Using synthetic lethality and integration of genomic data for finding spindle migration genes

Daniel Schöner*,[1,5], Markus Kalisch*,[2], Christian Leisner*,[3] , Lukas Meier*,[2], Marc Sohrmann[3,5], Mahamadou Faty[4], Yves Barral[3], Matthias Peter[3,5], Wilhelm Gruissem[1,5] and Peter Bühlmann[2,5]

[1]Institute of Plant Science, ETH Zurich, Universitaetsstr. 2, 8092 Zurich, Switzerland
[2]Seminar for Statistics, ETH Zurich, Leonhardstr. 27, 8092 Zurich, Switzerland
[3]Institute of Biochemistry, ETH Zurich, Schafmattstr. 18, 8093 Zurich, Switzerland
[4]Friedrich Miescher Institute, Maulbeerstrasse 66, Basel, Switzerland
[5]Competence Center for Systems Physiology and Metabolic Diseases (CC-SPMD)

Email: Daniel Schöner*- dhs@ethz.ch; Markus Kalisch*- kalisch@stat.math.ethz.ch; Christian Leisner*- christian.leisner@bc.biol.ethz.ch; Lukas Meier*- meier@stat.math.ethz.ch; Marc Sohrmann - marc.sohrmann@bc.biol.ethz.ch; Mahamadou Faty - mahamadou.faty@unibas.ch; Yves Barral - yves.barral@bc.biol.ethz.ch; Matthias Peter - matthias.peter@bc.biol.ethz.ch; Wilhelm Gruissem - wgruissem@ethz.ch; Peter Bühlmann - buhlmann@stat.math.ethz.ch;

*Corresponding author

Abstract

Background: Large scale screens for synthetic lethality are widely used in yeast genetics to systematically search for genes that are involved in specific biological processes. Often the amounts of data resulting from single screens far exceed the capacities of experimental characterization of every target found. Thus, computational tools are required to select promising candidates from a screen in order to reduce the number of experiments to a manageable size.

Results: We use an unsupervised statistical method for the analysis of yeast synthetic lethality data that integrates information from other biological data sources, such as gene expression measurements, phenotypic profiling, RNA degradation and sequence similarity. By virtue of a Multivariate Gaussian Mixture Model, we determine the best combination of features that result in a grouping of the genetic interactions into two parts. An analysis of synthetic lethality data from two screens performed with *arp1* and *jnm1*, two genes involved in the migration of the mitotic spindle, yields a small group of statistically significant candidate genetic interactors that we propose as potential members in this biological process. Preliminary experimental testing of a subset of these candidates confirms their role and yields novel genes involved in spindle migration.

Conclusions: We demonstrate, using statistical significance and biological validation, that multivariate Gaussian Mixture Modeling can be used to select candidate genetic interactions for experimental characterization from synthetic lethality datasets. For the given example, integration of different data sources contributes to the identification of genetic interactors of *arp1* and *jnm1* that play a role in the same biological process.

Background

One of the major challenges of computational biology is the extraction of relevant information from the increasing amounts of data resulting from large scale experimentation. While the data quality of single high-throughput assays has often been challenged, there is great promise that the reliability and precision of the outcomes can be increased through integration and combination of multiple data sources.

We applied statistical modeling, based on data integration, for finding yeast genes involved in spindle migration from two synthetic lethality screens performed with *arp1* and *jnm1*.

Synthetic lethality data

Synthetic lethality is the phenomenon of observing a dead phenotype when two otherwise viable gene deletions are combined in one cell [1]. Since the yeast collection of deletion mutants has been available, this procedure can be carried out on a global scale to uncover interactions between non-essential genes. This technology is called the Synthetic Genetic Array (SGA) [2]. By using well known genes as query genes and crossing them into the deletion set, one can systematically search for target genes that are synthetically lethal with the query gene. As a conclusion, these targets, together with the query gene can be placed in a common functional context in the cell. SGA is a powerful genetics tool for studying vital biological processes and for finding new components involved in them. However, as with large scale experimentation in general, not all of the targets identified by such a screen are specific to the biological scenario that is investigated [3]. For instance, a lethal phenotype can occur as a non-specific effect when both single deletions already have reduced fitness. Also, two very distantly related genes can show synthetic lethality because a gene deletion does not singly represent the loss of a gene, but rather a whole cellular response to it, possibly affecting many pathways. Finally, some target genes might simply represent noise inherent to the experiment. Of primary interest, however, are genetic interactions that occur within the same biological scenario because they indicate a close functional relationship between query and target gene, that can be readily examined by additional experimental efforts.

Several approaches confronted the problem of characterizing synthetic lethality by relying on additional biological information.

In [4] a supervised approach with decision trees was used to predict synthetic lethal interactions on the basis of genomic and proteomic features such as localization, mRNA expression, physical interaction, protein function, and characteristics of network topology. Instead of predicting genetic interactions, an interesting attempt to characterize synthetic lethality was undertaken by [5]. By scrutinizing the topology of global genetic and protein interaction networks, they found that synthetic lethality can occur either within or between pathways with the majority occurring between pathways. Their approach was restricted to protein interactions for the interpretation of synthetic lethality. Here, we investigate whether integrating additional genomic information facilitates the interpretation of synthetic lethal interactions found in high-throughput screens and serves to distill the close genetic relationships from the broad and indirect ones.

Our Approach

Our approach is closest to the goals in [5], but we integrate multiple data sources for characterizing the relationship that might exist between genetic interactors found by SGA. Moreover, instead of considering the whole genetic network in a global approach, we focus on the results of two synthetic lethal screens performed with *arp1* and *jnm1*, two genes involved in the migration of the mitotic spindle [6, 7]. This process is essential for high fidelity chromosome segregation and proper cell division in budding yeast and serves as a case study because some important regulatory elements are already known and can be used for reference (see Table 1). Both query genes used for the screens are involved in dynein-dependent spindle positioning. Due to its importance for the cell, spindle migration is a highly buffered process. We chose *arp1* and *jnm1* as query genes, because an SGA-screen will detect genes functioning in pathways that compensate for the loss of dynein-dependent spindle positioning and thus also play a role in spindle migration. Further and more importantly, a set of secondary assays is available that is adequate for testing the outcomes of a computational analysis in an experimental setting.

In a statistical model that integrates synthetic lethality data with information from multiple sources, we seek to set apart genetic interactions with genes involved in spindle migration from interactions with genes that are more distantly related or not related at all. Since it is much easier to characterize genetic interactions of the first kind in a follow-up experiment, which we have pursued after statistical modeling, we focus on finding a subset of genetic interactors that have a close relationship to *arp1* and *jnm1* assuming that they also have a function in spindle migration.

Our method does not require a curated data set as a gold standard, as in supervised learning approaches, but uses the structure found in the data for grouping in an unsupervised manner, using a Gaussian Mixture Model and the Expectation Maximization (EM) algorithm for estimating the corresponding parameters. To our knowledge, the presented approach is novel because it uses mixture modeling as a framework for characterizing synthetic lethal interactions and for integrating different data types. The following paragraphs describe how measurements of mRNA expression, phenotypic profiling, mRNA decay and sequence similarity were integrated and used for statistical modeling.

Included Datasets

When searching for target genes that are closely related to the query gene of an SGA, a direct protein interaction between both gene products is an obvious feature to look for. However, the protein interaction network measured on a large scale by two-hybrid or copurification techniques only features the products of less than 10% of the genes in the datasets analyzed in this work. Hence, this incomplete information could not be included in a reasonable way in our model. Instead, we focused on data sources with good genome coverage to ensure that comprehensive information is incorporated.

Genes involved in the same biological process are likely to show similar mRNA-expression profiles [8]. To include knowledge about gene co-expression we chose three gene expression data sets. In [9], 15 environmental and chemical stress conditions were tested. In [10], changes in mRNA expression in response to 300 gene deletions and drugs were monitored. In [11], the changes in mRNA expression were measured at 80 experimental conditions related to the cell cycle. For each of these data sets, we calculated the Pearson correlation coefficient of both query genes to all corresponding target genes. In the following, we will refer to these variables according to their source as *gasch.corr*, *hughes.corr* and *spellman.corr*, respectively.

In another microarray-based study, [12], the sensitivity of the yeast gene deletion library to various growth conditions was measured. Strains responding in the same or similar fashion to these cues are likely to carry deletions of genes that are functionally related. We therefore calculated the Pearson correlation coefficients of the sensitivity profiles of all target genes to their corresponding query genes. We will refer to this variable as *pheno.corr*.

To include information about posttranscriptional regulation in the analysis, we considered the degradation rates of mRNA-transcripts. In a systematic approach, [13], the mRNA decay rates of nearly all yeast ORFs were measured. To compare the rates of the different targets to the query genes we calculated the ratios of decay rates of the target genes and query genes and performed a log-transformation on them. We will refer to this variable as *logRNA.ratio*.

Genes carrying out similar biological or biochemical functions are likely to share common activity domains in their protein structure, coded in their protein sequence. To include such information we determined the sequence similarities between the query genes and the corresponding target genes using pblast [14]. The log-transformed percentage values of sequence similarity were included in the analysis as the variable *logseq.sim*.

In total, we use 6 different data sources in addition to the information from synthetic lethal screens.

Results and Discussion

The statistical analysis performed on synthetic lethality data with query genes *arp1* and *jnm1* and including the 6 additional data sources resulted in a small group of 5 genes that we propose to be closely related to *arp1* and *jnm1* in the sense that they have a function in the same biological process. Based on the analysis of already known genes (see Table 1), this group is enriched for genes involved in spindle migration. Initial experimental validation of the candidates previously uncharacterized in this context confirmed an involvement for the majority of the statistically significant genes.

Feature selection

We computed every combination of the variables derived from the 6 datasets ($2^6 - 1 = 63$) and used the Bayesian Information Criterion (BIC, see Methods) to evaluate the sets of features that best approximate the data. Since the aim was to integrate information from different datasets, single-variable models were not further considered.

The variables *hughes.corr*, *spellman.corr* and *pheno.corr* are part of the top-scoring combinations, whereas combinations containing *gasch.corr*, *logRNA.ratio* and *logseq.sim* yielded worse results (data not shown). We conclude from this that the former variables contribute to a structure that allows good separation of the data into two groups whereas the latter variables do not provide additional information. In a two-component Gaussian Mixture Model the two components are to be interpreted as either having a direct involvement in spindle migration or no direct involvement.

The best fit for two groups was achieved by a combination of the variables *hughes.corr* and *spellman.corr* and thus only relies on transcriptional information. As illustrated in Figure 1, there is a small group consisting of 7 data points (synthetic lethal interactions) that can be discerned from a bigger group containing the rest of the data. While the values of the variable *spellman.corr* do not differ much between the two groups, the variable *hughes.corr* is indicative of separation. All members of the small group share intermediate to high values of correlation (0.3-0.65) in the hughes-data set as opposed to the members of the big group that fall in a range between -0.4 and 0.3. Although, at first sight, only the information of *hughes.corr* seems to be important, the more detailed statistical analysis shows that this is not the case: The combination of *hughes.corr* and *spellman.corr* is clearly more suited with respect to the BIC score.

Statistical Assessment

Given unequal grouping, one would naturally consider the small group to comprise interesting information. The genetic interactors in the small group differentiate themselves from the remainder of the data set and thus represent promising candidates for thorough experimental testing. Moreover, the high positive correlation values for the variable *hughes.corr* show that these target genetic interactors are transcriptionally co-regulated with either *arp1* or *jnm1* and are therefore likely to be involved in the same biological process. For the grouping illustrated in Figure 1, the default cutoff for the posterior probability in the Gaussian Mixture Model was set to 0.5 to separate the small from the big group. Genes with a higher posterior probability were assigned to the small group and genes with lower values comprised the big group. Variation of this threshold shifts the quantitative proportion between the small and the big group. To judge the enrichment of the small group with genetic interactors known to be involved in spindle migration depending on group size, we used a reference list of 15 genes out of the 141 genes contained in the dataset that are known to be involved in spindle migration (see Table 1). Employing a hypergeometric test shows that the amount of known genes in the smaller group is significantly higher than would be expected by chance. Furthermore, we analyzed the enrichment when changing the size of the small group. First, we increased the cutoff of the posterior probabilities so that the small group contained only 5 genetic interactions of the total of 141. Then, we reduced the cutoff so that the small group contained 50 genetic interactions. In both cases, the small group showed a significant enrichment of known spindle migration genes (see Table 3).

The lists of genes resulting from both small groups can be considered as candidates for further biological experimentation. They are significantly enriched in genes known to be specific for spindle migration and thus it is likely that among the rest of the genes in these lists, which are unknown with respect to spindle migration, additional members of this biological process can be found.

Experimental Validation

In order to experimentally test whether the genetic interactions in the small group consisting of 5 candidates (see Table 3) are related to spindle migration we used time-lapse microscopy (see Methods). Proper spindle migration requires several cellular processes such as spindle integrity, spindle elongation and localization of Kar9. Thus, the genes in the list cannot be expected to show a unique phenotype that could be detected by a single assay. Therefore, we looked for perturbation in any of these processes. Ase1 has already been reported to be required for spindle integrity. In a deletion strain, the spindles fail to elongate during anaphase [15, 16] and Ase1 localizes to the spindle midzone, indicating that Ase1 is required for high fidelity chromosome segregation. Asymmetric localization of Kar9 is used as an assay to identifiy genes with a function in spindle migration [17].

From previous experiments, it was known that *tvp38*Δ cells show perturbed asymmetric Kar9 localization. Only 73% of the *tvp38*Δ cells have Kar9 localized in a strongly asymmetric manner as opposed to 83% in WT cells (see Figure 2). This indicates that Tvp38 plays a role in maintaining Kar9 asymmetry. For the reasons mentioned above, both genes were considered as required components for spindle migration and used in the reference gene list for statistical assessment.

The additional genes found in the small group of 5 genes were analyzed by using the same experimental methods.

For the *uba4*Δ mutant no effect was detected. However, it should be mentioned that Urm1, the only known substrate of Uba4 in urmylation, has mildly perturbed Kar9 localization when deleted. This suggests a role of Uba4 as a peripheral regulator of Kar9 asymmetry possibly through urmylation.

Deletion of the uncharacterized ORF, YHR127W, shows a Kar9 localization phenotype similar to that seen in *tvp38*Δ cells, although less pronounced, 76% asymmetry vs. 83% in WT (see Figure 2).

A deletion of she1 (YBL031W), a gene coding for a cytoskeletal protein of unknown function, resulted in a marked decrease in Kar9 asymmetry as well as a high frequency of broken spindles in anaphase cells (see Figure 2 and suppl. Figure 1). This gene has a phenotype in both assays used for validation, which gives strong support for an involvement in spindle migration.

These results demonstrate that all of 5 genes in the small group of our statistical model play a role in spindle migration. For one of the genes only a vague involvement can be assumed due to indirect evidence (*uba4*). The phenotypes of deletions in *tvp38* and YHR127W, though being relatively mild, suggest an immediate effect of both genes on spindle migration. For the other two genes strong evidence exists that they play a direct role in the proper functioning of spindle migration and chromosome segregation (*she1* and *ase1*). Interestingly, the role of YHR127W and *she1* in this context is a novel finding. In conclusion, all 3 of the additionally found candidates merit further experimental characterization in order to more precisely determine their mechanistic involvement in spindle migration.

Conclusions

We used and evaluated an unsupervised statistical method relying on data integration for separating synthetic lethal interaction partners of *arp1* and *jnm1* into those that are specific to spindle migration and those representing the unspecific or more distantly related remainder of interactions.

Multivariate Gaussian Mixture Modeling was applied to divide different subsets of a heterogeneous genomic data set into two groups. A combination of the two features *hughes.corr* and *spellman.corr* derived from two gene expression datasets resulted in the best model fit. For this combination, a small group of 5 genes was identified that was significantly enriched in known spindle migration genes. Moreover, biological testing of the three top scoring genes that were uncharacterized in this context (the other two genes were characterized already) yielded experimental confirmation for being involved in spindle migration. The fact that we obtained novel findings for two of the genes further underscores the usefulness of our method.

It is an interesting finding that a small model consisting of only two biological variables resulted in the best separation of genetic interactions. One would expect that including as much information as possible, which in this case would be a combination of all 6 variables, should result in the most conclusive outcome. Here, this is not the case. Apparently, correlating mRNA-expression for the genes considered in this study is more informative of a close relationship between them than is comparing their sequence similarity or their rates mRNA-decay. Obviously, including the latter variables obscures relevant information. Since our analysis is based on synthetic lethal screens, not all known spindle migration genes could be found. Essential genes and genes from the dynein-dependent spindle positioning pathway are missing in the analysis since they could not be detected by synthetic lethality, others, such as *kar9*, had to be excluded from the analysis due to missing data. In addition, the best combination of variables used to produce the results only relies on mRNA expression of the genes in the synthetic lethality dataset. Some genes important for spindle migration that are also included in the dataset might have been missed, if they are not transcriptionally co-regulated with *arp1* and *jnm1*, but regulated differently. Hence, our data integration approach does not comprehensively identify all members of this biological process. Yet, based on the features used for the model, we can assign probabilities to those interactions that have been identified by SGA according to whether they are closely related to the biological process of interest or not. This allows a ranking of the candidates and, based on the ranks, the construction of a list of genes that deserve further experimental characterization. Another point is that genes showing synthetic lethality with the two query genes *arp1* and *jnm1* do not necessarily have to be related to spindle migration since these two query genes are involved in many different processes where movement of microtubules is required, such as cell polarity, cell migration, vesicle transport, and the formation of membrane protrusions. Nevertheless, we focused on their role in spindle migration because it is the function that is experimentally described best, and because the experimental means for testing promising genes in that context were available. Indeed, experimental validation of a small group of candidate genes has supported the model-based predictions and adds initial biological evidence to the assumption that these genes are involved in the process that was studied. However, our model and preliminary experiments do not allow any ultimate conclusions about the exact mechanistic interplay of the validated genes in spindle migration. For this, thorough experimental characterization is indispensable. Still, the method contributes to the study of spindle migration by identifying primary candidates for further study.

We presume that among the 50 genetic interactions from the other statistically significant small group whose p-value is actually smaller than for the other small group (see Table 2) many genetic interactors can also be placed in the molecular environment of spindle migration.

The presented work shows that data-integration can be useful for analysis and characterization of synthetic lethality data. We demonstrate an efficient way, in terms of a statistical model, to reduce the list of target genes from a screen to a list of candidates with good prospects for further experimentation in the laboratory. Since the amounts and the quality of high-throughput data will increase in the future, more and better biological features can be expected to arise. Including them in our model will increase its predictive power and accuracy.

Methods

Mixture modeling

In order to divide the genetic interactions into two groups, we assumed that each sample was drawn with a given probability π_k from a p dimensional multivariate Gaussian distribution labeled by $k \in \{1, 2\}$. The goal is to recover the probabilities π_k and the parameters of the two multivariate Gaussian distributions from the given samples in order to quantify the groups. This can be more generally cast in the framework of Gaussian Mixture Models (GMM).

The probability density of a Gaussian Mixture Model with 2 groups can be written as

$$f(\mathbf{x}\,|\,\mu_1, \Sigma_1, \mu_2, \Sigma_2) = \sum_{k=1}^{2} \pi_k \phi_k(\mathbf{x}\,|\,\mu_k, \Sigma_k) \tag{1}$$

where $\phi(\mathbf{x}\,|\,\mu_k, \Sigma_k)$ is the probability density of the multivariate Gaussian distribution corresponding to group k:

$$\phi(\mathbf{x}\,|\,\mu_k, \Sigma_k) = (2\pi)^{-\frac{p}{2}} \det(\Sigma_k)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_k)^T \Sigma_k^{-1}(\mathbf{x}-\mu_k)\right)$$

The parameters can be found by the EM algorithm. The optimal set of variables of other data sources was found by minimizing the Bayesian Information Criterion (BIC), which is defined as:

$$BIC = -2\ln(L) + d\ln(n),$$

where $\ln(L)$ is the log likelihood, d is the number of parameters and n is the sample size. We used the `R`-package `mclust` [18] for all calculations.

All samples can then be assigned to one out of the two groups by inspecting the posterior probabilities $P[\text{Group} = k\,|\,\mathbf{x}], k \in \{1, 2\}$.

Experimental Procedures

SGA

Performed as described in [19]. In brief, query strains used were, arp1::kanMX his3 leu2 ura3 lys2 can1::MFA1pr-HIS3 and jnm1::kanMX his3 leu2 ura3 lys2 can1::MFA1pr-HIS3. Each strain was crossed into the Yeast Knock Out Collection (Open Biosystems), and double knock-out strains were scored for synthetic lethality. Each screen was performed once.

Yeast Strains

Deletion strains were generated for each predicted ORF in the presence of a chromosomal CFP-tub1:URA marker in the URA3 locus and a chromosomal kar9-YFP:NAT marker in the endogenous kar9 locus. Integration of CFP-tub1 and C-terminal tagging with YFP was done as described in (Liakopoulos et al., 2003).

Time-Lapse Fluorescence Microscopy

Fluorescence microscopy was performed on an Olympus BX50 fluorescence microscope equipped with a piezo motor, Polychrom IV monochromator as light source, a high speed CCD camera (Imago, TillPhotonics) and TILLVision software (TILLPhotonics, Martinsried, Germany). Dual color acquisitions were performed using a Chroma CFP/YFP dual band filter.

Images were acquired as stacks of 5 focal slices, 0.4 μm between each slice. Each time-lapse series was recorded with 10 time frames and presented as 10 maximum projections with 10 s intervals over the course of 100 s. The time-lapse series were acquired in a YFP and a CFP channel and fused as RGB movies in NIH Image J.

Spindle Integrity and Elongation

Cells with reduced spindle integrity may have difficulty with spindle elongation during anaphase [15, 16]. We quantified the fraction of WT vs. Δcells with compromised anaphase spindle integrity. Breaking and bending of the spindle was scored. The experiment was performed with two independent clones per strain, and n > 20 anaphase cells per strain. For single cell illustrations see supplement.

Kar9 Localization

Localization of spindle positioning protein Kar9 on the spindle pole body and the astral microtubules occurs only on the bud proximal side of the spindle. This asymmetric localization of Kar9 is essential for proper function of spindle migration. We scored for asymmetric Kar9 localization in all the predicted mutants. We quantified the fraction of cells with Kar9 localized in a strongly asymmetric, weakly asymmetric and symmetric manner. The experiment was performed with two independent clones per strain, n > 100 cells per strain. For single cell illustrations see supplement.

Yeast strains used in this study:

KAR9::YFP:NAT ura3::TUB1-CFP:URA3 gpd1::kanMX

ade2-101ura3-52 lys2-801 his3-Δ200 trp1-Δ63 leu2

KAR9::YFP:NAT ura3::TUB1-CFP:URA3 she1::kanMX

ade2-101ura3-52 lys2-801 his3-Δ200 trp1-Δ63 leu2

KAR9::YFP:NAT ura3::TUB1-CFP:URA3 tvp38::kanMX

ade2-101ura3-52 lys2-801 his3-Δ200 trp1-Δ63 leu2

KAR9::YFP:NAT ura3::TUB1-CFP:URA3 YHR127W::kanMX

ade2-101ura3-52 lys2-801 his3-Δ200 trp1-Δ63 leu2

KAR9::YFP:NAT ura3::TUB1-CFP:URA3

ade2-101ura3-52 lys2-801 his3-Δ200 trp1-Δ63 leu2

List of abbreviations

EM algorithm - Expectation Maximization algorithm

GMM - Gaussian Mixture Model

SGA - Synthetic Genetic Array

BIC - Bayesian Information Criterion

Authors contributions

Daniel Schöner, Markus Kalisch, Christian Leisner and Lukas Meier contributed equally to the work presented in this manuscript.

Acknowledgements

The authors would like to thank Nicolas Page for supporting in the synthetic lethality screens and also the Functional Genomics Center Zurich for its financial contribution to the Biomek-robot used for screening. Markus Kalisch has been partially supported by the Swiss National Science Foundation (No. 200020-113270). Christian Leisner and Mahamadou Faty have been supported by the Roche foundation. All authors acknowledge the ETH Zurich for financial support.

References

1. Guarente L: **Synthetic enhancement in gene interaction: a genetic tool come of age**. *Trends in Genetics* 1993, **9**(10):362–66.
2. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, et al.: **Global mapping of the yeast genetic interaction network**. *Science* 2004, **303**(5659):808–13.
3. Hartman JL, Garvik B, Hartwell L: **Principles for the Buffering of Genetic Variation**. *Science* 2001, **291**(9):1001–1004.
4. Wong SL, Zhang LV, Tong AHY, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey, et al.: **Combining biological networks to predict genetic interactions**. *Proc Natl Acad Sci USA* 2004, **101**(44):15682–15687.
5. Kelley R, Ideker T: **Systematic interpretation of genetic interactions using protein networks**. *Nature Biotechnology* 2005, **102**:561–66.
6. L Muhua KT, JA C: **A yeast actin-related protein homologous to that in vertebrate dynactin complex is important for spindle orientation and nuclear migration.** *Cell* 1994, **78**(4):669–79.

7. McMillan JN, Tatchell K: **The JNM1 gene in the yeast Saccharomyces cerevisiae is required for nuclear migration and spindle orientation during the mitotic cell cycle.** *Journal of Cell Biology* 1994, **125**(4):143–158.
8. DeRisi JL, Iyer VR, Brown PO: **Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale**. *Science* 1997, **278**(5338):680–686.
9. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11**(12):4241–57.
10. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett H, Coffey E, Dai H, He YD, et al.: **Functional discovery via a compendium of expression profiles**. *Cell* 2000, **23**(5):109–26.
11. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization.** *Mol Biol Cell* 2006, **12**(9):3273–3297.
12. Brown JA, Sherlock G, Myers CL, Burrows NM, Deng C, Wu HI, McCann KE, Troyanskaya OG, Brown JM: **Global analysis of gene function in yeast by quantitative phenotypic profiling**. *Mol Syst Biol* 2006, **2**.
13. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO: **Precision and functional specificity in mRNA decay**. *Proc Natl Acad Sci USA* 2002, **99**(9):5860–65.
14. **WU-Blast2, Protein Database Query** [http://www.ebi.ac.uk/blast2/].
15. Norden C, Mendoza M, Dobbelaere J, Kotwaliwale C, Biggins S, Barral Y: **The NoCut Pathway Links Completion of Cytokinesis to Spindle Midzone Function to Prevent Chromosome Breakage**. *Cell* 2006, **125**:85–98.
16. Schuyler SC, Liu JY, Pellman D: **The molecular function of Ase1p: evidence for a MAP-dependent midzone-specific spindle matrix**. *J. Cell Biol.* 2003, **160**(4):517–528.
17. Liakopoulos D, Kusch J, Grava S, Vogel J, Barral Y: **Asymmetric Loading of Kar9 onto Spindle Poles and Microtubules Ensures Proper Spindle Alignment.** *Cell* 2003, **112**(4):561–574.
18. Fraley C, Raftery A: *mclust: Model-based cluster analysis*. Dept. of Statistics, University of Washington 2006, [http://www.stat.washington.edu/mclust]. [R package version 2.1-12].
19. Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CWV, Bussey H, et al.: **Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants**. *Science* 2001, **294**(5550):2364–2368.
20. Cottingham FR, Hoyt MA: **Mitotic Spindle Positioning in Saccharomyces cerevisiae Is Accomplished by Antagonistically Acting Microtubule Motor Proteins**. *J. Cell Biol.* 1997, **138**(5):1041–1053.
21. Fujiwara T, Tanaka K, Inoue E, Kikyo M, Takai Y: **Bni1p Regulates Microtubule-Dependent Nuclear Migration through the Actin Cytoskeleton in Saccharomyces cerevisiae**. *Mol. Cell. Biol.* 1999, **19**(12):8016–8027.
22. Juang YL, Huang J, Peters JM, McLaughlin ME, Tai CY, Pellman D: **APC-Mediated Proteolysis of Ase1 and the Morphogenesis of the Mitotic Spindle**. *Science* 1997, **275**(5304):1311–1314.

Figures

Figure 1 - Best combination of data sources.

Scatterplot for the best combination of features, which consists of *hughes.corr* and *spellman.corr*. The default separation is shown with a cutoff posterior probability of 0.5. The small group is shown in red and the big group in black.

Figure 2 - Experimental validation of genes predicted to be involved in spindle migration.

A) The *she1*Δ, *tvp38*Δ and YHR127WΔ-strains show perturbed Kar9 localization. *gpd1*Δ, which is also included in the synthetic lethality dataset, but not in the small group, is shown as a negative control. B) Some *she1*Δ cells have broken or bent anaphase spindles suggesting compromised spindle integrity.

Tables

Table 1 - Genes in the datasets known to be related to spindle migration.

For the statistical assessment (hypergeometric test) of the Gaussian Mixture Modeling results we used a reference set of genes involved in spindle migration.

Reference gene list		
Gene name	ORF	Evidence
KIP3	YGL216W	[20]
MON1	YGL124C	unpublished data
ELP6	YMR312W	unpublished data
BNI1	YNL271C	[21]
YPT7	YML001W	unpublished data
PAT1	YCR077C	unpublished data
CCZ1	YBR131W	unpublished data
UBR1	YGR184C	unpublished data
CLB4	YLR210W	[17]
ASE1	YOR058C	[22]
TVP38	YKR088C	unpublished data

For some of the genes, their involvement in spindle migration has been published. The rest is known to be involved from the unpublished results of a previous Kar9-localisation screen (see Methods). Some key regulatory genes, such as *kar9* and *bim1*, though being in the original synthetic lethality dataset, are missing in the dataset used for the analysis (see Additional Files).

Table 2 - P-values for best model

Statistical assessment of the best combination of features {hughes.corr, spellman.corr}. The p-values based on the hypergeometric test are shown for two different group sizes.

P-values for small groups		
No. of genes in small group	Known genes	P-value
5	2	0.04
50	7	0.021

Table 3 - Experimental Validation

Phenotypes of the 5 members of the small group.

Experimental validation of small group members		
Gene name	ORF name	Phenotype
ASE1	YOR058C	compromised anaphase spindles
SHE1	YBL031W	broken spindle; perturbed Kar9-asymmetry
TVP38	YKR088C	perturbed Kar9-asymmetry
UBA4	YHR111W	no direct evidence; *urm1*Δ (Uba4 target) shows perturbed Kar9-asymmetry
YHR127W	YHR127W	weakly perturbed Kar9-asymmetry

Additional Files

Additional file 1 — Synthetic lethality data

The xls-file contains the standard and systematic gene names for all synthetic lethality interactions found in the systematic screen performed with *arp1* and *jnm1*.

Additional file 2 — Data matrix used for mixture modeling

The xls-file contains the dataset that was used for mixture modeling. Due to missing data in the source datasets biological variables were only calculated for genetic interactions where complete information was available.

Additional file 3 — Movie of spindle integrity and elongation in WT cell

Avi-file showing a WT cell in anaphase. The cell is expressing CFP-Tub1, which labels the elongated anaphase spindle. 10 images were captured every 10s revealing the relatively rigid structure of the spindle.

Additional file 4 — Movie of spindle integrity and elongation in mutant cell

Avi-file showing a she1Δ cell in anaphase. The cell is expressing CFP-Tub, which labels the elongated anaphase spindle. 10 images were captured every 10s showing the spindle breaking due to loss of spindle integrity in this mutant.

Additional file 5 — Movie of Kar9-localization in WT cell

Avi-file showing WT cell in metaphase. The cell is expressing CFP-Tub1 (red) and Kar9-YFP (green). 10 images were captured every 10s showing Kar9 localizing on the SPB and astral MT on bud-directed pole only (asymmetric).

Additional file 6 — Movie of Kar9-localization in mutant cell

Avi-file showing a she1Δ cell in metaphase. The cell is expressing CFP-Tub1 (red) and Kar9-YFP (green). 10 images were captured every 10s showing Kar9 localizing on the SPB and astral MT on both sides of the spindle (symmetric).