Supplemental Materials

Molecular Biology of the Cell

Hériché et al.

Supplementary information

Comparing different kernels on graph nodes for building process-specific siRNA libraries

(a) Global performance on single data sources

To test graph-based similarity measures, we represented biological information from various sources as undirected weighted graphs and used the corresponding adjacency matrices to compute the kernels. As a first test of the ability to predict new functional relationships between genes, we compared the performance of the different kernels to retrieve known functional relationships from the well documented Panther pathways database (Mi et al, 2005) . We mined six sources of data - protein interactions (PI), homology-inferred protein interactions (HIPPO), Gene Ontology biological process (BP), text mining (TM), a gene expression network from aggregation of many gene expression data sets (MEMP) and ab-initio predicted protein interactions from co-occurring domain architectures (CODA) – and examined which kernels gave the best performance for each data source. We focused on three kernels as different ways of measuring similarity between graph nodes taking into account indirect connections between nodes:

- the von Neumann diffusion kernel (VN), which counts all possible paths joining two nodes with a free parameter that sets penalties for longer paths,

- the commute-time kernel (CT) so called because it is derived from the computation of the number of steps needed to go from one node to another and back to the starting node in a random walk,

- the random forest kernel (RF) which can be interpreted in terms of probabilities of reaching a node in a random walk with a random number of steps.

Using all Panther pathway genes one at a time as a query to find similar genes, we measured kernel sensitivity by the fraction of other genes from the same pathway that are ranked in the top of the returned list for thresholds up to 10% of all ranked genes. To estimate false positive rates, we generated pathways composed of random genes and processed them in the same way with the assumption that randomly picked genes are unlikely to be functionally related. To assess overall performance, we first plotted sensitivity versus estimated false positive rate and used the area under the curve (AUC) up to 25% false positive rate as a measure of performance. We found that the commute time (CT) kernel gave the best AUC value for all tested data sources (Figure 2A). The RF kernel performance varied between data types with performance similar to the commute time kernel on protein interactions and lower for other data sets. In many cases, the Von Neumann diffusion kernel with a small value of its parameter (VN_{low}) also performed well while the worst performance was obtained with the original matrix (A) followed by the Von Neumann diffusion kernel at the upper limit of its parameter range (VN_{max}). Of note, the VN_{max} curve was close to the curve obtained with the degree-based similarity matrix (DB), which suggests that for high values of its parameter, the Von Neumann diffusion kernel ranks genes based on their number of connections. While our results highlight the fact that not all kernels adequately capture functional relationships between genes across all data types, they suggest that the commute time kernel is a robust and, since parameter free, easy to use way of measuring functional similarity between genes.

Of note, we found that all kernels tested performed poorly on the aggregated gene expression data suggesting that either these kernels are not a good representation of this data or that this data set doesn't contain significant information on functional relationships between genes.

(b) Evaluating kernel performance on a preset number of genes

The AUC does not provide information on the percentage of genes necessary to obtain a given sensitivity value. However, to use gene predictions to guide experiments like RNAi screening, it is pragmatic to be able to evaluate kernel performance by the number of true positives in a preset number of ranked genes, as the experimental capacity is limited to screen for example 100 genes and it is often acceptable to trade some false positives for a significant gain in true positives. A convenient way to do this is to look at the number of false positives and true positives of different kernels as a function of ranked gene number up to the top 10% of all ranked genes. This lets the experimentalist appreciate whether screening more genes would produce substantial gains in true positives. We illustrate this for the protein interaction data set (Supplementary Figure 1A, other data sets shown in Supplementary Figure 1). For this particular data set, using the random forest (RF) as the best kernel on this data, screening the top 5% of all ranked genes (e.g. about 600 genes) would already return, on average across all pathways, 63% of all true positives, while performing twice the number of experiments and screening the top 10% (e.g. 1200 genes) would increase this fraction to only 71% of all true positives.

Combining the best kernels from several data sources improves function retrieval

Several schemes to integrate multiple data sources using kernels have been described by other groups. They involve learning weights of a linear (Lanckriet et al, 2004, De Bie et al, 2007) or even non-linear (Diosan et al, 2008) combination of kernels or weights of a linear combination of graph adjacency matrices (Mostafavi et al, 2008; Mostafavi & Morris, 2010). So to keep the method simple and generic, we chose not to learn weights but instead use the straightforward approach of summing the kernels. Since the kernels have been normalized, each kernel has the same importance in the combination. Choosing to include or exclude kernels (or graphs) from the combination can be seen as a binary form of weighting, with the end user deciding which data set is relevant. For example, we chose to exclude the MEMP (aggregation of gene expression data) and CODA (predicted interactions) data sets because the tests reported above showed that, at least with the kernels we used, these data sources did not capture known functional relationships.

We compared different ways of integrating the data. Because we converted each data set into a graph, one way of integrating data is to combine all graphs into a single graph where two genes are connected if there is an edge between them in at least one of the source graphs (combined binary graph) and then applying the commute-time kernel to this graph. Another approach is to normalize the edge weights in each graph and form the combined graph by setting its edge weights as an average of the corresponding edges in the source graphs followed by computation of the commute-time kernel or random forest kernel as in the GeneMANIA algorithm. The integration methods using combined weighted graphs as inputs all give similar results and are better than combining data in an unweighted graph (Supplementary Figure 2A). However, instead of combining the graphs, combining the best kernel for each data set (e.g. RF for protein interactions and CT for all other data) clearly improves performance over the best single data source, protein interactions (Supplementary Figure 2B). With the combined kernels, screening the top 5% of the genes would now return 77% of true positives, compared to 63% if using RF on protein interactions (PI) alone.

Chromosome condensation phenotypes of S. pombe mutants

When pmt1, the only DNA methyltransferase found in the S. pombe genome and ortholog of DNMT3B, is deleted, the distance between the marked loci starts to decrease earlier than in wild-type cells and does so at a slower pace over a longer time (Figure 6B, duration and timing). This slower, premature chromosome condensation is reminiscent of the DNMT3B knock-down phenotype in HeLa cells where condensation takes longer than in control cells (Supplementary Figure 5A).

Cid14 is the S. pombe ortholog of PAPD5. Its deletion leads to a strong reduction in the number of cells entering mitosis. However, the few cells that enter mitosis also do so with premature chromosome condensation as indicated by the increased duration and timing of compaction (Figure 6B), similar to the lengthened prophase observed in HeLa cells for the PAPD5 knock-down (Supplementary Figure 5B).

In a top1 deletion mutant, condensation takes slightly less time, starting from a more condensed state and reaching a less condensed state than in wild-type cells (Figure 6B). Duration of condensation is also shorter than in wild-type cells which is consistent with the knock-down of TOP1 in HeLa cells leading to a shorter prophase (Supplementary Figure 5C).

Clr6 is the S. pombe ortholog of HDAC1. Because a clr6 deletion turned out to be inviable, we used a temperature-sensitive mutant. At the restrictive temperature, this mutant enters mitosis with significantly more compacted chromosomes and reaches a compaction level similar to wild-type slightly faster than wild-type cells (Figure 6B), consistent with the shorter prophase observed when HDAC1 function is reduced in HeLa cells (Supplementary Figure 5D).

References

De Bie T, Tranchevent LC, van Oeffelen LM, Moreau Y (2007) Kernel-based data fusion for gene prioritization. *Bioinformatics* **23**: i125-132.

Diosan L, Rogozan A, Pécuchet JP (2008) Optimizing multiple kernels for SVM by Genetic Programming. In EvoCOP, van Hemert J & Cotta C (eds.) pp 230-2241. Heidelberg: Springer.

Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble W (2004) A statistical framework for genomic data fusion. *Bioinformatics* **20**: 2626-2635.

Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* **33**: D284-8.

Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* **9** Suppl 1: S4.

Mostafavi S, Morris Q (2010) Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics* **26**: 1759-1765.

Supplementary figure legends

Supplementary figure 1: False positive and sensitivity curves as a function of rank threshold for the data sources tested.

Solid lines represent sensitivities and dashed lines represent estimated false positive rates. Each colour represents a different kernel. A: protein interactions, B: protein interactions from other species mapped to human, C: gene co-expression network from the MEM aggregation tool, D: CODA-predicted interactions, E: iHOP-generated interactions, F: semantic similarities across GO biological processes

Supplementary figure 2: Performance of different data integration schemes.

A- Comparison of different integration schemes

B- Comparison of the best integration scheme with the best single data source (protein interactions)

Solid lines represent sensitivities and dashed lines represent false positive rates. Each colour represents a different kernel or graph combination.

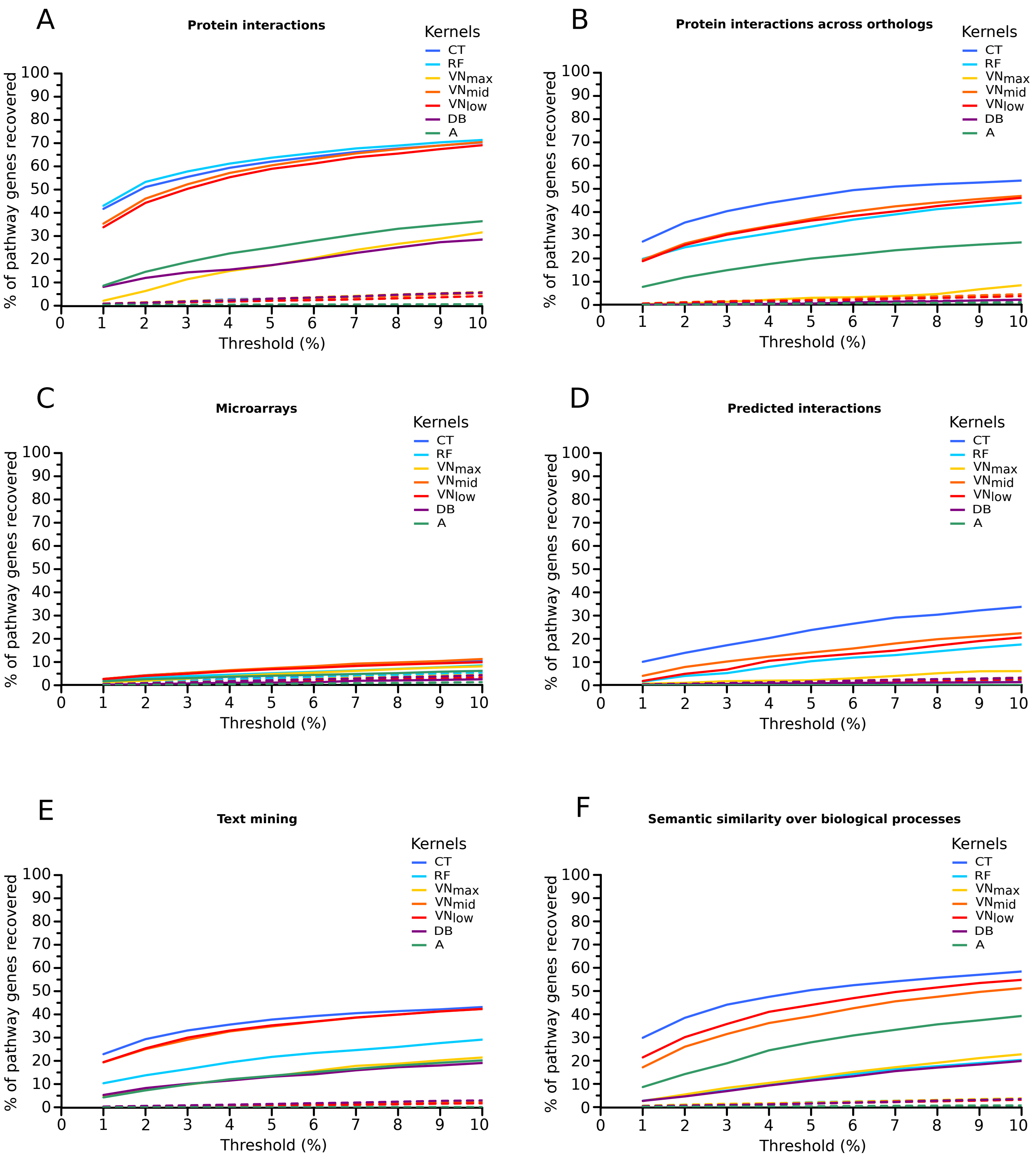
Supplementary figure 3: Score distribution of the top 1000 genes predicted to be involved in chromosome condensation.

Supplementary Figure 4: Curves from representative cells showing strong phenotypes.
NCAPD3 and TRAF3IP1 knockdowns show transient chromatin decondensation at the time of nuclear envelope breakdown (arrow) while MCPH1 and DNMT3B knockdowns show slower condensation starting earlier than in control cells.

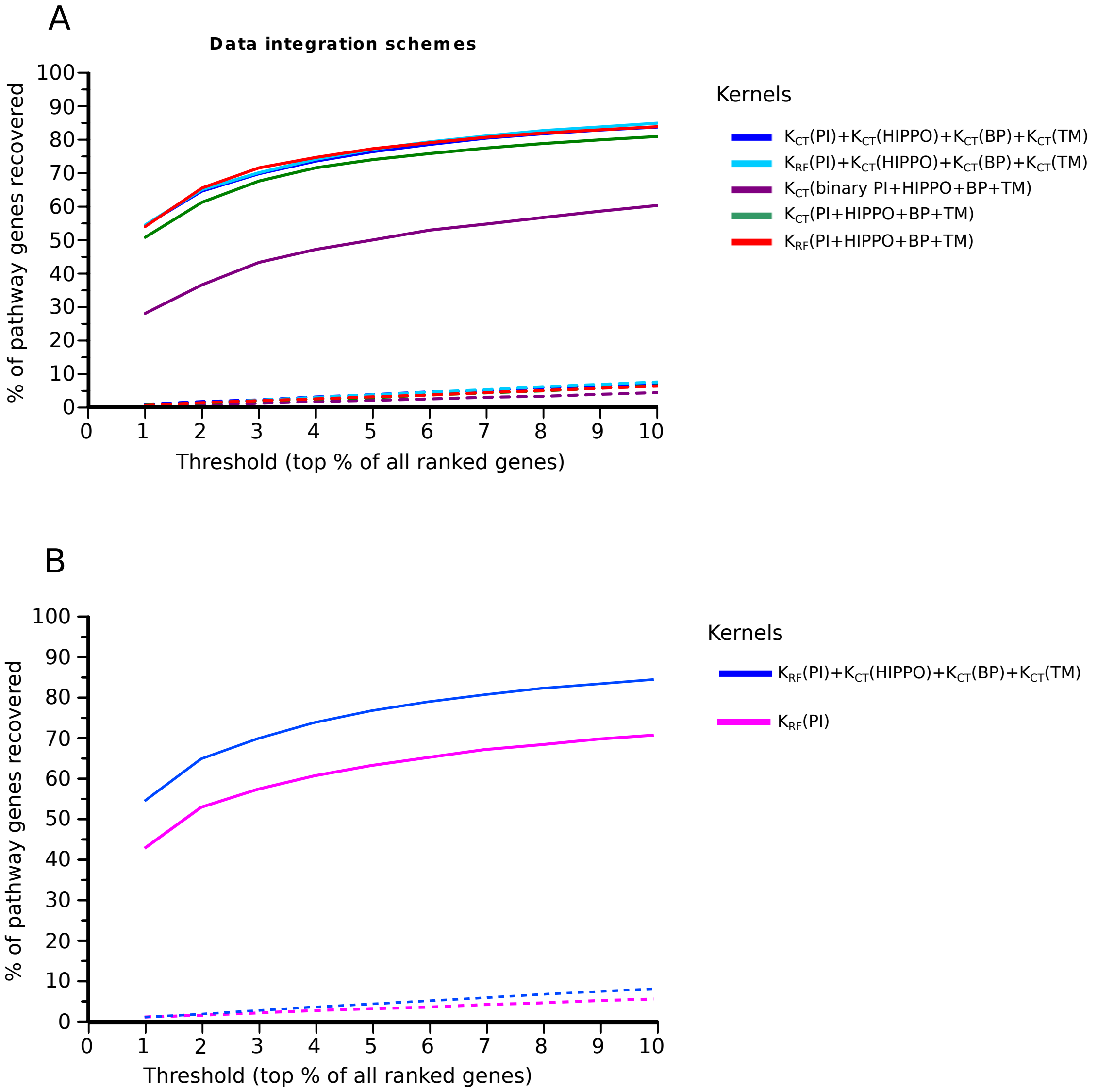
Supplementary figure 5: Distribution of prophase lengths for the hits whose orthologs were tested in S. pombe.
The distribution of prophase length in negative control cells is shown in magenta, the distribution for the siRNA treatment is shown in green, overlaps appear in grey. A – DNMT3B (siRNA s4223), B – PAPD5 (siRNA s34602), C – TOP1 (siRNA s14304), D – HDAC1 (siRNA s74).

Supplementary figure 6: Example of inadequate curve fitting to a strong MCPH1 knockdown phenotype.
Dots represent normalized chromatin volume from confocal images of a cell with long prophase. The blue line represent the result of fitting the chosen sigmoidal function to the points.

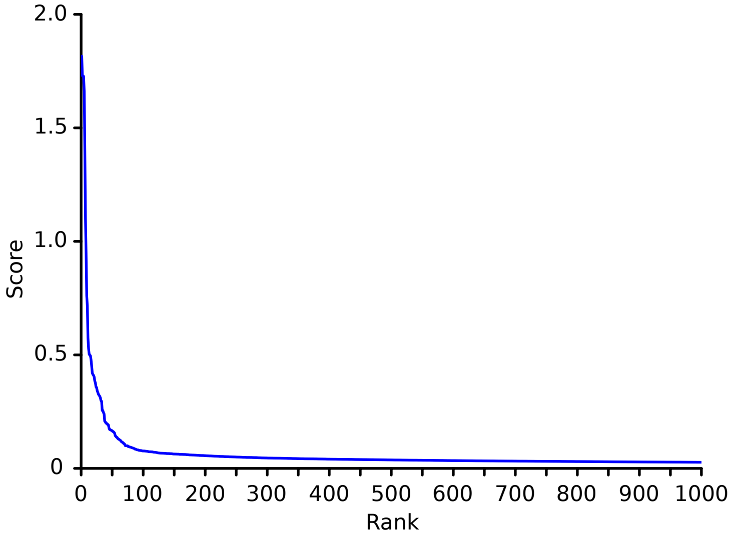
Supplementary figure 1



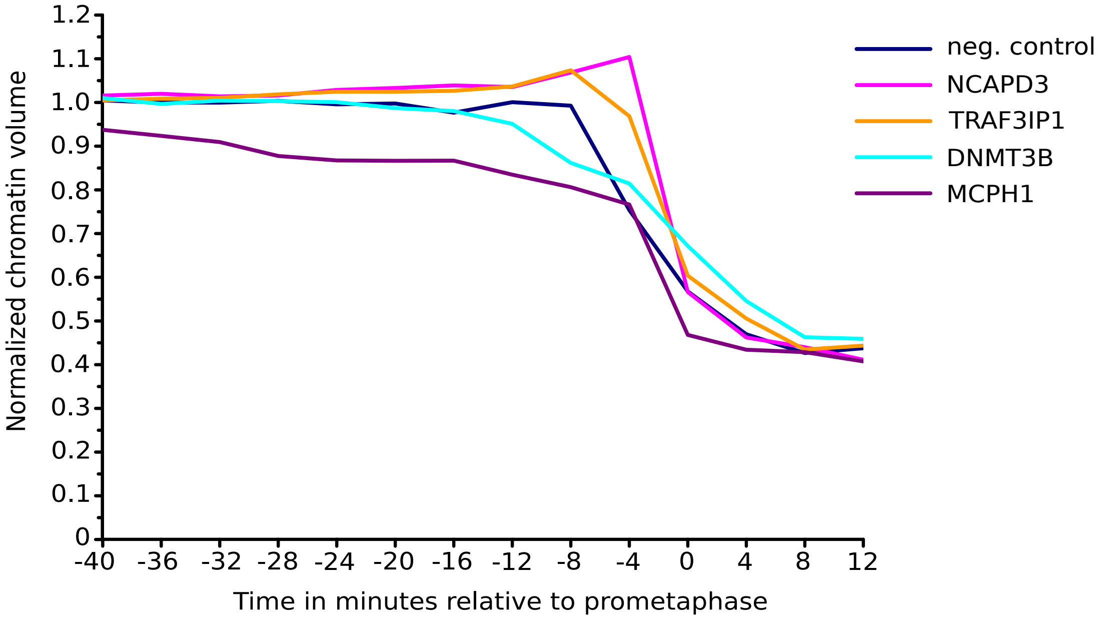
Supplementary Figure 2



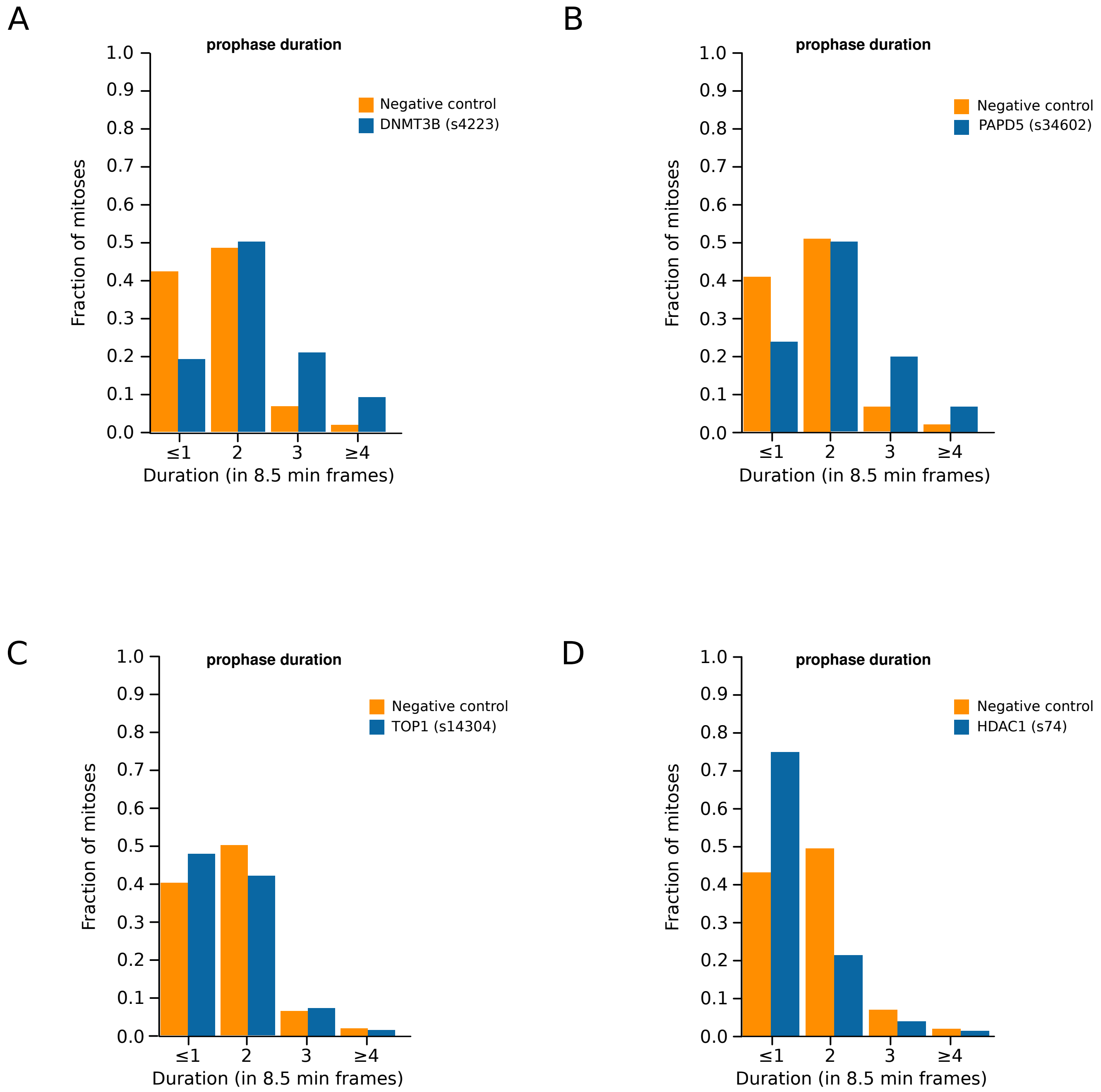
Supplementary figure 3



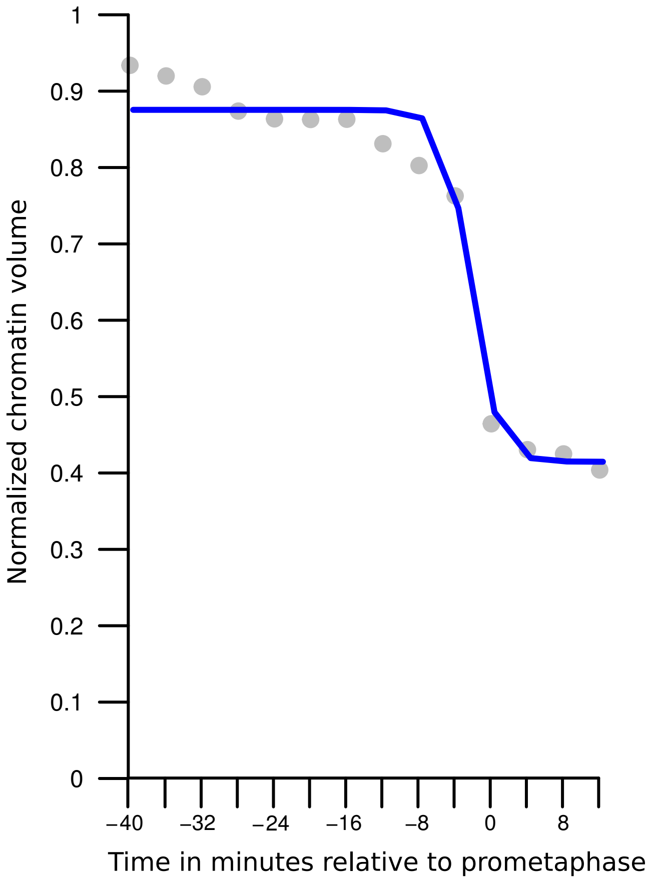
Supplementary Figure 4



Supplementary figure 5



Supplementary figure 6


Normalized chromatin volume
Time in minutes relative to prometaphase