Molecular Systems Biology 4; Article number 172; doi:10.1038/msb.2008.13
Citation: ***Molecular Systems Biology*** **4:172**

www.molecularsystemsbiology.com

molecular systems biology

REPORT

A systems-level approach to mapping the telomere length maintenance gene circuitry

Rafi Shachar[1], Lior Ungar[2], Martin Kupiec[2,*], Eytan Ruppin[1,*] and Roded Sharan[1,*]

[1] School of Computer Science, Tel Aviv University, Ramat Aviv, Israel and [2] Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Ramat Aviv, Israel

* Corresponding author. M Kupiec, Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Haim-Levanon, Ramat Aviv 69978, Israel. Tel.: 6406170; Fax: 6409028; E-mail: martin@post.tau.ac.il and E Ruppin or R Sharan, School of Computer Science, Tel Aviv University, Haim-Levanon, Ramat Aviv 69978, Israel. E-mails: ruppin@post.tau.ac.il (or) roded@post.tau.ac.il

Received 21.1.08; accepted 28.1.08

The ends of eukaryotic chromosomes are protected by telomeres, nucleoprotein structures that are essential for chromosomal stability and integrity. Understanding how telomere length is controlled has significant medical implications, especially in the fields of aging and cancer. Two recent systematic genome-wide surveys measuring the telomere length of deleted mutants in the yeast *Saccharomyces cerevisiae* have identified hundreds of telomere length maintenance (TLM) genes, which span a large array of functional categories and different localizations within the cell. This study presents a novel general method that integrates large-scale screening mutant data with protein–protein interaction information to rigorously chart the cellular subnetwork underlying the function investigated. Applying this method to the yeast telomere length control data, we identify pathways that connect the TLM proteins to the telomere-processing machinery, and predict new TLM genes and their effect on telomere length. We experimentally validate some of these predictions, demonstrating that our method is remarkably accurate. Our results both uncover the complex cellular network underlying TLM and validate a new method for inferring such networks.
Molecular Systems Biology 4 March 2008; doi:10.1038/msb.2008.13
Subject Categories: computational methods; signal transduction
Keywords: network inference; telomere

Introduction

Telomeres are specialized DNA–protein structures at the ends of eukaryotic chromosomes, whose overall length is highly regulated (Blackburn, 2000; Blackburn *et al*, 2000). Telomeres are essential for chromosomal stability and integrity, as they prevent chromosome ends from being recognized as broken molecules. Telomere stability is conferred by a protein–DNA structural cap that protects the chromosomal ends and is conserved from yeast to human cells (reviewed in de Lange, 2005). Telomeric DNA is synthesized by the enzyme telomerase, which is expressed at the early stages of development, but not in most somatic cells. Telomeres shorten with replicative age, leading eventually to cellular senescence. Replenishing telomeres by an activated telomerase is one of the few essential steps that a normal human fibroblast cell must take on its route to become malignant. Thus, understanding how telomere length is monitored has significant medical implications, especially in the fields of aging and cancer. Telomere size is maintained through a complex and delicate balance between activities that negatively and positively affect the activity of telomerase, of certain nucleases and of still to be uncovered additional mechanisms of telomere length maintenance (TLM) (Verdun and Karlseder, 2007).

Recently, two systematic genome-wide surveys measuring the telomere length of deleted mutants in the yeast *Saccharomyces cerevisiae* were conducted (Askree *et al*, 2004; Gatbonton *et al*, 2006). The two screens jointly identified 272 non-essential genes whose deletion mutants show alterations in telomere length, hence termed TLM genes (Supplementary Table I). The deletion of some of these TLM genes results in telomeric DNA sequence that is longer than the wild type (113 genes), whereas mutations in other genes

result in shorter telomere length (159 genes). The identified TLM genes span over 5% of the ~4800 deletion mutants tested in the two screens. The real number of non-essential genes involved in TLM may even be higher, given the degree of overlap between the results obtained by the two screens (Materials and methods). Taking into account the potentially large number of essential genes that could also affect telomere length, the overall number of telomere-related genes is likely to be fairly large, spanning a large variety of processes and localizations. Yet, telomere length is tightly and precisely regulated. How do so many genes interact and accurately regulate telomere homeostasis?

Previous studies aiming at addressing such questions and inferring cellular networks based on large-scale phenotypic data have mostly been targeted toward the identification of gene regulation networks from large-scale perturbation data measuring gene expression profiles (reviewed in Tegner and Bjorkegren, 2007). Other studies have identified protein–protein interaction (PPI) subnetworks where the phenotypically identified proteins were statistically over-represented (Begley *et al*, 2002; Calvano *et al*, 2005), had a specific cellular localization (Begley *et al*, 2004) or had a characteristic topological property (Said *et al*, 2004). Recently, Yeang *et al* (2004), Yeang and Vingron (2006) and Ourfali *et al* (2007) devised probabilistic models for inferring physical pathways that explain gene expression changes in response to knockout data. In contrast to these previous investigations, the novel method presented here has been developed to address the challenge of identifying a task-specific PPI subnetwork from pertaining phenotypic gene knockout data. Its end result is the first chart of the cellular subnetwork controlling telomere length.

Results

Characterizing topological and functional properties of TLM genes

We compiled a comprehensive list of 250 TLM genes (Askree *et al*, 2004; Gatbonton *et al*, 2006) for which we had PPI information (leaving out 22 genes with no such information, Supplementary Table II). We defined a small group of telomere-binding proteins, including telomerase subunits and telomerase-interacting proteins as the 'telomerase machinery' (Supplementary Table IV), serving as the end point of the various TLM-related PPI signaling pathways, which we aim to identify. We applied a series of quantitative measures to characterize various topological and functional properties of the corresponding TLM protein set, and compared them with those of random sets of the same size containing proteins encoded by either essential or non-essential genes (Materials and methods, Table I and Supplementary Table III). Like Krylov *et al* (2003), we find that the topological and functional properties of essential proteins are significantly different from those of non-essential proteins. Interestingly, the properties of the TLM proteins are distinct from those of other non-essential genes, and lie in the mid-range between those of non-essential and essential proteins (analogous to the results reported by Said *et al*, 2004). In addition, we find that proteins within known complexes tend to uniformly affect the telomere length ($P<0.001$; see Materials and methods). This result reflects the intuitive notion that complexes tend to function in a coherent manner and accordingly, that the knockout of their components would tend to lead to similar functional effects.

Constructing the telomere length-regulating network

We developed a framework for elucidating TLM pathways connecting TLM genes to telomere-binding proteins (Materials and methods). We applied our algorithm to the 250 TLM genes of the combined data set of Askree *et al* (2004) and Gatbonton *et al* (2006), supplemented by 23 TLM-related genes reported in the literature, which were not identified by either screen (Supplementary Table I). The algorithm reconstructed pathways for 180 TLM proteins inducing a telomere length-regulating subnetwork (TRS) with 327 proteins (Supplementary Figure I and Supplementary Table V). On the TRS network, 54 of the 180 TLM proteins lie in between other TLM proteins and the telomere-binding proteins; the other 139 non-TLM proteins were required for connecting the TLM proteins to the telomere-binding proteins. In total, 71 of the non-TLM proteins were non-essential and 68 were essential. We validated the reconstructed pathways by computing their functional coherency according to the gene ontology (GO) biological process annotation. The pathways were found to be significantly coherent ($P<4.5e{-}3$; see Materials and methods).

Functional characterization of TRS proteins

Mutations in TLM genes may have various effects on telomere length. They can lead to slight, moderate or large alterations (short or long) (Askree *et al*, 2004; Gatbonton *et al*, 2006). We used these fine phenotypic descriptions to test the extent to which TLM proteins that occupy internal nodes on the TRS affect telomere length. We observed that these proteins have greater effect on telomere length than TLM proteins occupying end nodes (hypergeometric $P<3.5e{-}4$; Supplementary Table VI). One explanation for the observed relation between a protein's location on the TRS and its functional TLM effect may be due to its distance in the PPI network from telomere-binding proteins. To test this hypothesis, we computed a partial correlation index that factors out the distance (Materials and methods). Remarkably, the results show that internal TLM proteins affect telomere length more than other TLM proteins independently of their distance from the telomere-binding proteins ($P<1e{-}4$).

What hence further determines the functional effects of TLM proteins on telomere length? Interestingly, we find that the likelihood scores of the pathways connecting a TLM protein to the TLM end targets (Materials and methods) correlate with the magnitude with which the knockout of TLM proteins affects telomere length (Spearman correlation of 0.16, $P<0.04$). These results suggest that TLM proteins that have larger effect on telomere length do play a more central role in

Table I Topological and functional characteristics of the TLM proteins

Features	Mean value (s.d.)[a]		
	Essential proteins	Non-essential proteins	TLM proteins
Global network characteristic[b]			
Degree	**35 (27)**	**11.8 (29.3)**	18.6 (23.7)
Expected degree	**18.5 (29)**	**7 (27)**	11.2 (16.2)
Compactness	**301 (5.4)**	**359 (7.7)**	322
Network-based characteristics relating proteins to the telomerase machinery set[c]			
Path length	**2.25 (0.6)**	**2.57 (0.6)**	2.4 (0.6)
Path probability	**0.87 (0.25)**	**0.65 (0.35)**	0.77 (0.3)
Betweenness centrality	36.5 (4.9)	**9.9 (3)**	30
Functional characteristics[d]			
Complex-based monochromaticity			81 %[e]
Propensity for gene loss (PGL)	**0.08 (0.1)**	**0.15 (0.12)**	0.12 (0.11)

[a]Boldfaced values indicate that the reported value is significantly different compared to the respective value for the TLM protein set. For the complete table, including *P*-values see Supplementary Table III.
[b]*Degree* and *expected degree* measure protein node degrees, either unweighted or weighted by the reliabilities of the incident edges. *Compactness* measures the size of the minimal connected component, which includes a given protein set (Materials and methods).
[c]*Path length* and *path probability* are the minimal edge distance and the probability of the most reliable path between a source protein and the target telomere-binding proteins, respectively. Given a source protein set and a target protein set, the *betweenness centrality* measures the relative number of pathways from the source proteins to the target proteins that pass through a given node (Materials and methods).
[d]*Monochromaticity* measures the coherency in which protein complex members affect telomere length (Materials and methods). *PGL* (Krylov *et al*, 2003) measures the propensity for gene loss of a given protein (evolutionary conservation of a protein).
[e]The mean monochromaticity rate for randomized coloring permutations was 0.29 (s.d. = 0.1), $P < 0.001$.

TRS connectivity, and are connected to the telomerase machinery set by pathways with higher likelihood scores.

Experimental testing of predicted TLM proteins

The TRS includes, in addition to TLM proteins, non-essential proteins that are not known to be in the TLM set (NTLM proteins) as well as essential proteins. While the genome-wide screens were comprehensive, it is possible that NTLM proteins *do* affect TLM, but were not included among the TLMs either because they were absent from the deletion collection, gave inconclusive results or affected telomere length in a subtle way that was difficult to observe. We hence re-evaluated the telomere length screens of Askree *et al* (2004) for 20 strains deleted for NTLM genes (see Materials and methods). In total, 14 out of 20 mutants deleted for NTLM genes exhibited defects in telomere length (9 were short and 5 exhibited elongated telomeres). On the basis of the TRS, for 11 of these mutants we could predict the expected phenotype (Materials and methods). In 8 out of the 11 cases, the observed telomere length matched the prediction. In two additional cases, no telomere length defect was observed (*sml1*Δ and *tor1*Δ), and in one case (*vrp1*Δ), long telomeres were seen instead of the predicted short phenotype (Table II).

A similar analysis was carried out for 12 temperature-sensitive mutants of essential genes within the TRS, which were not included in the original screens (see Materials and methods). All mutants were grown at 30°C, a semi-permissive temperature, and telomere length was measured (Table II and Figure 1). In total, 8 out of the 12 strains analyzed exhibited telomere length defects. In six out of seven cases in which a prediction could be made regarding telomere length, the expected phenotype was observed. The only exception was *mak21-1*, which exhibited very short telomeres instead of the expected long telomere phenotype (Figure 1).

Table II Experimental testing of predicted TLM proteins

Mutant allele	Observed telomere phenotype	Predicted telomere phenotype[a]
Non-essential genes		
*vps41*Δ	Short	Short
*pcl6*Δ	Short	Short
*vps53*Δ	Short	Short
*sif2*Δ	Short	Short
*vps20*Δ	Short	Short
*vps27*Δ	Short	Short
*vps21*Δ	Short	Short
*fzo1*Δ	Long	Long
*pub1*Δ	Long	NA
*mrpl3*Δ	Short	NA
*pir1*Δ	Long	NA
*tfp1*Δ	Short	NA
*tif2*Δ	Long	NA
*vrp1*Δ	Long	Short
*msh2*Δ	Wild type	NA
*hpr5/srs2*Δ	Wild type	NA
*sgs1*Δ	Wild type	NA
*sml1*Δ	Wild type	Short
*slt2*Δ	Wild type	NA
*tor1*Δ	Wild type	Short
Essential genes		
rfa1-t11	Short	Short
cdc45-27 and *cdc45-1*	Short	Short
pol30-59	Long	NA
mak21-1	Short	Long
cdc33-1	Wild type	NA
smc3-1	Wild type	NA
rad3-2	Wild type	NA
cdc48-1	Wild type	NA
pre7-1	Short	Short
kae1-1	Short	Short
kre33-1	Short	Short
rpc10-1	Short	Short

[a]NA, not applicable.

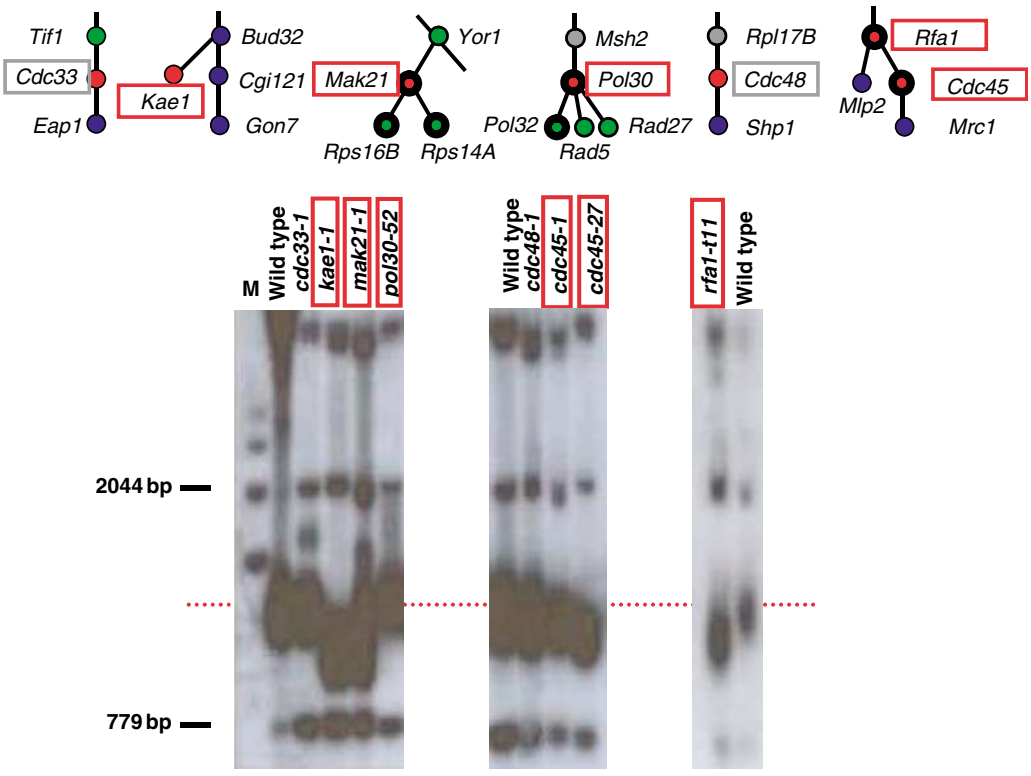

Figure 1 Telomere Southern blot of mutants in NTLM and essential proteins in the TRS. Green circles: mutants with elongated telomeres; blue circles: mutants with short telomeres; gray circles: NTLM mutants; red circles: mutants of essential genes. DNA was digested with *Xho*I and probed with telomeric sequences and with unique genomic sequences used as markers (Askree *et al*, 2004). A red line marks the telomere size of the wild-type strain. Mutants exhibiting expected phenotypes are marked with a red square, those that do not show telomere phenotype are marked with a gray square.

Biological significance of the TRS network

Our approach to reconstruction of the TRS was based on the assumption that if a protein affects telomere length then it must be connected by some pathway to the telomerase machinery. The pathways uncovered are signaling-regulatory pathways that may contain protein complexes. Expectedly, the TRS network shows branches that are enriched both for specific biological processes and for specific complexes. Deletion of genes in these branches results in uniform telomere length effects. For example, mutations in any of the genes composing the vacuolar transport and in the degradation of membrane proteins (ESCRT complexes) result in short telomeres (Rog *et al*, 2005). The yeast vacuole is the functional analog of the mammalian lysosome, the major site of degradation of both exogenous and endogenous macromolecules. This branch co-locates with two other branches, one containing the COMPASS (Set1C) complex, which methylates histone H3 on lysine 4 and is required for transcriptional silencing near telomeres (Krogan *et al*, 2002) and another one composed of the Bre1 and Rad6 proteins, which ubiquitinate histones H2B (Wood *et al*, 2003). Interestingly, the activity by Rad6/Bre1 on histone H2B depends on the prior activity of COMPASS on histone H3 (Dover *et al*, 2002) (Figure 2A). The convergence of these three pathways by their interaction with histones, and the uniformity of the phenotype, suggests that the vacuolar pathway affects telomere length through histone modification. It is possible that the activity of one or more histone-modifying enzymes is regulated by degradation or modification in the vacuole.

However, not all branches affect telomere length uniformly. For example, components of three related RNA polymerase regulators, the mediator, the Paf1 complex and the Tho complex, cluster in the TRS; however, they affect telomere length in different ways: whereas mutations in components of the Paf1 complex lead to short telomeres, mutations in some components of the RNA polymerase holoenzyme or mediator produce mostly elongated telomeres (with the exception of *srb5* mutations that lead to short telomeres). Similarly, most proteins in the THO complex shorten telomeres, but *hpr1* mutants show longer than normal telomeres (Figure 2B). These results suggest that these large complexes do not act as a single monolithic unit, and that deletion of different regions may change regulation in subtle, and still unclear, ways.

The TRS model may also predict the interaction of seemingly unrelated proteins. During the course of this work, it was found that the Bud32, Cgi121, Gon7 and Kae1 form a complex, which acts on transcription and affects telomere length homeostasis (Downey *et al*, 2006; Kisseleva-Romanova *et al*, 2006). The genes encoding three of the four proteins in this complex were identified as TLM proteins (Figure 2C); the fourth (Kae1) is encoded by an essential gene. We have validated that mutations in this gene also cause telomere shortening (Figure 1 and Table II). The KEOPS complex in the TRS is linked to Est1, a protein known to bind the RNA moiety of telomerase, and may help activate it via Kre33, a protein of

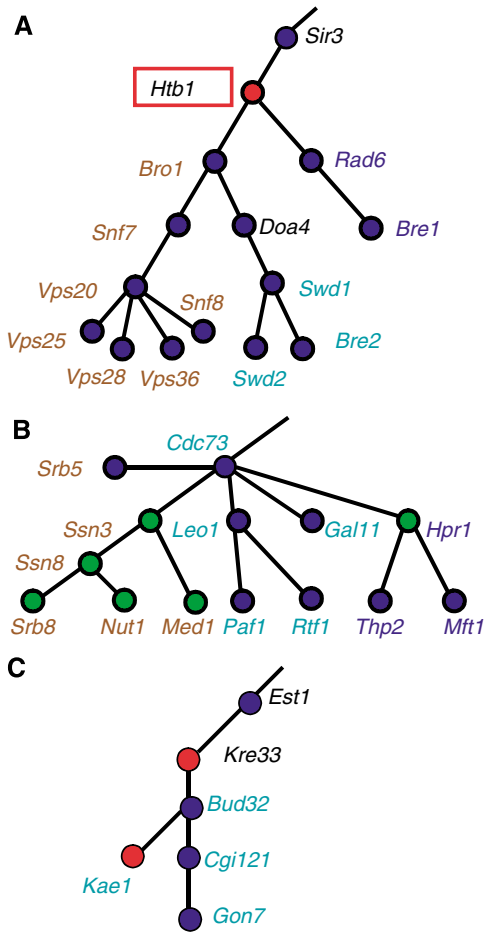

Figure 2 Pathways predicted by the TRS model. Green circles denote mutants with elongated telomeres; blue circles denote mutants with short telomeres and red circles denote mutants of essential genes. (**A**) Proteins belonging to the vacuolar transport pathway (brown letters), the COMPASS complex (turquoise letters) and the Rad6/Bre1 complex (blue letters) converge at the histone H2A node. Mutations in any of the genes encoding these proteins lead to short telomeres. (**B**) Proteins belonging to the Pol II holoenzyme and the mediator complex (brown letters), to the Paf1 complex (turquoise letters) and to the THO complex (blue letters) interact but mutations in these units lead to different phenotypes. (**C**) The KEOPS complex (turquoise letters) is connected to the telomere machinery *via* Kre33.

unknown function. Our model predicts that Kre33 may play a role in mediating the access of KEOPS to the telomere machinery (Figure 2C).

This study presents a reconstruction of the TRS network from the combined data sets of two knockout genome-wide assays of TLM in the yeast, utilizing the pertaining PPI data. To this end, we have developed a new probabilistic approach to identify cellular signaling pathways based on large-scale phenotypic data. Connecting TLM proteins with the telomerase machinery proteins has required the inclusion of intermediate proteins with no experimentally observed telomere length phenotypes in the TRS. The pathways identified were then further investigated and validated in both a large-scale computational manner and in a smaller scale experimental manner. The pathways identified were shown to be functionally coherent. Remarkably, we found that the likelihood of a pathway correlates with the magnitude of its effect on telomere length. Moreover, TLM proteins that have a larger effect on telomere length play a more central role in connecting TLM proteins to the target telomere-binding proteins. Finally, the method for phenotypic data analysis and network identification presented here is general and is likely to have many applications in the identification of protein networks underlying numerous cellular functions, and in more complex organisms.

Materials and methods

Data acquisition

We collected and integrated PPI data from DIP (Salwinski *et al*, 2004; April 2005 download) and from two recent assays (Gavin *et al*, 2006; Krogan *et al*, 2006). The combined data set comprises over 39 936 interactions involving 5414 proteins. To assign confidence scores to these interactions, we used the logistic-regression-based scheme employed in Sharan *et al* (2005b). Briefly, true-positive and true-negative interactions were used to train a logistic regression model, which assigns each interaction a reliability score based on the experimental evidence for this interaction, including the type of experiments in which the interaction was observed, and the number of observations in each experimental type.

The TLM gene set was created by merging two genome-wide screens measuring the effect of deletion mutants on telomere length (Askree *et al*, 2004; Gatbonton *et al*, 2006). The unified set contained 272 non-essential telomere length-affecting genes. For the purpose of pathway identification, we complemented the combined data set with 23 genes from the literature reaching a total of 295 genes. In total, 6 of the additional 23 genes were essential genes. Of the 295 genes in the merged data set, 273 encode proteins from the PPI network.

Estimating the total number of TLM genes

Assuming that the two screens are independent and that false-negative results occur at random, the expected number of TLM proteins was estimated as follows: denote by n_A and n_B the number of genes that were identified by screens A and B, respectively, and by n_{AB} the number of genes identified by both A and B. Let n denote the unknown number of TLM proteins. Then n can be derived from the equation $n_{AB}/n = n_A/n \cdot n_B/n$. In our case, 171 genes were identified by Askree *et al* (2004), 152 by Gatbonton *et al* (2006) and 51 genes were identified in both screens. These numbers yield an estimated number of 510 TLM genes, implying that indeed, many TLM genes are yet to be discovered.

Network measures

Computation of shortest paths was conducted using the breadth first search algorithm. Most probable paths were computed using the classical Dijkstra algorithm, where interaction probabilities were transformed to their absolute log value.

The betweenness centrality index quantifies the importance of a vertex in a graph to the connectivity of other vertices (Freeman, 1977; Brandes, 2001). Denote by $\sigma(u)$ the number of shortest paths connecting TLM protein u and the telomere-binding proteins. Denote by $\sigma(u,n)$ the number of shortest paths passing through n and connecting TLM protein u with the telomere-binding proteins. The relative betweenness of protein n with respect to the set T of TLM proteins is

$$C_r(n) = \sum_{u \in T} \frac{\sigma(u,n)}{\sigma(u)}$$

The betweeness centrality of the entire TLM set was computed as the number of TLM proteins with non-zero betweeness centrality score.

We defined the compactness of a protein set as the minimal number of additional protein nodes that are required to make it connected (i.e. admit a path between every two proteins in the set). The compactness

computation gives rise to a node-weighted Steiner tree problem, which can be efficiently approximated (Klein and Ravi, 1995).

Statistical significance testing

P-values were computed empirically based on 10 000 randomized protein sets containing equal number of proteins as the TLM protein set. The randomized sets were uniformly selected from the set of proteins in the PPI network.

Monochromaticity testing

We 'colored' proteins according to the effects that their respective mutants exhibit on telomere length (short or long). We defined the monochromaticity rate of a collection of known protein complexes from the MIPS database (Mewes *et al*, 2006) as the fraction of complexes with at least two TLM proteins that are colored uniformly. Statistical significance was computed by comparing the observed rate to the rate obtained in 10 000 randomized colorings of the complex members, preserving the number of 'long' and 'short' proteins.

Pathway identification

We filtered the original PPI network to remove low confidence interactions (with probability <0.5) and nonspecific protein interactors (with more than 100 interactions), yielding a network with 5324 vertices and 12 521 edges.

Our algorithm enumerates all paths in the PPI network that are at most five edges long, originating at telomere-binding proteins and ending at the identified TLM proteins. Each discovered path is assigned a score. Finally, the algorithm returns for each TLM protein the highest scoring path; these are then were combined to form a network.

To simplify pathways visualization and analysis, we reduced the resulting network to a tree rooted at the telomere-binding proteins. This was carried out by computing the most probable paths in the reduced network from telomere-binding proteins to each TLM protein using the Dijkstra algorithm, and merging the obtained paths.

A probabilistic model for TLM pathways

Following the approach developed by Sharan *et al* (2005a), we computed the likelihood of a path to connect a TLM protein and a telomere-binding protein. The likelihood score of a path, $L(p)$, has two components: an edge-based score and a vertex-based score, which we describe below. To normalize for the length of a path, we multiplied the score by a penalty factor, favoring short paths over long ones. The final score of a path of length l was: $W(p)=L(p)\cdot e^{-cl}$, where c is a free parameter.

The edge-based score of a path measures the fit of a path to a path model versus the likelihood that it arises at random. The path model assumes that each pair of consecutive proteins along the path interacts. The null model assumes that the PPI network was randomly chosen from the collection of all networks with the same degree sequence. This induces a probability that a pair of vertices interacts, which depends on their degrees. The likelihood ratio is computed as in Sharan *et al* 2005a), taking into account the reliability of each interaction.

The vertex-based score is again a likelihood ratio score. The path model, M_p, assumes that each vertex on the path corresponds to a TLM protein. The null model, M_n, assumes that each protein on the path, except the two ends, is a random selection from the network's vertices. Denote by T_v the event that protein v is a TLM protein and by F_v that it is not. Denote by O_v the experimental evidence on the telomeric phenotype of protein v. For a path $p=(p_1,\ldots,p_k)$, using the law of complete probability we get

$$P(Op_1,\ldots,Op_k|M_p)=\prod_{p_i\in p}P(Op_i|M_p)=\prod_{p_i\in p}P(Op_i|Tp_i)$$

Similarly, we derive the probability of observing p according to the null model

$$P(Op_1,\ldots,Op_k|M_n)=\prod_{p_i\in p}P(Op_i|M_n)=\prod_{p_i\in p}P(Op_i)$$

The vertex-based score of p is thus

$$\mathrm{LV}(p)=\sum_{p_i\in p}\log\frac{P(Op_i|Tp_i)}{P(Op_i)}=\sum_{p_i\in p}\log\frac{P(Tp_i|Op_i)}{P(Tp_i)}$$

where the last equality follows from Bayes theorem.

It remains to estimate the probabilities $P(T_v)$ and $P(T_v|O_v)$. Using the available biological experimental observations, we estimated the probability that protein v is a TLM: 1 if it was identified in a small-scale study. On the basis of estimation of the false detection rate (Askree *et al*, 2004), this probability was set to 0.9 for proteins that were detected in a single screen. Following the assumption that the two genome-wide screens are independent, this probability was set to 0.99 for proteins that were identified in both genome-wide screens. Given the overlap between the two screens, we estimated the probability that an NTLM protein was falsely classified by the two screens as 0.05. Using this last result and given the frequency of essential proteins identified in the pertaining small-scale studies, we estimated the probability that an essential protein is a TLM as 0.13. Finally, based on these probability estimations we could estimate the prior probability $P(T_v)$.

Setting the length penalty parameter

To select a value for the length penalty parameter c, we performed a search over a range of non-negative real numbers. For each value of c, we executed the pathway reconstruction algorithm twice: once applied to the TLM proteins identified by Askree *et al* (2004) and a second time applied to the TLM proteins identified by Gatbonton *et al* (2006). Next, we computed the hypergeometric enrichment of the TLM proteins from one data set within the set of vertices on pathways inferred from the other data set. We chose the value c that minimized the product of the two enrichment *P*-values.

GO enrichment analysis

Functional coherency values of pathways were computed as in Shlomi *et al* (2006) based on GO process annotations for their proteins. These values were compared to those attained by random pathways of the same length. Random paths were generated such that their one end was a randomly selected telomere-binding protein and they had an equal path length distribution as the identified pathways. To determine if the annotation enrichment distributions of two path sets were significantly different, we used the non-parametric Wilcoxon test.

Spearman partial rank correlation coefficient

The Spearman partial rank correlation coefficient between two random variables A and X, given the fact that both A and X are correlated to random variable Y, denotes the correlation between A and X, when Y is kept constant. It is calculated as follows

$$r_{AX,Y}=\frac{r_{AX}-r_{XY}r_{AY}}{\sqrt{(1-r_{XY}^2)(1-r_{AY}^2)}}$$

Here, r_{AX}, r_{XY} and r_{AY} represent the Spearman correlation coefficients between A and X, X and Y, and A and Y, respectively. The significance level is given by

$$D_{AX,Y}=\frac{1}{2}\sqrt{N-4}\ln\left(\frac{1+r_{AX,Y}}{1-r_{AX,Y}}\right)$$

$D_{AX,Y}$ has a normal distribution with zero mean and variance one. N represents the size of the data set.

Telomere length measurement

Telomeric Southern blots were used to measure telomere length as described in Askree *et al* (2004). PCR fragments containing telomeric sequences and a genomic region that hybridizes to two bands (2044 and 779 bp) were used as probes.

Telomere length effect prediction and validation

The telomere length effect of a protein was predicted to be short or long if the telomere length phenotype of the proteins immediately upstream to it was the same as for those downstream to it, in the pertaining TRS branch on which the protein lies. Otherwise, a prediction based on the monochromatic assumption could not have been made and it was classified as not applicable.

The list of NTLM and essential genes validated was chosen in a two-step process: first, we exhaustively identified all non-essential, non-TLM genes and all essential genes in the TRS. In the second step, we narrowed down this list to include (a) all the non-essentials for which we had already carried out a screening test previously and for which DNA was available, 20 such cases overall and (b) the essential genes for which we had temperature-sensitive strains in our lab collection and could be further tested (12 overall).

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

MK's research was supported by grants from the Israel Science Foundation (FIRST and Convergent Technologies) and the Israel Cancer Research Fund. ER and RS were supported by a France–Israel grant from the Israeli Ministry of Science and Technology and by the Converging Technologies Program of the Israel Science Foundation (grant no. 1748/07).

References

Askree SH, Yehuda T, Smolikov S, Gurevich R, Hawk J, Coker C, Krauskopf A, Kupiec M, McEachern MJ (2004) A genome-wide screen for *Saccharomyces cerevisiae* deletion mutants that affect telomere length. *Proc Natl Acad Sci USA* **101:** 8658–8663

Begley TJ, Rosenbach AS, Ideker T, Samson LD (2002) Damage recovery pathways in *Saccharomyces cerevisiae* revealed by genomic phenotyping and interactome mapping. *Mol Cancer Res* **1:** 103–112

Begley TJ, Rosenbach AS, Ideker T, Samson LD (2004) Hot spots for modulating toxicity identified by genomic phenotyping and localization mapping. *Mol Cell* **16:** 117–125

Blackburn EH (2000) Telomere states and cell fates. *Nature* **408:** 53–56

Blackburn EH, Chan S, Chang J, Fulton TB, Krauskopf A, McEachern M, Prescott J, Roy J, Smith C, Wang H (2000) Molecular manifestations and molecular determinants of telomere capping. *Cold Spring Harb Symp Quant Biol* **65:** 253–263

Brandes U (2001) A faster algorithm for betweenness centrality. *J Math Sociol* **25:** 163–177

Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, Miller-Graziano C, Moldawer LL, Mindrinos MN, Davis RW, Tompkins RG, Lowry SF (2005) A network-based analysis of systemic inflammation in humans. *Nature* **437:** 1032–1037

de Lange T (2005) Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev* **19:** 2100–2110

Dover J, Schneider J, Tawiah-Boateng MA, Wood A, Dean K, Johnston M, Shilatifard A (2002) Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6. *J Biol Chem* **277:** 28368–28371

Downey M, Houlsworth R, Maringele L, Rollie A, Brehme M, Galicia S, Guillard S, Partington M, Zubko MK, Krogan NJ, Emili A, Greenblatt JF, Harrington L, Lydall D, Durocher D (2006) A genome-wide screen identifies the evolutionarily conserved KEOPS complex as a telomere regulator. *Cell* **124:** 1155–1168

Freeman L (1977) A set of measures of centrality based on betweenness. *Sociometry* **40:** 35–41

Gatbonton T, Imbesi M, Nelson M, Akey JM, Ruderfer DM, Kruglyak L, Simon JA, Bedalov A (2006) Telomere length as a quantitative trait: genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS Genet* **2:** e35

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M *et al* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440:** 631–636

Kisseleva-Romanova E, Lopreiato R, Baudin-Baillieu A, Rousselle JC, Ilan L, Hofmann K, Namane A, Mann C, Libri D (2006) Yeast homolog of a cancer-testis antigen defines a new transcription complex. *EMBO J* **25:** 3576–3585

Klein P, Ravi R (1995) A nearly best-possible approximation algorithm for node-weighted Steiner trees. *J Algorithms* **19:** 104–115

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B *et al* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440:** 637–643

Krogan NJ, Dover J, Khorrami S, Greenblatt JF, Schneider J, Johnston M, Shilatifard A (2002) COMPASS, a histone H3 (lysine 4) methyltransferase required for telomeric silencing of gene expression. *J Biol Chem* **277:** 10753–10755

Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* **13:** 2229–2235

Mewes HW, Frishman D, Mayer KF, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res* **34:** D169–D172

Ourfali O, Shlomi T, Ideker T, Ruppin E, Sharan R (2007) SPINE: a framework for signaling-regulatory pathway inference from cause–effect experiments. *Bioinformatics* **23:** i359–i366

Rog O, Smolikov S, Krauskopf A, Kupiec M (2005) The yeast VPS genes affect telomere length regulation. *Curr Genet* **47:** 18–28

Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD (2004) Global network analysis of phenotypic effects: protein networks and toxicity modulation in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **101:** 18006–18011

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res* **32:** D449–D451

Sharan R, Ideker T, Kelley B, Shamir R, Karp RM (2005a) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J Comput Biol* **12:** 835–846

Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005b) Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA* **102:** 1974–1979

Shlomi T, Segal D, Ruppin E, Sharan R (2006) QPath: a method for querying pathways in a protein–protein interaction network. *BMC Bioinformatics* **7:** 199

Tegner J, Bjorkegren J (2007) Perturbations to uncover gene networks. *Trends Genet* **23:** 34–41

Verdun RE, Karlseder J (2007) Replication and protection of telomeres. *Nature* **447:** 924–931

Wood A, Krogan NJ, Dover J, Schneider J, Heidt J, Boateng MA, Dean K, Golshani A, Zhang Y, Greenblatt JF, Johnston M, Shilatifard A (2003) Bre1, an E3 ubiquitin ligase required for recruitment and substrate selection of Rad6 at a promoter. *Mol Cell* **11:** 267–274

Yeang CH, Ideker T, Jaakkola T (2004) Physical network models. *J Comput Biol* **11:** 243–262

Yeang CH, Vingron M (2006) A joint model of regulatory and metabolic networks. *BMC Bioinformatics* **7:** 332