The Protein Composition of Mitotic Chromosomes Determined Using Multiclassifier Combinatorial Proteomics

Shinya Ohta,1,6 Jimi-Carlo Bukowski-Wills,1,2,6 Luis Sanchez-Pulido,4 Flavia de Lima Alves,1 Laura Wood,1 Zhuo A. Chen,1 Melpi Platani,1 Lutz Fischer,1 Damien F. Hudson,1,5 Chris P. Ponting,4 Tatsuo Fukagawa,3 William C. Earnshaw,1,* and Juri Rappsilber1,*

1Wellcome Trust Centre for Cell Biology
2Centre for Systems Biology at Edinburgh, School of Biological Sciences
University of Edinburgh, Edinburgh EH9 3JU, UK
3Department of Molecular Genetics, National Institute of Genetics and The Graduate University for Advanced Studies, Mishima, Shizuoka 411-8540, Japan
4MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK
5Present address: Murdoch Children’s Research Institute, Royal Children’s Hospital, Melbourne, Victoria 3052, Australia
6These authors contributed equally to this work

*Correspondence: bill.earnshaw@ed.ac.uk (W.C.E.), juri.rappsilber@ed.ac.uk (J.R.)
DOI 10.1016/j.cell.2010.07.047

SUMMARY

Despite many decades of study, mitotic chromosome structure and composition remain poorly characterized. Here, we have integrated quantitative proteomics with bioinformatic analysis to generate a series of independent classifiers that describe the ∼4,000 proteins identified in isolated mitotic chromosomes. Integrating these classifiers by machine learning uncovers functional relationships between protein complexes in the context of intact chromosomes and reveals which of the ∼560 uncharacterized proteins identified here merits further study. Indeed, of 34 GFP-tagged predicted chromosomal proteins, 30 were chromosomal, including 13 with centromere-association. Of 16 GFP-tagged predicted nonchromosomal proteins, 14 were confirmed to be nonchromosomal. An unbiased analysis of the whole chromosome proteome from genetic knockouts of kinetochore protein Ska3/Rama1 revealed that the APC/C and RanBP2/RanGAP1 complexes depend on the Ska complex for stable association with chromosomes. Our integrated analysis predicts that up to 97 new centromere-associated proteins remain to be discovered in our data set.

INTRODUCTION

As cells enter mitosis, chromosomes undergo a remarkable series of physiological and structural transformations known as chromosome condensation. This process involves individualization of the chromosomal territories to create the characteristic mitotic chromosome morphology and maturation of the kinetochores so that chromosomes can align and segregate on the mitotic spindle. Our understanding of the mechanisms underlying chromosome condensation is still fragmentary. These processes can be fully understood only when all components of mitotic chromosomes have been identified and functional relationships between them determined. We have developed a new approach that we term multiclassifier combinatorial proteomics (MCCP) to do this.

The current list of mitotic chromatin proteins reported in proteomic studies is surprisingly short. Early analyses described 62 and 79 proteins, respectively, in mitotic chromosome scaffolds (Morrison et al., 2002; Gassmann et al., 2005). A later study identified > 250 proteins that bound to sperm chromatin in Xenopus egg extracts in vitro, revealing the kinetochore protein Bod1 (Porter et al., 2007). Other studies identified ∼240 proteins, subsequently corrected to roughly 50 bona fide putative structural proteins (Uchiyama et al., 2005; Takata et al., 2007). In a targeted study, 98 proteins were identified as shared in isolated telomeres from wild-type and ALT cells (Dejardin and Kingston, 2009). Despite these efforts, currently available proteomics reports miss a significant fraction of known mitotic chromosomal proteins, particularly kinetochore components.

Biochemical analysis of important chromosome substructures such as kinetochores is extremely challenging. The kinetochore is one of the most complex cellular substructures (Cheeseman and Desai, 2008), with over 120 constituents described by a range of approaches (Earnshaw and Rothfield, 1985), including targeted proteomic studies (Obuse et al., 2004; Foltz et al., 2006; Okada et al., 2006; Hori et al., 2008; Amano et al., 2009). Biochemical dissection of kinetochores is complicated by the fact that it is not known to what extent the constituent protein complexes can be recovered in soluble form from chromosomes with their relevant intermolecular associations intact. As described here, those problems can be circumvented by purifying and analyzing whole mitotic chromosomes.
Purifying large cellular structures or organelles free of contaminants is virtually impossible. Genuine components have been distinguished from contaminants in such preparations by subtractive (Schirmer et al., 2003) or quantitative (Foster et al., 2003) proteomics by determining the difference between two near-identical fractions, one enriched and the other depleted of the target structure. In protein correlation profiling, a set of known components was used to define a common intensity profile across neighboring biochemical fractions from sucrose gradients during purification of organelles and this was used to select other proteins that show a similar profile (Andersen et al., 2003). These methods do not recognize cellular background proteins that adhere to the structure of interest due to nonspecific hydrophobic or electrostatic interactions. This type of contamination is particularly relevant for vertebrate mitotic chromosomes.

In the present study, we identified approximately 4000 polypeptides in highly purified chromosomes. We developed a statistical approach for analysis of proteomic data to confirm which known and uncharacterized proteins from this long list are chromosomal. An experimental test of our method led to the identification of 32 chromosomal proteins, including 13 kinetochore-associated proteins. Key to our analysis is the innovative use of subtractive (Schirmer et al., 2003) or quantitative (Foster et al., 2003) proteomics by determining the difference between two near-identical fractions, one enriched and the other depleted of the target structure. In protein correlation profiling, a set of known components was used to define a common intensity profile across neighboring biochemical fractions from sucrose gradients during purification of organelles and this was used to select other proteins that show a similar profile (Andersen et al., 2003). These methods do not recognize cellular background proteins that adhere to the structure of interest due to nonspecific hydrophobic or electrostatic interactions. This type of contamination is particularly relevant for vertebrate mitotic chromosomes.

RESULTS

Identification of the Proteome of Isolated Mitotic Chromosomes

We isolated mitotic chromosomes from chicken DT40 cells by a refinement of the polyamine method (Lewis and Laemmli, 1982) (Figures 1A and 1B). These preparations are negative for porin in immunoblots (Figure 1C), indicating that mitochondrial contamination (common in chromosome preparations) is low.

Proteomic analysis (Cox and Mann, 2008; de Godoy et al., 2008) of 250 μg of total chromosomal protein identified 4029 proteins in 28 functional categories (Figure 1D and Table S1 available online), including essentially all previously described chromosomal proteins.

When vertebrate cells enter mitosis, the nuclear envelope breaks down and chromosomes are newly exposed to cytoplasmic proteins, organelles and cellular membranes. Since highly positively charged histones contribute ~38% of the chromosome mass and an equivalent amount is highly negatively charged DNA, many charged nonchromosomal proteins exhibit strong adventitious binding to chromosomes. These “hitchhikers” differ from conventional contaminants (e.g., mitochondria), as they are physically associated with the chromosomes before cell lysis and apparently cannot be separated by conventional purification protocols. Thus, the 1331 cytoskeletal, cytoplasmic, mitochondrial, membrane and receptor proteins found in our preparations, may be physically associated with chromosomes following nuclear envelope disassembly, but many not be functionally relevant.

A Classifier Approach to Identify Genuine Mitotic Chromosome Proteins

The presence of hitchhiker proteins complicates the definition of what constitutes a “true” chromosomal protein, as well as the design of biochemical control experiments. For example, comparing mitotic to interphase chromatin is of limited use, since the latter is shielded from the cytoplasm by the nuclear envelope. In such a comparison, cytoplasmic hitchhiker proteins would be scored as mitosis-specific chromosomal proteins.

Here, we describe an approach to study the complex chromosomal proteome that both identifies proteins that merit further study and reveals functional relationships between all chromosomal proteins. We quantify the chromosomal association of each protein in a series of quantitative proteomics experiments, mostly using SILAC technology (Ong et al., 2002). Each experiment provides an independent measure of a protein’s association with mitotic chromosomes, which we term a “classifier.” Integration of the data obtained with all classifiers enables us to detect patterns in the behavior of groups of proteins that reveal shared membership in protein complexes as well as functional dependency relationships.

The experimental protocols that define the five proteomic classifiers are shown in Figure 1E and all classifiers are summarized below.

Classifier I: Abundance estimation

To estimate the amounts of individual proteins in mitotic chromosomes, we used an established protocol (Rappsilber et al., 2002; Ishihama et al., 2005, 2008) to calculate a scaled protein abundance index based on the number of peptides observed and the number of times that each peptide is observed (spectral count) for each protein. This calculation and its validation are discussed in Extended Experimental Procedures.

In the conventional pie chart of Figure 1D, all proteins are weighted equally, independent of their actual abundance in isolated chromosomes. A more informative representation of chromosome composition is obtained by normalizing each class by its mass, obtained by multiplying the estimated abundance by the predicted molecular mass of each protein (Figure 1F). As expected, histones comprise the bulk of mitotic chromosomal protein (48%). Overall 68% of the protein mass is annotated as chromosomal.

Classifier II: Enrichment in Chromosomes

We expected core chromosomal components like histones or structural proteins would be more abundant in isolated chromosomes than in cytoplasmic extracts. The reverse would be true of background proteins. We therefore mixed isolated chromosomes from mitotic DT40 cells grown in light medium with an equal mass of protein from post-chromosomal extracts of parallel cultures grown with heavy SILAC medium (Figure 1E). Classifier II was calculated as the ratio of light/heavy peaks for each protein. Among the 20% most enriched proteins, chromosomal proteins outnumbered nonchromosomal proteins by 3 to 1. Conversely, among the 20% least enriched proteins, background proteins outnumbered chromosomal proteins 6 to 1.

Classifier III: In Vitro Exchange on Chromosomes

We ranked proteins based on their ability to stably bind to chromosomes during an incubation in cytosol. A crude light chromosome fraction obtained by gentle centrifugation was mixed with
an excess of heavy post-chromosomal extract and incubated to allow proteins to exchange at 14°C for 30 min (Figure 1E). The chromosomes were then subjected to rigorous purification. All heavy proteins identified must have bound to the chromosomes during the incubation in vitro. Classifier III is the light/heavy ratio for each protein identified in this experiment.

The most stable chromosomal-associated proteins were histones (average classifier III = 63), topoisomerase IIα (classifier III = 61) and condensin I (average classifier III = 59). Interestingly, condensin II was more exchangeable (average classifier III = 14). In contrast, the ratios for ~75% of ribosomal proteins were in the range from 0.45 to 3.0 (Figure 1G), indicating significant binding to chromosomes from cytosol during the incubation.

Although we sought to optimize purity rather than preserve functionality of chromosomes, the exchange experiment revealed that at least one aspect of kinetochore function was retained in purified chromosomes. Similar to one recent study (Kulukian et al., 2009), kinetochores of the purified chromosomes can recruit spindle checkpoint proteins Mad2, Bub3, and BubR1 from cytosol (Figures 1H and 1I) but not Mad1 (Figure 1J).

**Classifier IV: SMC2 Dependency**

We used a conditional genetic knockout of SMC2 in DT40 SMC2ON/OFF cells (Hudson et al., 2003) to compare the composition of mitotic chromosomes formed in the presence or absence of condensin, which is required for structural integrity of mitotic chromosomes. DT40 SMC2ON cells were cultured in SILAC heavy medium. To obtain chromosomes depleted of condensin, cells grown in SILAClight medium were cultured with doxycycline for 30 hr to shut down SMC2 expression prior to the nocodazole block (SMC2OFF). Equal numbers of mitotic cells from the two different populations were mixed and mitotic chromosomes

**Figure 1. Proteomic Analysis of Mitotic Chromosome Proteins**

(A) Outline of chromosome isolation procedure. (B) Silver-stained gel of isolated chromosomes. (C) Immunoblot with markers for mitochondria (porin) and chromosomes. (D) The 28 classes of proteins found in chromosomes. (E) Experimental designs to prepare samples for determination of classifiers I (abundance), II (enrichment in chromosomes versus post-chromosomal extract), III (exchange onto chromosomes from post-chromosomal cytosol), IV (dependency on SMC2/condensin) and V (dependency on Ska3/RAMA1). (F–J) (F) Estimated percentages of total chromosomal protein mass in the major classes of proteins. Sample spectra used to calculate classifier III (exchange) values for (G) Ribosomal protein SA (a hitchhiker protein), (H) MAD2, (I) MAD1. Spectra are plotted as the peak intensity (a measure of abundance) against the mass divided by charge. Peaks of light peptides from chromosomes are indicated by blue bars and heavy peptides bound from cell extracts during the incubation are indicated by red bars. See also Figure S1.
isolated (Figure 1E). Classifier IV is the heavy/light ratio (SMC2<sup>ON</sup>/SMC2<sup>OFF</sup>) for each of the proteins identified in this experiment.

SMC2-depleted chromosomes contained 5.8% of the wild-type level of SMC2. It is not possible to isolate chromosomes from cultures completely lacking SMC2 as SMC2 is an essential gene and dead cells do not accumulate in mitosis. These chromosomes were similarly depleted of all condensin I and II subunits.

Classifier V: Ska3/Rama1 Dependency
To demonstrate the targeted analysis possible with our approach, we compared the association of kinetochore proteins with chromosomes in cells with or without Ska3/Rama1/C13orf3. Ska3/Rama1, which was identified in this analysis as a chromosomal protein, was described in several recent publications (Daum et al., 2009; Gaitanos et al., 2009; Raaijmakers et al., 2009; Theis et al., 2009; Welburn et al., 2009).

To obtain chromosomes depleted of Ska3 complex, we isolated a genetic knockout of the Ska3/Rama1/C13orf3 gene (Figures S5A and S5B). Ska3/Rama1<sup>−/−</sup> cells (homozygous knockout is viable) were grown in SILAC light medium (Figure 1E). Equal numbers of Ska3/Rama1<sup>−/−</sup> and wild-type DT40 (cultured in SILAC heavy medium) mitotic cells from the two different populations were mixed and mitotic chromosomes isolated. Classifier V is the heavy/light ratio (wild-type/Ska3/ Rama1<sup>−/−</sup>) for each protein identified in this experiment.

Classifier VI: Domain Analysis
We added an additional nonproteomic classifier to our analysis using the protein domains found in chromosomal and nonchromosomal proteins (red/pink and green wedges in Figure 1D). This made use of bioinformatic analysis in order to segregate chromosomal from nonchromosomal proteins, but importantly did not consider a protein’s relevance to mitosis. We counted how often each domain was observed in chromosomal and nonchromosomal proteins and assigned it this frequency as a score (Table S2). Classifier VI was then determined for each protein based on the sum of its domain scores.

Multiclassifier Combinatorial Proteomics
Traditional one-dimensional analysis (e.g., sorting the various proteins according to their value for each classifier) was of limited utility, as the data lacked a clear boundary between chromosomal (red/pink in Figure 2A) and nonchromosomal proteins (green in Figure 2A) for each classifier (Figure 2A and Figures S2A–S2E).

By contrast, when classifiers were combined, our ability to identify chromosomal proteins was vastly improved. For example, an enrichment of centromeric or chromosomal proteins relative to nonchromosomal proteins was obtained when classifiers I (abundance) and II (enrichment) were plotted (Figures 2C–2E). The clustering of protein complex subunits in this plot (Figure 2D) reflects both their relative stoichiometry...
(x axis), and the similar degree to which subunits in a complex are all present either on or off chromosomes (y axis). Members of the APC/C, Ndc80 and Mis12 complexes form closely knit clusters. It is important to note that this was achieved in the context of entire chromosomes and without requiring solubilization of the complexes.

We used random forest (RF) analysis, a machine learning approach, to progress beyond two-dimensional analyses and integrate the information present in all proteomics classifiers. This analysis offered two powerful benefits. First, it enabled us to work with data sets that contain missing values. This is a significant advantage in proteomics studies where not every protein is observed in every experiment, as seen in Figure 2A and Figure S2F. Second, RF analysis allowed us to use any descriptor of our proteins as a classifier and integrate it into our overall analysis. Here, we also included a bioinformatic analysis of the distribution of protein domains in our analysis distinguishing chromosomal from nonchromosomal proteins (classifier VI).

In brief, RF is a decision tree analysis that separates data sets into “true” and “false” groups. The decision trees are trained on defined data sets and randomly built to optimize the separation between them. Analysis of the experimental data set then occurs by running each protein through all trees and adding up its overall RF score (i.e., the fraction of trees that scored it as “true”). RFs perform much better on training data than application data, so their performance is evaluated by ten-fold cross-validation. The training data are split into random sections of 90% for training and 10% for evaluation, so that successively the entire set is used for evaluation. Here, the two training data sets chosen were “nonchromosomal” (green wedges in Figure 1D) and “chromosomal” (red + pink wedges in Figure 1D), and the RF score for a given protein is the fraction of trees that scored it as “chromosomal.”

RF analysis readily discriminated chromosomal from nonchromosomal proteins. In the RF rug plot of Figure 2A, which represents the ranked list of proteins generated by RF analysis, the left side is predominantly red, while the right side is predominantly green. To reach the 500th chromosomal protein on the RF-ranked list only 229 nonchromosomal proteins are included (Figure 2B). In contrast, 410-671 nonchromosomal proteins are rejected. RFs perform much better on training data than application data, so their performance is evaluated by ten-fold cross-validation. The training data are split into random sections of 90% for training and 10% for evaluation, so that successively the entire set is used for evaluation. Here, the two training data sets chosen were “nonchromosomal” (green wedges in Figure 1D) and “chromosomal” (red + pink wedges in Figure 1D), and the RF score for a given protein is the fraction of trees that scored it as “chromosomal.”

RF analysis readily discriminated chromosomal from nonchromosomal proteins. In the RF rug plot of Figure 2A, which represents the ranked list of proteins generated by RF analysis, the left side is predominantly red, while the right side is predominantly green. To reach the 500th chromosomal protein on the RF-ranked list only 229 nonchromosomal proteins are included (Figure 2B). In contrast, 410-671 nonchromosomal proteins are rejected.

The advantage of combining classifiers can be statistically expressed by ROC curves (Figure S3A), with increased area under the curve (AUC) for our combined analysis when compared to each of the individual classifiers (AUCRF(L-VI) = 0.81, AUCCRF(L-V) = 0.41-0.76). The combined classifiers assigned 88.8% of our gold standard, the 125 centromere proteins, correctly as chromosomal, at the cut-off that minimizes misassignment of chromosomal versus nonchromosomal proteins. This specificity was further improved when bioinformatic domain analysis (cVI) was integrated with our proteomic classifiers (AUCRF(L-VI) = 0.97; identification of the first 500 chromosomal proteins yields 17 nonchromosomal proteins, Suppl. Figures S3B and S3C, proteins lacking known domains are excluded from this boost, Figure S3D). Now, 92.4% of the centromere proteins were assigned as chromosomal. In summary, RF analysis provides us with a tool for productively combining the outcomes of our individual proteomics classifier experiments and further empowering our analysis by including data from other sources.

If results of a random forest based on the five proteomic classifiers plus the bioinformatics-based classifier VI were plotted against those from the initial random forest analysis (Figure 3A), a near-perfect separation of the training data was achieved. Only a single chromosomal protein of the training set and two nonchromosomal proteins were misassigned when placing manually a separation line.

Using ten-fold cross-validation, we found that 118 centromere proteins positioned right and only 7 left of the separation line (Figure 3B). This compares to 14 and 9 centromere proteins being missed when using the one-dimensional ranked lists by classifiers I-V and I-VI, respectively. Accepting the line as a threshold returns known centromere proteins with a yield of 94.4% and all other chromosomal proteins with 93.1% success. In contrast, 83.1% of nonchromosomal proteins are rejected. Thus, the classifier approach is sufficiently powerful to suggest chromosomal proteins from among hitherto uncharacterized proteins.

Identification of New Chromosomal Proteins

To test the predictive power of our RF analysis, we cloned and tested the location in mitosis of 50 previously uncharacterized proteins including 15 without known domains: 34 predicted to be and 16 predicted not to be on chromosomes in mitosis.

Reasoning that important proteins would be conserved, we expressed GFP-tagged human homologs of these chicken proteins in U2OS cells. Remarkably, 30 of 34 cloned proteins from the chromosomal region were confirmed as chromosomal, contrasting with only 2 of the 16 predicted to be nonchromosomal. This confirms the power of our analysis and indicates a success rate of 88%, with 44 of 50 tagged proteins localized as predicted. Of 50 newly cloned proteins, 13 were associated with kinetochores in mitosis, 12 had a more general distribution on mitotic chromosomes and 7 others were perichromosomal, a class whose new members we propose to term chromosome periphery proteins (cPERPs A-G) (Figures S4B–S4G and Table S3). The chromosome periphery (perichromosomal layer) is enriched in ribosomal and nucleolar hitchhiker proteins, and is of unknown function (Van Hooser et al., 2005).

The new centromere proteins all appeared to localize to the outer kinetochore, relative to CENP-C and HEC1 as standards (Figures 4B–4J, Figure S4A). In keeping with established nomenclature, we propose to name these proteins CENP-Y, CENP-Z and CENP-T through CENP-33. Beyond ‘Z’, the 26th letter of the basic modern Latin alphabet, we propose to designate the new proteins with numbers starting with CENP-27.

Functional Analysis of New Kinetochore-Associated Proteins

We focused our initial functional analysis on kinetochore proteins. Clustering analysis (Gentleman et al., 2004) allowed us to combine data for proteins identified by all classifiers, and
look for informative groupings. This revealed a striking tendency for functionally related proteins to form clusters, as exemplified by members of the NDC80, CPC, Nup107-160 and APC/C complexes (Figure 4). Interestingly, our clustering sorted CENP-27 as a component of the APC/C. This was confirmed and the protein named APC16 in three recent reports (Hutchins et al., 2010; Kops et al., 2010; Hubner et al., 2010). To further test the predictive value of cluster analysis for proteomic data sets, we examined two kinetochore proteins in greater detail.

**Ska3/Rama1 and Functional Analysis of Kinetochore Subcomplexes**

C13orf3 was located adjacent to Ska2 in the cluster analysis. This protein, now known as Ska3/Rama1, has been suggested to be involved in microtubule attachment to kinetochores (Gaitanos et al., 2009; Raaijmakers et al., 2009; Welburn et al., 2009) or coordination of the spindle checkpoint response (Daum et al., 2009; Theis et al., 2009). We analyzed the kinetochore proteome in the presence or absence of Ska3/Rama1 (defined as classifier V) in order to determine the role of this protein in kinetochore structure (Figure 5).

A map of the Ska3/Rama1 locus in DT40 is shown in Figures S5A and S5B, together with a targeting strategy for inactivating the gene. The Ska3/Rama1 gene is not essential for life in DT40 cells (Figure S5 C). However, these cells struggle to achieve a normal chromosome alignment, and show a ~3x increase in mitotic index (Figure S5D and S5E), a ~3x increase in the percentage of apoptotic cells and a ~6x increase in the number of bi-nucleated cells.

Proteomic analysis of isolated chromosomes revealed that loss of Ska3/Rama1 was accompanied by the loss of Ska1 and Ska2. Loss of the Ska complex caused no systematic changes in the chromosomal association of proteins of the constitutive centromere-associated network (CCAN), Knl-1/Mis12/Ndc80 (KMN), Mis18, Ndc80, CPC, and Nup107-160 kinetochore subcomplexes. However, striking changes were seen in the levels of the APC/C, RanBP2/RanGAP1, spindle checkpoint, Rod/Zw10/Zwilch (RZZ), and dynein/dynactin complexes. We confirmed that the RanBP2/RanGAP1 complex is indeed depleted from kinetochores when Ska3/Rama1 is deleted in HeLa cells (Figure S5 F and S5G). Attempts to confirm the specific kinetochore depletion of the APC/C were uninformative, as we were unable to reproducibly obtain kinetochore staining for the APC/C in HeLa cells using four independent antibodies.

We conclude that combining genetic and SILAC analysis provides a powerful new method for analysis of multicomplex protein superstructures.

**A Protein Involved in Chromosome Alignment and Spindle Organization**

A second new kinetochore-associated protein, CENP-32/C9orf114, was sorted on our kinetochore cluster diagram next to CLASP1 and CLASP2, two paralogues known to be involved in the regulation of microtubule dynamics. Like CLASPs, CENP-32/C9orf114 mapped to the outer kinetochore (Figure 6A), and its depletion caused a significant accumulation of cells in prometaphase (Figure S6A) with misaligned chromosomes (Figure 6B). These cells frequently had bipolar spindles, however 60% of those spindles exhibited remarkable abnormalities where centrosomes appeared to have detached from the poles (Figures 6C and 6D and Figure S6B). In one remarkable case, the centrosomes appeared at the midzone of a bipolar spindle (Figure 6C).
DISCUSSION

Multiclassifier Combinatorial Proteomics

The approach described here for analysis of the proteome of vertebrate mitotic chromosomes can be used to study any complex proteome. The key approach of combining classifier data can in principle be expanded indefinitely and can include nonproteomics data sets such as the bioinformatic protein domain analysis used here. Other classifiers that could be used in the future include microarray, protein interaction...
(e.g., two hybrid screens or pull-down), protein phosphorylation, and localization data and, indeed, data from any experimental approach in which the proteins of interest are sorted systematically.

We first showed that plotting pairs of classifiers against one another improved our ability to delineate potential chromosomal proteins. As that approach could not be generalized when the number of classifiers exceeded three, we adopted a random forest (RF) analysis approach. This allowed us to integrate information from all classifiers into decision trees on which known and unknown proteins could be classified. Importantly, RF analysis handles missing values systematically. This is crucial when not every protein is observed in every experiment. In contrast, cluster analysis, which has been used both in this study and in other recent work (Theis et al., 2009; Neumann et al., 2010), can only integrate data for proteins that have a value for every classifier.

**Integrity of the Isolated Chromosomes**

Our methods focused on optimizing the purity of the chromosomes. Thus, our list of proteins is likely to represent the minimal, stably associated components of mitotic chromosomes. Nonetheless, kinetochores of isolated chromosomes retain some function, as judged by their ability to recruit components of the mitotic checkpoint complex from cytoplasm. This may be because chromosomes were isolated from nocodazole-treated cells, with kinetochores actively engaged in spindle checkpoint signaling.

**New Insights into Kinetochore Functional Organization**

Remarkably, although no biochemical enrichment for centromeres was performed, our data set contained all known centromeric subcomplexes, with peptides from 125 reported centromere proteins (eight present as multiple isoforms) (Table S1). We identified all members of the CCAN, KMN, Mis12 and Mis18 complexes and all members of the RZZ complex except Zwint (which is not yet annotated in the chicken genome). Our success in identifying centromere and telomere proteins may be explained because 66 of the 78 chicken DT40 chromosomes are microchromosomes whose purification provides a natural enrichment for centromeres and telomeres since the chromosome arms are so short.

We combined genetics with whole proteome analysis in order to identify complexes and structural dependencies in their “native environment” (e.g., kinetochore proteins in actual kinetochores). Chromosomes lacking Ska3/Rama1 were depleted for the entire Ska complex, confirming that these three proteins are interdependent for chromosome binding. Similarly, depletion of key condensin subunit SMC2 caused a loss of all seven subunits of the condensin I and II complexes from chromosomes. Importantly, this analysis did not require tagging of any proteins or attempts to solubilize functional complexes from large subcellular structures.

In addition to these primary effects of depletion, loss of Ska3/Rama1 also caused a significant secondary depletion of the APC/C and RanBP2/RanGAP1 complexes from chromosomes but had no consistent effect on most other kinetochore proteins. Importantly, all members of the secondary-depleted complexes...
Figure 6. Initial Characterization of CENP-32

(A) Localization of GFP-CENP-32 (panel 2, green) in the kinetochore relative to DNA (panel 1, blue) and CENP-C (panel 3, red).

(B) CENP-32 RNAi causes problems with chromosome alignment. DNA, red; CENP-C, green; microtubules, white.

(C) CENP-32 RNAi causes detachment of centrosomes from bipolar spindles. DNA (panels 1,5,9,13), blue; pericentrin (panels 2,6) or $\gamma$-tubulin (panels 10,14), green; $\alpha$-tubulin (panels 3,7,11,15), red.

(D) Quantitation of spindle phenotypes following CENP-32 RNAi.

(E) Schematic representation of domain architectures for new kinetochore proteins tagged in this study (drawn to approx. scale). The locations of domains are according to Pfam and SMART family databases (Letunic et al., 2004; Finn et al., 2010), complemented by REP web server analysis (Andrade et al., 2000). The localization of the PEHE domain has been assigned from (Marin, 2003). Phylogenetic distributions of proteins are indicated in blue, yellow, green and violet for Metazoa, Fungi, Plantae and Archaea, respectively. CENP-35 has been truncated. Abbreviations: BCNT, Bucentaur or craniofacial development protein; Fbox, cyclin-F motif; G-Patch, Glycine-rich nucleic binding domain; JmjC, jumonji C domain; LRR, Leucine-rich repeats; NuA4, Nucleosomal acetyltransferase of H4; OB, Oligonucleotide/oligosaccharide Binding; PEHE, Pro-Glu-His-Glu conserved region; PHD, Plant HomeoDomain; RA, Ras associated (RaiGDS/AF-6) domain; SPOUT, spoU and trmD RNA methyltransferase; WD40, WD/beta-transducin repeats; Zf-AD, Zinc finger-associated domain; and Zf-C2H2, Classical zinc finger Cys(2)His(2).

See also Figure S6.
also behaved coordinately. Our results (e.g., the behavior of CENP-27/APC16 as a component of the APC/C) confirm the utility of single protein depletion analysis for the identification of protein complexes and determination of their mutual interdependencies for association with chromosomes.

Our data suggest that the Ska complex may provide a docking site for the APC/C in the outer kinetochore. Alternatively, sumoylation by RanBP2 may have a role in APC/C binding to chromosomes. The RanBP2-RanGAP1 complex is known to be involved in kinetochore-microtubule interactions and localization of several spindle checkpoint proteins (Joseph et al., 2004). We note that among the recent spate of publications on Ska3/Rama1, our observations appear to support a role in integration and regulation of the spindle checkpoint response (Daum et al., 2009; Theis et al., 2009).

Our whole-proteome analysis revealed that Cdc20 behaved like a member of the APC/C and was distinct from other components of the spindle checkpoint pathway with respect to its Ska complex dependency. Spindle checkpoint components associate with one another in cytoplasm as a mitotic checkpoint complex (MCC), containing BubR1, Bub3, Cdc20 and Mad2. Our data suggest that once the MCC associates with chromosomes, Cdc20 stably associates with the APC/C.

**What Classes of Kinetochore Proteins Remain to Be Discovered?**

We identified 13 kinetochore-associated proteins among previously uncharacterized proteins and, as discussed below, we predict that many more remain to be described. We therefore asked whether there is any functional relationship between these new proteins. That is, what sorts of kinetochore proteins had been missed in the many previous genetic and biochemical screens? An interesting answer has emerged.

Since the new kinetochore proteins were identified solely based on their occurrence in chromosomes, they could potentially represent a wide range of functions. Nevertheless, it is striking that five of the new centromere proteins (namely, CENP-28, –29, –31, –35, and –36) are subunits of complexes that modify and/or bind histones (Figure 6E). Yeast orthologs of two of these proteins (namely, CENP-28/C1orf149 and CENP-29/CDFP1) contribute to NuA4 histone acetyltransferase (HAT) and SWR1 ATP-dependent chromatin remodelling complexes, respectively. These complexes are known to share components (Wu et al., 2005) and together stimulate the exchange of histone H2A for H2A.Z, following acetylation of H2A or H4 (Altaf et al., 2010). A third centromere- associated protein, CENP-36/MSL1v1, is necessary for the activity of MOF, another HAT, on nucleosomal H4 (Li et al., 2009). Finally, CENP-31/PHD6 and CENP-35/PHF2 each contain PHD (plant homeodomain) zinc fingers, which are usually associated with chromatin-mediated transcriptional regulation. The PHD of CENP-35, which also contains a JmjC (likely histone demethylase) domain, appears to be required for demethylation of H3 at the promoters of ribosomal RNA genes (Wen et al., 2010).

Why were these proteins not discovered earlier as kinetochore-associated? One likely explanation is that they may have essential functions in other chromatin regions as well. Thus, mutations might have pleiotropic phenotypes not recognized as specific for mitosis. Furthermore, their association might depend on a fully assembled kinetochore and thus be lost when attempting other than whole chromosome analysis.

**Characterization of a Kinetochore Protein**

CENP-32 is required both for chromosome alignment and for association of the centrosomes with the poles of the bipolar spindle during metaphase. This latter phenotype is very similar to an unusual spindle morphology phenotype seen in Drosophila cells following depletion of the CLASP homolog Mast/Orbit (Maiato et al., 2002). Indeed, in our analysis, CENP-32 clusters with CLASP1 and CLASP2. A yeast homolog of CENP-32 interacts with CBF5, an enzyme involved in the posttranscriptional modification of rRNA, that was shown to bind to budding yeast centromeres and microtubules (Jiang et al., 1993). Bioinformatics analysis suggests that CENP-32 is a member of the SPOUT family of methyltransferases but is atypical in possessing a possible RNA-binding OB fold inserted into its catalytic domain (Tkaczyk et al., 2007). It is tempting to speculate that CENP-32 may function at kinetochores by interacting with an as-yet unknown RNA.

**How Complex Is the Kinetochore?**

MCCP analysis allows us to predict how many more kinetochore proteins remain to be identified in our data set. In the plot of Figure 3, where the chromosomal proteome is displayed in two dimensions, we found 35% of novel tagged proteins from region R and 6% from region L to associate with the kinetochore during mitosis. These regions have 224 and 287 as-yet uncharacterized novel proteins, respectively. Assuming no bias in the proteins we cloned, this suggests that approximately 97 more kinetochore proteins remain to be discovered. Taking into account the 13 kinetochore-associated proteins confirmed in our work, this roughly doubles the currently known protein complexity of the kinetochore during mitosis, confirming it as one of the most complex cellular substructures.

**Conclusions**

Multiclassifier combinatorial proteomics and the data sets described here open the door to the identification of all functional components of mitotic chromosomes despite the adventitious binding of cellular background proteins during mitosis. Furthermore, MCCP can be extended by adding additional classifiers to delineate protein complexes and define functional dependencies between them in the context of intact mitotic chromosomes. This will serve both as a starting point for systematic determination of the full range of functions involved in mitotic chromosome segregation, and as a basis for the development of detailed structural and functional interaction maps of key chromosomal subdomains. MCCP should also prove useful for the analysis of other cellular structures that lack defined boundaries, e.g., membrane associated complexes like the post-synaptic density.

**EXPERIMENTAL PROCEDURES**

**Preparation of Mitotic Chromosomes**

DT40 cells were incubated with Nocodazole for 13 hr, resulting in a mitotic index of 70%–90%. Mitotic chromosomes were isolated in the polyamine-EDTA buffer system optimized for chicken DT40 cells (Lewis and Laemmli, 2010).
Preparation of Chromosome-Free Mitotic Cell Extracts

Nocodazole blocked DT40 cells were douence-homogenized under hypotonic conditions. Mitotic chromosomes were removed by centrifuging the supernatant twice at 10,000 g and discarding the pellets to prepare a cell extract free of chromosomes.

To measure the ratios between chromosomal and nonchromosomal proteins, SILAC based mass spectrometry was performed with 150 µg of labeled cell extract from 7.0 × 10^9 cells. This mixture was incubated at 14 °C for 30 min. Finally, we re-isolated the chromosomes as described above.

Mass Spectrometry Analysis

Proteins were separated into a high and a low molecular weight fraction by SDS-PAGE, in-gel digested using trypsin (Shevchenko et al., 2006), and fractionated into 30 fractions each using SCX. The individual SCX fractions were desalted using StageTips (Rappsilber et al., 2003) and analyzed via a nanoelectrospray ion source. The six most intense ions of a full MS acquired in the orbitrap analyzer were fragmented and analyzed in the linear ion trap. The MS data were analyzed using MaxQuant (Cox and Mann, 2008) and proteins identified by searching MS and MS/MS data using the MASCOT search engine (Matrix Science, UK). For more details, see the Extended Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, and five tables and can be found with this article online at doi:10.1016/j.cell.2010.07.047.

ACKNOWLEDGMENTS

We dedicate this paper to Uli Laemmli on the occasion of his 70th birthday. We thank Mayumi Takahashi for assistance with preparation of the Ska3/Rama1 knockout, Frauke Melchior for anti-RanBP2 and David Tollervey, Margarete Heck, Robin Allshire, Iain Cheeseman, Kumiko Samejima, Susana Ribeiro and Jan Bergmann for critical reading of the manuscript. This work was supported by a European Community Marie Curie Excellence Grant (JR), the MRC (CPP), EMBO (LSP) and the Wellcome Trust (WCE, JR). WCE and JR are Principal and Senior Research Fellows of The Wellcome Trust, respectively.

Received: December 7, 2009
Revised: May 20, 2010
Accepted: July 14, 2010
Published: September 2, 2010

REFERENCES

Altaf, M., Auger, A., Monnet-Saksouk, J., Brodeur, J., Piquet, S., Cramet, M., Bouchard, N., Lacoste, N., Utley, R.T., Gaudreau, L., et al. (2010). NuA4-dependent acetylation of nucleosomal histone H4 and H2A directly stimulates incorporation of H2A.Z by the SWR1 complex. J. Biol. Chem., in press.
Amano, M., Suzuki, A., Horii, T., Backer, C., Okawa, K., Cheeseman, I.M., and Fukagawa, T. (2009). The CENP-S complex is essential for the stable assembly of outer kinetochore structure. J. Cell Biol. 186, 173–182.
Andersen, J.S., Wilkinson, C.J., Mayor, T., Mortensen, P., Nigg, E.A., and Mann, M. (2003). Proteomic characterization of the human centrosome by protein correlation profiling. Nature 426, 570–574.
Andrade, M.A., Ponting, C.P., Gibson, T.J., and Bork, P. (2000). Homology-based method for identification of protein repeats using statistical significance estimates. J. Mol. Biol. 298, 521–537.
Cheeseman, I.M., and Desai, A. (2008). Molecular architecture of the kinetochore-microtubule interface. Nat. Rev. Mol. Cell Biol. 9, 33–46.
Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372.
Daum, J.R., Wren, J.D., Daniel, J.J., Sivakumar, S., McAvoys, J.N., Potapova, T.A., and Gorbsky, G.J. (2009). Ska3 is required for spindle checkpoint silencing and the maintenance of chromosome cohesion in mitosis. Curr. Biol. 19, 1467–1472.
de Godoy, L.M., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Frohlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 455, 1251–1254.
Dejean, J., and Kingston, R.E. (2009). Purification of proteins associated with specific genomic Loci. Cell 136, 175–186.
Earnshaw, W.C., and Rothfield, N. (1985). Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. Chromosoma (Berl) 91, 313–321.
Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al. (2010). The Pfam protein families database. Nucleic Acids Res. 38, D211–D222.
Foltz, D.R., Jansen, L.E., Black, B.E., Bailey, A.O., Yates, J.R., and Cleveland, D.W. (2006). The human CENP-A centromeric nucleosome-associated complex. Nat. Cell Biol. 8, 458–469.
Foster, L., De Hoog, C.L., and Mann, M. (2003). Unbiased quantitative proteomics of lipids reveals high specificity for signaling factors. Proc. Natl. Acad. Sci. USA 100, 5813–5818.
Gaitanos, T.N., Santamaria, A., Jeyaprakash, A.A., Wang, B., Conti, E., and Nigg, E.A. (2009). Stable kinetochore-microtubule interactions depend on the Ska complex and its new component Ska3/C130rI3. EMBO J. 28, 1442–1452.
Gassmann, R., Henzing, A.J., and Earnshaw, W.C. (2005). Novel components of human mitotic chromosomes identified by proteomic analysis of the chromosome scaffold fraction. Chromosoma 173, 385–397.
Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Goe, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.
Hori, T., Amano, M., Suzuki, A., Backer, C.B., Welburn, J.P., Dong, Y., McEwen, B.F., Wang, W.H., Suzuki, E., Okawa, K., et al. (2008). CCAN makes multiple contacts with centromeric DNA to provide distinct pathways to the outer kinetochore. Cell 135, 1039–1052.
Hubner, N.C., Bird, A.W., Cox, J., Spallettsesser, B., Bandilla, P., Poser, I., Hyman, A., and Mann, M. (2010). Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. J. Cell Biol. 189, 739–754.
Hudson, D.F., Vagnarelli, P., Gassmann, R., and Earnshaw, W.C. (2003). Condensin is required for nonhistone protein assembly and structural integrity of vertebrate mitotic chromosomes. Dev. Cell 5, 323–336.
Hutchins, J.R., Toyoda, Y., Hegemann, B., Poser, I., Heriche, J.K., Sykora, M., Augsburg, M., Hudecz, O., Bushchorn, B.A., Bulkescher, J., et al. (2010). Systematic analysis of human protein complexes identifies chromosome segregation proteins. Science 328, 593–599.
Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol. Cell. Proteomics 4, 1265–1272.
Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kern, J.M., and Frishman, D. (2008). Protein abundance profiling of the Escherichia coli proteome. BMC Genomics 9, 102.

Jiang, W., Middleton, K., Yoon, H.J., Fouquet, C., and Carbon, J. (1993). An essential yeast protein, CBF5p, binds in vitro to centromeres and microtubules. Mol. Cell. Biol. 13, 4884–4893.

Joseph, J., Liu, S.T., Jablonski, S.A., Yen, T.J., and Dasso, M. (2004). The RanGAP1-RanBP2 complex is essential for microtubule-centromere interactions in vivo. Curr. Biol. 14, 611–617.

Kops, G.J., van der Voet, M., Manak, M.S., van Osch, M.H., Naini, S.M., Brear, A., McLeod, I.X., Hentschel, D.M., Yates, J.R., 3rd, van den Heuvel, S., and Shah, J.V. (2010). APC16 is a conserved subunit of the anaphase-promoting complex/cyclosome. J. Cell Sci. 123, 1623–1633.

Kulukian, A., Han, J.S., and Cleveland, D.W. (2009). Unattached kinetochores catalyze production of an anaphase inhibitor that requires a Mad2 template to function. Dev. Cell 16, 105–117.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P., and Bork, P. (2004). SMART 4.0: towards genomic data integration. Nucleic Acids Res. 32, D142–D144.

Lewis, C.D., and Laemmli, U.K. (1982). Higher order metaphase chromosome structure: evidence for metallloprotein interactions. Cell 29, 171–181.

Li, X., Wu, L., Corsa, C.A., Kunkel, S., and Dou, Y. (2009). Two mammalian MOF complexes regulate transcription activation by distinct mechanisms. Mol. Cell 36, 290–301.

Maiato, H., Sampaio, P., Lemos, C.L., Findlay, J., Carmena, M., Earnshaw, W.C., and Sunkel, C.E. (2002). MAST/Orbit has a role in microtubule-kinetochore attachment and is essential for chromosome alignment and maintenance of spindle bipolarity. J. Cell Biol. 157, 749–760.

Marin, I. (2003). Evolution of chromatin-remodeling complexes: comparative genomics reveals the ancient origin of “novel” compensosome genes. J. Mol. Evol. 56, 527–539.

Morrison, C., Henzing, A.J., Jensen, O.N., Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics 1, 376–386.

Porter, I.M., McClelland, S.E., Khouzouli, G.A., Hunter, C.J., Andersen, J.S., McAlinsh, A.D., Blow, J.J., and Swedlow, J.R. (2007). Bod1, a novel kinetochore protein required for chromosome biorientation. J. Cell Biol. 179, 187–197.

Raaijmakers, J.A., Tanenbaum, M.E., Maia, A.F., and Medema, R.H. (2009). RAAM1 is a novel kinetochore protein involved in kinetochore-microtubule attachment. J. Cell Sci. 122, 2436–2445.

Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal. Chem. 75, 663–670.

Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. Genome Res. 12, 1231–1245.

Schirmer, E.C., Flores, L., Guan, T., Yates, J.R., 3rd, and Gerace, L. (2003). Nuclear membrane proteins with potential disease links found by subtractive proteomics. Science 301, 1380–1382.

Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V., and Mann, M. (2006). In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat. Protoc. 1, 2856–2860.

Takata, H., Uchiyama, S., Nakamura, N., Nakashima, S., Kobayashi, S., Sone, T., Kimura, S., Lahmers, S., Granzier, H., Labeit, S., et al. (2007). A comparative proteome analysis of human metaphase chromosomes isolated from two different cell lines reveals a set of conserved chromosome-associated proteins. Genes Cells 12, 269–284.

Theis, M., Slabicki, M., Junqueira, M., Paszkowski-Rogacz, M., Sontheimer, J., Kitzler, R., Heninger, A.K., Glatter, T., Kruismaa, K., Poser, I., et al. (2009). Comparative profiling identifies C13orf3 as a component of the Ska complex required for mammalian cell division. EMBO J. 28, 1453–1465.

Tkaczuk, K.L., Dunin-Horkawicz, S., Purta, E., and Bujińcki, J.M. (2007). Structural and evolutionary bioinformatics of the SPOUT superfamily of methyltransferases. BMC Bioinformatics 8, 73.

Uchiyama, S., Kobayashi, S., Takata, H., Ishihara, T., Hori, N., Higashi, T., Hayashihi, K., Sone, T., Higo, D., Nirasawa, T., et al. (2006). Proteome analysis of human metaphase chromosomes. J. Biol. Chem. 280, 16994–17004.

Van Hooser, A.A., Yuh, P., and Heald, R. (2005). The perichromosomal layer. Chromosoma 114, 377–388.

Welburn, J.P., Grishchuk, E.L., Backer, C.B., Wilson-Kubalek, E.M., Yates, J.R., 3rd, and Cheeseman, I.M. (2009). The human kinetochore Ska1 complex facilitates microtubule depolymerization-coupled motility. Dev. Cell 16, 374–385.

Wen, H., Li, J., Song, T., Lu, M., Kan, P.Y., Lee, M.G., Sha, B., and Shi, X. (2010). Recognition of histone H3K4 trimethylation by the plant homeodomain of PHF2 modulates histone demethylation. J. Biol. Chem. 285, 9322–9326.

Wu, W.H., Alami, S., Luk, E., Wu, C.H., Sen, S., Mizuguchi, G., Wei, D., and Wu, C. (2005). Swc2 is a widely conserved H2AZ-binding module essential for ATP-dependent histone exchange. Nat. Struct. Mol. Biol. 12, 1064–1071.
Cell Culture
DT40 cells with the SMC2 conditional knockout or Skα3/RAMA1 knockout were maintained in RPMI 1640 medium supplemented with 10% (v/v) FBS, 1% CS, 100 μ g/ml penicillin, 100 μ g/ml streptomycin (GIBCO-BRL) at 39°C and 5% CO₂ in a humid incubator. For labeling Lysine and Arginine by carbon 13, cells were maintained in RPMI without L-Lysine and L-Arginine (GIBCO-BRL) supplemented with 10% (v/v) FBS dialyzed 10,000 molecular weight cut-off (Sigma), 100 μ g/ml U-¹³C₀-L-Lysine:2HCl, 30 μ g/ml U-¹³C₀-L-Arginine:HCl, 100 μ g/ml penicillin, 100 μ g/ml streptomycin (GIBCO-BRL) at 37°C and 5% CO₂ in a humid incubator.

For RNAI, U2OS or HeLa cells in exponential growth were seeded onto coverslips and grown overnight at 37°C in RPMI/10% FBS (GIBCO-BRL), 5% CO₂. 200 ng of plasmid DNA was administered to the cells at 30%–40% confluence by transfection with FuGENE6 (Roche) in complete medium without antibiotics. Cells were maintained in this medium for 24 hr.

Preparation of Mitotic Chromosomes
DT40 cells at densities of 8–10 × 10⁸/ml were incubated with 0.5 μ g/ml Nocodazole for 13 hr, resulting in a mitotic index of 70%–90%. Mitotic chromosomes were isolated in the polyamine-EDTA buffer system optimized for chicken DT40 cells (Lewis and Laemmli, 1982). 19.3 OD₂₆₀ units were obtained from pooling the material of four independent preparations totaling 7.5 × 10⁹ DT40 cells, sonicated for 15 min, the proteins precipitated with TCA and the pellet was solubilized in SDS-polyacrylamide gel electrophoresis (SDS-PAGE) sample buffer.

Preparation of Chromosome-Free Mitotic Cell Extracts
Nocodazole blocked cells were treated with hypotonic conditions using 40mM KCl and dounced with 10 vigorous strokes. Mitotic chromosomes were removed by centrifuging the supernatant twice at 10,000 x g and discarding the pellets to prepare a cell extract free of chromosomes.

To measure the ratios between chromosomal and nonchromosomal proteins, SILAC based mass spectrometry was performed with 150 μ g of labeled cell extract from 7.0 × 10⁶ cells and 150 μ g of nonlabeled proteins contained in isolated chromosomes from 2.0 × 10⁶ cells.

To measure the exchange ratio, we isolated mitotic chromosomes from roughly 1.0 × 10⁶ cells. Mitotic chromosomes were pelleted after centrifuging at 3,000 x g and mixed into 10 ml cell extract that were made from 3.0 × 10⁹ cells. This mixture was incubated at 14°C for 30 min. Finally, we re-isolated the chromosomes as described above.

Control Experiments
Control experiments in which beads loaded with DNA were exposed to cytoplasm did not clearly differentiate between chromosomal proteins and contaminants (data not shown). Although those experiments did identify many obvious contaminants that can bind to DNA, they also yielded many true chromosomal proteins (e.g., histones and topoisomerases) whose presence in the cytoplasmic fractions was likely due to unavoidable nuclear damage during extract preparation. Likewise, although procedures have been published for the proteomic analysis of interphase chromatin from C. elegans sperm (Chu et al., 2006) and from rice (Tan et al., 2007), comparing mitotic to interphase chromatin is of limited use. Interphase chromatin is shielded from the cytoplasm by the nuclear envelope. Thus in such a comparison, many cytoplasmic hitchhikers proteins would be scored as mitosis-specific chromosomal proteins.

Sample Preparation and Fractionation for Mass Spectrometry
300 μg protein were separated into a high and a low molecular weight fraction on a 1mm 4%–12% tris-glycine NuPAGE gel in MOPS buffer (Invitrogen), fixed with 50% methanol, 5% acetic acid, and stained with Invitrogen’s Colloidal Blue kit. The two fractions were excised, and proteins were digested using trypsin at an enzyme-to-protein ratio of 1:50 as described (Shevchenko et al., 1996). In brief, proteins were reduced in 10 mm DTT for 30 min at 37°C, alkylated in 55 mm iodoacetamide for 20 min at room temperature in the dark, and digested overnight at 37°C with 12.5 ng/μl trypsin (proteomics grade, Sigma). The digestion medium was then quenched with phosphoric acid (Sigma-Aldrich) until pH 3.0 was reached. The sample was then diluted with Buffer A (5 mM KH₂PO₄, 10% acetonitrile, pH 3.0) to 20% acetonitrile and fractionated using strong cation exchange chromatography (2.1-mm × 20-cm polysulfethyl A column, Poly LC, Columbia, MD) on an UltiMate 3000 HPLC instrument (Dionex UK Ltd., Surrey, UK). Peptides were separated using Buffer A, Buffer B (Buffer A with 1 M KCl), a flow rate of 200 μl/min, and a 50 min gradient as follows: 0–1 min, 0% Buffer B; 1–5 min, raise to 0.5% Buffer B; 5–14 min, raise to 1.0% Buffer B; 14–41 min, raise to 70% Buffer B (curve 7 equation, CHROMEOLEON software version 6.80, Dionex UK Ltd.); 41–42 min, 70% Buffer B. Fractions were collected every 1 min. All the fractions were diluted with 0.1% TFA at a ratio of 1:1 and desalted using StageTips (Rappsilber et al., 2003) for subsequent LC-MS/MS analysis.

Mass Spectrometric Analysis
An LTO-Orbitrap mass spectrometer (Thermo Fischer Scientific) coupled on line to an Agilent 1100 binary nanopump and an HTC PAL autosampler (CTC Analytics) was used to analyze the fractionated peptides. The analytical column with a self-assembled particle frit (Ishihama et al., 2002) was prepared by packing C18 material (ReproSil-Pur C18-AQ 3 μm, Dr. Maisch, GmbH) into a spray
emPAI (Rappsilber et al., 2002; Ishihama et al., 2005, 2008) in slightly modified form. Briefly, emPAI is calculated based on the number of peptides observed (Table S4) and the number of times that each peptide is observed (spectral count) for each protein as follows:

\[ \text{emPAI} = 10^{\left(\frac{\text{spectral count}}{\text{protein-specific normalization}}\right)} - 1. \]

Protein molecular weight was used here as a specific normalization, to adjust for differences between proteins in the number of observable peptides.

In order to estimate absolute abundance of proteins based on emPAI we have made three amendments to emPAI. First, a proportionality constant \( a \) was introduced, reflecting the fact that emPAI might in the best case be directly proportional to protein abundance. Second, a scaling constant \( b \) was introduced. The values obtained by emPAI have been shown to estimate stoichiometry within a factor of 1.7 ± 0.8 accuracy for proteins varying 50-fold in their abundance (Ishihama, 2005 #1013). However, it is unclear how this accuracy would scale when considering stoichiometries that span several orders of magnitude. These adjustments give the following formula:

\[ \text{estimated copy number} = a \left(10^{b \left(\frac{\text{spectral count}}{\text{protein-specific normalization}}\right)} - 1\right). \]

Third, calibrating emPAI by using internal standards for which the absolute abundance is known (or can be reliably estimated) and that span the dynamic range in protein abundance allows minimizing the scaling effects and determination of the proportionality and

Data Processing and Protein Identification

MS data were processed using MaxQuant 1.0.11.2 (Cox and Mann, 2008) using the default settings (slice peaks: true; recalibrate: true; correct mass errors: true; max. peptide FDR: 0.01; max. peptide FPR: 0.1; max. protein FDR: 0.01; min. peptide length: 6; peptides used for protein quantitation: razor). Searches were conducted using Mascot version 2.1.3 (Matrix Sciences) for database searches against the chicken IPI database (version 3.49) in target/decoy configuration. Parameters used for the Mascot search were Trypsin/P for cleavage, Oxidation (M) and Acetyl (Protein N-term) for variable modifications, Label:13C(6) (R) and Label:13C(6) (K) for SILAC labels, monoisotopic mass, 7 ppm for MS tolerance and 0.5 Da for MS2 tolerance. Protein identification required two peptides, of which at least one was required to be unique in the database. A minimum of two peptide SILAC ratios was required to report a protein ratio.

Proteins that share some but not all of the observed peptides were combined for the purpose of abundance estimation into a group, with a summed, unique spectral count and summed mass of all group members minus sequence redundancies provided that they were at least the length of a tryptic peptide. The group was then split following a tree based on sequence similarities between proteins. At the first branch of the tree, for each subgroup the count was estimated again based on the unique spectral count and unique mass. The abundances of the subgroups were accepted if their sum was within a two-fold window of the total group’s abundance. This was continued until each protein obtained an estimation of its own abundance or until the sum of abundances deviated from the group estimate on the previous level by more than two-fold (“out of range”).

To obtain the name and functions of each protein, we referred to the following online databases: International Protein Index (IPI; http://www.ebi.ac.uk/IPI/IPIhelp.html), Ensembl (http://www.ensembl.org), Uniprot (http://beta.uniprot.org), Entrez Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db= gene) and InterPro (http://www.ebi.ac.uk/interpro). Over 3,000 individual searches were performed to generate the correspondence between IPI numbers and gene names listed in Table S1.

Replicate Analyses

All experiments minimized the impact of biases introduced due to fluctuations in the preparation by using pooled material from four independent preparations of mitotic chromosomes, as described above. In addition, the experiments were performed in duplicate and thus classifier values are averages of two measurements. Finally, no protein measurement is considered alone but only in conjunction with measurements of the same protein for other classifiers in a multidimensional analysis. Alternatively, members of the same complex are considered together. Both strategies provide a safeguard mechanism against artifacts resulting from outliers despite pooling and duplicate analysis.

Classifier I: Abundance

In order to estimate the abundance of proteins in mitotic chromosomes based on mass spectrometric data we rely on emPAI (Rappsilber et al., 2002; Ishihama et al., 2005, 2008) in slightly modified form. Briefly, emPAI is calculated based on the number of peptides observed (Table S4) and the number of times that each peptide is observed (spectral count) for each protein as follows:

\[ \text{emPAI} = 10^{\left(\frac{\text{spectral count}}{\text{protein-specific normalization}}\right)} - 1. \]

Protein molecular weight was used here as a specific normalization, to adjust for differences between proteins in the number of observable peptides.

In order to estimate absolute abundance of proteins based on emPAI we have made three amendments to emPAI. First, a proportionality constant \( a \) was introduced, reflecting the fact that emPAI might in the best case be directly proportional to protein abundance. Second, a scaling constant \( b \) was introduced. The values obtained by emPAI have been shown to estimate stoichiometry within a factor of 1.7 ± 0.8 accuracy for proteins varying 50-fold in their abundance (Ishihama, 2005 #1013). However, it is unclear how this accuracy would scale when considering stoichiometries that span several orders of magnitude. These adjustments give the following formula:

\[ \text{estimated copy number} = a \left(10^{b \left(\frac{\text{spectral count}}{\text{protein-specific normalization}}\right)} - 1\right). \]
Abundance Estimate Calculation

Example Calculations for the Classifiers

Classifier VI: Domain Analysis

We included prior knowledge in the form of domain annotations as part of our analysis. In order to obtain value from domain annotations, we first counted how many chromosomal and nonchromosomal proteins each domain was found in (Table S2), and expressed these as a proportion of all the training-set proteins the domain was found in (i.e., so that these two proportions add up to 1). For each protein, we then summed the proportions for all domains found in that protein, giving us a sum of chromosomal domain proportions and a sum of nonchromosomal domain proportions. We used these two values as variables for each protein in classifier I. We included prior knowledge in the form of domain annotations as part of our analysis. In order to obtain value from domain annotations, we first counted how many chromosomal and nonchromosomal proteins each domain was found in (Table S2), and expressed these as a proportion of all the training-set proteins the domain was found in (i.e., so that these two proportions add up to 1). For each protein, we then summed the proportions for all domains found in that protein, giving us a sum of chromosomal domain proportions and a sum of nonchromosomal domain proportions. We used these two values as variables for each protein in classifier I. We allowed the relative importance of these columns to be determined by the random forest itself.

Example Calculations for the Classifiers

Abundance Estimate Calculation

Spc24 (IP100576606) was identified 36 independent times in our reference experiment. This means that 36 individual fragmentation spectra were matched to peptide sequences from Spc24. This number is known as the spectral count. The protein abundance index (PAI) attempts to normalize for the size of the protein, since larger proteins have a greater chance of being seen because they can generate more peptides. In this case we divide the spectral count by the molecular weight of this protein in kDa, 22.831, to give a PAI of 48%. Since histones are so abundant, their estimation involves multiplication of by far and away the largest numbers. Thus even a minor scaling error could easily skew the proportion of histones versus all other proteins, particularly since the remainder of chromosomes is composed of many low abundance proteins. To illustrate this point, taking emPAI values without our scaling factor and proportionality constant returns histones as 100% of the chromosomal protein. In contrast, using spectral count alone would see histones at 10% of the chromatin protein mass.

It should also be noted that we predict 24 copies of CENP-A to be present at each of the 156 DT40 centromeres. This compares to 25-40 copies of CENP-A as measured at DT40 kinetochores recently by quantitative super-resolution microscopy (Ribeiro et al., 2010). The close coincidence of these two values measured by two independent approaches offers further support for our classifier I.

For approximation scaling of emPAI we used three proteins for which quantities had either been measured or estimated in chicken cells: CENP-H (4680 molecules/cell measured by quantitative fluorescence – (Ribeiro et al., 2009)), topoisomerase IIa (400,000 molecules/mitotoc cell measured in similar chicken cell lines by quantitative immunoprecipitation) and histone H4 (5 × 10^7 molecules/cell estimated based on a diploid genome size of 2.4 × 10^9 base pairs and 190 base pairs per nucleosome – J. Allan, personal communication). The proportionality and scaling constants in our modified equation for emPAI were approximated to provide values for the three reference proteins that approximate most closely the reported or predicted values. This resulted in the proportionality constant a = 285097 and the scaling constant b = 0.01055. Taking for the three proteins the reported/predicted and emPAI calculated abundance (Table S1) reveals a linear relationship over several orders of magnitude (Figure S1A). This shows that the two constants can be chosen to fit all three data points. Abundance determined for all proteins observed in our analyses was reproducible when using the modified emPAI equation with the approximated constants, as shown by plotting the results from two independent experiments against one another (Figure S1B). In this analysis, 60% of the determinations were within a factor of ± 2x and 96% were within ± 10x. For comparison, the reproducibility of all classifier experiments is plotted in Figure S1C.
**Domains Calculation**

For example, INCENP (IPI00823144) is annotated with 3 domains, an Aurora-related kinase (ARK) binding domain (PF03941), a tropomyosin signature domain (PR00194) and a glutamic acid-rich region (PS50313). Table S5 shows the number of TRUE (chromosomal) and FALSE (nonchromosomal) proteins that each domain is associated with in our dataset. We calculated the proportion of true and false proteins that each domain was associated, e.g., the tropomyosin domain was found in 20 TRUE proteins and 42 FALSE protein, so the proportions are 20/62 = 0.32 and 42/62 = 0.68 for TRUE and FALSE respectively. We then summed these proportion for all domains in each protein, as shown at the bottom of the table.

These two sums provide the two columns for this classifier.

**Combination of Classifiers by Random Forest Analysis**

Combination of classifiers was performed using the WEKA implementation of random forest (RF) (Breiman, 2001; Frank et al., 2004; Hall et al., 2009). We defined two classes of proteins whose assignment as chromosomal or nonchromosomal was in general nonambiguous: centromere, telomere, chromosomal periphery, other chromosomal, replication and repair, transcription factors, RNA polymerases, histones and histone modifying proteins were labeled as true and cytoskeletal, cytoplasmic, mitochondrial, membrane and receptor proteins were labeled as false. Note that the choice of training data sets to establish the random forest is crucial. Our choice of the training data was validated in retrospect by the characterization of the 50 novel cloned proteins as shown in Figure 3A. We asked the RF to train on the training data and give values on all of our identified proteins. However, additional steps were taken to improve the value of the dataset prior to combination. Each classifier was subject to k-nearest-neighbors (k-NN) (Cover and Hart, 1980) using in-house software, and these values were added to the raw classifier data to be used in the RF. The k-NN was optimized at 50 nearest neighbors using receiver operator characteristic (ROC) across a range of values from 10 to 100. In addition to k-NN, we also added the number of missing data points in a protein as another variable to the random forest.

Random forests work by constructing a series of decision trees, the aim of each of which is to discriminate different classes of data (e.g., true and false). The trees are built randomly, so at each node (decision point) in the tree any classifier could be used. Then the data are divided on this classifier, each decision giving rise to another node where another random decision is made. Eventually, a node will adequately segregate the data on which it operates into two classes, and no more nodes are needed. Each tree is capable of distinguishing different classes of data, but a single tree does not perform definitively on data that was not in the training set, so multiple trees are used. The summation of all trees provides a “vote” on each data point to decide which class (true/chromosomal, false/nonchromosomal) the data point (e.g., a protein) is likely to be.

The data were subjected to RF using WEKA, via Java. The number of trees used was optimized using 10-fold cross-validation and ROC. The optimal number of trees was found to be 450. The depth of the trees was not limited. 10-fold cross-validation is a standard method of evaluating RF, since RF are susceptible to over-training and evaluation on the training set gives deceivingly good results, not likely to be representative of how the RF will perform on new data.

Random forests were trained using WEKA on the training set (in this case 772 chromosomal and 1331 nonchromosomal proteins) adequately represents the unknown proteins that we would like to classify. WEKA does not enable output of the 10-fold cross-validation data itself, so to obtain these data for proteins in our training set, we randomly split our training set into 10 portions of roughly equal size (within 10%) and use 9 of the ten portions to train the RF, which was then used to produce values for the tenth portion. This was repeated until each protein had a value from an RF in which it was not a member of the training set. All the RF data we show on chromosomal (true) and nonchromosomal (false) proteins, except in Figure 3A, are derived from these cross-validation data.

For example, INCENP (IPI00823144) is annotated with 3 domains, an Aurora-related kinase (ARK) binding domain (PF03941), a tropomyosin signature domain (PR00194) and a glutamic acid-rich region (PS50313). Table S5 shows the number of TRUE (chromosomal) and FALSE (nonchromosomal) proteins that each domain is associated with in our dataset. We calculated the proportion of true and false proteins that each domain was associated, e.g., the tropomyosin domain was found in 20 TRUE proteins and 42 FALSE protein, so the proportions are 20/62 = 0.32 and 42/62 = 0.68 for TRUE and FALSE respectively. We then summed these proportion for all domains in each protein, as shown at the bottom of the table.

These two sums provide the two columns for this classifier.

**Combination of Classifiers by Random Forest Analysis**

Combination of classifiers was performed using the WEKA implementation of random forest (RF) (Breiman, 2001; Frank et al., 2004; Hall et al., 2009). We defined two classes of proteins whose assignment as chromosomal or nonchromosomal was in general nonambiguous: centromere, telomere, chromosomal periphery, other chromosomal, replication and repair, transcription factors, RNA polymerases, histones and histone modifying proteins were labeled as true and cytoskeletal, cytoplasmic, mitochondrial, membrane and receptor proteins were labeled as false. Note that the choice of training data sets to establish the random forest is crucial. Our choice of the training data was validated in retrospect by the characterization of the 50 novel cloned proteins as shown in Figure 3A. We asked the RF to train on the training data and give values on all of our identified proteins. However, additional steps were taken to improve the value of the dataset prior to combination. Each classifier was subject to k-nearest-neighbors (k-NN) (Cover and Hart, 1980) using in-house software, and these values were added to the raw classifier data to be used in the RF. The k-NN was optimized at 50 nearest neighbors using receiver operator characteristic (ROC) across a range of values from 10 to 100. In addition to k-NN, we also added the number of missing data points in a protein as another variable to the random forest.

Random forests work by constructing a series of decision trees, the aim of each of which is to discriminate different classes of data (e.g., true and false). The trees are built randomly, so at each node (decision point) in the tree any classifier could be used. Then the data are divided on this classifier, each decision giving rise to another node where another random decision is made. Eventually, a node will adequately segregate the data on which it operates into two classes, and no more nodes are needed. Each tree is capable of distinguishing different classes of data, but a single tree does not perform definitively on data that was not in the training set, so multiple trees are used. The summation of all trees provides a “vote” on each data point to decide which class (true/chromosomal, false/nonchromosomal) the data point (e.g., a protein) is likely to be.

The data were subjected to RF using WEKA, via Java. The number of trees used was optimized using 10-fold cross-validation and ROC. The optimal number of trees was found to be 450. The depth of the trees was not limited. 10-fold cross-validation is a standard method of evaluating RF, since RF are susceptible to over-training and evaluation on the training set gives deceivingly good results, not likely to be representative of how the RF will perform on new data.

Random forests were trained using WEKA on the training set (in this case 772 chromosomal and 1331 nonchromosomal proteins) adequately represents the unknown proteins that we would like to classify. WEKA does not enable output of the 10-fold cross-validation data itself, so to obtain these data for proteins in our training set, we randomly split our training set into 10 portions of roughly equal size (within 10%) and use 9 of the ten portions to train the RF, which was then used to produce values for the tenth portion. This was repeated until each protein had a value from an RF in which it was not a member of the training set. All the RF data we show on chromosomal (true) and nonchromosomal (false) proteins, except in Figure 3A, are derived from these cross-validation data.

The output of the random forest, in this context, is a number between 0 and 1 that represents the probability that a protein is chromosomal as opposed to nonchromosomal. It does not consider any other class of protein, and makes the implicit assumption that our training set (in this case 772 chromosomal and 1331 nonchromosomal proteins) adequately represents the unknown proteins that we would like to classify.

**RNAi and Indirect Immunofluorescence Microscopy**

100 pmol of siRNA (Control: AACGATCGGGAATCTCCGAAdT; Ska3/Rama1: ATGGATATAGTCCAGTGTCAAdT – Hs_C13orf3_5; CENP-27: AAGCCCGTGAAGTAAAGAdT – Hs_C10orf104_7; CENP-23: TGCAGGACCCTGACACAAAdT – Hs_C9orf114_3, QIAGEN) was administered to U2OS or HeLa cells at 30%–40% confluence by transfection with OligoFectamine (Invitrogen) in complete medium without antibiotics. Cells maintained in this medium for 48 (Ska3/Rama1 and CENP-32), 72 (CENP-27) hours were fixed for 5 min. with 4% (v/v) paraformaldehyde (Electron Microscopy Services) in PBS or for 20 min. with Methanol in –20°C. After permeabilization with 0.15% (v/v) Triton X-100 in PBS, coverslips were blocked with 1% (v/v) BSA in PBS. Cells were probed with antibodies against: CENP-C (1:600, rabbit 554), ACA (1:200, human), α-tubulin (B512, 1:2000, mouse; Sigma), RanBP2 (1:1000, goat) (Hutten et al., 2008), RacGAP1 (N-19, 1:100, goat, Santa-Cruz Biotechnology), SBP (1:300, mouse 20), Pericentrin (ab4448, 1:1000, rabbit, Abcam), γ-tubulin (AK15, 1:1000, rabbit; Sigma). Cells were washed three times with cytoskeleton buffer for 5 min, fluorescence-labeled secondary antibodies were applied at 1:200, and the DNA was counterstained with DAPI at 0.1 μg/ml.

**Bioinformatics Analysis**

Multiple sequence alignments were produced with T-Coffee (Notredame et al., 2000) using default parameters, slightly refined manually and viewed with the Belvu program (Sonnhammer and Hollich, 2005). The amino acid coloring scheme indicates average BLOSUM62 scores (correlated with amino acid conservation) for each alignment column: red (greater than 3), violet (between 3 and 1.5) and light yellow (between 1.5 and 0.5).
Sequence names (and UniProt accession codes in parentheses) for the analysis of CENP-32 are as follows: hCENP32 (Q5T280) Homo sapiens; Fly (Q8MSZ6) Drosophila melanogaster; Worm (Q10950) Caenorhabditis elegans; Plant (Q6NLH7) Arabidopsis thaliana; YGR283C (P53336) Saccharomyces cerevisiae; YMR310C (Q04867) S. cerevisiae; S. pombe (O13641) Schizosaccharomyces pombe; 1KJR (O26109) Methanobacterium thermoautotrophicum protein of known structure.

The ribbon representation and electrostatic surface potential map for the homolog of the published crystal structure of the CENP-32 homolog from M. thermoautotrophicum (Zarembinski et al., 2003) and obtained using MODELER (Marti-Renom et al., 2000). Images and electrostatic potentials were created with Pymol (http://pymol.sourceforge.net) and APBS (Baker et al., 2001).

Among the other novel kinetochore proteins, CENP-30/FBOX28 contains an F-box domain characteristic of proteins that link substrates with E3 ubiquitin ligases. Skp1, a member of the SCF complex (Skp1-Cullin-F-box) regulates the stability of kinetochore protein Ctf13 in budding yeast (Kaplan et al., 1997), and with Sgt1 (another component of an E3 ubiquitin ligase) is involved in checkpoint regulation (Kitagawa et al., 1999). Thus, CENP-30 may be involved in regulating protein turnover within the kinetochore.

SUPPLEMENTAL REFERENCES

Andrade, M.A., Ponting, C.P., Gibson, T.J., and Bork, P. (2000). Homology-based method for identification of protein repeats using statistical significance estimates. J. Mol. Biol. 298, 521–537.

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., and McCammon, J.A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. Proc. Natl. Acad. Sci. USA 98, 10037–10041.

Breiman, L. (2001). Random Forests. Mach. Learn. 45, 1.

Chu, D.S., Liu, H., Nix, P., Wu, T.F., Ralston, E.J., Yates, J.R., 3rd, and Meyer, B.J. (2006). Sperm chromatin proteomics identifies evolutionarily conserved fertility factors. Nature 443, 101–105.

Cover, T.M., and Hart, P.E. (1980). Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, IT-13(1):21–27, January 1967. In Pattern Recognition, K.S. Fu, ed. (Hong Kong, Chinese University Press).

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al. (2009). The Pfam protein families database. Nucleic Acids Res. 36, D211–D222.

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I.H. (2004). Data mining in bioinformatics using Weka. Bioinformatics 20, 2479–2481.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutermann, P., and Witten, I.H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations 11, 1.

Hutten, S., Flotho, A., Melchior, F., and Kehlenbach, R.H. (2008). The Nup358-RanGAP complex is required for efficient importin alpha/beta-dependent nuclear import. Mol. Biol. Cell 19, 2300–2310.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomes by sequencing of sequenced peptides per protein. Mol. Cell. Proteomics 4, 1265–1272.

Ishihama, Y., Rappsilber, J., Andersen, J.S., and Mann, M. (2002). Microcolumns with self-assembled particle frits for proteomics. J. Chromatogr. A 979, 233–239.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kermer, M.J., and Frishman, D. (2008). Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics 9, 102.

Kaplan, K.B., Hyman, A.A., and Seger, P.K. (1997). Regulating the yeast kinetochore by ubiquitin-dependent degradation and Skp1p-mediated phosphorylation. Cell 91, 491–500.

Kitagawa, K., Skowrya, D., Ellledge, S.J., Harper, J.W., and Hieter, P. (1999). SGT1 encodes an essential component of the yeast kinetochore assembly pathway and a novel subunit of the SCF ubiquitin ligase complex. Mol. Cell 4, 21–33.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P., and Bork, P. (2004). SMART 4.0: toward genomic data integration. Nucleic Acids Res. 32, D142–D144.

Lewis, C.D., and Laemmli, U.K. (1982). Higher order metaphase chromosome structure: evidence for metalloprotein interactions. Cell 29, 810–821.

Marin, I. (2003). Evolution of chromatin-remodeling complexes: comparative genomics reveals the ancient origin of “novel” compensasome genes. J. Mol. Evol. 56, 527–539.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophysics. Biomol. Struct. 29, 291–325.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 298, 2391–2398.

Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. Genome Res. 12, 1231–1245.

Ribeiro, S.A., Gattin, J.C., Dong, Y., Joglekar, A., Cameron, L., Hudson, D.F., Farr, C.J., McEwen, B.F., Salmon, E.D., Earnshaw, W.C., et al. (2009). Condensin regulates the stiffness of vertebrate centromeres. Mol. Biol. Cell 20, 2371–2380.

Ribeiro, S.A., Vagnarelli, P., Dong, Y., Hori, T., McEwen, B.F., Fukushima, T., Flors, C., and W.C. Earnshaw, W.C. (2010). A super-resolution map of the vertebrate kinetochore. Proc. Natl. Acad. Sci. USA, in press.

Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V., and Mann, M. (2006). In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat. Protoc. 1, 2856–2860.
Sonnhammer, E.L., and Hollich, V. (2005). Scoredist: a simple and robust protein sequence distance estimator. BMC Bioinformatics 6, 108.

Tan, F., Li, G., Chitteti, B.R., and Peng, Z. (2007). Proteome and phosphoproteome analysis of chromatin associated proteins in rice (Oryza sativa). Proteomics 7, 4511–4527.

Zarembinski, T.I., Kim, Y., Peterson, K., Christendat, D., Dharamsi, A., Arrowsmith, C.H., Edwards, A.M., and Joachimiak, A. (2003). Deep trefoil knot implicated in RNA binding found in an archaeobacterial protein. Proteins 50, 177–183.
Figure S1. Proteomic Analysis of Mitotic Chromosome Proteins Gave Reproducible Classifier Values, Related to Figure 1

(A) Calibrating the abundance estimation. Shown is a modified version of the emPAI equation, incorporating proportionality and scaling constants. We optimized these constants on three proteins to give the minimum error. The equation is capable of achieving a near linear relationship between known abundance and the abundance we estimate for the three standard proteins.

\[ n_p = a \times (10^{a \cdot b} - 1) \]

- $n_p$: protein copy number
- $a$: proportionality constant
- $b$: scaling factor
- $n_S$: spectral count
- $S_p$: protein-specific normalization

(Here used: molecular weight)

\[ a = 285097 \]
\[ b = 0.019553 \]

(B) A scatter-plot comparing abundance between two independent analyses, using different preps, shows that the data have an approximately $x = y$ correlation. Two-thirds of proteins have abundance values within 2-fold of each other and only 1% vary by $>10$-fold.

(C) The values from two independent analyses of each classifier, using different preps, and divided one by the other to give a fold-error. For abundance, enrichment, exchange and Smc2 dependency, we found that 66%, 73%, 68% and 83% of proteins, respectively, had values falling within a factor of 2 for the two experiments and 90%, 92%, 93% and 97% were within a factor of 4.
Figure S2. Smoothed Histograms Showing the Distribution of Proteins Across Classifiers and RF Analysis, Related to Figure 2

(A–E) Classifiers I - V have been scaled between 0 and 1 on a log scale. The classifier is plotted on the x axis, while the y axis shows the number of proteins in each tenth of the classifier.

(F) The equivalent plot for the result of the ten-fold cross-validation of the random forest analysis shows that the proteins are spread more widely and is expected to provide better enrichment of chromosomal proteins at higher scores than the other classifiers.
Figure S3. The Performance of the Random Forest, Related to Figure 3

(A) A receiver operating characteristic (ROC) plot graphs sensitivity versus specificity. The area under the ROC curve demonstrates the power of each classifier. The ROC curve for all proteomic classifiers combined using the random forest is taken from a ten-fold cross-validation, indicating the robustness of the system when applied to data not included in the training set. This shows that the combination of classifiers is expected to outperform them individually.

(B) Addition of a bioinformatic classifier, consisting of two variables: chromosomal domain score and nonchromosomal domain score. Integrating proteomic and nonproteomic classifiers by random forest provides the best performance as indicated by the area under curve (AUC = 0.97).

(C) True over false positives when including domain data. A random forest integrating all six classifiers reports with the first 500 true positives only 17 false positives.

(D) Success of Classifier VI (domains) used for training a random forest to separate chromosomal and nonchromosomal proteins. Note that this is the data for the training set (similar to Figure 2A) and not ten-fold cross-validation, i.e., showing the ability of the random forest to be trained using the training data and not the ability of the random forest to predict localization. The classifier is plotted on the y axis, while the x axis shows the result of the ten-fold cross-validation of the random forest analysis using Classifiers I - V. Classifier VI (domains) assigns a high score to many chromosomal training proteins and a low score to many nonchromosomal training proteins. For these proteins a random forest can be trained to separate the training set. However, for a significant group of proteins (roughly found at y = 0.42) no separation is achieved. For these proteins the random forest cannot be trained. This is not surprising as proteins lacking known domains are found on and off chromosomes and can thus not be assigned to either localization based on known domains or lack thereof. Note that these proteins are segregated into predicted chromosomal and nonchromosomal efficiently using random forest analysis and Classifiers I - V.
Figure S4. Chromosomal Localization of Novel Proteins Identified in This Study, Related to Figure 4

(A–G) Novel or previously uncharacterized chromosomal proteins expressed as GFP fusion protein (panels 2, green) in mitotic U2OS cells stained for DNA (panels 1, blue). A(4) and B–G(3) show merged images. (A) CENP-Y/CCDC75 plus CENP-C (panel 3, red), (B) cPERP-A/C1orf131, (C) cPERP-B/CCDC137, (D) cPERP-C/KIAA0020, (E) cPERP-D/DEAD box 18, (F) cPERP-E/Autosomal recessive 1A, and (G) cPERP-F/DDX27.
Figure S5. Generation of a Knockout Cell Line and Phenotypic Analysis of Ska3/RAMA1, Related to Figure 5

(A) Restriction maps of the chicken Ska3/RAMA1 locus, gene disruption constructs, and targeted loci. Black boxes indicate the positions of exons. ApaL1, EcoRV and EcoRI restriction sites are shown. There is a EcoRI polymorphism in this locus; a EcoRI site is absent from the 1st round targeted allele. The position of the probe used for Southern hybridization is indicated. Novel 8.5 and 7.7 kb EcoRI fragments hybridize to the probe if targeted integration of the construct occurs.

(B) Restriction analysis of genomic DNAs from cells with targeted integration of a Ska3/RAMA1 disruption construct.

(C) Representative growth curves for the indicated cell cultures (Wild-type and Ska3/RAMA1 knockout cell line clones 12-17, 12-12 and 12-22).

(D) Quantitation of aberrant wild-type or Ska3/RAMA1 12-17. We scored the number of mitotic cells, bi-nucleate cells, and apoptotic cells.

(E) Abnormal mitotic morphologies of Ska3/RAMA1 knockout cells. Cells were stained with α-tubulin (green), and DNA was counterstained with DAPI (blue).

(F) Depletion of RanBP2 (panels 2,6; red) from kinetochores of Ska3/Rama1-depleted cells. DNA (panels 1,5), blue; human anti-centromere antibody (panels 3,7), green; merge (panels 4,8).

(G) RanGAP1 (red – panels 2,6) is depleted from kinetochores of Ska3/Rama1-depleted cells. DNA (panels 1,5), blue; human anti-centromere antibody (panels 3,7), green; merge (panels 4,8).
Figure S6. Characterization of CENP-32, Related to Figure 6

(A) Depletion of CENP-32 yields a significant increase in prometaphase cells and decrease in later mitotic stages.

(B) CENP-32 RNAi causes detachment of centrosomes from bipolar spindles. DNA, blue (panels 1,5,9,13); γ-tubulin (panels 2,6) or pericentrin (panels 10,14), green; α-tubulin (panels 5,8,12,16), red.