Supporting Online Material for:

A Genome-wide siRNA Screen Reveals Diverse Cellular Processes and Pathways that Mediate Genome Stability

Renee D. Paulsen,[1] Deena Soni,[1] Roy Wollman,[1] Angela T. Hahn,[1] Muh-Ching Yee,[1] Anna Guan, [1] Jayne A. Hesley,[2] Steven C. Miller,[2] Evan F. Cromwell,[2] David Solow-Cordero,[3] Tobias Meyer,[1] Karlene A. Cimprich[1]

[†]To whom all correspondence should be directed.

Email: cimprich@stanford.edu

Table of Contents

I. Description of Supplemental Figures and Tables

Supplemental Figures

Supplemental Tables

II. Materials and Methods

A. siRNA Screening Methods

siRNA library preparation

One copy (0.25 nmoles of desiccated siRNA pools targeting 21,122 genes in 267 X 96-well plates, 80 siRNA pools/plate) of the siARRAY whole human genome siRNA library from ThermoFisher Scientific (formerly Dharmacon; Cat# G-005000-025) was diluted to 2 μM with 125 μl of 1X siRNA buffer (20 mM KCl, 6 mM HEPES-pH 7.5, and 0.2 mM, $MgCl_2$) using a Velocity11 VPrep with a 96 tip disposable tip head. Plates were immediately heat-sealed using an integrated PlateLoc and BenchCel (Velocity11) and then frozen at -20 ^{o}C. After 24 hours, the plates were thawed and 10 μL was transferred with the V-prep to single-use 96-well polypropylene U-bottom plates. The "daughter" plates were immediately sealed and placed at -20 ^{o}C until use.

siRNA library plate mapping

The 96-well information for the siRNA pools and individual siRNA duplexes was stored in an ORACLE database using MDL Plate Manager. *In silico* remapping to 384-well assay plates was also performed using MDL Plate Manager and locations of hits were determined using MDL Report Manager.

Transfection protocol

10 μL of the positive control Chk1 siRNA pool (ThermoFisher Cat# M-003255-02-0020) and 10 μL of the negative non-targeting control siRNA pool (ThermoFisher Cat# D-001206-13-20) were added to two wells each of the siRNA daughter plates. 150 μL of

Optimem was then added to all the wells of the 96-well siRNA daughter plates. Using the Velocity11 VPrep with a 96 high-volume head and 200 µL disposable tips, the siRNA plates were then mixed and 10 µL was aliquoted to a quadrant of duplicate 384-well black-walled clear bottom polystyrene plates. Four 96-well daughter siRNA plates were used to fill each 384-well plate enabling replicates or triplicates to be in separate assay plates and resulting in the incorporation of eight separate wells of both negative and positive controls per 384-well plate. The siRNA daughter plates were sealed and frozen so remaining siRNA could be used in future experiments. 10 µL of the transfection reagent Dharmafect 1 (0.03 µL/well) in Optimem was then added to the 384-well assay plates and the siRNAs were allowed to complex for one hour at room temperature. To start the transfection reaction, 30 µL of HeLa cells (25,000/mL in DMEM, 10% fetal calf serum, and 2 mM L-glutamine) were added to the assay plates. Thus, a final concentration of 25 nM siRNAs was used to transfect 750 cells/well. The plates were then placed in a CO_2 incubator for 72 hours. All reagent additions were performed using a Matrix WellMate with small bore tubing cassette and integrated stackers.

Fixing and staining

72 hours post-transfection, the reaction was stopped by manually tossing the media and fixing the cells with 30 µL ice-cold 90% methanol. The plates were immediately placed at -20 ^{o}C for 20 minutes. The methanol was manually tossed, and the plates were blocked by adding 30 µL blocking buffer (1X TBS, 2 % BSA). After 30 minutes, the blocking buffer was removed and 10 µL of phospho-H2AX antibody (blocking buffer and 1:500 pH2AX antibody, Cell Signaling Cat# 2577) was added and plates were

incubated at 4 ^{o}C overnight. The plates were then washed three times with 1X TBS and the final volume manually tossed. 15 μL of the secondary antibody (blocking buffer and 1:1000 goat-anti-rabbit antibody Alexa-488 conjugated, Invitrogen Cat# A11008) was added, and the plates were incubated at room temperature for one hour. The plates were then washed five times and 25 μL of a 3 X RNAse A/propidium iodide solution (1X TBS, 30 μg/mL RNAse A, Sigma-Aldrich Cat# R4642, 0.3 μg/mL propidium iodide, Calbiochem Cat# 537059) was added to 50 μL of TBS in each well to cleave the RNA and stain the nucleus. Plates were heat-sealed with a Velocity11 PlateLoc and stored at 4 ^{o}C until imaged. All reagent addition steps were performed using a Multidrop dispenser or a Matrix WellMate. All plate washing steps were performed using a Bio-Tek EL405x plate washer with an integrated plate stacker

Imaging

The IsoCyte™ (MDS Analytical Technology) laser scanning platform was equipped with a 20mW 488nm laser and set up for two channels of acquisition. γH2AX-Alexa-488 fluorescence was acquired using a 510-540 nm band pass (green) filter, and PI fluorescence was acquired using a 600 nm long pass (red) filter. Image acquisition was done at nominally 10x10 μm^2 sampling. The entire 384-well plate was scanned with a complete image of each well acquired and analyzed in a single 3 minute scan cycle. The IsoCyte™ was integrated with a Twister II Microplate Handler (Caliper) to allow automated imaging of up to 80 plates at one time.

Image analysis was done concurrently with data acquisition as follows: the fluorescence intensity images were flattened and background corrected using a rolling-ball algorithm with a characteristic length of 150 μm. A region of interest (ROI) of approximately 90% of the area of each well was defined. A threshold algorithm based upon the pixel intensity histogram in the ROI was applied to these processed images in the PI channel, and contiguous pixels above threshold were grouped into individual objects. Intensity and area filters were applied to select those objects consistent with individual cells and the total fluorescence intensity of each object, or cell, was calculated for both channels by integrating the pixel values associated with the cell and subtracting the average background intensity of the well.

B. Data analysis and statistical analysis

Here we describe the statistical procedure used to analyze the raw intensity data per cell and to determine which siRNA treatments showed γH2AX staining increases that were statistically significant. We first give an overview of the analysis procedure and then provide further details to demonstrate some of the analysis decisions using appropriate diagnostic plots.

First, we identified two sources of variability in the data, per-plate and per-day. The analysis as described below takes both types into account. Each plate had its own characteristics (e.g. staining level in control wells) so we identified several parameters on a per-plate basis to determine the strength of the response in each well. The experiments were performed in weekly runs, with between 20 to 40 plates in each day. We noticed

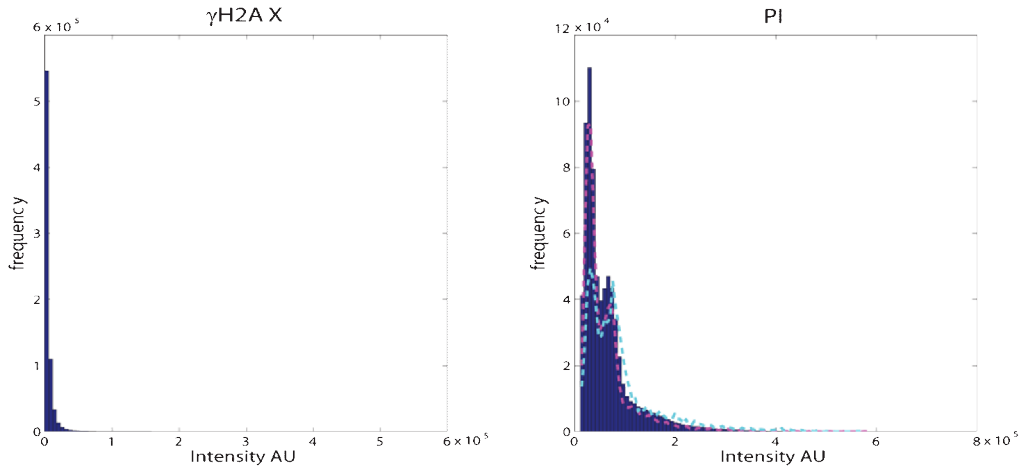
the strength of the γH2AX signal observed in the negative and positive controls varied between treatment days. Therefore, we also estimated another set of parameters to describe the per-day variability of the data. The results of this dual step procedure is that the strength of the response, which is derived on a per-plate basis, does not translate directly into p-value since the significance was derived based on probabilistic estimate that took the per-day variability into account. This can sometimes result in unintuitive results where the ranking of p-value is not in agreement with the ranking based on signal strength.

The following steps describe the overall data analysis:

1. **Per plate data transformation.**
   1.1. Transform into log scale
   1.2. Estimate linear regression model for γH2AX as a function of PI
   1.3. Correct γH2AX signal for DNA content
2. **Calculation of signal strength per well.**
   2.1. Fit Gaussian Mixture (GM) model to positive controls, and identify responsive cell intensity Gaussian.
   2.2. Use non-parametric kernel smoothing to identify distribution of negative controls.
   2.3. Using the two distributions identified above, define optimal intensity threshold.
   2.4. Using the intensity threshold, calculate number of responsive cells per well (k) and total number of cells (n) per well
3. **Assignment of p-value for each well**
   3.1. Estimation of within-day and between-day variability of controls
   3.2. Derivation of mathematical expression for the probability to get k responsive cells out of n trials given known between day/within day variability.
   3.3. Assignment of p-value for each treatment.
4. **Correction for multiple testing to assign significant categories.**

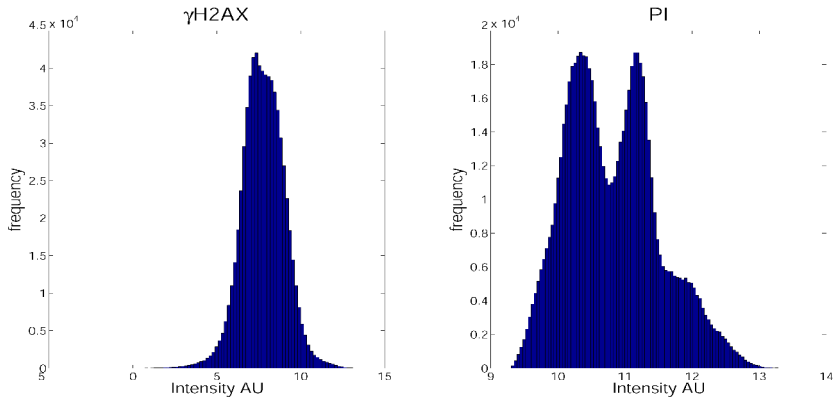
1. Per plate data transformation

The raw data included the measured intensity of γH2AX and propidium iodine (PI) per cell. For both PI and γH2AX, the intensity distribution had a heavy right tail as can be seen in the histogram of all cells in a single plate (Methods Fig. 1).



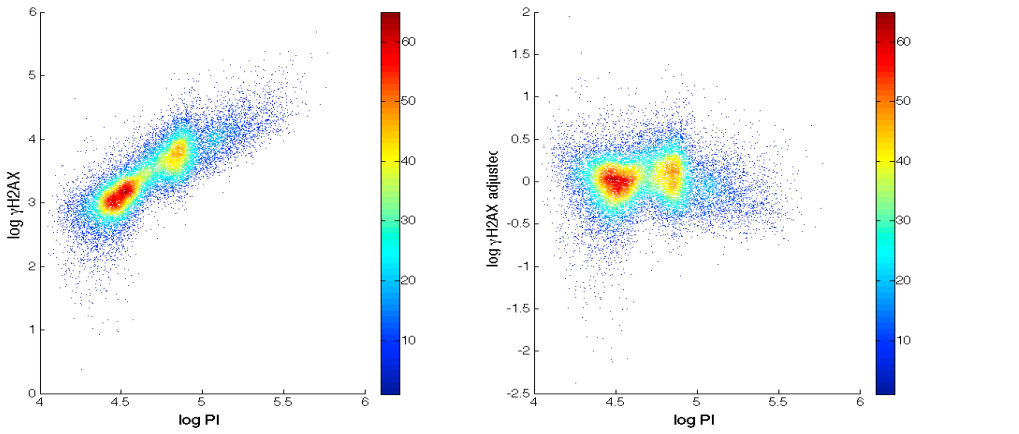
Methods Figure 1. Histograms of measured raw cellular γH2AX and propidium iodide intensities. Left panel shows the γH2AX distribution of all cells within a representative plate. Right panel histogram show the propidium iodide (PI) distribution of all cells within the same representative plate. The dashed cyan and dashed magenta lines represent the PI distribution for the positive and negative controls respectively. Intensity is measured in arbitrary units (AU). Approximately 6*10^5 cells were scored.

Therefore, the first step was to log transform the data to reduce the heavy right tail. This step was especially critical for the γH2AX staining. After initial transformation, the histograms show the distribution is more balanced and the PI is clearly bimodal as expected from a DNA content dye (Methods Fig. 2).



Methods Figure 2. Histograms of log transformed cellular γH2AX (left panel) and propidium iodide (right panel) intensities.

We also needed to consider the fact that a correlation between the γH2AX and the PI stain was evident. The following scatter plots show the negative control cells from all the wells of a single plate before (left panel) and after (right panel) correction (Methods Fig. 3).



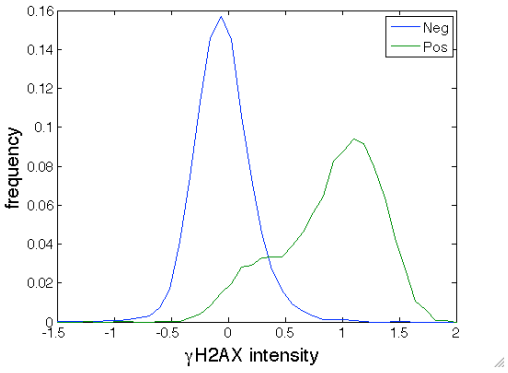
Methods Figure 3. Scatter plots of negative control cells before (left panel) and after (right panel) correction for DNA content. Frequency density is represented by the shown pseudocoloring ranging from dark blue (low density) to dark red (high density). Approximately 12,000 cells were scored.

In this plot, each point's color represents the number of points at this position. It is obvious from this plot, as from previous biological knowledge (Mirzoeva and Petrini,

2003), that as DNA content increases, the intensity of γH2AX staining increases as well. To correct for the DNA content in each cell, we decided to use the negative control cells from each plate. For each plate, we fit a simple linear regression model of γH2AX as a function of PI. Then, for each cell in the plate, we subtract the expected value based on this linear model from the observed γH2AX intensity. For the rest of the analysis, we only considered a single measurement per cell, its adjusted log γH2AX (y-axis in right panel of Methods Fig. 3). For simplicity, we will still use the term γH2AX intensity, but from here on we are referring to the adjusted log transformed γH2AX intensity.

2. Calculation of signal strength per well

After data transformation, we further investigated the response of the siChk1 positive control wells. The following histogram, which is representative of all the plates, shows the distribution of the positive and negative controls:



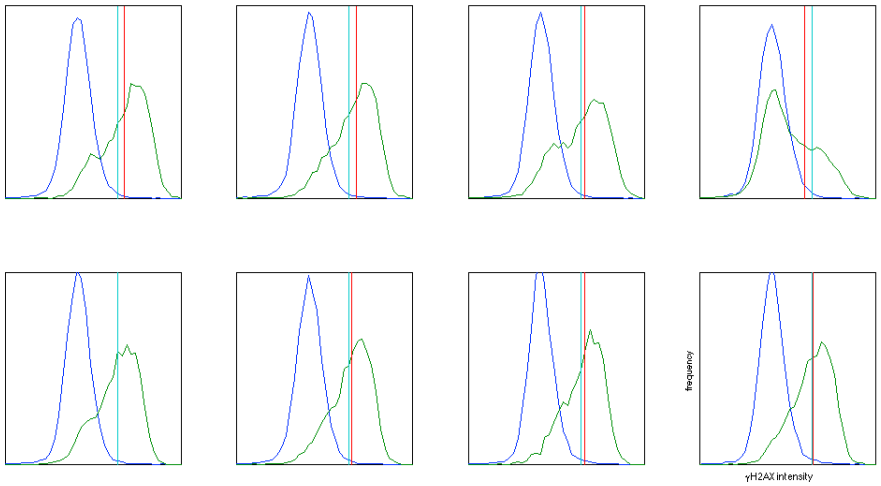
Methods Figure 4. Histogram comparing the negative and positive control γH2AX staining distributions. The blue line indicates the negative control which was a non-targeting siRNA pool and the green line indicates the positive control which was a siRNA pool targeting the Chk1 kinase. The distributions represent all cells scored for a representative plate.

The negative control showed a single unimodal Gaussian whereas the positive controls showed a strong bimodal distribution where one peak is close to the negative control (~-0.2) and another strong peak is evident at ~1.2. One interpretation of the data is that in each well some percent of the population is affected by the treatment and shifts its intensity to a new distribution centered around 1.2. Therefore, we decided to divide each well into two populations: the non-responding and responding cells, and use the percent of responding cells as the signal from each well. However, since estimating the two populations seperately on each well is not robust enough for such large dataset, we opted on estimating a single threshold value per plate based on the controls. Then, the statistics for each well was based off of the percent of cells above this threshold.

To determine the threshold for each plate, we estimated the probability distribution for the negative and positive controls and used those parametric distributions to determine the cutoff. To estimate the distribution of the negative control (non responsive) cells, we used a simple, single Gaussian fit of normal distribution. For the positive controls, we first fitted, using the Expectation Maximization (EM) algorithm, a Gaussian Mixture (GM) distribution with two Gaussians to the positive control wells. Then, we took the mean and standard deviation of the higher mixture and used a single Gaussian normal distribution with these parameters as an approximation of the distribution of the responsive cells. We then determined the cutoff by using two Gaussians, the negative control Gaussian and the Gaussian of the responsive cells in the positive control, e.g. the right Gaussian in the Gaussian mixture. The cutoff for a γH2AX positive cell was determined as the point in this probability where the responsive cell distribution is higher

than the non-responsive distribution. The result of this procedure was a single threshold below which a cell was considered to be non-responsive (negative) and above which the cell was considered to be responsive (positive).

The following plots show the positive (green) and negative control (blue) distributions with the per-plate threshold (red) and the average of all thresholds (cyan).



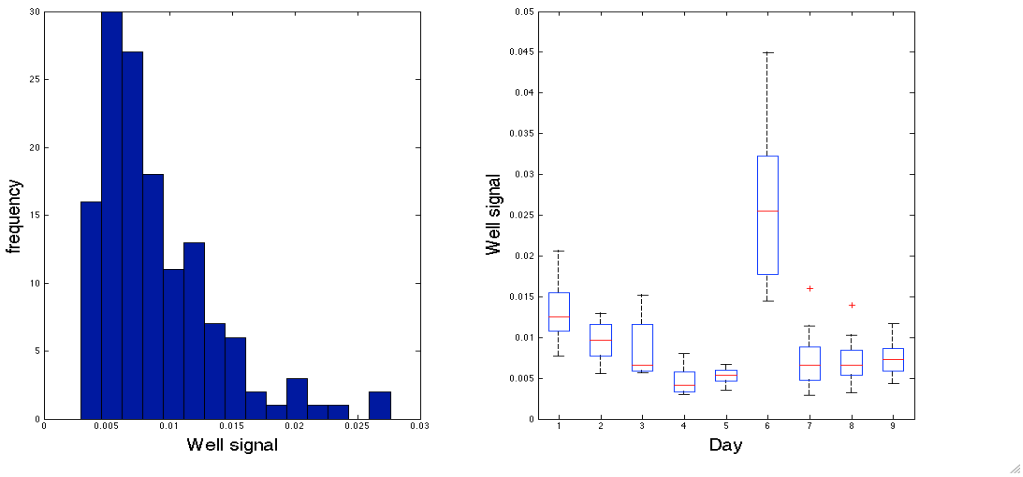
Methods Figure 5. Representative comparative histograms from random plates assayed. The blue line indicates the negative control which was a non-targeting siRNA pool and the green line indicates the positive control which was a siRNA pool targeting the Chk1 kinase. Each graph represents an individual plate in which the distributions represent all control cells scored within that plate. Vertical lines indicate calculated thresholds, with red showing that plate's individual threshold for what was considered to be an γH2AX positive cell and cyan indicating the average threshold for all plates analyzed. Approximately 12,000 cells were scored for the negative control and 6,500 cells were scored for the positive control for each graph shown

3. Assignment of P-value for each well

Given the signal strength from each well, we now turned to the question of statistical significance or assignment of P-values. By definition, a P-value is the probability to obtain a certain result assuming a null hypothesis. Therefore, we needed to

specify the null hypothesis in detail and use it to determine the P-value for each well. The higher the variance of the null hypothesis was, the more likely we were to obtain high signal wells by chance alone and therefore the higher (less significant) the P-value of these seemingly strong wells would be.

The following two plots analyze the variability of the negative controls from all the plates in the screen. The left panel shows the histogram of the average signal strength per plate of the negative control wells (8 replicates) in that plate. The histogram shows non-negligible variability suggesting that one needs to consider differences of signal strength in plates. The right panel shows a box-plot of different days. The red line is the median signal strength of all the negative control wells done in that day (across multiple plates). This shows that not only was there well to well variability in the signal strength of the controls but more importantly that this variability changed between different days, both of which needed to be accounted for within our analysis.



Methods Figure 6. Demonstration of the types of signal variability observed. Left panel shows the distribution of the average γH2AX signal strength (percent γH2AX positive population) of the negative controls for each plate assayed. Right panel shows a box plot of the variation seen on a per day basis with each box representing the variation seen between plates within that day's set. The red line indicates the median γH2AX

signal strength (percent γH2AX positive population) of all of the negative control wells assayed within that day. The upper and lower limits of the boxes represent the 25th and 75th percentile of the median signal strength of the same data set. The error bars indicate the maximum and minimum median signal with crosses indicating outliers within that day's analysis.

Since we treated each cell as having only possible two states, we can think of each well as a binomial distribution and ask what is the probability of finding K responsive cells out of N possible cells given the probability to belong to the responsive population q. Determining q will give us a full description of the null hypothesis in a parametric form and will allow us to assign P-values. The big question is what is q? The above analysis of plate-to-plate and day-to-day variability shows that this might not be a trivial question. In the next section, we describe an analytical expression that we developed to allow us to estimate the P-values without knowing q explicitly, but assuming a parametric form of the distribution of q.

Mathematical derivative of analytical expression for the P-value

Let the probability of a cell in well $w_{i,j}$ to show positive γH2AX staining be $q_{i,j}$. We assume each cell follows a simple Bernoulli distribution. We assume cells within a well are independent of one another so that the probability of seeing k positive cells in a well with n cells is given by the binomial distribution:

$$P(positive\#=k)=\binom{n_{i,j}}{k_{i,j}}q_{i,j}^{k}\left(1-q_{i,j}\right)^{n-k}$$

where i indexes all plates and j indexes all wells within plate i. To estimate the P-value, the probability of seeing more than k positive cells in a well, we can look at the complimentary probability, the probability of observing k or lower cells:

$$P-value = 1-P(positive\# \le K) = 1-\sum_{k}\binom{n_{i,j}}{k_{i,j}} q_{i,j}^{k}\left(1-q_{i,j}\right)^{n-k}$$

Therefore, if we know *q* for each well, we can estimate the p-value. However, as we showed earlier, $q_{i,j}$ is not constant and has non-negligible variability that must be taken into consideration. Below we show how this variability in q was taken into account using a hierarchical statistical model.

In our model, a cell can be classified into the responsive class either if it really belongs to that class or if it was misclassified as such. We define λ as the probability of belonging to the responsive class. Note, λ is different from q in that it represents the probability of being classified as responsive. What is then the relationship between q and λ? To answer this question let us define the intensity distribution I of a single well to be a mixture of two distributions:

$$I \sim \begin{cases} \lambda & F_{responsive} \\ 1-\lambda & F_{non-responsive} \end{cases}$$

where $F_{responsive}$ follows a normal distribution with parameters μ and θ and $F_{non\text{-}responsive}$ follows some arbitrary distribution. Based on these probability assumptions, we can write the probability that a cell is above threshold to be (q) in terms of the probability to belong to the responsive class (λ) and the two conditioned probabilities:

$$q = \lambda \cdot P\left(I > T_i \mid responsive\right) + \left(1-\lambda\right) \cdot P\left(I > T \mid non-resposive\right)$$

Given the variability in well signal strength in the negative controls from different wells and the day-to-day variability demonstrated in the previous figure, we assume that λ has a log-normal distribution with parameters ν_d, θ_d where ν_d is the average and θ_d is the standard deviation for all negative control wells from plates done in day d. Since q is a function of λ, which is a random variable, without prior knowledge of the specific λ value for a specific well, we take the expectation over all values of λ:

$$q = q(\lambda) = E\left(q(\lambda)\right) = \int q(\lambda) f(\lambda) d\lambda$$

where f(λ) is the probability density for λ. Substituting the expression for q and the two responsive / non responsive intensity distributions we get:

$$q_{i,j} = \int_0^{Inf} \left(\lambda \cdot P\left(I > T_i \mid \ responsive\right) + \left(1-\lambda\right) \cdot P\left(I > T_i \mid \ non-responsive\right)\right) f(\lambda) d\lambda =$$

$$\int_0^{Inf} \left[\lambda \int_{T_i}^{Inf} \frac{1}{\sqrt{2\pi}\theta_i} e^{\frac{-(I-\mu_i)^2}{2\sigma_i^{\ 2}}} dI + \left(1-\lambda\right) \int_{T_i}^{Inf} F_{non-responsive} dI \right] \cdot \left(\frac{1}{\lambda\sqrt{2\pi}\theta_d} e^{\frac{-(\ln\lambda - \nu_d)^2}{2\theta_d^{\ 2}}} \right) d\lambda$$

This integral can be solved numerically. The distribution parameters are estimated based on several sources: μ_i, σ_i are based on the higher Gaussian in the Gaussian Mixture fit to

the positive control, ν_d, θ_d are based on the negative controls from all the plates done in day d, T_i is the threshold for this plate, and $F_{non\text{-}responsive}$ is estimated from same plate negative controls using a kernel smoothing procedure. Given the numerical solution of this integral, the estimation of the p-value uses the binomial distribution as explained above.

4. Correction for multiple testing and assignment of significant categories

We applied the procedure described above to 46,400 different wells. When performing so many statistical tests, one must correct for multiple testing. We used a FDR (false discovery rate) correction to adjust for a 5% false positive rate and define two P-value cutoffs, the traditional alpha 0.05 and an FDR-corrected alpha for each replica. We then defined three levels of significance. Group 4 includes all the genes where both replicas were lower than the FDR corrected alpha (0.0042). In group 3, one replica was below the FDR corrected alpha and the other replica was below the traditional alpha (0.05). In group 2, one of the replicas was below the traditional alpha. Genes were also required to have a γH2AX signal within the top 25% of the genome to belong to one of the aforementioned groups. Because genes in group 2 did not have P-values below the FDR-corrected alpha, a false positive rate was estimated using a cross validation approach where all but one negative control was used to determine the threshold for what was considered to be a positive cell. Then the omitted negative control was treated as an experimental well, and we used the number of times it scored with a P-value less than 0.05 in our analysis to estimate the FDR for group 2. In addition, we acknowledge that due to the stringent statistical procedure described above, we are likely to loose true

biological 'hits' due to high variability and lack of statistical power. Therefore, we added a final group, group 1, of all the genes not in groups 2-4, which had a signal strength in the upper 0.05 percentile.

Significance Group	Number in Group	False Discovery Rate (FDR)
4	581	5%
3	206	5%
2	1451	10%*
1	164	N/A

Methods Figure 7. Significance group gene numbers and false discovery rate. Genes in the most significant level (group 4) have a p-value for both replicates lower than the FDR corrected level (p-value < 0.0042) for a 5% false positive rate. Genes in the next level (group 3) have p-values that are significant at the FDR level for one replicate and that fall below the traditional level of p = 0.05 for the second replicate. Genes in the third level (group 2) have a p-value < 0.05 for only one replicate and were calculated to have a false positive rate of 10%. We also created a final group containing those genes which have a strong, albeit statistically insignificant signal. This group includes all genes not included in other groups with an average γH2AX signal in the top 5% of the genome (group 1). For this group, a FDR is not applicable as these genes were assigned no statistical significance. * indicates the FDR was calculated using cross validation.

C. Additional Experimental Methods

Deconvolution

The individual siRNA oligos for the 350 siRNA pool hits were tested in the deconvolution assay. For this screen, each unique siRNA oligo in a 96-well plate was resuspended in 125 µl of 1X siRNA buffer (2 µM). Plates were heat-sealed using the PlateLoc (Velocity11) and frozen at -20 ^{o}C. After 24 hours the plates were thawed and 20 µL was transferred to 96-well polypropylene U-bottom plates with 88 siRNAs per plate. 10 µL from these daughter plates was transferred to a new daughter plate that was then used for the screen in the same manner as described above. Image aquisition for

these siRNAs was performed as above and resulting data (percent γH2AX staining of cells) for each of the four individual siRNAs was determined by setting an intensity cut-off approximately three times greater than the average γH2AX staining observed in the siControl-treated wells.

Z' Factor Calculation

The Z' Factor was defined from the means and standard deviations of both the positive and negative controls on each plate according to Zhang, 1999.

Propidium iodide cell cycle analysis

To assign cell cycle stage values, all cells within a given well were compiled and a histogram of the PI intensities per cell was plotted. To eliminate nuclear fragments and cellular doublets, a nuclear area parameter gate was set. To assign stages of the cell cycle and percent $\gamma H2AX^+$ per stage, the first four non-targeting control siRNA wells of each individual plate were compiled and a cumulative histogram was plotted. According to external experiments, we calculated the average cell cycle profile for siControl-treated HeLa cells to be 40% G1, 30% S, and 20% G2 when the 5^{th} and 95^{th} percentile of cells were eliminated. Thus, we set PI intensity parameters within the compiled control samples to fit this model. To account for the effect of cell number variance on PI intensity, the siRNA targeted histogram was overlaid on the histogram from the nontargeting control, and PI intensities were adjusted according to the min and max values of the control eliminating the 5^{th} and 95^{th} percentile. Then the same PI intensity cutoffs for each cell stage were applied to the siRNA targeted histogram. Analysis was

done in duplicate and averages with SD were compiled. The percent γH2AX positive population (% $\gamma H2AX^+$) was calculated by applying an intensity cutoff approximately three times greater than the average H2AX intensity of all siControl treated wells. This value was then subdivided to calculate the % $\gamma H2AX^+$ population per cell cycle stage by categorizing each positive cell based off of its corrected PI intensity and where that fell in relation to the cell cycle stage PI intensity cutoffs.

Cell culture and transfection

HeLa cells (ATCC, VA) were cultured in DMEM (Invitrogen, CA) with 10% FBS (Invitrogen) and supplemented with penicillin, streptomycin and glutamine (Invitrogen). GJB1 patient cell lines, GM04593 and GM04595 (Coriell Cell Repositories) were cultured in MEM with Earl's salts and nonessential amino acids (Invitrogen, CA) with 15% FBS and supplemented with penicillin, streptomycin and glutamine. For transfection, cells were grown in media without antibiotics. The siGenome siRNAs (Thermofisher Scientific) used for silencing candidate screen hits can be found in supplemental tables S9 and S10. Cells were transfected with Dharmafect Reagent #1 (Dharmacon) diluted in Optimem (Invitrogen). All siRNA were used at 25nM final concentration during transfection.

Antibodies

Antibodies to Chk1 (G-4, #sc-8408, Santa Cruz Biotechnology); phospho-Chk1 (Ser 345), and phospho-histone H2A.X (Ser139) (Cell Signaling Technology); BrdU and

Hsp90 (Becton- Dickinson); MPM-2 and phospho-histone H3 (Ser10) (Upstate Biotechnology Inc), and 53BP1 (Chemicon) are all commercially available.

Enrichment analysis and bioinformatics

PANTHER (Protein Analysis through Evolutionary Relationships): Genes within the highest significance group within the original screen (i.e. group 4) were uploaded into in the application. Each gene identifier was mapped to its corresponding gene object in the PANTHER database. Biological processes were assigned for each protein and visualized using the Pie chart function built into the application. The top eleven categories with the most genes assigned within their designations are shown in Figure 2A. These defined categories were then subdivided into subcategories by the application as demonstrated in Figure 2B.

Functional analysis: Genes within the highest significance group within the original screen (i.e. group 4) were uploaded into David bioinformatics database (http://david.abcc.ncifcrf.gov/) and Ingenuity pathway analysis (Ingenuity Systems, www.ingenuity.com) to identify the biological functions and/or diseases that were most significant in the data set. The genes were categorized according to GO terms (biological process, cellular complex, molecular function), protein information resource keywords, or the OMIM/Genetic Association disease datasets. Fischer's exact test was used to calculate a p-value determining the probability that each biological function and/or disease assigned to that data set is due to chance alone. Select enrichments are shown in Figure 2C with the full list of significantly enriched functions listed in TableS4.

Networks: After a functional category had been identified, the genes within each category were uploaded into the IPA software to look for protein interaction networks. Each gene identifier was mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base. These genes, called focus genes, were overlaid onto a global molecular network developed from information contained in the Ingenuity Pathways Knowledge Base. Networks of these focus genes were then algorithmically generated based on their connectivity. Once a network had been defined, genes within the remaining significance groups (ie. 1-3), as well as the genes validated with 1 or more siRNAs upon deconvolution, were mined to look for other interacting proteins which also fell into the original functional category. Networks were modified by hand to reflect additional literature defined direct protein interactions where appropriate.

Screening Results Comparison: Genes that gave a significant γH2AX signal within the original screen (i.e. groups 2-4), plus genes that gave a γH2AX signal greater than 2SD from the control signal with one or more siRNAs within the deconvolution experiment were used for cross comparison with relevant published screening results. For published human studies, gene names were converted to the most recent identifier using either the NCBI gene database or the Biobase Proteome platform and used for comparison. The significance of enrichment was calculated using a Chi-squared test. For screening results from other species, the functional orthologs (if applicable) were identified using the Biobase Proteome platform and used for comparison purposes. Significance of enrichment was not calculated for cross-species comparisons.

Immunofluorescence

53BP1: 72 hours following siRNA knockdown, cells were preextracted with 0.25% NP-40 on ice for 10min, fixed with 4% para-formaldehyde for 20min, blocked in 2% BSA/PBS for 30 min, and incubated in primary anti-53BP1 antibody for 1 hour at room temperature (1:500). γH2AX: cells were fixed with 4% para-formaldehyde for 20min, permeabilized with 0.2% Triton-X/PBS, blocked in 2% BSA/PBS for 30 min, and incubated in primary γH2AX antibody overnight at 4°C (1:500). In both cases, cells were then washed three times in 2%BSA/PBS, incubated in Alexa 488 (1:1000) secondary antibody for 1 hour, and then washed three times with PBS. DAPI was added to stain nuclei. Cells were imaged on an automated epifluorescent microscope, and data was analyzed using Metamorph imaging software.

HeLa RNAseH H2AX Assay

The RNAseH1 construct was obtained from Open Biosystems and subcloned in the pBMN blastocidin resistant retroviral vector containing an HA and FKBP L106P degron to obtain conditional protein expression via the addition of the small molecule Shld as described previously (Banazynski, 2006). Stable RNAseH1 expressing HeLa pools were selected and utilized for subsequent experiments. After 24 hours knockdown 1μM Shld was added for the subsequent 48 hours after which cells were fixed and stained for γH2AX. In the absence of Shld minimal destabilization of the protein was observed, so the γH2AX change was compared to the parental HeLa cell line. Percent change in

γH2AX expression was calculated as follows: [(% $\gamma H2AX^+$ cells RNAseH HeLa - % $\gamma H2AX^+$ parental HeLa) / (% $\gamma H2AX^+$ parental HeLa)] *100.

G2/M checkpoint assay

HeLa cells transfected with 25nM of a single siRNA for 48h were irradiated with 10 Gy of IR and allowed to recover for 1 h. Thereafter, the growth medium was replaced with medium containing 100 ng/ml of nocodazole (Sigma, MO). Eight hours later the cells were harvested and fixed with 90% methanol and processed for flow cytometry. The mitotic index and the DNA content were determined with Mpm2 antibody (1:250) (Upstate Biotechnology Inc, NY) and propidium iodide (Sigma) staining respectively. Samples were analyzed on a FACSCalibur (BD Biosciences, CA). The listmode data was analyzed using WinList 3D 5.0 (Verity, ME).

Homologous Recombination Assay

To test the effect of knockdown of the CMT genes on homologous recombination (HR) (Figure 6), we measured the repair of an I-SceI endonuclease-generated DSB as described in (Pierce, 1999) with the following modifications. We generated a pooled HeLa cell line stably transduced with a fragment of the YFP gene missing the last 35 bp that had an I-SceI target site at its 3' end. Approximately 150,000 of the 5'-YFP/I-SceI target site cells were transfected in one well of a 6-well tissue culture plate with an siRNA (25nM) using Dharmafect transfection reagent. Twenty-four hours post-knockdown, these cells were transiently-transfected with two plasmids - one encoding the

I-SceI enzyme and another plasmid containing both the 3' fragment of YFP missing the first 97bp of the YFP gene and mCherry (a red fluorescent protein) (3' YFP/mCherry).

Successful homologous recombination events occurred in cells that had fluorescence in both the red (to verify successful transfection of the 3' YFP/mCherry plasmid) and YFP (homologously recombined 5' and 3' YFP fragments) channels. Transfection of the plasmids was performed on trypsinized siRNA transfected cells with 6 μL Fugene6 (Roche) + I-SceI enzyme plasmid (0.5μg) + 3' YFP/mCherry plasmid (0.5 μg). Cells were divided into three populations while still in the transfection mix and plated into 3 separate wells of a 6 well plate to avoid overcrowding. Cells were harvested by trypsinization four days after plasmid transfection and replating and analyzed by flow cytometry using a BD FACSCalibur. Approximately 50,000 events were acquired and the percent of homologous recombination events was calculated thusly: 1) cells were gated by forward- and side-scatter readings to eliminate dead or clumped cells; 2) live cells were gated for positive red fluorescence; 3) live, red cells were scored for yellow fluorescence. Successful HR events occurred in live cells that were both red and yellow, and percents were calculated as (# of YFP^{+}Red cells)/(# of Red cells) and were normalized to the luciferase values.

For the HR efficiency measured after splicing gene knockdown (Figure 5), we used the method described above with the following modifications. Approximately 300,000 HEK293M 2.1 reporter cells which harbor a stably-integrated I-SceI target site disrupting the YFP gene were transfected with the siRNA (25nM) using Dharmafect transfection

reagent. Twenty-four hours post-knockdown, these cells were transiently-transfected with pdYFP (0.8μg), pYFP-97bp (0.8μg) containing the YFP gene missing the first 97bp or pYFP-97bp (0.8μg) plus the I-SceI endonucease (1.6μg) using Lipofectamine 2000 (Invitrogen, CA). Cells were harvested at 72 h and single cell populations were analyzed by flow cytometry using a BD FACSCalibur. Approximately 250,000 events were acquired and the percent YFP positive cells was determined. The percent of recombinants was calculated as follows: % YFP positive in pYFP-97bp+SceI transfectants - % YFP positive in pYFP-97bp transfectants. The recombination frequency was calculated by normalizing to the transfection efficiency determined from pdYFP transfected cells.

Real Time PCR

Real time PCR was performed following manufacturer's instructions using iTaq SybrGeen Supermix with ROX (BioRad) on an ABI 7900HT system.

DNA Damage Sensitivity Assay

To assess damage sensitivity to various DNA damage agents the following protocol was adapted from Smogorzewska, 2007. eYFP was cloned into the pBMN retroviral vector with hygromycin resistance. Stable YFP expressing HeLa pools were selected and utilized for subsequent experiments. Approximately 150,000 YFP expressing cells were transfected in one well of a 6 well tissue culture plate with the siRNA (25nM) of interest using Dharmafect transfection reagent. An equal population of unfluorescent HeLa were transfected with siLuciferase. After 48 hours, the two populations were mixed in a one to

one ratio, plated at 20% density, damaged four hours post-replating (IR 2 Gy or aphidicolin 100nM), and allowed to grow for the following 5 days. Media was changed every two days to ensure maximal cell growth. Cells were then trypsinized and analyzed by FACS for the ratio of YFP^{+}/ YFP^{-} cells. Ratios were corrected for the growth rate effect of each siRNA alone. $[(\%YFP^{+}siRNA_{X}{}^{\text{Damage condition}} * (\%YFP^{+}siRNA_{Luc}{}^{NT} / \%YFP^{+}siRNA_{X}{}^{NT}))/ \%YFP^{+}siRNA_{Luc}{}^{\text{Damage condition}}]*100$

BrdU FACS Staining

Cells were pulse-labeled with BrdU (20μM) for 30 min, washed in PBS, and fixed in 70% ethanol for flow cytometric analysis. After fixation, cells were permeabilized with 0.2% Triton X/PBS for 15 min on ice, denatured with 2M HCl for 15min, blocked in 2% BSA/PBS for 10 min, and incubated in primary BrdU antibody (1:100) for 1.5 hours. Cells were then washed three times in 2% BSA/PBS, incubated in Alexa 488 secondary antibody for 1 hour, and then washed three times with PBS. PI/RNAse was added to determine DNA content and cells were analyzed on a BD FACSCalibur.

Cell extract preparation

To prepare whole-cell extracts, cells were harvested by trypsinization, washed in PBS, resuspended in modified TGN buffer (50mM Tris pH7.5, 150mM NaCl, 50mM β-glycerophosphate, 10% glycerol, 1% Tween 20, 0.2% NP-40, 1mM NaF, 1mM NaVO4, 1mM DTT, and 10μg/mL leupeptin, pepstatin and aprotinin) for 20 min on ice and cleared by centrifugation (10 min, 20,000g).

III. Supplemental References

Mirzoeva, O. K., and Petrini, J. H. (2003). DNA replication-dependent nuclear dynamics of the Mre11 complex. Mol. Cancer Res. *1*, 207-218.

Zhang J.H. et al. (1999). A simple statistical parameter for use in evaluation and validation of high throughput screening assays. J Biomol Screen. 4(2): 67-73.

Banazynski L.A. et al. (2006). A rapid, reversible, and tunable method to regulate protein function in living cells using synthetic small molecules. Cell. Sep 8;126(5):995-1004

Pierce A.J. et al. (1999). XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. Genes Dev. Oct 15; 13(20) 2633-8.

Smogorzewska A. et al. (2007). Identification of the FANCI protein, a monoubiquitinated FANCD2 paralog required for DNA repair. Cell. Apr 20; 129(2):289-301

IV. Supplemental Figures

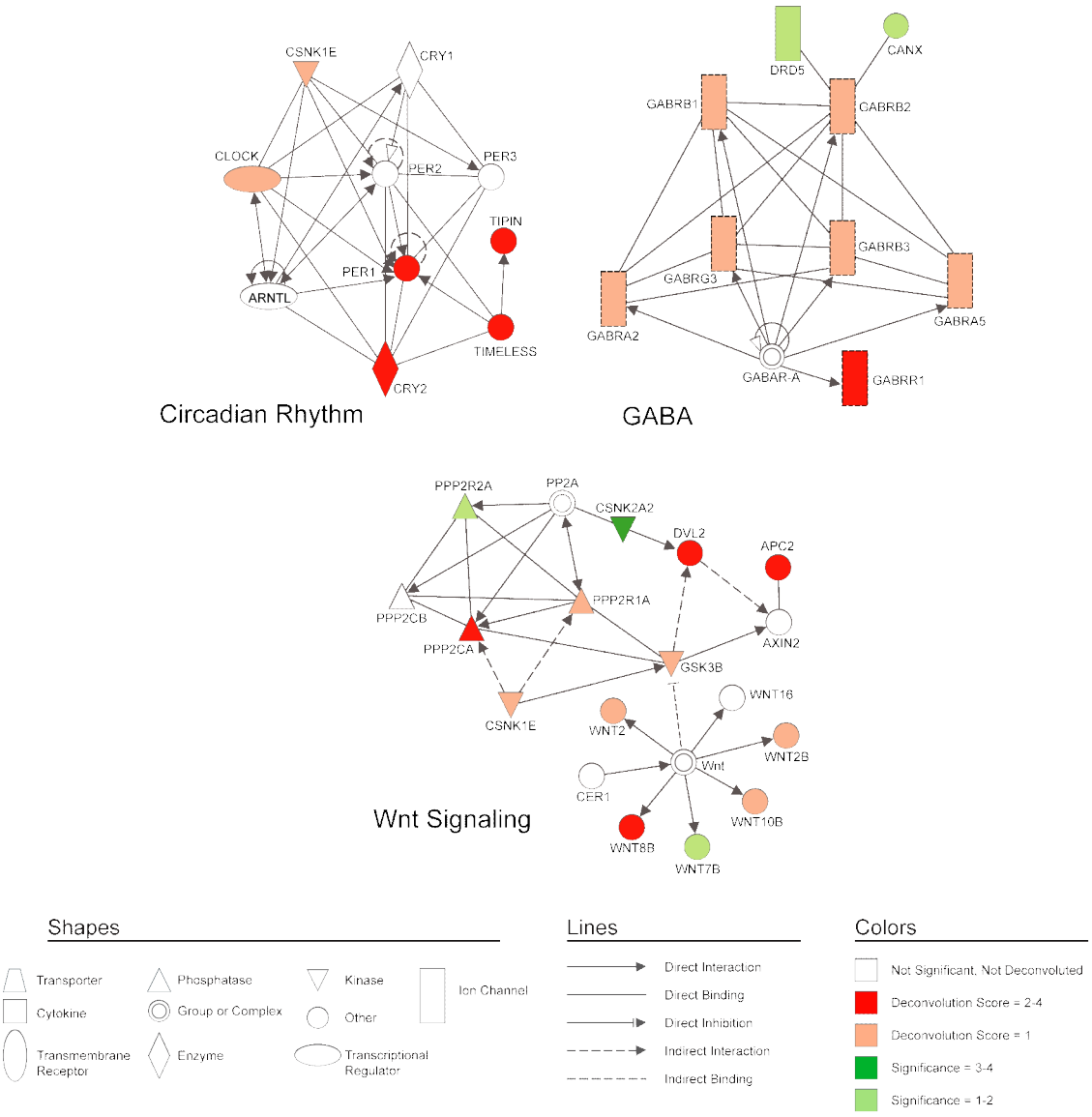


Figure S1. Network modeling of screen hits identifies new functional groups linked to genome maintenance. Pathways validated upon siRNA deconvolution include proteins involved in circadian rhythm, GABA signaling, and Wnt signalling. Protein-protein interactions were mapped with Ingenuity pathway analysis.

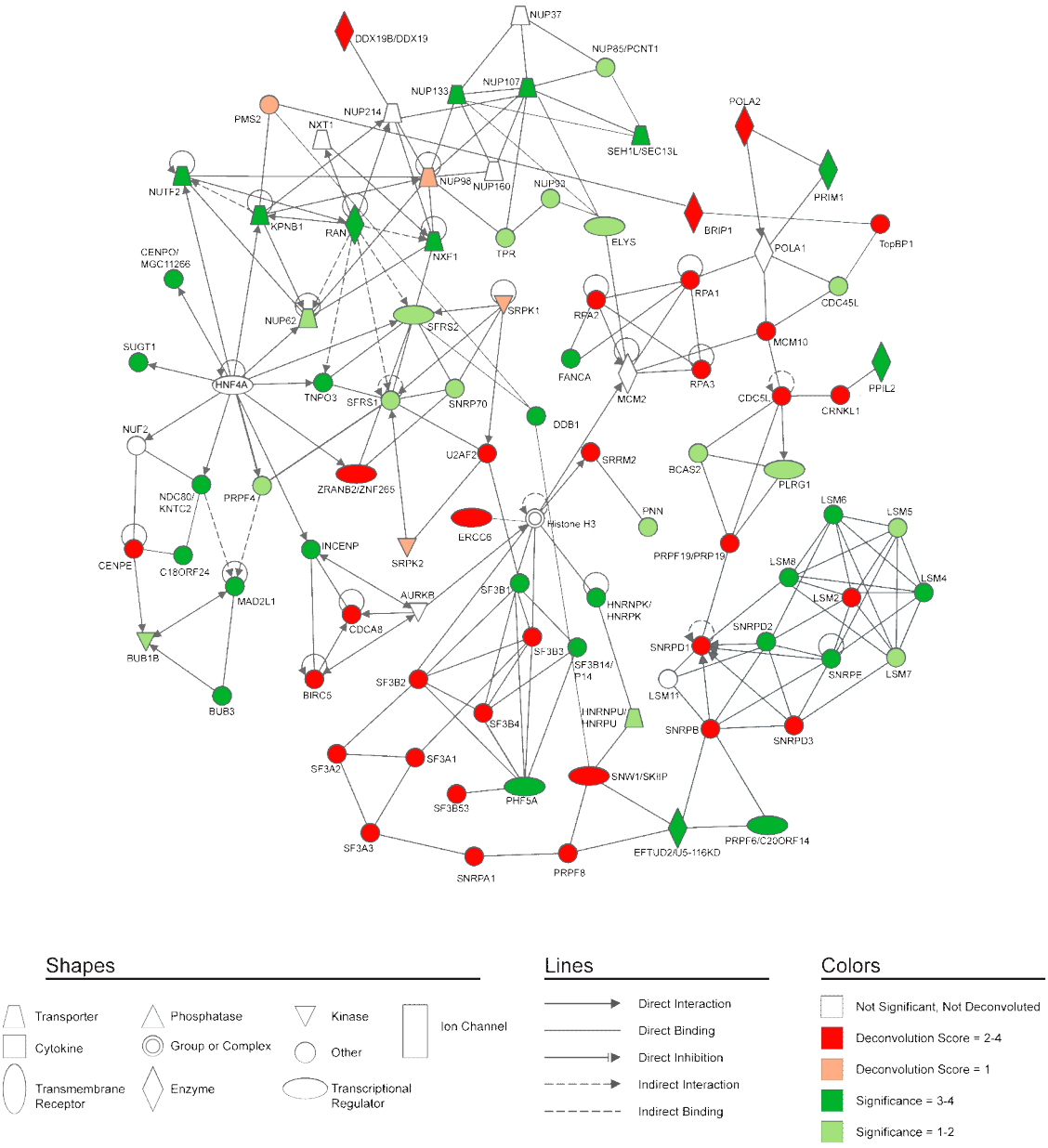


Figure S2. Interconnections between functional groups linked to genome maintenance. The functional groups: Pericentric chromatin binding, nuclear pore, mRNA processing, DNA replication and DNA repair were examined for direct protein-protein interactions using Ingenuity pathway analysis.

Description of Supplemental Tables

1. **Supplemental Table ST1..... download**
   γ-H2AX signal and cell cycle distribution for the genome
   The following list shows the percentage of γ-H2AX positive cells, cell number and cell cycle distribution for all genes tested within the original screen. Genes are arranged in order of descending average γ-H2AX percentage. Where no cell cycle information is available, the PI intensity varied beyond that which could be corrected in our analysis program because of insufficient cell numbers or variation between duplicates. Average control values for all plates are located at the top of the list, before the genome data.

2. **Supplemental Table ST2..... download**
   List of genes with significant γ-H2AX values
   The following list shows genes that had a γ-H2AX signal significantly above the calculated noise of the screen. The significance column designates the category of assigned confidence. The confidence groupings ranged from 4 to 1, with the highest level of confidence in group 4. See supplemental details and text for further description of these confidence levels.

3. **Supplemental Table ST3..... download**
   Genes that cause extensive cell death
   The following is a list of genes that when knocked down lead to widespread cell death. This was defined based on the number of cells/well in the original genome-wide screen left 72 hours after siRNA transfection. For these genes, less than 400 cells were left/well after transfecting the corresponding siRNA, a value less than 50% the number of cells originally transfected. The gene/protein product localization and function were categorized using Ingenuity Pathway Analysis (Ingenuity Systems, www.ingenuity.com). These genes were eliminated from subsequent enrichment analyses.

4. **Supplemental Table ST4..... download**
   Categories of enrichment as determined by the DAVID bioinformatic database and Ingenuity pathway analysis
   The genes in significance group 4 were analyzed for enrichments in GO terms (biological process, cellular complex, molecular function), protein information resource keywords, or the OMIM/Genetic Association disease datasets. Significantly enriched categories (p value < 0.05) are shown, along with the corresponding genes. Highlighted categories are shown in Figure 2C.

5. **Supplemental Table ST5..... download**
   List of deconvoluted genes and their corresponding γ-H2AX values
   The genes for which we individually tested four different siRNAs are shown, along with the percentage of cells considered γ-H2AX positive and the number of viable cells/well for each siRNA tested. Duplicate measurements for γH2AX signal are represented by H2AX 01A and 01B, and duplicate measurements for

cell number are represented by Cell# 01A and 01B. Averages of both values are also shown together with standard deviation. The average control values for all plates are located at the top of the list, before the deconvolution data.

6. **Supplemental Table ST6..... download**
   Individual components of gene modules and networks enriched amongst screening hits
   The individual modules identified by DAVID Bioinformatic Database and linked by Ingenuity Pathway Analysis are shown in Figures 4 and S1. Components of these modules are shown here, along with their gene ID numbers.

7. **Supplemental Table ST7..... download**
   γ-H2AX inducing mRNA processing genes
   The following genes were identified as RNA processing genes by the GO biological process term 0006396. The functional assignments were designated by literature mining and use of the Biobase Knowledge Library.

8. **Supplemental Table ST8..... download**
   Charcot-Marie-Tooth Gene classification and effects on γ-H2AX in original screen and deconvolution analysis
   CMT genes and classifications were identified by literature mining. HeLa deconvolution results were taken from ST5 or added from further gene testing found in ST10. U2OS individual siRNA γH2AX percent positive values are not shown.

9. **Supplemental Table ST9..... download**
   γ-H2AX values from retesting mRNA processing genes and the effect with RNAseH treatment
   The average γH2AX value after gene knockdown is shown with the standard error calculated from three replicates. Genes highlighted in gray were shown to have consistent decreases in γH2AX after RNAseH treatment. A consistent decrease was classified as any gene that showed a decrease with at least three siRNAs.

10. **Supplemental Table ST10..... download**
    Individual CMT gene and siRNA information
    CMT siRNAs were retested for γH2AX induction in HeLa cells along with verification of gene knockdown by Q-PCR. Individual siRNA were tested for homologous recombination defects and damage sensitivities. Phenotypes are listed below. If more than one siRNA was tested, both phenotypes are listed with the siRNA shown in Figure 6 designated as (shown).

11. **Supplemental Table ST11..... download**
    Table of genes identified in this screen as compared to other screening results
    Hits from the screens shown below or their orthologs were compared to the genes found in our screen within our significance groups or which were confirmed by deconvolution with at least 1 siRNA.