Fundamental principles governing sporulation efficiency: A network theory approach

Camellia Sarkar[1], **Saumya Gupta**[2], **Himanshu Sinha**[2,3,4,*], **and Sarika Jalan**[1,5,*]

[1]Centre for Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Khandwa Road, Simrol, Indore 453552, India
[2]Department of Biological Sciences, Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai 400005, India
[3]Department of Biotechnology, Indian Institute of Technology Madras, Chennai 600036, India
[4]Initiative for Biological Systems Engineering, Indian Institute of Technology Madras, Chennai 600036, India
[5]Complex Systems Lab, Discipline of Physics, Indian Institute of Technology Indore, Khandwa Road, Simrol, Indore 453552, India
[*]Corresponding authors: Himanshu Sinha (sinha@iitm.ac.in); Sarika Jalan (sarikajalan9@gmail.com)

ABSTRACT

Using network theory on an integrated time-resolved genome-wide gene expression data, we investigated the intricate dynamic regulatory relationships of transcription factors and target genes to unravel signatures that contribute to extreme phenotypic differences in yeast, *Saccharomyces cerevisiae*. We performed a comparative analysis of the gene expression profiles of two yeast strains SK1 and S288c which are known for high and low sporulation efficiency, respectively. The results based on various structural attributes of the networks, such as clustering coefficient, degree-degree correlations, and betweenness centrality suggested that a delay in crosstalk between functional modules can be construed as one of the prime reasons behind low sporulation efficiency of S288c strain. A more hierarchical structure in the late phase of sporulation in S288c indicated an attempt of this low sporulating strain to obtain modularity, which is a feature of early sporulation phase. Further, the weak ties analysis revealed that mostly meiosis-associated genes were the end nodes of the weak ties for the high sporulating SK1 strain, while for the low sporulating S288c strain these nodes were mitotic genes. This again was a clear indication of the delay in regulatory activities in the S288c strain, which are essential to initiate sporulation. Our results demonstrate the potential of this framework in identifying candidate nodes contributing to phenotypic diversity in natural populations with application prospects in drug target discovery and personalized health.

Introduction

Phenotypes arise from complex molecular interaction networks, hence comparing transcriptional regulation of a phenotype across multiple genetic backgrounds can provide comprehensive insights into the network regulating the phenotype. Comparing transcriptional networks of developmental or temporal processes allows identification of differences in chronological order of events regulating a process. For a developmental process such as yeast sporulation, which comprises of meiosis and ends with spore formation, several genome-wide transcriptome analyses have been done to understand the complete cascade of transcriptional regulation.[1] This has led to identification of stages of gene regulation during sporulation with *IME1* and *NDT80* regulating transition between these stages.[2] To understand how different genetic backgrounds modulate this transcriptional cascade, yeast strains from different genetic backgrounds have been compared, such as SK1 (a high sporulating *S. cerevisiae* strain with over 90% sporulation efficiency in 48h) and W303 (a moderate sporulating *S. cerevisiae* strain with 60% efficiency). This comparison between the two strains carried out across each time point showed that while the patterns of gene expression profiles remained similar, only 60% genes overlapped in expression levels.[3] While this is a useful method to compare across time scales,[4] it does not allow comprehensive comparison between the expression profiles of all genes across time.

Various methods have been suggested to study gene expression profiles across strains such as clustering methods,[5] bootstrapping clustering,[6] four-stage Bayesian model,[7] Gaussian mixture models with a modified Cholesky-decomposed covariance structure,[8] biclustering algorithm,[9] etc. Clustering proved to be a successful initial approach for analyzing gene expression data and allowed biologists to identify groups of potentially meaningful genes that have related functions or are co-regulated.[5] However, these studies are gene centric with aim to assign functions to functionally unknown genes, for instance, clustering methods predict function of genes based on their clustering. Moreover, these studies did not provide information about inter-cluster functional interactions in the genes.[10] Thus, in order to assess reliability of results obtained from clustering methods in a statistically quantifiable manner, a bootstrap method was proposed.[6] However, considering the generic nature of clustering methods and in particular their inability to incorporate system-specific prior information, a model-based approach was proposed in which experiments were indexed by ordered variables such as time, temperature, or the dose level of a toxin and the trajectory was modelled as a function of the ordering variable (e.g. time) and a gene-specific set of parameters.[7] Although classical model-based clustering continues to extend into new application areas, none of the models had a covariance structure specifically designed for the analysis of longitudinal data. This feature accounting for the relationship between measurements at different time points of the longitudinal data was introduced in Gaussian mixture models with a modified Cholesky-decomposed covariance structure.[8] But all these methods tend to overlook local patterns where these genes are similar based on only a subset (subspace) of attributes (for example expression values). This led to implementation of a pattern similarity based biclustering approach to gene expression data that could find bi-clusters among co-regulated genes under different subset of experimental conditions.[9] The next step in interpreting gene expression profiles would be to go beyond the gene-centric techniques and employ a more holistic approach in order to acquire an integrative understanding of how a gene expression profile on the whole is specifically related to the genomic regulatory circuitry of the genome[11] and network theory offers this platform. Network theory provides an efficient framework through which the behavior of complex systems can be explained in terms of the structural and functional relationships between different molecular entities.[12–14] The basic structural properties of networks are dependent on how the networks evolve, the inherent interdependencies of the nodes as well on the architectural constraints. While on one hand, these network measures help in identifying important nodes of the network, on the other hand they also enable realization of the impact of interactions on the behavior of the underlying system. Hence, studying network parameters have expanded our understanding of biological processes, for instance by identifying important genes for diseases,[15] elucidating the mechanism behind human diseases by analyzing relationships between disease phenotypes and genes,[16] and deciphering the common genetic origin of multiple diseases.[17]

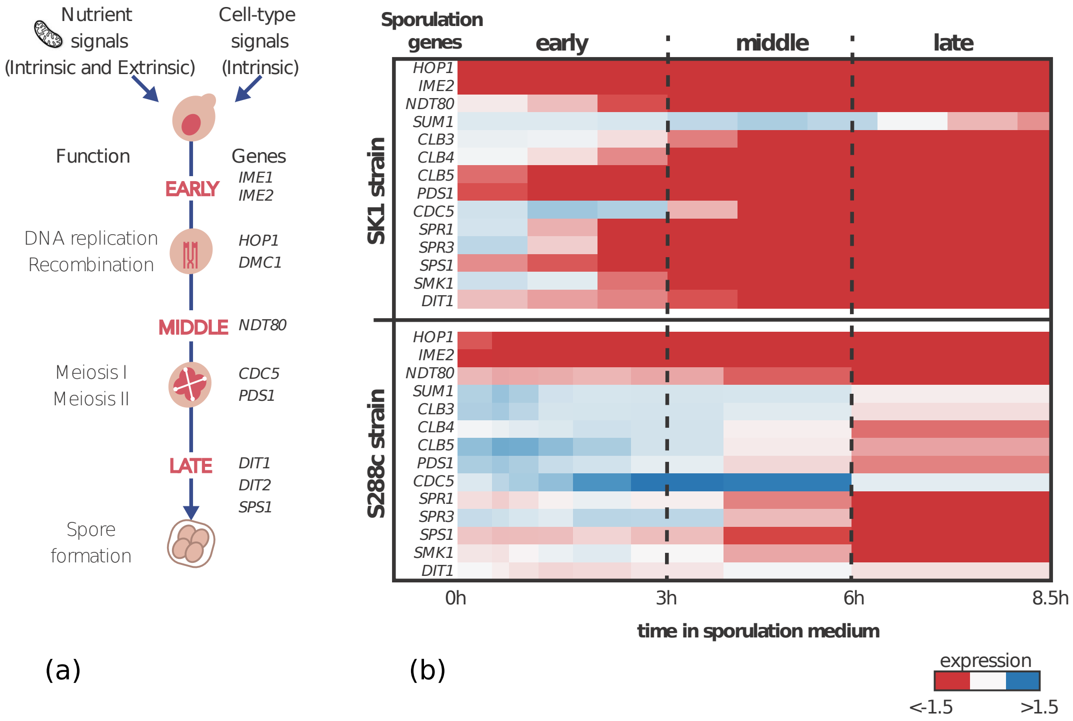

Figure 1. (a) Schematic diagram for the sporulation process. (b) Heatmap of gene expression profiles of crucial meiosis regulators in early, middle and late phases of sporulation.

In the current work, we used the network theory approach to investigate how transcriptional regulatory networks differ between two genetic backgrounds leading to extreme phenotypes in the same environment. We studied two genetically divergent *S. cerevisiae* strains that differ adversely in terms of sporulation efficiency: SK1 (90% efficiency in 48h) and S288c (10% efficiency). Various cellular activities such as DNA replication, recombination and repair, RNA transcription and translation, intracellular trafficking, enzymatic activities of general metabolism, and mitochondrial biogenesis being conserved from yeast to humans,,[18] 46% human proteins having homologues in yeast proteome, 30% to 40% sequence identity of yeast genes with human disease-associated genes[19] has rendered yeast as an irreplaceable model organism for approaching the molecular basis of humans. To the best of our knowledge this is the first study using network theory carried out on time-resolved gene expres-

Table 1. Structural attributes of SK1 sporulation networks. N, N_c and N_{TF} respectively denote the network size, number of connections and number of transcription factors in a particular time point. Catalogue of high degree nodes (degree mentioned in braces) and degree of *NDT80* gene are provided for each time point of SK1 sporulation networks.

Time points	N	N_c	N_{TF}	High degree nodes	Degree of *NDT80*
T_1	360	529	13	*BAS1* (283), *RIM101* (39)	-
T_2	791	2107	34	*BAS1* (480), *MSN4* (300), *KAR4* (132), *FKH1* (109), *AFT2* (75)	-
T_3	1096	4033	46	*ASH1* (612), *BAS1* (560), *MSN4* (363), *HMS1* (168)	47
T_4	1495	6540	63	*ACE2* (1250), *ASH1* (770), *MSN4* (428), *AFT1* (369), *HMS1* (209)	69
T_5	1640	7720	73	*ASH1* (926), *MSN4* (499), *AFT1* (437), *GCR1* (399), *HMO1* (248)	87
T_6	1900	8978	77	*ASH1* (1074), *MSN4* (580), *AFT1* (486), *ISW2* (292)	101
T_7	1928	8012	75	*ASH1* (1118), *AFT1* (504), *ISW2* (309)	102
T_8	1886	8281	76	*ASH1* (1094), *AFT1* (484), *STE12* (432), *FHL1* (392), *ISW2* (299), *HMO1* (289)	95
T_9	2024	9213	74	*ACE2* (1700), *ASH1* (1039), *AFT1* (461), *STE12* (421), *FHL1* (362), *ISW2* (266)	87
T_{10}	1880	8410	69	*ACE2* (1579), *ASH1* (955), *AFT1* (435), *STE12* (402), *FHL1* (322), *ISW2* (243)	79
T_{11}	1711	7188	66	*ACE2* (1429), *ASH1* (870), *AFT1* (428), *STE12* (386), *FHL1* (291), *HMS1* (222)	69
T_{12}	1745	7206	61	*ACE2* (1472), *ASH1* (886), *AFT1* (481), *GAL4* (299), *HMS1* (246), *SIN4* (216), *INO4* (177)	77

sion data drawn from two *S. cerevisiae* strains lying on extreme ends of the sporulation efficiency. The framework used in this study considering *S. cerevisiae* as a model organism also allowed us to integrate transcriptional regulation information on the time-resolved gene expression data in order to comprehensively understand how the sequence of events differ between these two genetic backgrounds. We used several network parameters to perform comparative analysis of gene expression profiles corresponding to different sporulation time points of both SK1 and S288c strains in order to unravel the complexity of the sporulation process and to predict the nodes implicated in high sporulation efficiency of SK1 strain as compared to the S288c strain.

Results and discussion

We constructed a dynamic transcriptional regulatory network of SK1 strain during sporulation (see Methods) and noted the early, middle and late phases of sporulation (Fig. 1(a)) by comparing the appearance of crucial meiosis regulators in the network (Fig. 1(b)) with their expression profiles described in previous literature.[2] For instance, *NDT80* gets activated in the early-mid phase of sporulation, around 2-3h in sporulation medium.[20] Concordantly, we observed *NDT80* appearing in the time span from T_3 to T_{12} (from 3h till 12h in sporulation) in the dynamic sporulation network of SK1 strain (Supplementary Fig. S1). Interactions of *NDT80* increased from T_3, reaching a maximum in T_7 and then decreased as time progressed in sporulation (Supplementary Fig. S1). Based on the appearance of *NDT80* in the dynamic sporulation network, we classified T_1-T_3 (1-3h in sporulation) as the early sporulation phase, T_4-T_6 (4-6h in sporulation) as the middle sporulation phase and T_7 onwards (7h onwards in sporulation medium) as the late sporulation phase in the SK1 strain. The regulators of *NDT80* constituted the high degree nodes in SK1 strain (Supplementary Fig. S1 and Table 1) such as *MSN4* (stress responsive transcriptional activator), *AFT1* (regulator of iron homeostasis) and *FHL1* (regulator of ribosomal protein transcription).

While *NDT80* is the prime initiator of sporulation in SK1, most of the cells of the low sporulating strain S288c did not enter meiosis at all and were arrested in the stationary phase (G1/G0 phase).[21] Therefore, the transcriptional network of SK1 and S288c would show differences during sporulation. Since this could help in understanding why the two strains showed sporulation efficiency variation, we sought to determine the reason behind the observed differences during sporulation. We began with comparing the general network properties of the two strains during sporulation (Fig. 2, Supplementary Tables S1 and S2). Please note that the the early, middle and late phases of SK1 were compared to the corresponding time-points (in hours) with S288c in order to draw a fair comparison between the gene expression profiles of both the strains. Hence, T_1-T_5 (30m to 2h30m in sporulation medium) , T_6-T_7 (3h50m to 5h40m in sporulation medium) and T_8 (8h30m in sporulation medium) time points of S288c corresponded to the early, middle and late phases of sporulation, respectively.

SK1 strain exhibited a wider range of network sizes across different sporulation time points as compared to S288c strain (Tables 1 and 2), indicating in regulatory changes in the former strain. In order to follow these changes, we investigated the early, middle and late phases of sporulation in SK1 independently and compared them to the corresponding phases in S288c strain which are known to exhibit markedly different sporulation events.[22] We found that there was a drastic increase in the number of genes having significantly high or low expression values in the consecutive time points at the onset of sporulation in both the strains (Tables 1 and 2) which could be due to cells transitioning from mitotic growth to initiate meiosis. Keeping in view this massive reprogramming of gene expression in the early sporulation phase for preparing the cells to enter meiotic cell

Table 2. Structural attributes of S288c sporulation networks. N, N_c and N_{TF} respectively denote the network size, number of connections and number of transcription factors in a particular time point. Catalogue of high degree nodes (degree mentioned in braces) and degree of *IME1* gene are provided for each time point of S288c sporulation networks.

Time points	N	N_c	N_{TF}	High degree nodes	Degree of *IME1*
T_1	434	605	15	*BAS1* (356), *HAP4* (92), *TUP1* (48)	10
T_2	706	1327	24	*BAS1* (523), *STE12* (158), *HAP4* (134), *CUP2* (117), *TUP1* (72)	14
T_3	691	1280	26	*BAS1* (524), *HAP4* (136), *CUP2* (114), *TUP1* (76), *PUT3* (70)	13
T_4	569	1025	25	*BAS1* (442), *HAP4* (108), *RLM1* (93)	13
T_5	582	1049	24	*BAS1* (455), *HAP4* (112), *SWI5* (92)	15
T_6	415	733	23	*HMO1* (138), *HAP4* (125), *SWI5* (98)	16
T_7	409	686	18	*HMO1* (131), *HAP4* (124), *SWI5* (97)	19
T_8	513	1195	23	*MSN4* (261), *HSF1* (246), *HMO1* (129), *SWI5* (95), *INO4* (75)	21

division,[23] our analysis revealed increased involvement of genes during this phase in both the strains was not very surprising. In the later phases of sporulation, the rate of change of network size subsides. Despite changes in the early sporulation phase in both the strains, the ratio of number of differentially expressed transcription factors (N_{TF}) and target genes remains almost constant across all the time points (Tables 1 and 2). The proportion of regulatory genes remaining constant throughout the sporulation indicates that it may be an intrinsic property of the sporulation process.

A change in the number of connections modulates the intrinsic properties of a network.[12] We investigated the impact of this change for both the strains during various sporulation phases. Similar to the network size, the number of connections (N_c) increased drastically in the early time points of sporulation in both the strains. However, this rate of increase in the number of connections was much higher in the case of SK1 strain as compared to the S288c strain. For instance, while S288c exhibited a two-fold increase in the number of connections in the early sporulation, SK1 exhibited a four-fold increase in the same phase (Tables 1 and 2). Note that change in the number of connections will only be possible if either old nodes (genes) disappear or/and new nodes arise in the networks, since all interactions for both the strains are taken from the same repository base network. A higher rate of increase in the number of connections in SK1 strain as compared to the rate of increase in their size can be attributed to the appearance of more number of high degree nodes in the second time point (Table 1). In the middle phase of sporulation, associated with processes involved in meiotic divisions,[1] the number of connections did not show considerable change for both the strains since we find that more than 75% of the genes remain same across the different time points in the middle phase in the individual strains. However, again in the late sporulation phase, there was a change in the number of connections in S288c strain while for the SK1 strain this number remained almost constant compared to the middle phase. For instance, towards the mid-late phase, there was a fall in the number of connections in S288c strain. Incidentally, this decrease in the number of connections could be due to the disappearance of the high degree node *BAS1*, a Myb-related transcription factor involved in amino acid metabolism and meiosis.[24] Interestingly, *BAS1* contributes to approximately 50% of the connections in the early phase of S288c (Table 2) though this gene is not one of the known regulators of sporulation,[22] and its disappearance in the middle phase is reflected in the number of connections. What is more intriguing is that this gene is involved in the regulatory processes only in the early phase of sporulation and disappears in the middle phase in both the strains. On one hand, this indicates the specific significance of this gene intrinsic to the early phase of sporulation, while on the other hand it reflects the drastic changes in the regulatory activities from the early to the middle phase. Furthermore, in the late sporulation phase of S288c strain, the number of connections almost doubled. This difference arose due to the appearance of *MSN4* and *HSF1*, known stress-responsive regulators[25] showing high degrees only in the last phase of sporulation in S288c (Table 2). While *MSN4* appeared as a high degree node in the early phase of SK1 strain, *HSF1* did not appear as a high degree node at all in the SK1 sporulation network (Table 1), indicating the interesting possibility that either the late appearance or absence of these nodes could be involved in decreasing sporulation efficiency and maintaining stationary phase in the S288c strain. The differences in the number of connections between the strains in the early, middle and late phases of sporulation, motivated us further to compare the general principles of regulatory interactions during sporulation between these strains.

So far, we focused only on the number of genes and the interactions in the networks. To understand how the interacting patterns impacted the overall structure of the underlying networks, we investigated the degree-degree mixing of the connected nodes across different sporulation phases in the two strains. (Dis)assortativity is a parameter that measures the correlation in the degrees of the nodes in a network and provides understanding of the (dis)likelihood in connectivity of the underlying systems.[26] In gene regulatory networks, highly connected nodes avoid linking directly to each other and instead connect to proteins with only a few interactions, thus exhibiting disassortative topology.[27] This behavior of the nodes leads to a reduction in crosstalk between different functional modules and increase in the robustness of the networks by localizing the effects of

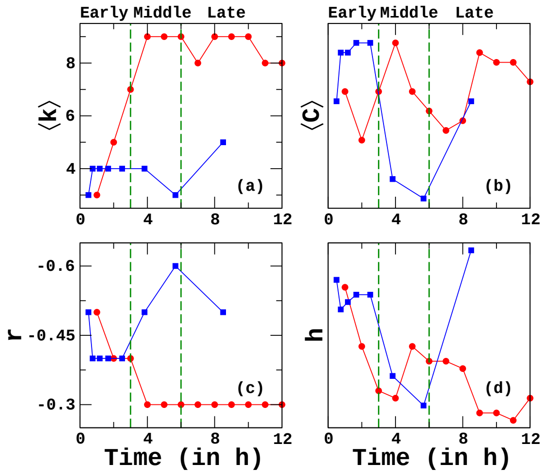

Figure 2. Structural properties of sporulation networks of SK1 strain (red circles) and S288c strain (blue squares) in early, middle and late phases of sporulation. (a) Average degree ($\langle k \rangle$), (b) average clustering coefficient ($\langle C \rangle$), (c) Pearson degree-degree correlation coefficient (r) and (d) global reaching centrality (h) are plotted as a function of time in sporulation medium (in hours) for both the strains.

deleterious perturbations.[28] The Pearson (degree-degree) correlation coefficient (r) was calculated for the networks at all time-points in each of the strains (see Methods). As expected for gene regulatory networks, sporulation networks in both SK1 and S288c strains exhibited disassortativity at all time points (Fig. 2). A high value of this property was observed in both the strains during the early phase of sporulation, suggesting that the strain required being more resilient to perturbations while carrying out early sporulation transcriptional events.[28] Post early phase, disassortativity values in SK1 strain reached a steady state at middle sporulation phase, while those of S288c still showed fluctuations (Fig. 2). Taken together, these observations implied that the necessary crosstalk between functional modules occurred early and then stabilized in SK1, while they were still going on or were random and unstable in the middle and late phases in S288c strain.

After analyzing the global properties of the sporulation networks, we next investigated the local properties of the networks, which were expected to reveal the impact of local architecture on the phenotypic profiles of the two strains. Clustering coefficient is one such local property that measures the local cohesiveness between the nodes.[29] A high value of clustering coefficient of a node depicts high connectivity among the neighbors of that node. For SK1 and S288c strains, we evaluated the average value of clustering coefficient ($\langle C \rangle$) for each time point (see Methods). As expected for various biological networks,[12] a high value of $\langle C \rangle$ was observed for the networks at all time points in both the strains as compared to their corresponding random networks (Fig. 2, Supplementary Tables S1 and S2) as expected.[29] Furthermore, keeping in view the manner in which we constructed the sporulation networks, a high $\langle C \rangle$ meant that many of the neighbor target genes of a transcription factor also acted as transcription factors for the other neighbor target genes of that same transcription factor. On comparing the average value of clustering coefficient ($\langle C \rangle$) between the strains, a sharp increase in $\langle C \rangle$ was observed thrice for the SK1 strain coinciding with the early, middle and late phases of sporulation, while for the S288c strain only two such transitions were observed for this property (Fig. 2). Moreover, while the transitions between the three peaks were rapid in the SK1 strain, a slower transition between the first and second peak was observed for S288c strain. Since high clustering in cellular networks is known to be associated with the emergence of isolated functional modules,[14] these results pertaining to average clustering coefficient suggested that the increased duration of time taken by S288c strain for forming functional modules can be considered essential for transferring information from early to middle phases in sporulation, and thus could be involved with the lower sporulation efficiency shown by this strain.

In order to further unravel the differences of the sporulation process in the two strains, we investigated how number of neighbors of nodes denoted by node degree is associated with their neighbor connectivities (interactions between the neighbors of the node of interest) evaluated in terms of clustering coefficient (see Methods). All the networks in SK1 and S288c strains exhibited negative degree-clustering coefficient correlation (Supplementary Figs. S2 and S3) as also witnessed in various other real world networks, indicating the existence of hierarchy in the underlying networks.[14] A hierarchical architecture implies that sparsely connected nodes are part of highly clustered areas, with communication between the different highly clustered neighborhoods being maintained by a few hubs. We quantified this hierarchy (h), also termed as global reaching centrality in the networks[30] (see Methods) and found that in both the strains, the networks were more hierarchical at the beginning of sporulation (Fig. 2). A high value of hierarchy has been associated with modularity in the network, for instance in case of

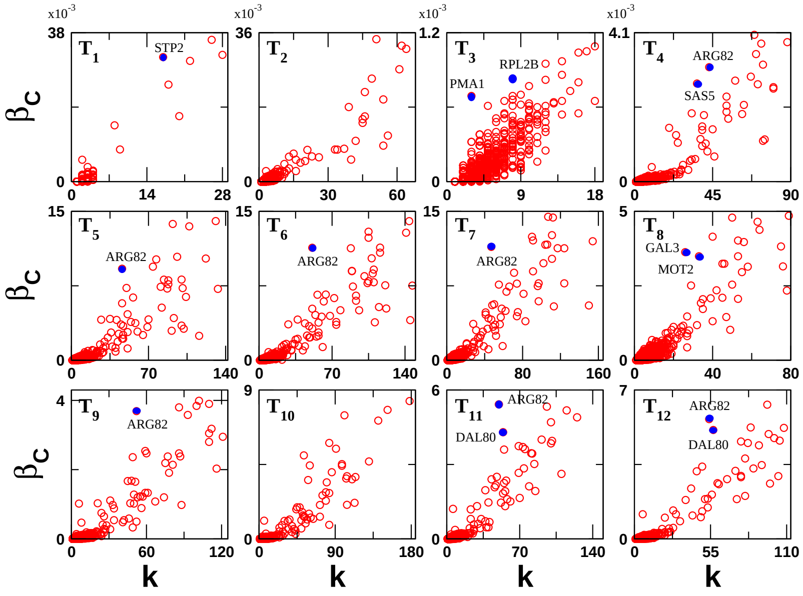

Figure 3. Degree (k) - Betweenness centrality (β_C) correlation in SK1 networks. The points marked in blue correspond to the genes having low degree but high betweenness centrality in respective time points.

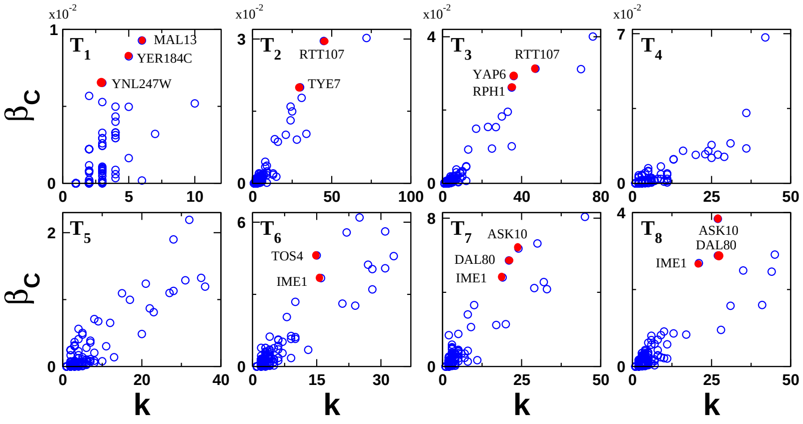

Figure 4. Degree (k) - Betweenness centrality (β_C) correlation in S288c networks. The points marked in red correspond to the genes having relatively low degree but relatively high betweenness centrality in respective time points.

metabolic networks, hierarchical structure indicates that the sets of genes sharing common neighbor are likely to belong to the same functional class.[31] A low value of h indicates more random interactions in the underlying networks. A decrease in hierarchy was observed until the middle phase of sporulation in both the strains. While SK1 continued to exhibit diminishing hierarchy in the late phase, in S288c there was an increase in the hierarchy at the last time point, again suggesting that the strain was attempting to achieve modularity in late phases of sporulation. These results implied that since both the strains showed high values of disassortativity, average clustering coefficient ($\langle C \rangle$) and h early in sporulation, the nature of genes involved in transferring information from the early phase to middle and late phases of sporulation would be important for us to understand the phenotypic difference between them. Therefore, next we identified the genes that would directly or indirectly be involved in bringing about the phenotypic differences in both the strains as sporulation progresses.

For a network, betweenness centrality (see Methods) is a measure of network resilience[32] and it estimates the number of shortest paths (the minimum number of edges traversed between each of the pairs of nodes) that will increase if a node is removed from the network.[33] Usually, nodes with high betweenness centrality are known to bridge different communities in the network. High degree nodes have high betweenness centrality (Figs. 3 and 4). But there are few nodes, which have low degree, yet high betweenness centrality (Figs. 3 and 4), i.e. these genes, despite having very less number of interactions, appear in multiple pathways and hence can be expected to have special significance in the underlying networks. Hence, removal of these nodes can bring about major breakdown in the pathways controlling the sporulation process. We identified

Table 3. Pairs of interacting genes which have low overlap (O) and high link betweenness centrality (β_L) in SK1 networks. Their corresponding indices as given in Fig. S4.

Time points	Gene pair
T_1	*STP2-BAS1*, *HCM1-BAS1*, *BAS1-RIM101*
T_9	*ACE2-MGA1*, *ACE2-SIR2*, *DAL81-ACE2*, *CDC14-ACE2*
T_{10}	*CDC14-ACE2*, *ACE2-ARG82*, *DAL81-ACE2*
T_{11}	*DAL81-ACE2*, *ACE2-MGA1*, *CDC14-ACE2*
T_{12}	*YPL056C-ARG82*, *DAL80-ACE2*, *RIF1-TPK1*

a few important sporulation genes showing this property of low degree and high betweenness centrality in both SK1 and S288c strains. In the SK1 networks, these were known markers of respiratory stress and starvation, namely *STP2*,[34] *PMA1*[35] and *RPL2B*,[36] while in S288c these were *IME1* and *TOS4* which are involved in initiation of meiosis and DNA replication checkpoint response, respectively. Interestingly, these sporulation genes appeared to show this property in the early phase of SK1 strain and during the middle to late meiotic transition in S288c strain. These results suggested that this late appearance of important early sporulation genes as bridges that could transfer information between regulatory modules during sporulation, might be the cause for sporulation not proceeding in the S288c strain. The above analyses helped us to identify the influential genes underlying the sporulation process. We next identified a few interactions that might be instrumental in regulating the sporulation process by considering an important proposition of sociology, Granovetter's 'Weak ties hypothesis'. This hypothesis states that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another.[37] In the networks, the ties having low overlap in their neighborhoods (i.e. less number of common neighbors) are termed as the weak ties.[38] The weak ties that have high link betweenness centrality (see Methods) are the ones known to bridge different communities.[39] Such weak ties revealed through our analysis of different sporulation networks are listed in Tables 3 and 4. Interestingly, we found repetitive occurrence of the same weak ties in consecutive time points for both the strains indicating their phase-specific importance in yeast sporulation. For instance, *BAS1-RTT107*, *BAS1-TYE7*, *YAP6-BAS1* and *ASK10-HMO1* were repetitive weak ties with high link betweenness centrality in consecutive time points of S288c networks while so were *DAL81-ACE2* and *CDC14-ACE2* in SK1 networks. In order to assess the functional importance of these weak ties, we investigated the characteristic properties of the end nodes of these weak ties. Unlike social networks where the end nodes of weak ties are low degree nodes,[40] in the sporulation networks of both the strains, the nodes forming weak ties were high degree nodes. An example of this was again *BAS1*, which as discussed above, is a Myb-related transcription factor involved in amino acid metabolism and meiosis.[24] In addition to *BAS1*, other important sporulation regulatory genes were identified in SK1, such as *RIM101*, a pH-responsive regulator of an initiator of meiosis;[41] *IME2*, a serine-threonine kinase activator of *NDT80* and meiosis;[42] *CDC14*, a protein phosphatase required for meiotic progression;[43] *HCM1*, an activator of genes involved in respiration.[44] Whereas in S288c, apart from *BAS1*, genes associated with mitotic functions such as *TYE7* for glycolytic gene expression,[45] *YAP6* for carbohydrate metabolism,[46] *RTT107* for DNA repair,[47] *ASK10* for glycerol transport[48] and *HMO1* for DNA structure modification[49] were identified. These results showed that while in SK1 strain meiosis-associated genes formed important bridges, in S288c strain bridges were formed by genes involved in mitotic functions. This implied how differences in weak ties in regulatory networks can help us understand the dramatic differences observed in phenotypes. Moreover, *DAL81*, a nitrogen starvation regulator[50] and *ACE2*, a regulator of G1/S transition in the mitotic cell cycle,[51] were identified as end nodes of repetitive weak ties in SK1, suggesting their probable regulatory role in the sporulation process that

Table 4. Catalogue of links having low overlap yet relatively high link betweenness centrality in each time point for S288c. Their corresponding indices as given in Supplementary Fig. S5.

Time points	Nodes with low O high β_L
T_1	*BAS1-RTT107*, *BAS1-TYE7*
T_2	*BAS1-RTT107*, *BAS1-TYE7*
T_3	*BAS1-RTT107*, *BAS1-RPH1*, *YAP6-BAS1*
T_4	*YAP6-BAS1*
T_6	*ASK10-HMO1*, *HMO1-XBP1*
T_7	*ASK10-HMO1*, *YAP6-HMO1*
T_8	*HSF1-IME1*, *ASK10-HMO1*

requires further investigation.

Conclusion

This study presents a novel framework for assessing the molecular consequences of genetic variation across strains. This framework can help reveal the characteristic signatures of the phenotype of interest and identify novel candidate genes contributing to phenotypic variation. Using this framework for the dynamic yeast sporulation network, we showed that the comparative analysis of parameters measuring the network connectivity and degree-degree mixing were the best in identifying differences between two yeast strains showing diverse sporulation efficiency. Comparing the basic structural attributes of the dynamic sporulation networks of the two strains revealed that a delayed crosstalk between functional modules of the low sporulating S288c strain might be the plausible reason behind its low sporulation efficiency. The end nodes of the repetitive weak ties, which are instrumental in bridging communities, were meiosis-associated genes for SK1 strain while these nodes in S288c were involved in mitotic functions, thus outlining the importance of this parameter in unraveling the differences between the two strains.

Model organisms have provided remarkable amount of information to construct the core molecular network. However, identifying the nodes within this network causing variable phenotypic consequences in a natural population, still remains a major challenge. Application of genome-wide strategies to elucidate the molecular networks in multiple genetic backgrounds provides us with the opportunity to understand the impact of natural variation. These strategies can be used to incorporate this new molecular information such as a different type of interaction between molecules, in the already known . This notion is encouraging, especially for understanding complex diseases such as cancer and diabetes, as it provides with a cohort of key nodes governing phenotypic behaviors that get altered by underlying genetic defects. Given the large scale availability of gene expression data, the framework proposed in this study and the network parameters used, can find application in personalized medicine and drug target discovery by carrying out comparative investigations on individuals showing phenotypic diversity.

Methods

Network construction

For constructing the transcriptional regulatory sporulation network, the known static regulatory interactions were overlaid on the time-resolved transcriptomics data of the two strains. This created the dynamic integrated sporulation network. The static network known for yeast contains all the known regulatory interactions between all the yeast transcription factors (TF) and their target genes (TG). These interactions were obtained from YEASTRACT database,[52] a curated repository of regulatory associations in *S. cerevisiae*, based on more than 1,200 literature references.

Gene expression data for yeast strains SK1[53] and S288c[21] was obtained from previously published studies. These datasets contained gene expression of 6,926 genes across 13 different time points in linear scale (0h to 12h with 1h intervals termed as T_0 to T_{12}, respectively) in SK1 and 9 different time points in logarithmic scale (0h, 30m, 45m, 1h10m, 1h40m, 2h30m, 3h50m, 5h40m, 8h30m termed as T_0 to T_8, respectively) in S288c. Gene expression analysis was performed as described previously.[21] In brief, all time points were normalized together using *vsn*[54] and the time-resolved data of each strain was smoothed separately using $locfit$.[55] Gene expression data for each strain was base-transformed, by calculating the fold differences, for all the time-points from t = 0h (t_0), as follows:

$$Y'_{SK1(t_n)} = Y_{SK1(t_n)} - Y_{SK1(t_0)} \tag{1}$$

such that Y is the expression value of a transcript for a strain (SK1) at a specific time point n and Y' is the transformed expression value.

Over-expressed and repressed genes were identified at each time point, by setting the threshold value on the fold differences as 1. This threshold was selected to include known sporulation genes such as *IME2*, *NDT80* and *DIT1*[56] in the network. For each time point, we obtained all the genes showing fold difference greater than 1 (over-expressed) or lower than -1 (repressed).

The dynamic sporulation network was constructed by overlaying the experimentally determined yeast sporulation-specific gene expression values on the yeast static network. For each time point of each strain, only those TF-TG pairs were considered, which both showed either over-expression or repression. These pairs were included in the subnetwork for that specific time point. In such a manner, subnetworks for each time point were constructed for each strain. For comparison of the gene names obtained from YEASTRACT and the sporulation gene expression data, aliases were obtained from Saccharomyces Genome Database.[57]

Data availability

The adjacency matrices of the networks constructed using time-resolved sporulation data drawn from SK1 and S288c strains, the corresponding gene indices and transcription factors are freely available online at Figshare.[58]

Structural parameters

Several statistical measures are proposed to understand specific features of the network.[12,13] The number of connections possessed by a node is termed as its degree. The spread in the degrees is characterized by a distribution function $P(k)$, which gives the probability that a randomly selected node has exactly k edges. The degree distribution of a random graph is a Poisson distribution with a peak at $P(\langle k \rangle)$. One of the most interesting developments in our understanding of complex networks was the discovery that for most large networks the degree distribution significantly deviates from a Poisson distribution. In particular, for a large number of networks, including the World Wide Web, the Internet or the metabolic networks, the degree distribution has a power-law tail $P(k) \sim k^{-\gamma}$. The inherent tendency of social networks to form clusters representing circles of friends or acquaintances in which every member knows every other member, is quantified by the clustering coefficient.[29] Clustering coefficient of a node i denoted as C_i, is defined as the ratio of the number of links existing between the neighbors of the node to the possible number of links that could exist between the neighbors of that node[59] and is given by

$$C_i = \frac{2\sum_{j_2=1}^{k_i}\sum_{j_1=1}^{k_i}(A_{ij_1}A_{j_1j_2}A_{j_2i})}{k_i(k_i-1)} \tag{2}$$

where i is the node of interest and j_1 and j_2 are any two neighbors of the node i and k_i is the degree of the node i. The average clustering coefficient of a network corresponding to a particular condition ($\langle C \rangle$) can be written as

$$\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{3}$$

We define the betweenness centrality of a node i, as the fraction of shortest paths between node pairs that pass through the said node of interest.[33]

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}} \tag{4}$$

where n_{st}^i is the number of geodesic paths from s to t that passes through i and g_{st} is the total number of geodesic paths from s to t.

We quantify the degree-degree correlations of a network by considering the Pearson (degree-degree) correlation coefficient, given as[26]

$$r = \frac{[M^{-1}\sum_{i=1}^{M} j_i k_i] - [M^{-1}\sum_{i=1}^{M}\frac{1}{2}(j_i+k_i)^2]}{[M^{-1}\sum_{i=1}^{M}\frac{1}{2}(j_i^2+k_i^2)] - [M^{-1}\sum_{i=1}^{M}\frac{1}{2}(j_i+k_i)^2]}, \tag{5}$$

where j_i, k_i are the degrees of nodes at both the ends of the i^{th} connection and M represents the total connections in the network.

Link betweenness centrality is defined for an undirected link as

$$\beta_L = \sum_{v \in V_s}\sum_{w \in V/v} \sigma_{vw}(e)/\sigma_{vw} \tag{6}$$

where $\sigma_{vw}(e)$ is the number of shortest paths between v and w that contain e, and σ_{vw} is the total number of shortest paths between v and w.[38]

The overlap of the neighborhood of two connected nodes i and j is defined as[38]

$$O_{ij} = \frac{n_{ij}}{(k_i-1)+(k_j-1)-n_{ij}} \tag{7}$$

where n_{ij} is the number of neighbors common to both nodes i and j. Here k_i and k_j represent the degree of the i^{th} and j^{th} nodes.

Hierarchy can be defined as the heterogeneous distribution of local reaching centrality of nodes in the network. The local reaching centrality, (C_R), of a node i is defined as[30]

$$C_R(i) = \frac{1}{N-1}\sum_{j:0<d(i,j)<\infty}\frac{1}{d(i,j)}, \tag{8}$$

where $d(i,j)$ is the length of the shortest path between any pair of nodes i and j. The measure of hierarchy (h), termed as global reaching centrality is given by

$$h = \frac{\sum_{i \in V}[C_R^{max} - C_R(i)]}{N-1} \tag{9}$$

Acknowledgements

SJ acknowledges Department of Science and Technology (DST), Govt. of India grant EMR/2014/000368 and Council of Scientific and Industrial Research (CSIR), Govt. of India grant 25(0205)/12/EMR-II. HS was supported by intra-mural grant from Tata Institute of Fundamental Research (12P-0120).

Author contributions statement

SJ conceived the idea. SJ and HS designed and supervised the project. CS constructed the networks and analyzed the structural properties. SG and CS analyzed the functional properties. All the authors wrote and approved the manuscript.

Additional Information

Supplementary Information

accompanies this paper at http://www.nature.com/naturescientificreports

Competing financial interests statement

The authors declare no competing financial interests.

References

1. Chu, S. *et al*. The Transcriptional Program of Sporulation in Budding Yeast. *Science* **282**, 699-705 (1998).

2. Neiman, A.M. Ascospore formation in the yeast Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.* **69**(4), 565-584 (2005).

3. Primig, M. *et al*. The core meiotic transcriptome in budding yeasts. *Nat. Genet.* **26**, 415-423 (2000).

4. Bar-Joseph, Z., Gitter, A. & Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13**, 552-564 (2012).

5. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *PNAS* **95**(25), 14863-14868 (1998).

6. Kerr, M. K. & Churchill, G. A. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *PNAS* **98**(16), 8961-8965 (2001).

7. Wakefield, J. C., Zhou, C. & Self, S. G. Modelling Gene Expression Data over Time: Curve Clustering with Informative Prior Distributions in *Bayesian Statistics*, Vol. 7 (eds Bernardo J. M. *et al.*) 721-732 (Oxford University Press, 2003).

8. McNicholas, P. D. & Murphy, T. B. Model-based clustering of longitudinal data. *The Canadian Journal of Statistics* **38** (1), 153-168 (2010).

9. Roy, S., Bhattacharyya, D. K. & Kalita, J. K. Cobi: pattern based co-regulated biclustering of gene expression data. *Pattern Recogn. Lett.* **34**(14), 1669-1678 (2013).

10. Sontag, E., Kiyatkin, A. & Kholodenko, B. N. Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics* **20**(12), 1877-1886 (2004).

11. Huang, S. Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *J. Mol. Med.* **77**(6), 469-480 (1999).

12. Albert, R. & Barabási, A. L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**(1), 47 (2002); Strogatz, S. H. Exploring complex networks. *Nature* **410**(6825), 268-276 (2001).

13. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. U. Complex networks: Structure and dynamics. *Phys. Rep.* **424**(4), 175-308 (2006) and references therein.

14. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**(2), 101-113 (2004).

15. Rai, A., Menon, A. V. & Jalan, S. Randomness and preserved patterns in cancer network. *Sci. Rep.* **4** (2014); Shinde, P., Yadav, A., Rai, A. & Jalan, S. Dissortativity and duplications in oral cancer. *EPJB* **88**(8), 1-7 (2015); Pujana, M.A. *et al.* Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.* **39**(11) 1338-1349 (2007).

16. Alon, U. Biological networks: the tinkerer as an engineer. *Science* **301**(5641), 1866-1867 (2003); Yu, D., Kim, M., Xiao, G. & Hwang, T. H. Review of biological network data and its applications. *Genomics & informatics* **11**(4), 200-210 (2013).

17. Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. & Barabási, A.L. The human disease network. *PNAS* **104**(21), 8685-8690 (2007).

18. Foury, F., & Kucej, M. Yeast mitochondrial biogenesis: a model system for humans?. *Current opinion in chemical biology* **6**(1), 106-111 (2002); Barrientos, A. Yeast models of human mitochondrial diseases. *IUBMB life* **55**(2), 83-95 (2003).

19. Lander, Eric S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**(6822), 860-921 (2001); Bassett, D. E., Boguski, M. S., & Hieter, P. Yeast genes and human disease. *Nature* **379** 589-590 (1996).

20. Tsuchiya, D., Yang, Y. & Lacefield, S. Positive Feedback of NDT80 Expression Ensures Irreversible Meiotic Commitment in Budding Yeast. *PLoS Genet.* **10**(6), 1-15 (2014).

21. Gupta S. *et al.* Temporal Expression Pro- filing Identifies Pathways Mediating Effect of Causal Variant on Phenotype. *PLoS Genet.* **11**(6), 1-23 (2015).

22. Neiman A.M., Sporulation in the budding yeast Saccharomyces cerevisiae. *Genetics* **189**(3), 737-765 (2011).

23. Gupta, S. *et al.* Meiotic Interactors of a Mitotic Gene TAO3 Revealed by Functional Analysis of its Rare Variant. *G3 Genes|Genomes|Genetics* doi:10.1534/g3.116.029900 (2016).

24. Mieczkowski, P. A. *et al.* Global analysis of the relationship between the binding of the Bas1p transcription factor and meiosis-specific double-strand DNA breaks in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **26**(3), 1014-1027 (2006).

25. Görner, W. *et al.* Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. *Genes & development* **12**(4), 586-597 (1998); Brandman, O. *et al.* A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* **151**(5), 1042-1054 (2012).

26. Newman, M. E. Assortative mixing in networks. *Phys. Rev. Lett.* **89**(20), 208701 (2002).

27. Yook, S. H., Radicchi, F. & Meyer-Ortmanns, H. (2005). Self-similar scale-free networks and disassortativity. *Phys. Rev. E* **72**(4), 045105.

28. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910-913 (2002).

29. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**(6684), 440-442 (1998).

30. Mones, E., Vicsek, L. & Vicsek, T. Hierarchy measure for complex networks. *PloS one* **7**(3), e33799 (2012).

31. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**(5586), 1551-1555 (2002).

32. Newman, M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64**, 016132 (2001); Wang, H., Hernandez, J. M. & Van Mieghem, P. Betweenness centrality in a weighted network. *Phys. Rev. E* **77**(4), 046105 (2008).

33. Newman, M. E. J. The Structure and Function of Complex Networks. *SIAM Rev.* **45**(2), 167-256 (2003).

34. Merz, S. & Westermann, B. Genome-wide deletion mutant analysis reveals genes required for respiratory growth, mitochondrial genome maintenance and mitochondrial protein synthesis in Saccharomyces cerevisiae. *Genome Biol.* **10**(9), R95 (2009).

35. Ding, J. *et al.* Tolerance and stress response to ethanol in the yeast Saccharomyces cerevisiae. *Appl. Microbiol. Biotechnol.* **85**(2), 253-263 (2009).

36. Davey, H. M. *et al.* Genome-wide analysis of longevity in nutrient-deprived Saccharomyces cerevisiae reveals importance of recycling in maintaining cell viability. *Environ. Microbiol.* **14**(5), 1249-1260 (2012).

37. Granovetter, M. S. The strength of weak ties. *Am. J. Sociol.* 1360-1380 (1973).

38. Onnela, J.P. *et al.* Analysis of a large-scale weighted network of one-to-one human communication. *New J. Phys* **9**(6), 179 (2007).

39. Szell, M. & Stefan, T. Measuring social dynamics in a massive multiplayer online game. *Soc. Net.* **32**(4), 313-329 (2010).

40. Sarkar, C., Yadav, A. & Jalan, S. Multilayer network decoding versatility and trust. *EPL* **113**(1), 18007 (2016).

41. Su, S. S. & Mitchell, A. P. Identification of functionally related genes that stimulate early meiotic gene expression in yeast. *Genetics* **133**(1), 67-77 (1993).

42. Honigberg, S. M. & Purnapatre, K. Signal pathway integration in the switch from the mitotic cell cycle to meiosis in yeast. *J. Cell Sci.* **116**(11), 2137-2147 (2003).

43. McDonald, C. M., Cooper, K. F. & Winter, E. The Ama1-directed anaphase-promoting complex regulates the Smk1 mitogen-activated protein kinase during meiosis in yeast. *Genetics* **171**(3), 901-911 (2005).

44. Rodriguez-Colman M. J. *et al*. The forkhead transcription factor Hcm1 promotes mitochondrial biogenesis and stress resistance in yeast. *J. Biol. Chem.* **285**(47), 37092-37101 (2010).

45. Sato, T. *et al*. The E-box DNA binding protein Sgc1p suppresses the gcr2 mutation, which is involved in transcriptional activation of glycolytic genes in Saccharomyces cerevisiae. *FEBS Lett.* **463**(3), 307-311 (1999).

46. Hanlon, S. E., Rizzo, J. M., Tatomer, D. C., Lieb, J. D. & Buck, M. J. The stress response factors Yap6, Cin5, Phd1, and Skn7 direct targeting of the conserved co-repressor Tup1-Ssn6 in S. cerevisiae. *PloS one* **6**(4), e19060 (2011).

47. Leung, G. P., Lee, L., Schmidt, T. I., Shirahige, K. & Kobor, M. S. Rtt107 is required for recruitment of the SMC5/6 complex to DNA double strand breaks. *J. Biol. Chem.* **286**(29), 26250-26257 (2011)..

48. Beese, S. E., Negishi, T. & Levin, D. E. Identification of positive regulators of the yeast fps1 glycerol channel. *PLoS Genet.* **5**(11), e1000738 (2009).

49. Murugesapillai, D. *et al*. DNA bridging and looping by HMO1 provides a mechanism for stabilizing nucleosome-free chromatin. *Nucleic Acids Res.* **42**(14), 8996-9004 (2014).

50. Marzluf, G. A. Genetic regulation of nitrogen metabolism in the fungi. *Microbiol. Mol. Biol. Rev.* **61**(1), 17-32 (1997).

51. Spellman, P. T. *et al*. Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **9**(12), 3273-3297 (1998).

52. Teixeira M. C. *et al.*. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. *Nucleic Acids Res.* gkt1015, 1-6 (2013).

53. Lardenois A. *et al*. Execution of the meiotic noncoding RNA expression program and the onset of gametogenesis in yeast require the conserved exosome subunit Rrp6. *PNAS* **108**(3), 1058-1063 (2011).

54. Huber W., Heydebreck von A., Sültmann H., Poustka A. & Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18** Suppl 1, S96-104 (2002).

55. Loader C. Locfit: Local regression, likelihood and density estimation. R package version 1.5-9.1 Merck, Kenilworth, N. J. (2013).

56. Kassir Y. *et al*. Transcriptional regulation of meiosis in budding yeast. *Int. Rev. Cytol.* **224**, 111–171 (2003).

57. Cherry, J. M. *et al*. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* gkr1029, 1-6 (2011).

58. Sporulation data available at Figshare.

59. Newman, M. E. J., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**(2), 026118 (2001).