Long-range order and short-range disorder in Saccharomyces cerevisiae biofilm

Vincent Piras¹, Adam Chiow², Kumar Selvarajoo³

¹Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Sud, Université Paris-Saclay, avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France
²Department of Pharmaceutical Engineering, Singapore Institute of Technology, 10 Dover Drive, Singapore 138683, Singapore
³Biotransformation Innovation Platform (BioTrans), Agency for Science, Technology & Research A*STAR, 61 Biopolis Drive, Proteos, Singapore 138873, Singapore

E-mail: kumar_selvarajoo@biotrans.a-star.edu.sg

Abstract: Biofilm, a colony forming cooperative response of microorganisms under environmental stress, is a major concern for food safety, water safety and drug resistance. Most current works focus on controlling biofilm growth by targeting single genes. Here, the authors investigated transcriptome-wide expressions of the yeast Saccharomyces cerevisiae biofilm in wildtype, and six previously identified biofilm regulating overexpression strains. Using statistical distributions for low expression filter (TPM > 5), Pearson auto- and cross-correlations reveal a strong transcriptome-wide invariance among all genotypes. The 50 highly expressed genes, however, differ significantly between the genotypes. Principal components analysis shows the global similarity between most overexpression strains. Thus, though single overexpression strains may show significant favourable local and acute expression changes (short-range disorder), the almost unperturbed global and collective structure between the genotypes indicate gradual adaptive response converging to original stable biofilm states (long-range order). Hierarchical clustering and gene ontology show 11 groups of local (e.g. mitochondria processes, amine and nucleotide metabolic processes) and 6 groups of global (e.g. transcription, translation and cell cycle) processes for all genotypes. The overall data indicate that there is a strong global regulatory structure that keeps the overall biofilm stable in all investigated strains.

1 Introduction

Biofilm is a major area of research in food, drug and dental industries, due to their persistence, emergent and adaptive responses [1]. Although biofilm can have positive attributes, most situations focus on the negative consequences. Previous studies have mainly focused on the extracellular polymeric substance formation or quorum sensing mechanisms on low-throughput platforms. Despite the progress made in understanding biofilm components and constituents, till today, there is no key solution in controlling its progression. Thus, there is a general need to venture into high-throughput methodologies to investigate further on the complex collective responses of biofilm.

Transcriptome-wide expression studies using RNA-Seq techniques have gained tremendous momentum in the last decade. The successes have mainly been due to deeper sequencing and wider dynamic ranges compared with its predecessors such as the microarray [2]. An overwhelming majority of works, however, is investigating a limited subset of genes with higher expression levels or fold changes to infer differential regulations, functional pathways or regulatory networks using common statistical approaches. This is often with the belief that lower expression transcripts are marred with technical/operator-induced noise or insignificant fold changing genes are not important. Although this is a necessary concern when we analyse and compare individual gene expressions, on a global transcriptome-wide scale, any arbitrary threshold cut-off may have negative outcomes. For example, the popular power-law distribution is only observed when we have a large number of entities within the datasets [3]. Taking a subset of a dataset can lead to the disappearance of the statistical structure [4]. Furthermore, information theory proposes that the expression level of an entity such as a gene, carry no real meaning without context or its probability [5, 6]. Thus, analysing high-throughput experimental data should not overly focus on the expression values alone (i.e. analysing only highly expressed or significant fold changing data) but look out for the principles used in information theory [7–9].

Living cells are complex, dynamical and dissipative systems considered to be in a state that is far from equilibrium [10, 11]. In other words, living cells are dynamically exchanging matter for their survival, and are able to evolve spontaneously (say, under any external perturbation) toward a critical point, without fine-tuning system parameters, for a phase transition [12, 13]. The formation of biofilm by certain microorganisms such as the bacteria Escherichia coli and yeast is a classic example [1, 14]. Such phase transformation is known to break the symmetry of the system leaving it invariant or in a collective mode [15].

Several recent studies have also shown the presence of strong equilibrium or attractor states in living cells, where even sorted heterogeneous daughter cells return to their parental distribution over time [16, 17]. Such observation or phenomenon where a dynamical system evolves and converges toward a specific or preferred state, set by the initial conditions, in physics, is called an attractor state. Here, to scrutinise the attractor states in biology, we require theoretical, physical and statistical concepts.

Previously, we have used information probability concepts and have shown for diverse mammalian cells, that pair-wise transcriptome-wide expressions correlations, with several orders of magnitude difference between their expression levels, are very high (>0.98) [18–23]. The strong invariance, across the large expression variability between genes, is a signature of order parameter organising the entire gene regulatory network imposed by the presence of a common attractor corresponding to the cell type [24, 25].

In this paper, to investigate the yeast Saccharomyces cerevisiae biofilm phase transition, we adopted information theory related statistical approaches to analyse the transcriptome-wide RNA-Seq expressions in wildtype (WT) and six overexpressed gene conditions (DIG1, SAN1, TOS8, ROF1, SFL1, HEK2) [26, 27]. Our analyses include scatter plots, distribution fitting, Pearson
correlation, principal components analysis (PCA), hierarchical clustering and gene ontology (GO) enrichment.

2 Results

We analysed the whole transcriptome of WT and six overexpressed strains (DIGI, SAN1, TOS8, ROF1, SFL1, HEK2), with four replicates for each condition, of biofilm modulators in the yeast S. cerevisiae strain F45, adapted from Cromie et al. [26, 27]. The selected genes are known to regulate the fluffy colony morphology of the yeast strain studied. There were a total of 11,236 unique read counts (genes) with non-zero expression in the datasets across all genotypes (Section 4). We next performed normalisation using transcripts per million (TPM) method [28].

2.1 Statistical frequency distributions show log normal as most probable

It has been shown on several occasions by others, and also by us, that gene expression distributions follow power-law or log-normal distributions due to their scale-free network organisation or complex transcriptional regulatory mechanisms [3, 29, 30]. Hence, we tested the frequency distribution of gene expressions in all genotypes (see Fig. 1 and Fig. S1) using five statistical distributions fits: (i) log normal, (ii) log logistic, (iii) Pareto (power law), (iv) Burr and (v) double Pareto log normal (DPLN) (Section 4). We observed that, in all datasets, the expression distributions fitted all five tested distributions above TPM > 1. The log normal has a slight advantage for the higher expressed genes (TPM > 5), while the DPLN fitted better at the lower expressed genes (TPM < 5).

This result suggests that the two tails of the gene expression distributions might be due to two different regulations [31]. The upper tail corresponds to higher gene expressions, from transcripts generated from highly stochastic (noisy) transcriptional events, and the measurements of these low expressions are more prone to technical errors [32, 33]. It will, therefore, be important to carefully determine a threshold to discriminate between genes that carry important information on the biological response from noisy data.

From the statistical structure, the threshold is at about TPM = 5, and we will scrutinise this threshold in the next section using correlations.

2.2 Strong transcriptome-wide invariance between biological replicates and different overexpression conditions

We computed pair-wise Pearson correlation, r, for the whole transcriptome (N = 11,236) between the replicates of each genotype (see Fig. S2). The result shows that the whole transcriptome correlation between each replicate is very close to 1 (min = 0.979, mean > 0.987, see Fig. 2a, left panels) for all genotypes. Such strong transcriptome-wide correlation invariance has also been observed for other cell types [18–23]. Next, comparing cross-genotype correlations, r_c, (correlations between genotypes), though the values are lower, they also remain near unity albeit slightly lower (mean between 0.945 and 0.972, see Fig. 2a, right panels). The data suggest that the global collective regulation is keeping the transcriptome-wide correlation high, despite the acute expression changes of a limited gene subset between the genotypes.

Fig. 1 Statistical distribution fitting. Gene expression distributions (in log scale) against (a) PDF, (b) CDF. Blue, red, orange and magenta curves indicate log normal, log logistic, power law (Pareto), Burr and DPLN fitting, respectively. The thick black curve corresponds to the experimental distribution in the WT condition (average of four replicates). Parameters of the fitted distributions are summarised in Table S1

Fig. 2 Correlation Plots. Pearson (auto- and cross-) correlations for (a) Whole transcriptome (N = 11,236), (b) Top 500 and top 50 most expressed genes in each condition, (c) Random gene extractions with sample sizes n = 500, 100 and 50. Filled shapes (on left panels): range (min, max) of values taken by Pearson correlations between replicates (r). Empty shapes (on right panels): range of Pearson correlations against other conditions (r_c), for each condition. Dashes: mean correlation values in the range. Dotted lines (repeated in every plot): whole transcriptome average for each condition. Note: cross-correlation values (min, max and mean) were obtained from all between-sample combinations (i.e each conditions four replicates against the four replicates of each of the six other conditions), (d) Mean r_c when selecting transcriptome size above (dashed) and below (solid line) an expression threshold (TPM, x-axis). For TPM < 5 (dashed blue line), mean r_c < 0.8. Individual replicates are indicated by thin coloured dashed lines
To observe the effect of highly expressed transcripts on the correlation values, $r$ and $r_c$ values were evaluated for the top ranked 50 and 500 highest expressed genes (see Fig. 2b). Notably, the $r$ values remained very high (mean between 0.960 and 0.988). However, the $r_c$ values were notably lowered. For example, the mean $r_c$ between WT and other conditions for the 50 highest genes expressed is 0.859, whereas it is 0.936 for 500 highest expressed genes and 0.955 transcriptome wide. Note that removing either 500 highest or lowest expressed genes from the transcriptome had little effect on $r$ and $r_c$ values compared with the whole transcriptome (see Figs. S3A and B).

Next, we evaluated the correlations for selections of 100 expressed genes with mean expression values around 100 TPM (highly expressed, see expression distributions, see Fig. 1), five TPM (in the middle of the distribution) and one TPM (lowly expressed). For genes with TPM>100, $r$ and $r_c$ values remained very high, close to correlations obtained for the whole transcriptome; however, both $r$ and $r_c$ values were significantly decreased for lower gene expressions: for genes with TPM>5, $r$ and $r_c$ values still range at positive values around 0.5, whereas for TPM~1, the correlations were close to 0 (see Figs. S3C and D).

This result confirms that genes with TPM < 5 carry high noise that distorts the correlation structure.

Finally, to test whether sample size impacts correlation values, we tested various sizes for random genes selections (Section 4) for each genotype 100 times and computed their $r$ values (see Fig. 2c). Notably, both $r$ and $r_c$ values are significantly higher than the highest expressed genes and similar to those of whole transcriptome values for all genotypes. Thus, the global or transcriptome-wide statistical structure is also revealed by the random selection of genes rather than choosing the highest expressed ones. In other words, a random selection of genes shows the fractal nature of transcriptome-wide gene expressions.

Overall, these data show that the highest or lowest expressed genes between different genotypes are differentially regulated causing lowering of their correlation values (see Fig. 2d): highly expressed genes show lower correlations because of their large variance in expression values (larger range of activation). In contrast, the decrease in correlations for lower expressed genes is due to non-informative noise. These data suggest that on a global scale, the technical and biological noise for the gene expressions is not sufficient enough to destroy the statistical structure of all investigated genotypes and replicates, for most of the transcriptome (TPM > 5, $N = 6328$). This result agrees with our previous work on neutrophil gene expression study in that the lowly expressed transcripts may not be wasteful but could play important roles in the global or transcriptome-wide gene expressions.

2.3 PCA and statistical clustering reveal similarity/diversity between overexpressed conditions

We evaluated the PCs for different sample sizes of sorted genes from the highest to lowest expression values ($N = 10, 30, 50, 100, 150, 200, 300, 500$ and 6328, i.e. TPM > 5, see Figs. S3a–c). PCs 1–3 constituted more than 76% variance (in the whole informative transcriptome, $N = 6328$, see Fig. S4A), and we plotted them for each genotype for the different $N$ sizes (see Fig. S4B). Notably, the PC loading scores (position) of each genotype became fixed when $N > 100$ (see Fig. S4C). That is, a minimum sample size of 100 highest expressed genes is sufficient for the convergence of PCs with the whole transcriptome data. If $N < 100$, the PC values became more variable and overlapped with other genotypes (see Figs. S4B–C). Looking closer at PCA loadings (see Fig. 3a, left), we observe PC1 is able to discriminate replicates in the TOS8 overexpression condition with no overlap with other conditions, while extreme loading scores for PC2 are obtained for the replicates of the WT condition, and PC3 extreme scores are for the ROF1 condition. Other PC vectors, in contrast, show low specificity with high overlap between replicates of different conditions. When plotting PC1 versus PC2 and PC3, we also observe clustering and seclusion of the WT replicates, as well for TOS8 replicates (see Fig. 3b, middle and right). Replicates of other conditions also tend to cluster together; however, with more overlap with other conditions. This is also confirmed by multidimensional scaling (MDS) over Euclidean distances between replicates and conditions (see Fig. 3b), which shows that the largest distances between conditions are between WT, TOS8 and SFL1. Together, PCA and MDS indicate that WT and TOS8 conditions have the most distinct variance (response) while other conditions are closer in terms of transcriptome-wide response. Hierarchical clustering of the expression matrix with (TPM > 5, see Fig. 3c), based on Pearson correlation as a distance metric, corroborates these results and indicated four major clusters: (i) all four WT replicates, (ii) TOS8, (iii) SFL1/DIG1 and (iv) ROF1/HEK2/SAN1. This result is also reflected from the $r_c$ values (see Fig. 2a, right panel, mean values of $r_c$ for WT and TOS8 are the lowest).

2.4 Characterisation of differentially and commonly expressed genes between overexpressed conditions

It is noteworthy that to identify differentially expressed genes between WT and the overexpressed conditions. One way to identify differentially expressed genes is through standardisation of gene expression values (Z-scores) [22]. In brief, the expression value ($x_i$) of each gene ($i$) in each experiment (replicate or condition, $j$) is normalised to the mean expression values across all experiments ($\mu_j$) and scaled by the standard deviation of expression values across all experiments ($\sigma_i$), that is: $Z_{ij} = (x_i - \mu_j)/\sigma_i$. We performed a log10 transformation of the data prior to normalisation to minimise outlier effects and averaged the $Z$-scores of all replicates (28 samples $\times 6328$ genes) for each condition, to obtain an averaged normalised matrix (7 conditions $\times 6328$ genes). To interpret Z-values, it is important to verify the normality of these distribution (see Fig. S5A): performing a qq plot showed most of the distribution approximates the standard normal distribution (mean = 0, standard deviation = 1), except for $|Z| > 2$, indicating significant outliers (corresponding $p < 0.05$ assuming normal distribution, see Fig. S5B). As verification, we checked the expression values and Z-scores of the different overexpressed genes in all conditions and confirmed: (i) all six overexpressed genes show TPM > 5 in all conditions and replicates (see Fig. S5C) and (ii) all six overexpressed genes are significantly overexpressed in their respective overexpression condition, with Z-scores > 2 (see Fig. S5D).

Hence, we identified differentially expressed genes with TPM > 5 and $|Z| > 2$ in at least one experiment, and found, out of the 6328 genes, only 296 genes corresponded to the acute transcriptome response to perturbation, that is, genes whose expression change is of similar order to the perturbed (overexpressed) genes (SAN1, TOS8 etc.). This data, therefore, suggest that only a small proportion of the transcriptome is directly regulated by the overexpressed genes, and a large majority are weakly or marginally affected.

2.5 Characterisation of the local acute and global collective transcriptome responses in the different overexpressed conditions

To characterise the genes that are dissimilar between WT and other conditions and their functions, we first performed hierarchical clustering on the Z-score matrix of the 296 differentially expressed genes that form the acute response to perturbation (TPM > 5 and $|Z| > 2$, see Fig. S7). As proof of principle, we confirmed the seven overexpressed genes (SAN1, TOS8, HEK2, SFL1, DIG1 and ROF1) were all present in the list. We obtained 52 super clusters that were then iteratively merged according to their similarity by minimising the correlation between clusters and maximising correlation within clusters (see Fig. S6A). After the procedure, we obtained 11 distinct clusters (see Figs. S6B and C), of which, seven consist of up-regulated genes in each condition and four consist of
down-regulated genes in WT, DIG1, SFL1 and TOS8 conditions. Note that down-regulated genes in WT indicate up-regulated genes in all overexpression experiments (and vice versa).

The two largest clusters correspond to up- and down-regulated genes in WT compared with all other conditions (up-regulated n = 69 genes, cluster Ac/C2 – acute common 2 and C; down-regulated n = 128, cluster Ac/C1, see Fig. 4a). In other words, from the 296 differentially expressed genes, 197 were common between all overexpression conditions (67%), indicating a core set of genes that define a non-specific response. Performing GO using AMIGO2 [34] showed the up-regulated genes of the non-specific acute response are involved in various mitochondrial processes, in particular, translation, while the down-regulated genes play a role in the nucleotide and amine metabolic processes (Table S2). On top of this, TOS8 specifically regulated 80 genes (21 up-regulated and 59 down-regulated, clusters Ac/S3 and Ac/S8 – acute specific 3 and 8, 27%). Notably, genes involved in the fungal-type vacuole membrane were down-regulated by TOSS8. In other conditions, very few other genes were specifically differentially expressed.

Other overexpressed conditions also regulated a small number of genes: DIG1 (alone, cluster Ac/S1) down-regulated six genes (cluster Ac/S7) and SFL1 regulated five other genes (one up, four down, clusters Ac/S3 and Ac/S9). HEK2 up-regulated another gene (cluster Ac/S6) and so did SAN1 (cluster Ac/S4). ROF1 did not show any specific acute response (cluster Ac/S2 only contained ROF1 gene). For these clusters, because of the low number of genes, GO could not be performed.

These results indicate that the acute response is very limited to only a small number of genes, that, for a large majority, belong to very generic and non-specific biological processes, e.g. to support the transcriptional and translational responses to external or internal stress. Notably, we have previously shown that genes with lower expression changes can also play an important role in shaping collective and coordinated transcriptional responses [19, 25]. It is, therefore, interesting to investigate the interplay between the acute and collective responses in the different overexpression conditions.

We, thus, expanded the pool of differentially expressed genes by lowering the Z-values threshold, that is, to include genes with much weaker (yet significant) gene expression changes compared with the acute response but that collectively respond to a perturbation [19, 25]. On the basis of the determined threshold (see Fig. S8 and Section 4, Section 4.6), we could include 2486 additional genes with 1.5 < Z ≤ 2 that form the collective global response (see Fig. S7). We again performed hierarchical clustering of the Z-values resulting in 48-super clusters (see Fig. S6D), and finally obtained six distinct expression profiles (see Fig. 4b and Figs. S5E and F). Similarly, to the acute response, we also observe a large group of genes that are commonly up- or down-regulated in all conditions (1260 genes, clusters Co/C2 and Co/C1 – collective common 1 and 2, see Figs. 4b and c). These genes belong to diverse biological processes involved in the collective transcriptionome specific response (transcription, translation, cell cycle, response to stress etc., Table S3).

We also observe a number of clusters that are specific to certain overexpressed conditions. However, in contrast to the acute response, where we found only one-condition-specific clusters, we see a more intertwined regulation of the clusters between the different conditions (see Fig. 4c). For example, 779 genes that are up-regulated in TOSS8 condition (cluster Co/S4 – collective specific 4) and involved in many metabolic processes are also weakly down-regulated by HEK2 (see Figs. 4b and c). Conversely, 266 genes (cluster Co/S3) enriched in processes such as protein folding are down-regulated by TOSS8 but up-regulated by HEK2 (see Table S3). A similar pattern occurs between DIG1 and ROF1 conditions for two smaller clusters (99 genes enriched in the fungal cell wall and 82 genes in RNA modifications, clusters Co/S2 and Co/S1). We also observe that SFL1 does not only down-regulate genes in the collective response but also weakly up-regulate genes that are up-regulated by TOSS8. (cluster Co/S3). Finally, the specific collective response seems absent when overexpressing ROF1.

Overall, the analysis indicates that, on top of a very limited acute response (296 genes, of which 197 are non-specific to any overexpression condition), there exists a weak but co-regulated
global collective transcriptome response that encompasses a large portion of the transcriptome (2486 genes, 1260 non-specific) (see Fig. 4c). Most of the differentially expressed genes, either from the acute or collective response, belong to general biological processes such as transcriptional, translational or metabolic processes that are collectively activated or repressed on external stress to adapt to diverse environmental changes (Tables S2 and S3). Thus, despite each overexpressed condition showing some specific and strong expression changes of a limited genes subset, the global transcriptome structure remains almost unaffected as a vast majority (59–76% depending on the condition, general biological processes, see GO terms in Table S4) of the expressed transcriptome is stable (Fig. 4c), and gradually adapts in order to converge to a phenotypic stable state.

3 Discussion

Single cell microorganisms are free-living system often yielding uncorrelated (disordered) responses [31–33]. However, when they are exposed to environmental or pathogenic threats, a large number of species are able to aggregate, often within a self-secreted extracellular polymeric substances matrix or through quorum sensing, resulting in a highly ordered structure called biofilm [1].

Self-organised criticality is a general property of dynamical systems, where the overall behaviour shows a critical point leading to a phase transition that will leave the system invariant or in a collective mode [12]. That is, at the critical point, all differences between individuals will be reduced to follow universal properties leading to long-range order. Below the critical point (subcritical), the system will be in disorder [35].

In this paper, we investigated the global properties of the biofilm formed by the *S. cerevisiae*. Numerous studies have investigated a number of strains that regulate the growth of the biofilm; however, the conclusions do not point to a clear winner [14, 30]. Here, we adopted a number of statistical metrics to analyse the transcriptome-wide RNA-Seq expressions in WT and six overexpressed gene conditions (*DIG1*, *SAN1*, *TOS8*, *ROF1*, *SFL1*, *HEK2*) of the yeast strain F45.

After accounting for technical noise, we observed log-normal gene distribution (see Fig. 1). Using this statistical structure for expression threshold cut-off, we analysed a total of 6328 gene transcripts. Pearson correlation reveals strong invariance not only between the same genotype replicates but also between all genotypes (see Fig. 2). The strong invariance was also recapitulated by a random selection of genes. However, when only the highly or lowly expressed genes were compared between genotypes, the correlations were notably lower. Thus, we observe
two modes of behaviours: a strong global correlation from low to high expression (long-ranger order), driven by stably expressed and collectively responding genes and a poorer local correlation at the highest gene expression changes (short-range disorder) that carry acute response (see Fig. S7, correlations of acute response genes only versus transcriptome without acute response genes).

Principal component and clustering (see Fig. 3) analyses showed that TOS8 was the most distinct response from WT and any other overexpressed condition. Other conditions were closer to each other in terms of response, while still being quite distinct from WT. These simple approaches, thus, reveal the similarity and diversity between the different overexpression and WT conditions.

Focusing only on selected highly expressed genes shows short-range disorder. It also revealed that a majority of genes (67% of the acute response and 51% of the collective response) are commonly regulated between all conditions and GO analysis reveals the distinct functions of the differentially regulated genes (see Fig. 4).

Overall, our analysis provided statistical evidence that biofilms are highly structured transcriptome wide (not only in their physical appearance) and single mutant/overexpression regulation such as the ones investigated here are only able to propagate response locally, that is, to a limited number of genes. Hence, these conditions will likely provide only transient success in biofilm control and may be subdued in the long run.

These data coincide with the notion of attractor states in biology [25, 33, 36], where there is a stable equilibrium state toward which a dynamical system tends to converge despite perturbation. Moving out of the attractor state requires the collective action of the entire system parameters, in this case, gene expressions [25, 37]. Thus, biofilm control using single-gene target may not be a viable option for successful long-term effects. Our data, therefore, ask for a deeper consideration for the understanding of the global statistical structure or organisation of dynamic living systems. We, therefore, urge the microbiologist communities to tackle biofilm challenge adopting multidisciplinary approaches that will shed light not only on the acute local effects but also on the collective global regulation and to identify novel targets for generating long-range disorder.

4 Materials and methods

4.1Datasets

Experimental data were obtained from datasets publicly available on the Gene Expression Omnibus Database (https://www.ncbi.nlm.nih.gov/geo/), under accession numbers GSE98079 and GSE85843 [26, 27]. The available fragments per kilobase million (FPKM) expression data was used and reprocessed into TPM expression units for this study [28] such as the TPM value of the ith gene is the ratio between the FPKM value of the ith gene and the sum of the FPKM values of all n genes, multiplied by a factor for 10^{6}:

\[ \text{TPM}_i = 10^6 \cdot \frac{\text{FPKM}_i}{\sum_{j=1}^{n} \text{FPKM}_j} \]

The used data contains 28 samples (4 replicates for 7 conditions). The total transcriptome size is 12,131 genes (distinct saccharomyces genome database (SGD) IDs), of which 11,236 were detected in at least one sample (TPM > 0). About 6328 genes were considered as expressed (TPM ≥ 5) in at least one sample, of which 3546 were considered as non-responsive (\( \sum|Z| < 1 \)), 2486 for the collective response and 296 for the acute response.

4.2 Statistical distributions fitting

Fitting of the distributions of experimental gene expression data was performed using the maximum-likelihood fitting method (fitdist of the R MASS package [38]) for Pareto, log-normal, log-logistic and Burr distributions. The DPLN distribution fitting was performed using a general simulate-annealing approach (GenSA package [39]) to reduce the error between experimental probability density function (PDF) curve and the simulated one.

The Pareto (power law) distribution [40] is characterised by two parameters, the exponent, \( \alpha \) that determines the slope of the upper tail of the distribution and a threshold \( x_m \) below which the power law does not apply. The PDF is defined as, for \( x > x_m \):

\[ f(x) = \frac{\alpha x^{\alpha - 1}}{x_m^{\alpha} \Gamma(\alpha)} \]

The log-normal distribution is defined for a random variable whose logarithm is normally distributed. It is characterised by 2 parameters, \( \mu \) and \( \sigma \), respectively, the (log) mean and standard deviation of the underlying normal distribution. Its PDF is defined as:

\[ f(x) = \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \]

The log-logistic distribution [40] is defined for random variable whose logarithm has a logistic distribution. It is characterised by two parameters, \( \alpha \) and \( \beta \), respectively, the scale and shape the underlying logistic distribution. Its PDF is defined as

\[ f(x) = \frac{\alpha \beta x^{\beta - 1}}{\alpha \beta + 1} \]

The DPLN distribution [41] arises out of a mixture of log-normal distributions shows Pareto power-law behaviour in both tails. It is characterised by four parameters, \( \alpha, \beta, r \) and \( \nu \), which stand for the shape parameters (exponents) of the two underlyng Pareto laws, and the mean and standard deviation of the underlying log-normal distribution. The PDF is defined as:

\[ f(x) = \frac{\alpha \beta x^{\beta - 1}}{\alpha \beta + 1} \]

\[ + \frac{\nu r^{\nu - 1} c^{-\nu} e^{-(\nu r^2)/2}}{\nu c^{-\nu} e^{-(\nu r^2)/2}} \Phi \left( \frac{\log x - \mu}{\sigma} \right) \]

where \( \Phi \) and \( \Phi^c \) are the cumulative density function (CDF) and complementary CDF of the normal distribution, \( N(0, 1) \).

4.3 Pearson correlation

The Pearson correlation coefficient \( r \) between two vectors (e.g. transcriptome in two different samples), containing \( n \) observations (e.g. gene expression values), is obtained by (for large \( n \))

\[ r(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y} \]

where \( x_i \) and \( y_i \) are the ith observation in the vectors \( X \) and \( Y \), respectively; \( \mu_X \) and \( \mu_Y \) are the average values of each vector; and \( \sigma_X \) and \( \sigma_Y \) are the corresponding standard deviations. Pearson correlation measures linear relationship between two vectors, where \( r = 1 \) if the two vectors are identical and \( r = 0 \) if there are no linear relationships between the vectors.

4.4 Hierarchical clustering

Hierarchical clustering builds a hierarchy of clusters using two methods: agglomerative and divisive algorithms. We used the former (Ward's) where each observation starts in its own cluster; and pairs of clusters are merged moving up the hierarchy. The
transcriptomes in different conditions. MDS lays the variables on a two-dimensional (2D) space such as the 2D-space Euclidean distances reproduces, with minimum error, the dissimilarity matrix can vary. Here, we used n-D Euclidean distances between M transcriptomes as dissimilarity matrix (here, M = 28 samples). The distance between two transcriptomes T_j and T_k (e.g. WT and SAN1 overexpression conditions) formed by n = 6328 gene expression values \( (j_1, ..., j_m, ..., j_N) \) is \[ d_{T_j, T_k} = \sqrt{\sum_{i=0}^{n} (t_j_i - t_k_i)^2} \]

On the 2D space (with x-y coordinates), the reproduced distance between transcriptomes \( T_j \) and \( T_k \) is defined by \[ d_{T_j, T_k} = \sqrt{(f_{x,j} - f_{x,k})^2 + (f_{y,j} - f_{y,k})^2} \]

MDS will iteratively rearrange the positions \( T_j \) and \( T_k \) on the 2D-plane to minimise the following function of a set of \( m \) transcriptomes:

\[ \min \sum_{j=1}^{M} \sum_{k=1}^{M} (d_{T_j, T_k} - d_{T_j, T_k})^2 \]

4.6 Determination of the collective response genes

To reliably determine from which threshold genes can still be considered as responsive, we first searched for the lowest between-replicate correlation value to determine the maximum correlation shift caused by random variations due to technical noise and biological variability. We found \( r = 0.979 \) to be the lowest value in our dataset among 42 possible between-replicate correlation values (six combinations for seven conditions) that occurs between two different HEK2 replicates. Thus, when comparing two experiments, only a correlation value lower than 0.979 would indicate a reliable difference or response between the samples.
[30] Beal, J.: ‘Biochemical complexity drives log-normal variation in genetic expression’, IET Eng. Biol., 2017, 1 (1), pp. 55–60
[31] Elowitz, M.B., Levine, A. J., Siggia, E. D., et al.: ‘Stochastic gene expression in a single cell’. Science, 2002, 297, (5584), pp. 1183–1186
[32] Del Giudice, M, Bo, S., Grigolon, S., et al.: ‘On the role of extrinsic noise in microRNA-mediated bimodal gene expression’, PLoS Comput. Biol., 2018, 14, (4), p. e1006663
[33] Selvarajoo, K.: ‘Understanding multimodal biological decisions from single cell and population dynamics’, Wiley Interdiscip. Rev. Syst. Biol. Med., 2012, 4, (4), pp. 385–399
[34] Balsa-Canto, E., Henriques, D., Gábor, A., et al.: ‘AMIGO2, a toolbox for dynamic modeling, optimization and control in systems biology’, Bioinformatics, 2016, 32, (21), pp. 3357–3359
[35] Tsuchiya, M., Giuliani, A., Hashimoto, M., et al.: ‘Self-organizing global gene expression regulated through criticality: mechanism of the cell-fate change’, PLoS One, 2016, 11, (12), p. e0167912
[36] Huang, S., Kauffman, S.: ‘How to escape the cancer attractor: rationale and limitations of multi-target drugs’, Semin. Cancer Biol., 2013, 23, (4), pp. 270–278
[37] Mojtahedi, M., Skupin, A., Zhou, J., et al.: ‘Cell fate decision as high-dimensional critical state transition’, PLoS Biol., 2016, 14, (12), p. e2000640
[38] Venables, W.N., Ripley, B.D.: ‘Modern applied statistics with S’ (Springer, New York, USA, 2002, 4th edn.)
[39] Xiang, Y., Guiban, S., Suomela, B., et al.: ‘Generalized simulated annealing for efficient global optimization: the GenSA package for R’, J. R., 2013, 5, (1), pp. 13–28
[40] Kleiber, C., Kotz, S.: ‘Statistical size distributions in economics and actuarial sciences’ (Wiley, Hoboken, NJ, USA, 2003)
[41] Reed, W.J., Jorgensen, M.: ‘The double Pareto-lognormal distribution – a new parametric model for size distributions’, Commun. Stat. Theory Methods, 2004, 33, (8), pp. 1733–1753
[42] Everitt, B.S., Landau, S., Leese, M.: ‘Cluster analysis’ (Oxford University Press, Inc., New York, Arnold, London, 2001, 4th edn.)
[43] Cox, T.F., Cox, M.A.A.: ‘Multidimensional scaling’ (Chapman and Hall, London, 2001, 2nd edn.)