from their continuous counterparts A and B. The authors, first, used Principal Component Analysis (PCA) to arbitrarily select the top k singular values and hence reduce the dimension of Eq. 3.23. A bilinear transformation is then used to compute the continuous network model A and perturbation intensity B:

$$A = \frac{2}{\Delta t}\frac{A_d - I}{A_d + I} \tag{3.26}$$
$$B = (A_d + I)^{-1} A_d B_d \tag{3.27}$$

where, I is the square identity matrix and Δt is the sampling interval.

Our method is independent of any pre-selection step or parameter fine-tuning, and is based on a purely data-driven approach. D2T2 is also capable of performing a statistical evaluation of model predictions, associating a confidence level (p-value) to each gene.

4. Dynamic flux balance analysis (dFBA)

Dynamic flux balance analysis (dFBA) was implemented as described previously (Reference 45 from main text) to produce time-dependent metabolic reaction rates (fluxes, $\boldsymbol{v}\equiv(v_1, v_2, \ldots, v_N)$) for *S. oneidensis*. In particular, we used the static optimization method which is more suited for genome-scale stoichiometric models. Briefly, dFBA performs iterations of flux balance analysis (Reference 46, 47 from main text) with flux bounds computed based on external metabolite concentrations updated at each time step. In this context, in addition to using fluxes ($\boldsymbol{v}$) measured per unit of biomass (as in regular FBA), we are also going to use fluxes that pertain to the total biomass ($BM=BM(t)$, measured in grams) present in the system ($\boldsymbol{V} = \boldsymbol{v} \cdot BM$). The dFBA formulation also requires keeping track of a vector $\boldsymbol{e}\equiv(e_1, e_2, \ldots, e_M)$ indicating the time-dependent concentrations of extracellular metabolites. All fluxes fall into one of two types: a biological flux is one that carries out some protein-mediated biological function (enzymatic reaction, transport), whereas an exchange flux, or source/sink flux, is one that balances the biological fluxes in the model. Formally, different types of reactions can be characterized by different sets of indeces:

$I_{bio} = \{\, j \mid \boldsymbol{V}_j$ is a biological flux $\}$ with $(|I_{bio}| = N\text{-}M)$
$I_{ex} = \{\, k \mid \boldsymbol{V}_k$ is an exchange flux $\}$ with $(|I_{ex}| = M)$

Note that exchange reactions in the current model are defined as sink fluxes (from inside

to outside of the cell). This is very useful for determining external metabolite availability under dynamically changing conditions. For example, if we write the reaction for the exchange of Lactate (with extracellular concentration e_{lac})as $\mathrm{Lac_{in}} \rightarrow \mathrm{Lac_{out}}$, with flux V_{lac}, then by constraining $-L \leq V_{lac} \leq \infty$, we effectively allow lactate to flow out of the cell at arbitrary rate, but impose an upper bound for lactate intake to a value L. Hence, in general, the maximal availability of external metabolites is translated into lower bound constraints on all exchange fluxes V_k, $k \in \mathrm{I_{ex}}$.

At each dFBA time step (of duration Δt), the lower bounds on the exchange fluxes are set based on to the current value of $\boldsymbol{e}$. Specifically, the maximal exchange flow possible in the system is given by the amount of material present, divided by the time elapsed. Hence, we impose on the biomass-dependent fluxes $\boldsymbol{V}$ the following constraints:

$$lb_i(t) = \boldsymbol{e}_i(t) / \Delta t \text{ for all } i \in I_{ex}$$

Note that the export of metabolites ($\boldsymbol{ub}$) is left unconstrained. In our simulations, we used $\Delta t = 1$ hour.

In order to consistently solve the problem for the total biomass available, at each time point (before performing the FBA step) we convert also the constraints on the biological fluxes from specific ($\boldsymbol{lb}^{(0)}$, $\boldsymbol{ub}^{(0)}$, defined per unit of biomass, mmol/g*h, as in standard FBA) to total ($\boldsymbol{lb}$, $\boldsymbol{ub}$,in mmol/h):

$$lb_i(t) = lb_i^{(0)}(t) \cdot BM \text{ for all } i \in \mathrm{I_{bio}}$$
$$ub_i(t) = ub_i^{(0)}(t) \cdot BM \text{ for all } i \in \mathrm{I_{bio}}$$

We then perform the FBA calculation, using a standard formulation. If S is the stoichiometric matrix, $\boldsymbol{c}$ the objective function, the FBA problem s formulated as:

$$\max(\boldsymbol{c}^T \boldsymbol{V}) \text{ s.t.:}$$
$$\boldsymbol{S}\boldsymbol{V} = 0 \text{ and}$$
$$lb_i \leq \boldsymbol{V}_i \leq ub_i$$

Using the new bounds, the media compositions $\boldsymbol{e}$ and the biomass BM are updated for the next time step:

$$BM(t + \Delta t) = BM(t) + \boldsymbol{V}_{\mathrm{biomass}} \Delta t$$
$$\boldsymbol{e_i}(t + \Delta t) = \boldsymbol{e_i}(t) + \boldsymbol{V}_i \Delta t \text{ for all } i \in \mathrm{I_{ex}}$$

The resulting variables $\boldsymbol{e}$ and BM thus record the metabolite levels in the medium and the amount of biomass over time.

The *S. oneidensis* MR-1 genome-scale metabolic model iSO783 (Reference 48 from main text) was used to predict growth of the bacteria in the LAC experimental condition. This model contains 774 reactions and 634 unique metabolites. Also, in keeping with the calculations made by (Reference 48 from main text) the non-growth rate dependent ATP

maintenance cost was set to 1.03 mmol gDW^{-1} hr^{-1} and the maximal uptake and secretion rate for L- and D- lactate was 4.08 mmol gDW^{-1} hr^{-1}. A death rate of 2% was also implemented. Other model parameters were made consistent with the experimental setup. This included both using the M4 medium formulation (Supplementary Table S1) with supplemented D- and L-lactate as the starting media composition; and refreshing the external O_2 levels to a constant concentration after each FBA cycle throughout the dFBA simulation (described in Strain, Cultivation and Sample Collections section of main text). To test the effect of O_2 on overall growth dynamics, a range of constant concentration values between 2 and 8 mM/hr were tested. A secondary objective function was implemented to maximize glycogen when biomass could not be produced. This was done by changing the objective function coefficients from $\boldsymbol{c}$(biomass) = 1 to $\boldsymbol{c}$(biomass) = 0.99 and $\boldsymbol{c}$(glycogen) = 0.01. When this combination was maximized, biomass was always favoured when the appropriate resources were available. For glycogen maximization to function properly, exchange and transport reactions from glycogen were added to the metabolic model. Shuttling excess glycogen outside the cell allows for accumulation over time to occur in the extracellular environment, simulating effectively the internal sequestration that happens *in vivo*.

5. Phenotypic analysis

In the following, we briefly summarize the strategy used to interrogate genome-wide fitness profiles published in (Reference 94 from main text). In our study, by integrating different source of data (i.e. expression and metabolite measurements) and computational frameworks (i.e. D2T2 and dFBA), we predict gene-triggers responsive upon three main external signals (nitrogen, oxygen and lactate limitations). In order to reinforce the validity of our predictions and provide complementary/independent evidences of a more general role of these genes in sensing such external perturbations we perform a comparative analysis on the data published in (Reference 94 from main text). The fitness for 3,355 nonessential mutants in 121 diverse conditions has been tested using DNA tag-based pooled fitness assay. The different conditions include aerobic/anaerobic experiments, different media compositions, among which different carbon (such as lactate) and nitrogen sources, and different stresses.

For each mutant, we used a two-sample t-test in order to identify significant fitness patterns in three main experimental conditions: anaerobic, lactate as sole carbon source, conditions where the nitrogen source is the only changed parameter. In Table S2, we report the results obtained for the predicted trigger-genes during batch growth in lactate minimal medium. Each column of Table S2 contains the t-test significance (i.e. p.-value) in each of the aforementioned conditions. Several genes show a significant (p.-value$\leq$0.01) phenotypic signature in at least one of the tested condition. This means that most of the predicted trigger-genes are indeed confirmed by independent data source to play a crucial role in interpreting and triggering the response to variations of specific external metabolic conditions (i.e. oxygen, nitrogen and lactate).

In order to test whether our list of trigger genes was significantly enriched for transcripts showing phenotypic patterns upon oxygen, lactate or nitrogen signals, we perform a permutation test, where we iteratively randomly select an equal subset of

putative trigger genes and count how many times a significant phenotypic patterns in each of these conditions was found. By performing this analysis (results are reported in the last row of Table S2), we highlight the enrichment in our predicted gene list for transcripts showing significant phenotypes under these 3 environmental conditions (p-values equal to 0.068, 0.091 and 0.025, respectively). In particular, there seems to be a prevalence of genes responding to nitrogen viability. As a judgment check we can clearly see that our predicted list of genes seems to have nothing to do with stress related signals (p-value=0.75), as one would expect given the environmental conditions in our experimental setup.

6. Data Collection

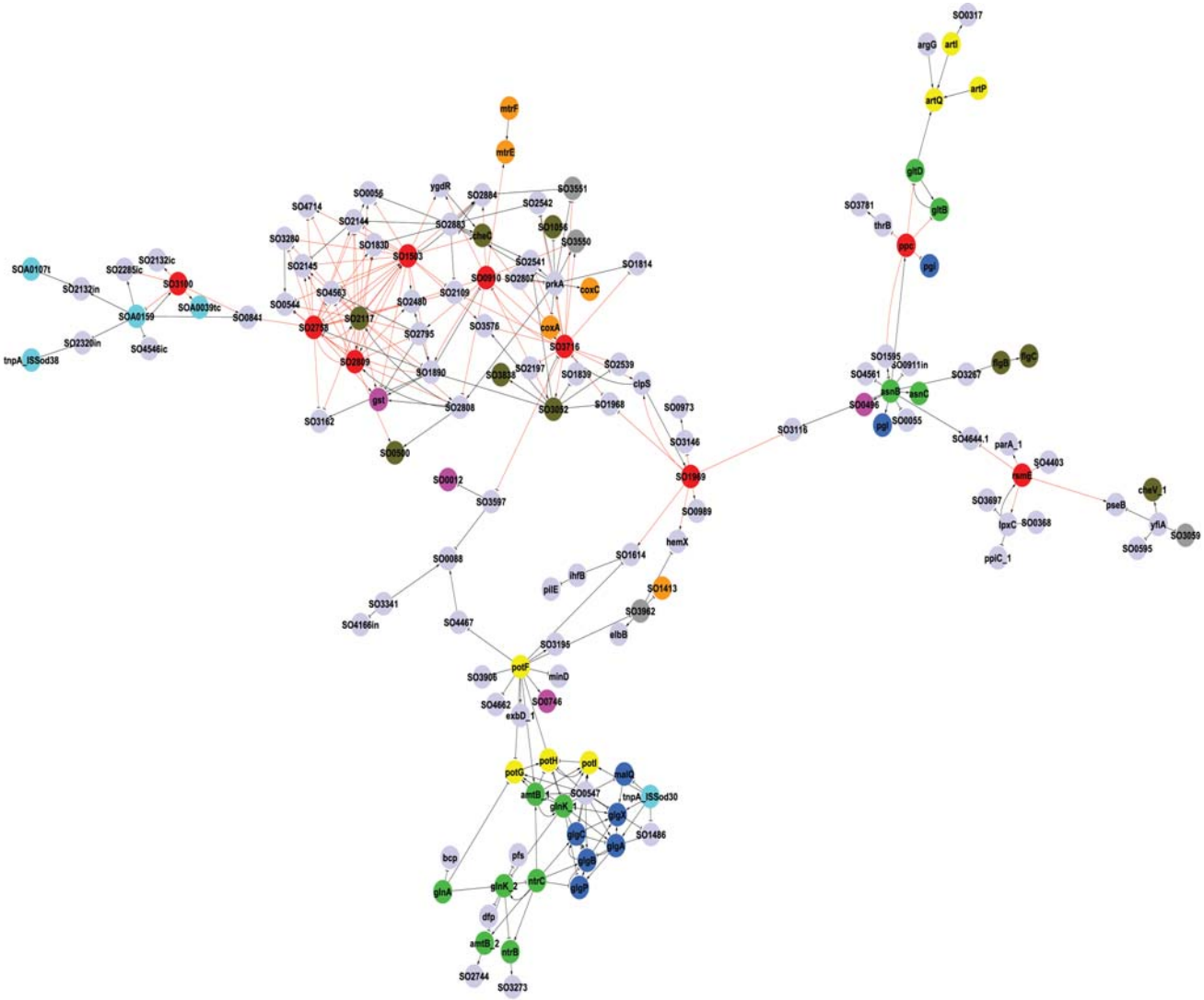
During analysis of chips data, we accessed the quality of each microarray chip using the quality control R-package (Bioconductor package: QC Report). The raw gene expression data was normalized using the Robust Multi Array analysis (RMA). In particular pairs of replicate chips were discarded if the correlation between two replicates was not among the top 10% of all possible pair-wise correlations between the replicate pair and all other chips in the data set. Datasets were downloaded from M3D (Many Microbe Microarrays Database, build E-coli-v4-Build-4) (Reference 39 from main text) for *S. oneidensis*, *E. coli* (to test D2T2) and *S. cerevisiae* (to test LNI and NLNI). Both datasets were preprocessed and normalized by RMA Irizzary et al 2003. We obtained TF-BS networks from the RegulonDB database, version 5.6, for *E. coli* (Regulon DB database: http://regulondb.ccg.unam.mx), and from (6) for *S. cerevisiae*. Here we use the M3D dataset (Reference 39 from main text) as a collection of steady state conditions. This assumption is itself a further simplification, adopted also in other reverse engineering algorithms. In particular, it is important to note that the steady state assumption is not exactly holding for all the experiments. In addition, the presence of some replicates affects the number of real independent observations. Given that the size of the dataset is a critical aspect in this type of algorithms we considered these deviations from an ideal dataset as an acceptable compromise.

7. Supplementary Table S1: Lactate M4 minimal medium used in the study.

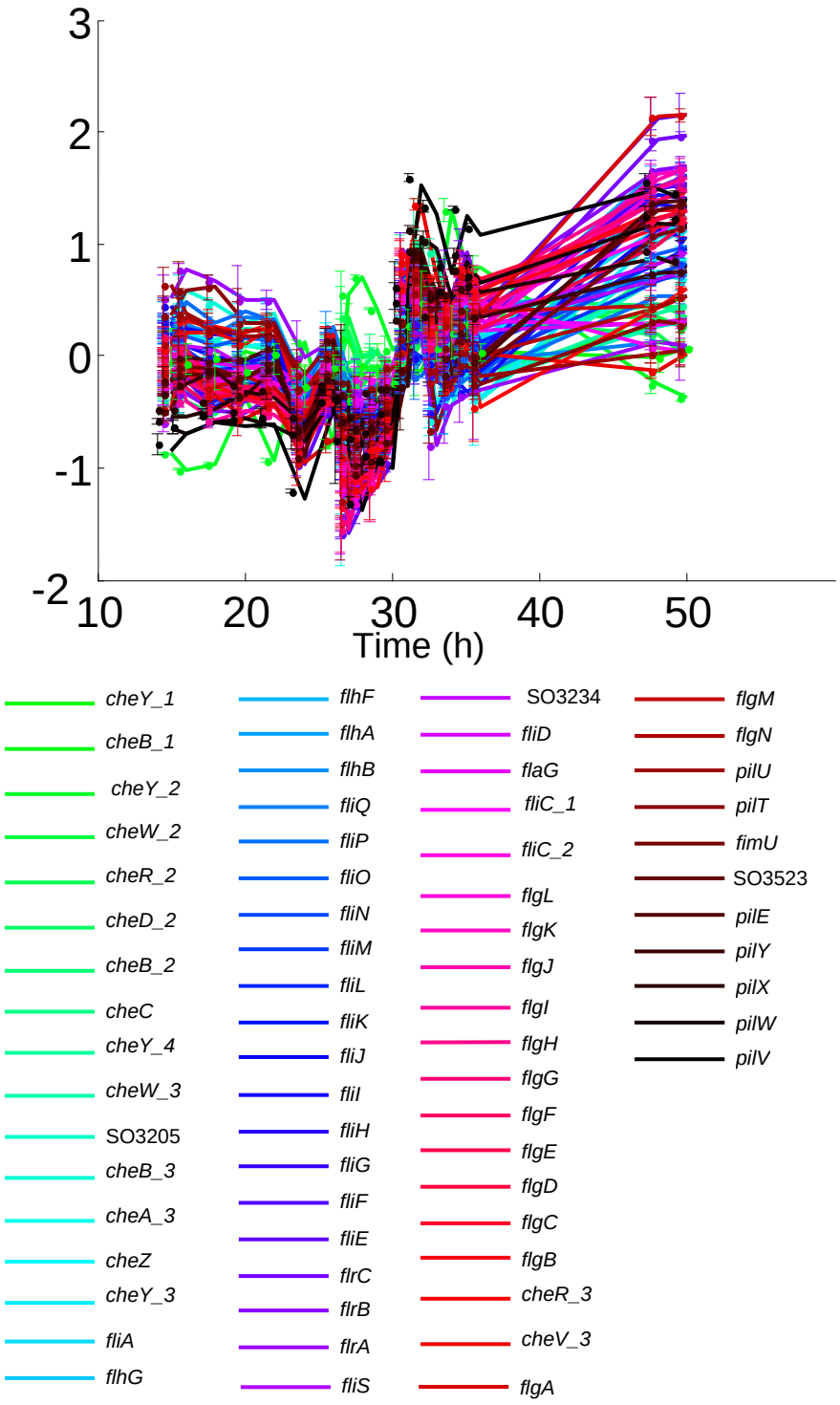
This medium consists of several parts that were mixed with each other only after autoclaving:

1. *Minerals solution*	
Component	Final concentration in mM
EDTA	0.0672
$MgSO_4.7H_2O$	1.0100
$MnSO_4.H_2O$	0.0013
$FeCl_2.4H_2O$	0.0054
$CoCl_2.6H_2O$	0.0050
$ZnSO_4.7H_2O$	0.0010
$CuSO_4.5H_2O$	0.0002
H_3BO_3	0.0566
$Na_2MoO_4.2H_2O$	0.0039
$NiCl_2.6H_20$	0.0050
Na_2SeO_4	0.0015
2. $CaCl_2.2H_2O$	0.485
3. *Buffer*	
K_2HPO_4	5.08
KH_2PO_4	2.92
$NaHCO_3$	8
$(NH_4)_2SO_4$	36
NaCl	500
Hepes	20
4. *Lactate*	50

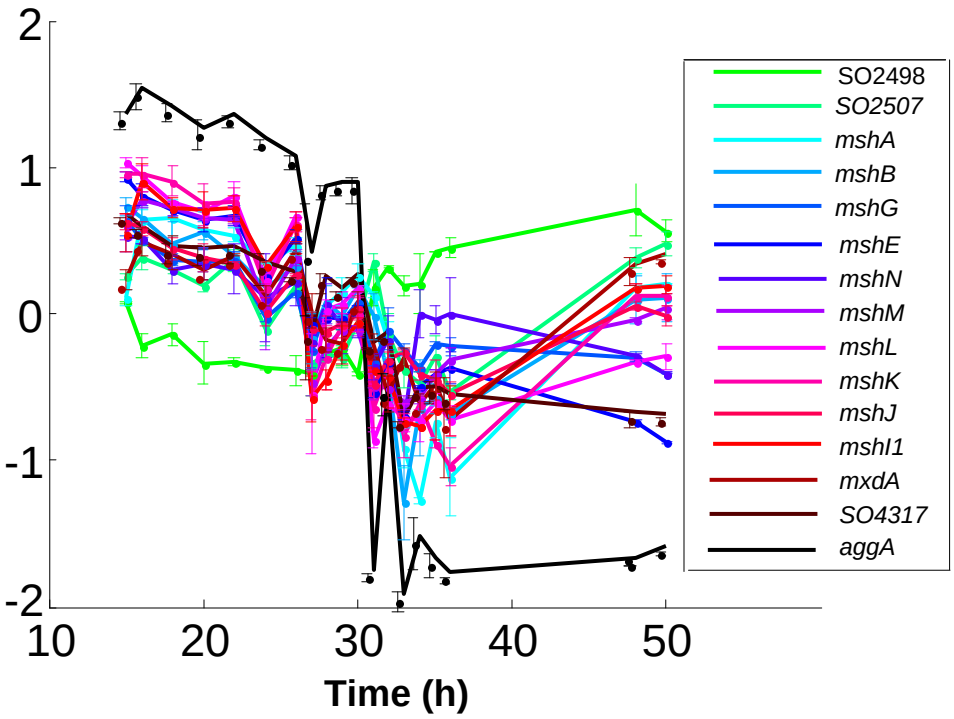
8. Additional Supplementary Figures S8-S10 as quoted in manuscript:



Supplementary Figure S8: Subnetwork of Glycogen and N_2 genes. This subnetwork has been generated by retaining only the most significant gene-gene interactions as predicted by the D2T2 algorithm. More precisely, by selecting only the top 1% among all the gene-gene interaction, the network is decomposed/fragmented in several connected components. Each component is characterized by a different number of genes (i.e number of nodes which are directly or indirectly connected to each other). Among all the components the one we here discuss has the largest number of genes. This subnetwork is enriched for genes involved in glycogen and nitrogen metabolism. Particularly relevant is the presence of *glnI, glnG* and *glnA* regulating the glutamate metabolism in low nitrogen conditions. Genes for ABC polyamine transporters and for polyamines (putrescine, spermidine and spermine) biosynthesis are also present as well. The predicted perturbed genes and their interactions are highlighted in red. The relationships directionality is represented by an arrowhead (teed for inhibitory interactions). The red genes in the figure might be considered as the entrance nodes sensing an environmental signal, which is further "processed" by the white nodes. So all together this subnetwork can be thought as the entrance pathway of external perturbations.



Supplementary Figure S9: Expression profiles ($\log_2$ on y-axis) as a function of time (x-axis) for flagella, chemotaxis and pili genes. The activation of several genes involved in flagellar assembly, pili biogenesis and chemotactic sensing is consistent with the observations in *E. coli*, where motility is induced upon starvation (7)



Supplementary Figure S10: Expression profiles (log_2 on y-axis) as a function of time (x-axis) for known biofilm-related genes. An additional level of regulation detected among the environmentally responsive genes include signalling with the molecule c-di-GMP. Previous work has identified c-di-GMP as a key regulator in biofilm formation by *S. oneidensis* MR-1 (8), a biological phenomena known to be linked to environmental conditions. MxdA, a diguanylate cyclase-like protein important for biofilm formation, was recently identified as controlling cellular levels of c-di-GMP (9). Interestingly MxdA and the biofilm-promoting protein (SO4317/ BpfA) are among the perturbed gene set in our lactate-growth gene expression data (Supplementary dataset 4). In addition this gene set also includes MSH pilus biosynthesis genes (*mshB, mshG, mshN*) and many other gene products known to be linked to *S. oneidensis* biofilm formation (10). Candidates for diguanylate cyclases for synthesis of di-c-GMP include the GGDEF domain-containing proteins (SO2498, SO2507)

9. Supplementary References

1. Everitt, B. and Dunn, G. (2001) *Applied multivariate data analysis*. 2nd ed. Arnold; Oxford University Press, London, New York.
2. Zheng, Q. and Wang, X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res*, **36**, W358-363.
3. Storey, J.D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 479-498.
4. Bansal, M. and di Bernardo, D. (2007) Inference of gene networks from temporal gene expression profiles. *IET Syst Biol*, **1**, 306-312.
5. Cosgrove, E.J., Zhou, Y., Gardner, T.S. and Kolaczyk, E.D. (2008) Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. *Bioinformatics*, **24**, 2482-2490.
6. Balaji, S., Babu, M.M., Iyer, L.M., Luscombe, N.M. and Aravind, L. (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol*, **360**, 213-227.
7. Chang, D.E., Smalley, D.J. and Conway, T. (2002) Gene expression profiling of Escherichia coli growth transitions: an expanded stringent response model. *Molecular Microbiology*, **45**, 289-306.
8. Thormann, K.M., Duttler, S., Saville, R.M., Hyodo, M., Shukla, S., Hayakawa, Y. and Spormann, A.M. (2006) Control of formation and cellular detachment from Shewanella oneidensis MR-1 biofilms by cyclic di-GMP. *Journal of Bacteriology*, **188**, 2681-2691.
9. Rakshe, S., Leff, M. and Spormann, A.M. (2011) Indirect modulation of the intracellular c-Di-GMP level in Shewanella oneidensis MR-1 by MxdA. *Appl Environ Microbiol*, **77**, 2196-2198.
10. Saville, R.M., Dieckmann, N. and Spormann, A.M. (2010) Spatiotemporal activity of the mshA gene system in Shewanella oneidensis MR-1 biofilms. *FEMS Microbiol Lett*, **308**, 76-83.