Nano Random Forests to mine protein complexes and their relationships in quantitative proteomics data

Nano Random Forests to mine protein complexes and their relationships in quantitative proteomics data

Luis F. Montaño-Gutierrez, Shinya Ohta*†, Georg Kustatscher*, William C. Earnshaw*§, Juri Rappsilber*‡§*

*Wellcome Trust Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom.
†Center for Innovative and Translational Medicine, Medical School, Kochi University, Kohasu, Oko-cho, Nankoku, Kochi 783-8505, JAPAN.

‡Chair of Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany.

§Corresponding authors: Bill.Earnshaw@ed.ac.uk, Juri.Rappsilber@ed.ac.uk

Supplemental Material can be found at:
http://www.molbiolcell.org/content/suppl/2017/01/02/mbc.E16-06-0370v1.DC1

Running head: NanoRF to mine complexes in proteomics

Abbreviations: RF, Random Forest; MCCP, Multi-Classifier Combinatorial Proteomics; NanoRF, Random forests trained with small training sets; MVP, Multivariate proteomic profiling; FP, Fractionation profiling; ICP, interphase chromatin probability; CCAN, Constitutive Centromere-Associated Network; Nup, Nucleoporin; SMC, Structural Maintainance of Chromosomes; SILAC, Stable Isotope Labeling by Amino acids in Cell culture

ABSTRACT

Ever-increasing numbers of quantitative proteomics datasets constitute a currently underexploited resource for investigating protein function. Multi-protein complexes often follow consistent trends in these experiments, which could provide insights about their biology. Yet, as more experiments are considered, a complex's signature may become conditional and less identifiable. Previously, we successfully distinguished the general proteomic signature of genuine chromosomal proteins from hitchhikers using the Random Forests (RFs) machine learning algorithm. In this technical note, we tested whether small protein complexes could define distinguishable signatures of their own, despite the assumption that machine learning needs large training sets. We show, with simulated and real proteomics data, that RFs can detect small protein complexes and relationships between them. We identified several complexes in quantitative proteomics results of wild-type and knock-out mitotic chromosomes. Other proteins covaried strongly with these complexes, suggesting novel functional links for later study. Integrating the RF analysis for several complexes revealed known interdependencies among kinetochore subunits, and a novel dependency between the inner kinetochore and condensin. Ribosomal proteins, although identified, remained independent of kinetochore subcomplexes. Together, these results show that this complex-oriented RF (NanoRF) approach can integrate proteomics data to uncover subtle protein relationships. Our NanoRF pipeline is available at https://github.com/EarnshawLab/NanoRF.

INTRODUCTION

Proteins influence many processes in cells, often affecting the synthesis, degradation and physicochemical state of other proteins. One strategy that diversifies and strengthens protein functions is the formation of multi-protein complexes. For this reason, identification of partners in complexes is a powerful first step to studying protein function. However, determination of membership to or interaction with protein complexes remains an arduous task, mainly achieved via demanding biochemical experimentation. The latter can be limited by the ability to overexpress, purify, tag, stabilise, and obtain specific antibodies for the proteins in complexes of interest. Thus, any methods that facilitate protein complex identification and monitoring (Gingras *et al.*, 2007; Kustatscher *et al.*, 2014; Skinner *et al.*, 2016) have the potential to accelerate the understanding of biological functions and phenotype. The vast amount of proteomics data already available represents a largely untapped resource that could potentially reveal features currently undisclosed by traditional analysis, such as condition-dependent links, inter-complex contacts and transient interactions.

To date, biochemical co-fractionation has been widely used to identify protein complexes. Members of a multiprotein complex typically co-elute with a single mass, charge, elution rate etc. in techniques such as chromatography, gel electrophoresis, and co-immunoprecipitation. Another common way to discover complexes is by combining chemical cross-linking, fractionation and mass-spectrometry, which covalently fixes proteins that interact (Leitner et al, 2016). However, biochemical fractions often contain contaminants, i.e. proteins that are not genuine subunits of the complex of interest, despite

having similar biochemical properties. One way to reveal bona-fide members is to combine several fractionation experiments, as well as perturbations (Moore and Lee, 1960). Members of a complex will behave coordinately, whereas contaminants will usually show a more independent behaviour. From a quantitative perspective, this translates into protein covariance - the covariance of proteins within a complex is stronger than that with contaminants. As additional biochemical fractionation conditions are considered, high covariance sets true members of a complex apart from contaminants or hitchhikers. This principle has been used recently in a large-scale effort that predicted 622 putative protein complexes in human cells by assessing the coordinated behaviour of proteins across several fractionation methods, among others (Havugimana et al., 2012; Michaud et al., 2012).

Covariance among members of protein complexes has been observed in several integrative proteomics experiments (Ohta *et al.*, 2010; Borner *et al.*, 2014) and even used to predict association with complexes (Andersen *et al.*, 2003; Borner *et al.*, 2014). This relies on the fact that the co-fractionation of proteins that are functionally interconnected will be affected by common parameters, such as knock-outs or varying biochemical purification conditions. However, performing covariance analysis using multiple quantitative proteomics datasets is non-trivial. First, experimental or biological noise hampers quantitation of protein levels. Second, only a fraction of the experiments may be informative for any given complex. Third, proteins may go undetected, leading to missing values. Fourth, the relationship between different protein groups may only be observed under specific circumstances. The power of multivariate analysis methods like Principal Component Analysis (PCA), hierarchical clustering or k-nearest neighbours (KNN) could be limited when a protein complex's signal in the data is affected in all these ways. Here we show that the supervised machine learning technique Random Forests can overcome these limitations, distinguish the covariance of small protein groups, and provide biologically sound, predictive insights to protein complex composition, relationships and function. We describe this approach using as an example the behaviour of multi-protein complexes in mitotic chromosomes.

RESULTS

Random Forests can detect small protein complexes in simulated organelle proteomics data

Proteins in multi-protein complexes have been shown to covary across quantitative proteomics experiments of organelles (Ohta *et al.*, 2010; Borner *et al.*, 2014). That is, the absolute or relative quantities of proteins that together form a complex increase or decrease in a coordinate manner. This concerted behaviour forms a potentially detectable 'signature' of the complex across sets of proteomics experiments. Other proteins that share the same signature may be functionally related to the complex.

We wondered how strong a complex's signature would need to be for its detection. The signature is an outcome of the resemblance of each protein's behaviour to each other and how much the group stands out from other groups. We reasoned that the strength of the signature could be modulated in two ways: a) by controlling the fraction of informative experiments (experiment subsets where the members of the complex correlate) and b) by

different amounts of noise. Less informative experiments should 'dilute' the complex's signal, whereas stronger noise would lead to fluctuations away from the common behaviour We therefore constructed artificial proteomics data in which we could independently control these two properties and evaluate their influence on detecting a hypothetical complex.

We generated artificial proteomics tables (Figure 1A) by populating random values into tables of similar sizes to our original dataset: 20 'experiment' columns by 5000 'protein' rows. In those tables, 12 'proteins', which were intended to represent a hypothetical protein complex, were constrained to have identical behaviour in a fraction X of columns, while leaving independent random values in the remaining experiments. This action imitated situations in which a complex covaried in only an informative subset of experiments (Figure 1A, middle panel). For example, if X=0.5, 10/20 'experiments' would contain the signature behaviour. Next, we jittered all the entries in the table by adding Gaussian noise of strength Y. Figure 1B illustrates the data generated by this approach and exemplifies visually how the number of informative variables and noise contribute to a protein group's signature behaviour.

We wondered first if the mean of pairwise correlations between proteins of a complex would suffice to reveal membership as levels of noise and informative experiments changed. As one would expect, when the noise was low and the fraction of informative experiments was high, protein correlation was high. However, it dropped rapidly with slightly weaker signatures (Figure 1D).

We then asked if the machine learning algorithm Random Forests (RF) (Breiman, 2001) would recognise stronger or weaker signatures in the behaviour of the hypothetical complex (for an introductory explanation of the algorithm, see methods). RFs have been used in several biological data contexts, including gene expression, protein-protein interactions and mass spectrometry (Fusaro *et al.*, 2009; Qi, 2012). Specifically, we asked whether RF could distinguish our hypothetical complex from an independent group of other proteins (negative class), composed of 365 rows in the random protein table (Figure 1A, middle panel). In two previous works from our group (Ohta *et al.*, 2010; Kustatscher *et al.*, 2014), we used RF because it a) samples combinations of experiments and attempts to draw a 'boundary' between a positive and negative class, b) it is non-parametric and can handle missing values (Qi, 2012) , and c) for every 'protein', RF would a score between 0 and 1 – the RF-score– indicating whether a 'protein' behaves as part of the hypothetical complex (Figure 1C). Proteins part of the positive and negative classes also obtain an unbiased score regardless of whether they belong to the training classes (See methods).

Figure 1E shows that the RF score of the hypothetical complex remained high even with few informative experiments, but fell significantly with higher noise. Therefore, if looking at the RF score alone, even small amounts of noise could lead to not recognising members of the true complex (false negatives), even when they initially had a fairly strong correlation. These results suggest that the RF score is, on its own, not robust to noisy conditions even when correlation in a complex is high.

We reasoned that a noise-induced decrease in RF scores could be tolerated as long as the scores of members of the hypothetical complex were overall higher than those of the negative class. Yet, levels of noise too high, and too few informative experiments, could lead to false positives. To strike a balance, we searched for a RF score that, if used as a boundary between the two classes, maximized separation quality – i.e. made the fewest class misassignments – between the hypothetical complex and the hypothetical contaminants. This can be assessed by the Matthews Correlation Coefficient (MCC - Exemplified in Figure 2A, lower panels). Figure 1F shows that class separation quality remains high for different levels of noise and small fractions of informative experiments. All measures showed the lowest values for the weakest signatures, where the complex can no longer be distinguished from randomly covarying groups. Altogether, we conclude that RF is able to distinguish significant signatures of a protein group in high noise and few informative experiments, even though the group could be as small as a protein complex. Because of the small training set size, we refer to this instance of Random Forests as NanoRF.

RF analysis can distinguish protein complexes from contaminants in proteomics experiments of mitotic chromosomes

Our group has both collected and published SILAC proteomics data of mitotic chromosomes isolated from chicken DT40 wild type and knockout cell lines (Ohta *et al*., 2010; 2016). The proteins targeted for knockouts belong to a range of mitotic chromosome complexes of two groups: Structural Maintenance of Chromosomes (SMC complexes, like condensin SMC2-4., cohesin SMC1-3 (Uhlmann *et al.*, 1999; Sonoda *et al.*, 2001; Mehta *et al.*, 2012) , SMC5-6 (Stephan *et al.*, 2011; Wu and Yu, 2012)) and the kinetochore (Ska3). We have previously used Random Forests to classify between large groups of 'true' chromosomal proteins and potential hitchhikers or contaminants. Given that RF could distinguish small covarying groups in simulated data, we asked whether it could detect known small protein complexes based on real data and if any other proteins shared the signature of the complexes.

The diagram in Figure 2A illustrates our strategy to detect protein complexes in mitotic chromosomes and retrieve proteins that may be functionally linked with them. First, we choose a protein complex (Figure 2, red dots), and a set of curated hitchhikers (Figure 2 blue dots (Ohta *et al.*, 2010), which serve as the negative class (Table S2). Then we use RF to distinguish the complex from the hitchhikers on the basis of our proteomics data. As every protein will get a RF score, we look for a 'boundary' cut-off that maximizes class separation quality – i.e. that most members of a complex are above it and the most contaminants below. True members that surpass the cut-off are said to be "identifiable". Other non-member, non-contaminant proteins above that cut-off together with true members are said to covary/associate with them (Figure 2A and 2B, orange dots). To find the boundary cut-off and its significance, we use the MCC (Figure 2A, bottom panel) as used in the previous section. A more 'traditional' way to evaluate the significance of this result is to consider a hypergeometric test. If we were to draw proteins at random from a bag in which true members and hitchhikers were mixed; such a draw is analogous to setting an RF cut-off. For a given draw, the larger the number of true members– and the lower the number of hitchhikers–, the lower the probability of such draw.

We analysed a number of different complexes with RF (Figure 2B). In particular, we performed NanoRF on the Constitutive-Centromere-Associated Network (CCAN), the KNL-Mis12-Ndc80 (KMN) complex, Nucleoporin 107-160/RanGAP, SMC 5/6, cohesin and ribosomal proteins. Figure 2B shows an overview of the RF result for each complex. Most complexes yielded a high separation quality and most bona-fide subunits were assigned higher RF scores than the contaminants. Other proteins (orange) distinguished themselves from hitchhikers, with scores as high as bona fide complex subunits. These significantly covarying proteins likely correspond to a mix of known, putative, and potentially spurious associations, which we will attempt to dissect in the following sections. The full list of proteins associated with each complex can be found in table S2.

A common concern in supervised machine learning is overfitting, which describes a situation in which the algorithm performs well for reasons other than the inherent properties of proteins in the data. This can be due, for example, to training set bias. A way to control for the latter is through out-of-bag (OOB) analysis, which overlaps functionally with cross-validation. For every tree constructed, the algorithm leaves a random number of training observations out. This internally avoids biasing the training towards particular observations (Breiman, 2001). As we built our trees in this way, the results shown are already controlled for training set bias.

An algorithm capable of overfitting may be able to identify any arbitrary group of proteins by leveraging noise and other irrelevant properties. Therefore, to further rule out overfitting, we ran RF on 5000 protein sets generated at random from our dataset. The size of those test sets (10 random positive-class proteins and 400 random negative-class proteins) was in the range of the chromosomal protein complexes we investigated, which ranged between 7 and 20.

In the bottom panel of figure 2B, we show the result of performing RF on a randomly picked target group of proteins rather than a complex, and another randomly picked group as a mock group of contaminants/hitchhikers. It can be seen that both classes intercalate; in other words, RF classification shows poor performance (there is no evidence for any discrete group in the data), no member of the mock target group is identifiable (RF score close to zero). As the RF classification (and MCC values) are of such poor quality, interpretation of the RF is undefined in these conditions. Our results with the random groups contrast starkly with the successful identification of proteins separating as protein complexes from the negative class (Figure 2B, upper panels).

We further evaluated the significance of our results using Receiver-Operating Characteristic (ROC) curves (Figure 2C) and the MCC values themselves (Figure 2D). Starting from the highest RF score, a ROC curve evaluates the fraction of positive class members recovered (true positives) on the vertical axis versus the negative class members recovered (false positives) on the horizontal axis. A ROC curve that climbs vertically is favourable because it means that the RF score is sensitive to the complex. Under these circumstances, the area under the ROC curve (AUC) is larger than 0.5. In contrast, if the RF score contained a poor signal, the positive and negative class would be retrieved randomly. In this case, the ROC curve climbs up the diagonal and has an area of around 0.5. In our analysis, all of the complex-specific RF retrieved roughly 70% of the complexes before any false positives were

collected (Figure 2C). All our complexes showed an AUC between 0.9 and 0.999 (Table S2), implying accurate classification. In contrast, ROC curves of the randomly selected groups (examples in Fig. 2C, black and grey lines) remained close to the diagonal.

Finally, for some complexes we studied, it could be a matter of chance that they separated well from the negative class. We sought to address how likely it is for a random group to obtain a high separation quality by chance in our dataset. Therefore, we evaluated the distributions of class separation quality (as quantified by the peak MCC values) for real complexes and for randomly sampled protein groups (Figure 2D). The highest MCC value obtained for the random classes was 0.543 ($P\approx0.0002$, N=5000), whereas the minimum MCC value for the complexes' separation was 0.71 ($P\approx0.002$, N=500). Altogether, these results support the hypothesis that the NanoRF can distinguish between protein complexes and contaminants in real data.

Integration of several complex-specific RF reveals known and novel interdependencies between protein complexes.

The covariance of each complex could be its unique signature or could overlap with that of other complexes, possibly implying conditional interdependency among complexes. We decided to test this hypothesis with kinetochore subcomplexes as there is significant contact among them. To this aim, we analysed 2D plots of RF for different complexes (Figure 3).

We categorized several possible interdependency scenarios between kinetochore complexes (Figure 3A, B). According to these scenarios, the CCAN and the Nucleoporin 107-160 /RanGAP complex (Figure 3C) appeared independent, i.e. they do not associate with each other. In contrast, the KMN network associated with both. We concluded that perturbations on both CCAN and Nup-107-160 have a hierarchical effect on KMN (i.e. their effects propagate to KMN but not vice versa), implying that the latter is involved in links between inner and outer kinetochore. These observations are consistent with current models of the kinetochore (Kwon *et al.*, 2007; Screpanti *et al.*, 2011). The other proteins associated with the CCAN, Nup-Ran or SMC5-6 complexes can be found in Figure S1.

Even though the CCAN RF prediction was rich in associated proteins – this might be expected from a crowded chromatin environment – the entire condensin complex associated with the CCAN. This dependency may imply a potential relationship between these complexes that merits further study. Finally, Figure 3D shows that the CCAN RF prediction is independent from the SMC 5/6 complex, and no CCAN protein co-fractionated with ribosomal proteins (Figure S2). Together, these results show that, by integrating the outcome of several complex-specific Random Forests, we can reconstruct known dependencies at the kinetochore and identify novel inter-complex dependencies. Notably, none of these relationships were directly addressed a priori by the experiments used.

We suggest that this strategy to infer protein functions and relationships training RF with small protein complexes be named NanoRF. The results presented here constitute a proof-of-concept demonstration of the method in the context of the kinetochore. A thorough usage in the context of SMC complexes, as well as experimental verification of NanoRF predictions, can be found in Ohta *et al.*(2016). Our code, as well as a step-by-step

guide on how to perform NanoRF, has been made available through the Github repository https://github.com/EarnshawLab/NanoRF (Montano-Gutierrez, 2016).

DISCUSSION

A recurrent goal in the post-genomic era has been to make sense of increasing amounts of underexploited data, including noisy and incomplete proteomics output. Our results show that, even with high noise and when few experiments are informative, small groups of strongly covarying proteins –i.e. complexes– can be recognised judging by their coordinated behaviour using Random Forests. (Figure 1 and 2). In data of this type, statistical measures such as the mean correlation (Figure 1C) or absolute RF score of members in a complex can drop considerably (Figure 1D). We have demonstrated that lower RF scores can be informative as long as complexes separate out from contaminants by their RF score (Figure 1F). By tolerating a decrease of the RF score while maximizing separation quality, we were able to predict highly specific associations with complexes (Figure 2B) and retrieve known inter-complex relationships in our dataset (Figure 3). As no experiment targeted all of the complexes detected, this strategy could potentially identify protein function in any combination of comparable proteomics results.

In simple terms, NanoRF attempts to find the strongest possible signature for a complex (if any exists) within a specific dataset. The premise of our work is that any other protein whose RF score is as high as that of bona-fide members, whilst being clearly distinguishable from contaminants, is essentially difficult to distinguish from the complex itself. Our results with real complexes suggest that the strongest statistical associations we have found have biological relevance.

Comparison between NanoRF and other methods

Two previous studies from our group have used Random Forests to attempt to find general trends shared by functional members of chromosomes (Ohta et al, 2016) or interphase chromatin (Kutstatscher et al, 2014) in proteomics data. The evidence presented in the current work suggests that the 'true chromosome class' is the integration of the signatures of multiple protein complexes covarying in specific, distinguishable ways. Because of strong, yet conditional complex-specific covariance, adding more than one complex to a training class may restrict the performance of RF. Compared to MCCP and fractionation profiling (Borner et al, 2014), our prediction would upgrade, for example, from "true chromosomal protein" to "protein dependent on complex A but not complex B". To use a previously unmentioned example, the polybromo-and-BAF-containing (PBAF) complex (ARID2, PBRM1, BRD7, SMARCB1 and SMARCE1) associated specifically with Nup107-160 but not with the CCAN (Figure S1A). Consistent with this prediction, another bromodomain-containing protein, CREBBP, has been found to interact with Nup98 in Nup107-160 complex and was linked to Nup98 oncogenicity (Kasper et al, 1999).

Methods like Fractionation Profiling (FP) (Borner *et al.*, 2014) and multivariate proteomic profiling (MVPP) (Borner *et al.*, 2012) are based on guilt-by-association analyses to similarly detect protein complexes and have cleverly dealt with the intricate nature of proteomics data –i.e. presence of missing values– but the conditional covariance of the

complex –i.e. a signal present in only a few experiments– has not been accounted for previously. We have shown that NanoRF finds such covariance, even when there is significant noise. Consequently, NanoRF has successfully predicted proteins with previously uncharacterized links to mitosis (Ohta *et al.*, 2016).

NanoRF is a supervised method that is ideal to deeply explore complexes or protein groups that are already of interest to the researcher. Such groups may be either true protein complexes or any potential protein group hypothesized to covary. We have demonstrated that known protein complexes show a strong detectable signature, but other kinds of protein groups may be detectable as well. For discovery of protein complexes, other methods, including unsupervised RF or clustering are more suitable.

Potential pitfalls and statistical considerations of NanoRF

It is not possible to conclude from computational analysis alone that the relationships predicted by NanoRF are direct physical interactions between the aforementioned protein complexes. Nevertheless, our results come strictly from protein-level dependencies (or indirect effects of these) rather than changing expression levels, so physical associations are likely. Further support for this comes from another study, where we used the algorithm to explore chromosome structure. NanoRF associated VRK1 and PTPN6 with CCAN and RZZ respectively. Fluorescence microscopy showed localisation of VRK1 to chromosomes and PTPN6 to microtubules (Ohta et al, 2016).

We believe that finding the objectively best separation quality lessens the burden to select an arbitrary significance cut-off for candidates, especially as more uninformative experiments are collected. We have intentionally avoided using a hypergeometric P-Value as a significance measure since a) the exact P-values we obtained for all of our complexes were of 10^{-11} or smaller (Table S2), b) P-Values were strongly influenced by the number of proteins in the complex, and c) were undefined for some some of the random group RF results, where none of the two classes were above the MCC threshold (Figure 2B, lowest panel).

Instead of direct P-Value usage, the significance of the predictions by NanoRF is subject to the probability of obtaining a high separation quality by chance for a given dataset. To minimise the risk of type I error, we recommend that the Peak MCC for a complex at the cut-off should be higher and non-overlapping with the MCC's obtained for randomly assigned protein groups in a data set. In our analysis, the probability of obtaining an MCC as high as that of real complexes by chance showed negligible –our sampled MCC distributions did not overlap (Figure 2D), but it may vary for other datasets. Naturally, a lower MCC may be accepted at the risk of more false positives.

For prediction of associations with a complex, the false discovery rate for each complex should be proportional to the fraction of negative-class proteins that surpass the classification threshold. A small negative class could lead to underestimating false positives as higher noise may increase the RF score of spurious proteins. Therefore, a large negative class may be essential for a realistic False Discovery Rate estimation (Tarca *et al.*, 2007) and a small one could be compensated with a more stringent prediction cut-off for the RF-score.

Potential applications of NanoRF

In the context of all the massive protein-protein interaction networks being identified, we face a lack of detail in the functionality, hierarchy, specificity and conditionality of these interactions. We have shown that NanoRF could satisfy these unmet needs by providing deep insight about protein complexes.

Experiments are informative if members of a complex covary in them (Figure 1A). Differentiating between informative and non-informative experiments (feature selection) could itself be a powerful tool for protein complex data mining. For example, a specific set of perturbations may break the stoichiometry (and hence the correlation) in a complex. In this direction, our NanoRF pipeline (Montano-Gutierrez, 2016) includes a calculation of each experiment's 'importance' for classification, though exploiting such importance may not be straightforward. This estimation employs the Gini importance, which compares classification performance with or without a given experiment. A thorough analysis of importance measures is provided by (Louppe *et al.*, 2013).

We speculate that NanoRF could be performed on the same complex multiple times, each time using a distinct subset of experiments. These subsets could correspond, for example, to different time points or biological conditions, such as drug treatments. Such analysis could potentially inform how a complex's identifiability changes with the experiments, or whether there is a difference in associated proteins from one condition to the next. Such changes in retrieval may provide insight about conditional binding partners, or the biology of specific conditions, drugs or diseases.

Importantly, NanoRF does not require proteins to remain physically attached to each other during analysis, which may be difficult for weakly interacting or insoluble protein complexes such as associated in chromatin or membranes.

Here we described NanoRF, which uses supervised machine learning to a) detect protein complexes of interest in noisy datasets with few informative experiments, b) predicts proteins that have functional associations with specific complexes and c) evaluates the relationship between complexes according to their behaviour. NanoRF enables hypothesis-driven data analysis from ever-increasing, underexploited quantitative proteomics data. It is generally assumed that machine learning requires large training sets to work. However, we have established that Random Forests can retrieve strikingly small protein complexes, their associated proteins and relationships between complexes from ordinary proteomics results. We anticipate NanoRF to complement experimental co-fractionation approaches such as immunoprecipitation.

MATERIALS AND METHODS

Cell culture, mitotic chromosome isolation and SILAC Mass spectrometry

The present analysis was done by collecting and integrating the data from 2 previously published works (Ohta et al 2010, 2016). All cell culture, mitotic chromosome extraction and mass spectrometric analysis procedures are detailed therein. In brief, chromosomes were extracted from with wild-type chicken DT40 cells (clone 18), as well as conditional knockouts for chromosome structure proteins SMC2, CAP-H, CAP-D3, Scc1, or SMC5 (Hudson *et al.*, 2003; Green *et al.*, 2012) (Ohta et al, 2016) and a genetic knockout of kinetochore protein Ska3 (Ohta et al, 2016). All strains were incubated with nocodazole for 13 hours to arrest the cells in metaphase. In the case of the conditional knockouts, cells were incubated with deoxicycline for 20-60 hours to inhibit target gene expression prior to nocodazole treatment. Mitotic chromosomes of wild type and knockout cell lines grown respectively in 'heavy' and 'light' medium and were then mixed in equal amounts judging by Picogreen quantification. In the Ska3 KO experiment, samples were equated using Histone H4 as a reference. 30 trypsin-digested fractions were desalted using StageTips (Rappsilber *et al.*, 2003) and analyzed by liquid chromatography-MS on a LTQ-Orbitrap (Thermo Fisher Scientific) coupled to high-performance liquid chromatography via a nanoelectrospray ion source. MS data were analyzed using MaxQuant 1.0.5.12 for generating peak lists, searching peptides, protein identification (Cox and Mann, 2008), and protein quantification against the UniProt database (release 2013_07).

Preparation of MS data for NanoRF

The SILAC ratios from the 'Protein groups' Maxquant output table were used directly. As for the Ska3 knock out experiment, SILAC ratio column values were directly taken from (Ohta *et al.*, 2010), and re-indexed according to the rest of the experiments. The ratio columns in table S1 were directly and only used for the analysis. The features included consist only of the SILAC ratio columns of our experiments – no other feature selection, engineering or combination was performed a priori.

All the raw MS and Maxquant output data, including those from the Ska3 experiment (Ohta *et al.*, 2010), are available via ProteomeXchange with identifier PXD003588. Missing values were substituted by the median value of each experiment, as is common practice in Random Forest applications. We reasoned that doing so would penalize the lack of observations by giving the same score to missing proteins of both positive and negative classes, which in turn increases the intersection between classes and thereby decreases separation quality.

Random Forest analysis

The analysis was done with a custom R pipeline based on the Random Forests algorithm of Leo Breiman and Adele Cutler's Random Forest™ algorithm (Breiman, 2001), implemented in R (Liaw and Wiener, 2002). All our scripts used are freely available through a Github repository (Montano-Gutierrez, 2016) and include a step-by-step R guide script to perform NanoRF on any particular dataset. The Random Forests algorithm attempts to find a series of

requirements in the data that are satisfied by the positive training class and not by the negative training class. All these decisions are performed sequentially, hence they become a decision tree. An example of a decision tree would be “proteins with values >x in experiments 1 and 2. Out of those, proteins with values < y in experiments 3 and 5”. As the best set and decision sequence is not known a priori, the best bet is to generate many decision trees at random (hence the name random forest). Each tree votes for all compliant proteins as members of the positive class. The clearer the difference between the two classes in the data, the larger the number of trees that will vote for the positive class as indeed positive. The RF score (calculated for each protein) is the fraction of trees that voted for a protein as positive. In order to get a score for the members of the positive class as well, during the generation of each tree, some of the members of the positive and negative class are left out and treated as unknown. This Out-of-bag (OOB) procedure intrinsically controls for training set bias.

We set the number of trees in the forest to 3000 in each run. The Matthews correlation coefficient was calculated by using the formula

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TF+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TP indicates true positives, FP false positives, TN true negatives and FN false negatives. For null values of any of the sums in the denominator, the MCC was defined as 0. To choose a particular RF-Score as a cut-off, we evaluated 100 possible cut-offs between RF-scores 0 and 1 and kept that which maximized the MCC. For cut-offs with the same maximum MCC, the smallest RF was chosen as a cut off to maximize sensitivity.

Simulation and random protein group size choices

The group size choices for simulated and randomly chosen protein groups used throughout the manuscript were chosen to correspond as closely as possible to the values in our original dataset. The original dataset consists of 4998 protein rows identified in the Maxquant analysis, and 12 columns (each corresponding to one of the experiments we used for this analysis). The manually curated dataset of hitchhikers and contaminants (Ohta et al, 2010), which we used as the negative class, comprised 367 out of the 4998 protein rows. With those actual parameters in mind we defined the number of rows, columns and controls to 5000, 20 and 400, respectively, for numerical simplicity.

We expect the trends found in our simulations (though not the specific numbers) to be general to different dataset sizes. Importantly, we took care to minimize the number of variables in this analysis that are dependent on dataset size. In particular: a) We study the proportion (rather than raw number) of informative variables in the dataset. b) The results in figures 1DEF are the mean of 5 independent hypothetical complexes in order to yield robust results. c) The simulated experiments were standard normal distributions, which should yield convergent results with larger numbers.

Informative experiment fraction VS noise analysis:

We arbitrarily generated matrices with ~5000 'protein' rows and 20 'experiment' columns (sizes similar to our SILAC ratio matrix) by sampling a standard normal distribution. In each matrix, 365 'proteins' were selected to be part of the negative class and 5 groups of 12 proteins were set to be identical within their group in 2,4 … 20 'experiments' (Figure 1D-F, horizontal axis). Next, all the values in each matrix were jittered with Gaussian noise with standard deviation of of .02, .04 … 2 (Figure 1D-F, vertical axis). Missing values were not added to the simulations on the basis that they would be filled with median values, which would add variance and thus have similar effect to noise addition. We then ran the RF analysis for the 5 groups versus the negative set. The values in Figure 1 DEF are the mean of means of the RF score and of highest MCC for each positive group. The correlation was the mean of intra-group correlations of all positive groups.

Definition of protein group covariance

The covariance between random variables is only defined pairwise, and as such, the 'mean correlation of a complex' as mentioned in the text could be seen as a matrix A where $A_{i\,j}$ is the correlation of protein *I* with protein *j.* Several proxies of a single group-covariance measure exist. For practical purposes, the average of the lower triangular entries of the correlation matrix was used as a proxy of covariance.

ACKNOWLEDGEMENTS

This work was supported by a Wellcome Trust four-year studentship [grant number 089396] to LFM, a grant from the Uehara Memorial Foundation and the Nakajima Foundation to SO, a Wellcome Trust Senior Research Fellowship [grant number 103139] to JR and a Wellcome Trust Principal Research Fellowship [grant number 107022] to WCE. The Wellcome Trust Centre for Cell Biology is supported by a core grant [numbers 077707 and 092076] and the work was also supported by Wellcome Trust instrument grant 091020.

REFERENCES

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.

Borner, G. H. H., Antrobus, R., Hirst, J., Bhumbra, G. S., Kozik, P., Jackson, L. P., Sahlender, D. A., and Robinson, M. S. (2012). Multivariate proteomic profiling identifies novel accessory proteins of coated vesicles. J. Cell Biol. *197* , 141–160.

Borner, G. H. H., Hein, M. Y., Hirst, J., Edgar, J. R., Mann, M., and Robinson, M. S. (2014). Fractionation profiling: a fast and versatile approach for mapping vesicle proteomes and protein–protein interactions. Mol. Biol. Cell *25* , 3178–3194.

Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 5–32.

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 Biotech *26*, 1367–1372.

Fusaro, V. A., Mani, D. R., Mesirov, J. P., and Carr, S. A. (2009). Prediction of high-responding

peptides for targeted protein assays by mass spectrometry. Nat Biotech *27*, 190–198.
Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. Nat Rev Mol Cell Biol *8*, 645–654.
Green, L. C. *et al.* (2012). Contrasting roles of condensin I and condensin II in mitotic chromosome formation. J. Cell Sci. *125*, 1591–1604.
Havugimana, P. C. *et al.* (2012). A census of human soluble protein complexes. Cell *150*, 1068–1081.
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.
Issaq, H. J., Conrads, T. P., Janini, G. M., and Veenstra, T. D. (2002). Methods for fractionation, separation and profiling of proteins and peptides. Electrophoresis *23*, 3048–3061.
Kustatscher, G., Hégarat, N., Wills, K. L. H., Furlan, C., Bukowski-Wills, J.-C., Hochegger, H., and Rappsilber, J. (2014). Proteomics of a fuzzy organelle: interphase chromatin. EMBO J. *33*, 648–664.
Kwon, M.-S., Hori, T., Okada, M., and Fukagawa, T. (2007). CENP-C is involved in chromosome segregation, mitotic checkpoint function, and kinetochore assembly. Mol. Biol. Cell *18*, 2155–2168.
Leitner, A., Faini, M., Stengel, F., & Aebersold, R. (2016). Crosslinking and MassSpectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends in Biochemical Sciences*. https://doi.org/10.1016/j.tibs.2015.10.008
Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2*, 18–22.
Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In: Advances in Neural Information Processing Systems 26, ed. C. J. C. Burges, ed. L. Bottou, ed. M. Welling, ed. Z. Ghahramani, and ed. K. Q. Weinberger, Curran Associates, Inc., 431–439.
Mehta, G. D., Rizvi, S. M. A., and Ghosh, S. K. (2012). Cohesin: a guardian of genome integrity. Biochim. Biophys. Acta *1823*, 1324–1342.
Michaud, F.-T., Havugimana, P. C., Duchesne, C., Sanschagrin, F., Bernier, A., Levesque, R. C., and Garnier, A. (2012). Cell culture tracking by multivariate analysis of raw LCMS data. Appl. Biochem. Biotechnol. *167*, 474–488.
Montano-Gutierrez, L. F. (2016). https://github.com/EarnshawLab/NanoRF.
Moore, B. W., and Lee, R. H. (1960). Chromatography of Rat Liver Soluble Proteins and Localization of Enzyme Activities. J. Biol. Chem. *235* , 1359–1364.
Ohta, S. *et al.* (2010). The Protein Composition of Mitotic Chromosomes Determined Using Multiclassifier Combinatorial Proteomics. Cell *142*, 810–821.
Ohta, S. *et al.* (2016). Proteomics analysis with a nano Random Forest approach reveals novel functional interactions regulated by SMC complexes on mitotic chromosomes. Mol. Cell. Proteomics.
Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble Machine Learning: Methods and Applications* (pp. 307–323). https://doi.org/10.1007/9781441993267_10
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.
Screpanti, E., De Antoni, A., Alushin, G. M., Petrovic, A., Melis, T., Nogales, E., and

Musacchio, A. (2011). Direct binding of Cenp-C to the Mis12 complex joins the inner and outer kinetochore. Curr. Biol. *21*, 391–398.
Skinner, O. S. *et al.* (2016). An informatic framework for decoding protein complexes by top-down mass spectrometry. Nat Meth *13*, 237–240.
Sonoda, E. *et al.* (2001). Scc1/Rad21/Mcd1 is required for sister chromatid cohesion and kinetochore function in vertebrate cells. Dev. Cell *1*, 759–770.
Stephan, A. K., Kliszczak, M., Dodson, H., Cooley, C., and Morrison, C. G. (2011). Roles of vertebrate Smc5 in sister chromatid cohesion and homologous recombinational repair. Mol. Cell. Biol. *31*, 1369–1381.
Tarca, A. L., Carey, V. J., Chen, X., Romero, R., and Drăghici, S. (2007). Machine Learning and Its Applications to Biology. PLoS Comput Biol *3*, e116.
Uhlmann, F., Lottspeich, F., and Nasmyth, K. (1999). Sister-chromatid separation at anaphase onset is promoted by cleavage of the cohesin subunit Scc1. Nature *400*, 37–42.
Wu, N., and Yu, H. (2012). The Smc complexes in DNA damage response. Cell Biosci. *2*, 5.

FIGURE LEGENDS

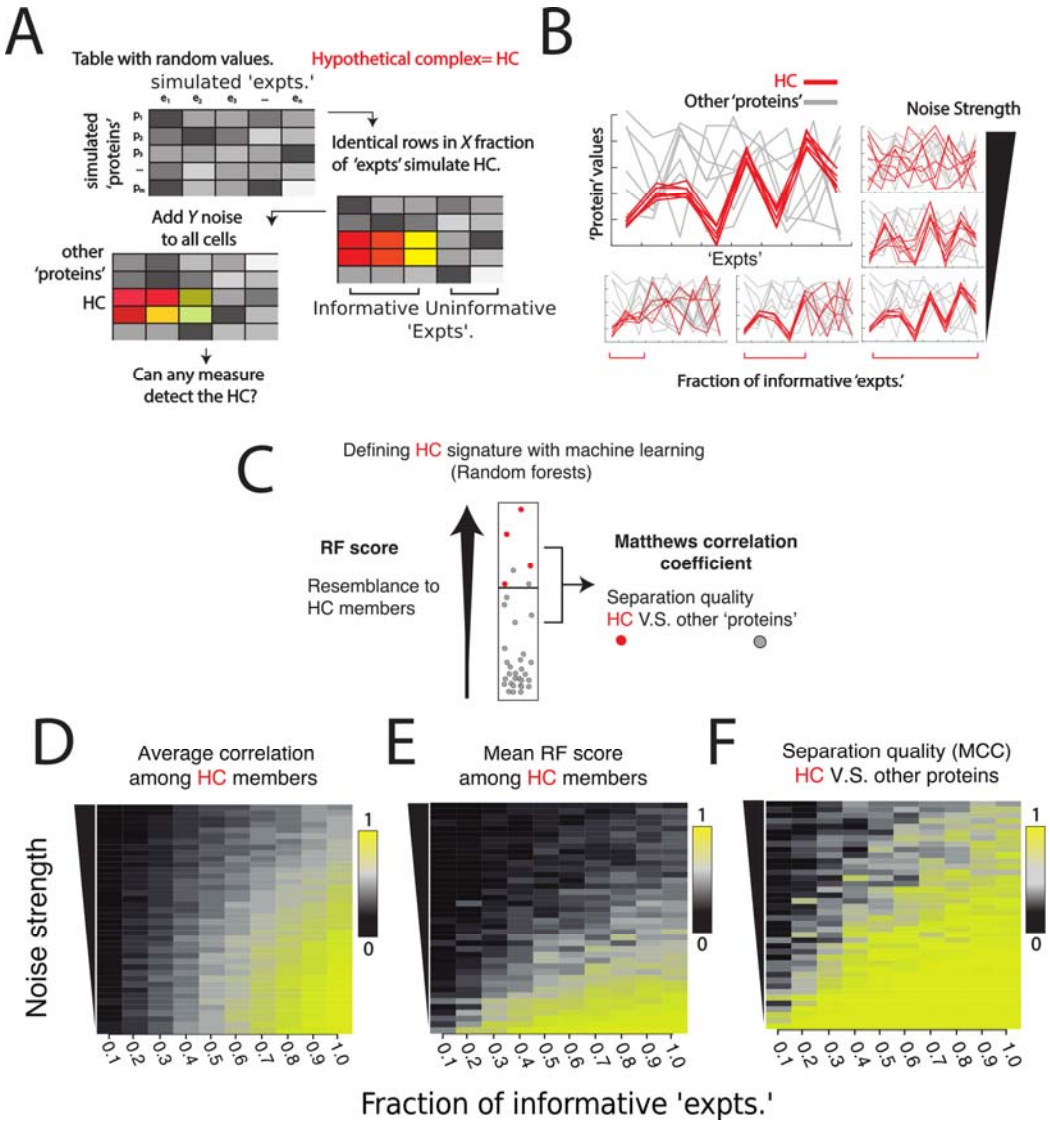

Figure 1. Supervised machine learning algorithm Random Forests can detect small, correlated protein groups in artificial proteomics data. A. Depiction of the procedure used to simulate proteomics data with 'protein' rows and 'experiment' columns. Some rows are made identical (red tones) in a fraction of experiments to simulate a hypothetical complex (HC) that correlates in some experiments, and Gaussian noise is then added element-wise to each table entry. B. Visual description of a hypothetical complex (red) versus other randomly generated proteins (grey) as the number of experiments (left-right) and the noise (bottom-up) affect the protein values in the experiments (all subpanels). C. Diagram to visualize the output from machine learning technique Random Forests. The RF score denotes how much a protein resembles the complex, while separation quality indicates how easily unrelated proteins covary with the complex. Red and grey dots depict the hypothetical complex and other proteins respectively. D,E,F. heat maps showing how the fraction of informative experiments (X axis) and the noise amount (Y axis) affect the Mean correlation (D) Random

Forest score (E) and separation quality (F) of proteins in a complex. In each square, the value projected is the mean of means of 5 independent groups.

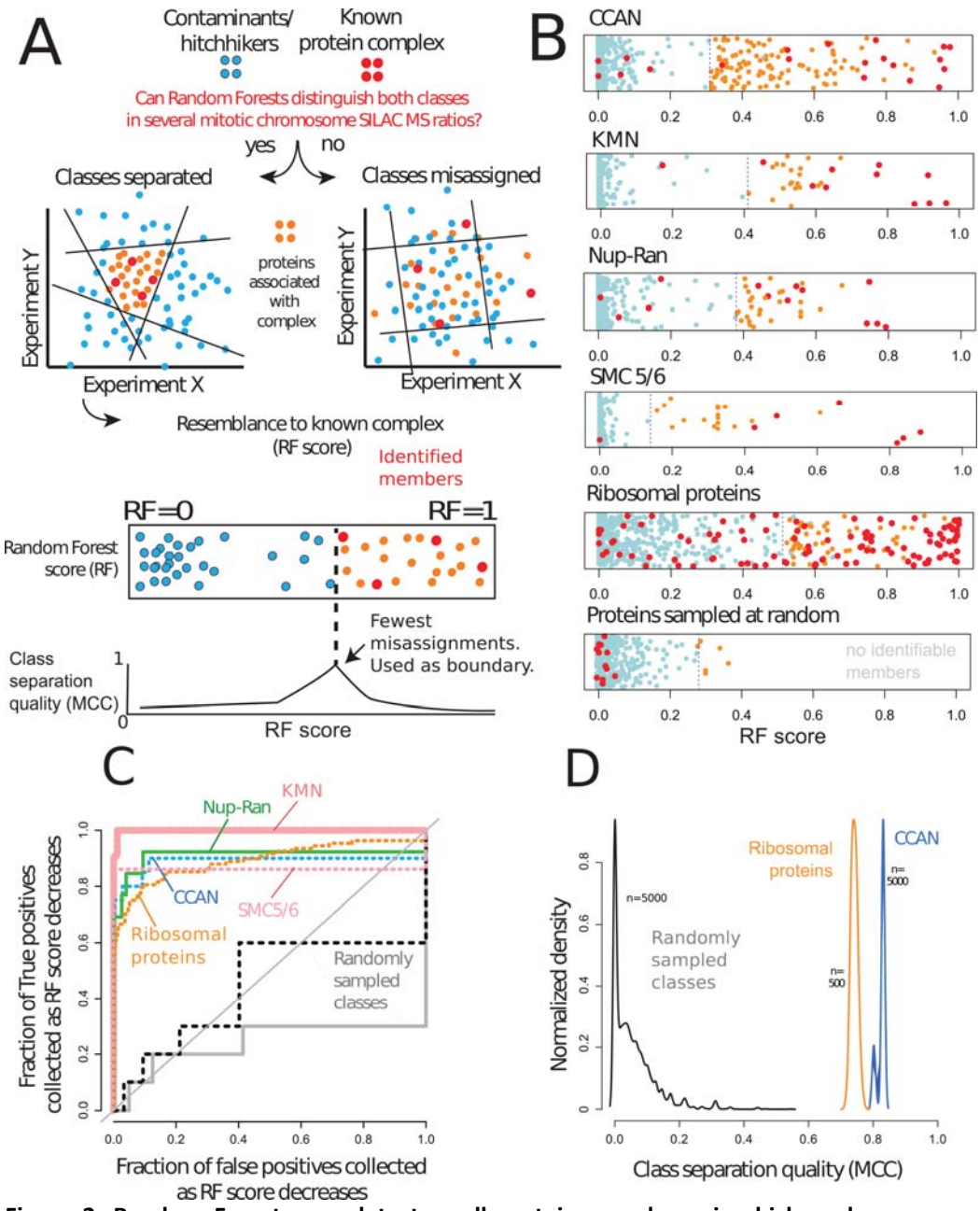

Figure 2. Random Forests can detect small protein complexes in chicken chromosome SILAC proteomics experiments. Entire figure: red-protein complex, blue tones-contaminants/hitchhikers. Orange dots: proteins showing high covariance with identifiable members of the complex (RF scores as high as the complex) that potentially associate with

it. A. Logic of the procedure to detect complexes with Random Forests. Groups separable in multiple dimensions (only 2 depicted) yield a higher MCC than inseparable groups. B. RF scores of multiple complexes versus the same set of contaminants/hitchhikers, and randomly selected groups from the table. "Bottom panel: Randomly chosen sets of proteins yield poor red/blue separation, implying that the members of the mock complex are unidentifiable. In this case, the next optimal cut-off is driven by exclusion of contaminants.
C. Receiver operating characteristic (ROC) performance curves of the RF as a classifier for each protein complex and for two randomly selected protein groups (grey, black). Diagonal shows the random assignment scenario. D. Kernel densities of MCC values for 500 random forest runs of each complex and 5000 runs for randomly assigned groups (black. Sample sizes: 10 for positive class and 425 for the negative class). All distributions were made of height 1 for visualization purposes.

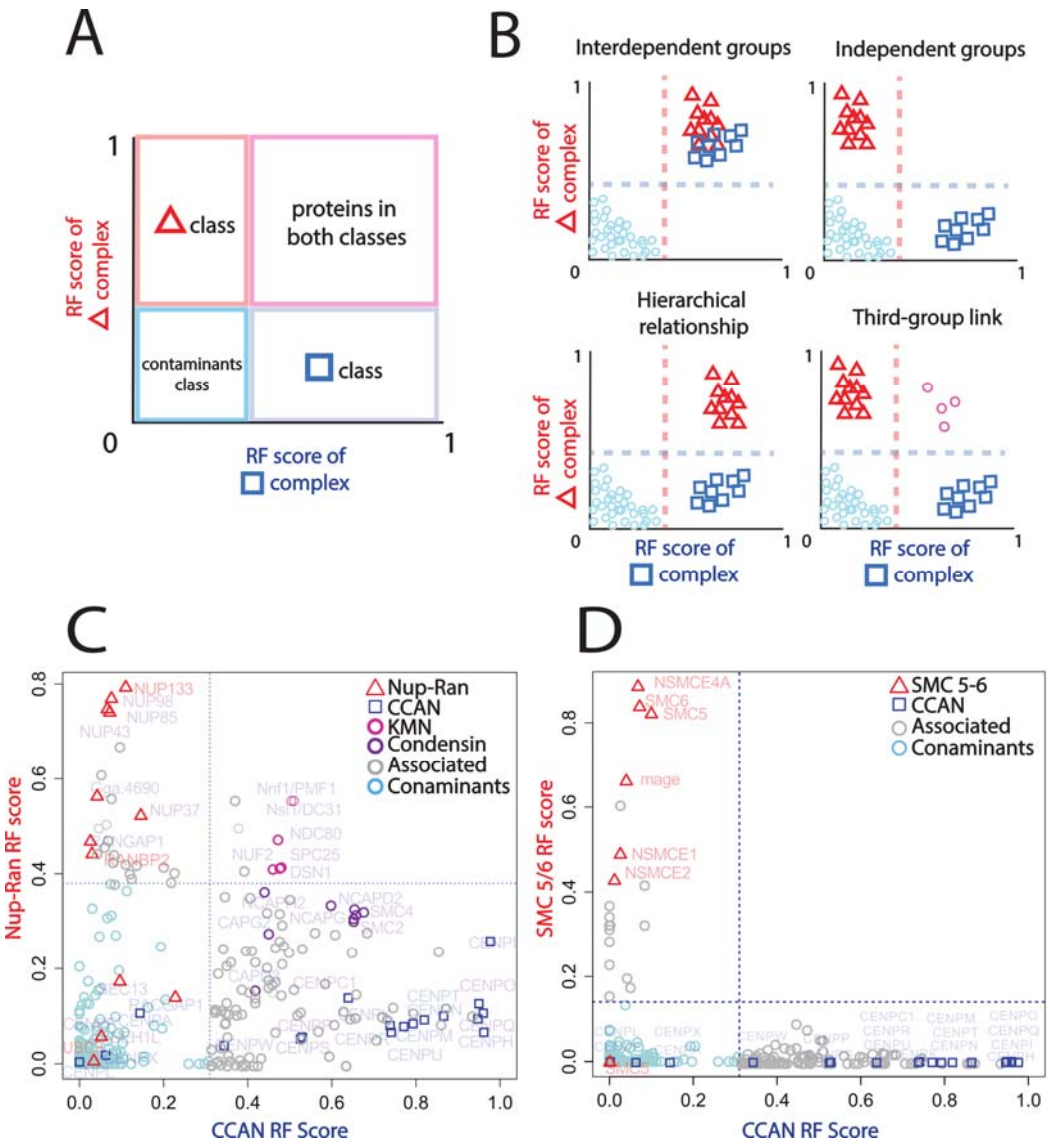

Figure 3. Known and novel interdependencies between complexes revealed by RF. A. Schematic of a 2D diagram to visualise intersections between Random Forests for different complexes. Highest separation quality thresholds are depicted by dotted lines. Proteins above both thresholds (pink quadrant) associate with both complexes whereas those above only one remain independent. B. Possible scenarios of interdependence between complexes inferred from 2D RF plots. A hierarchical effect (lower left) happens when the RF of one complex (squares) brings up the other complex (triangles) but not vice versa. A third-group link (lower right) is equivalent to a double hierarchical effect on the pink circle complex. Many three group relationships exist; in the case shown, the circle complex is the only link between the triangle and square complexes. C-D. 2D interdependence plot of the Constitutive Centromere-Associated network (CCAN, C and D, squares) versus the Nup107-160/RanGap complex (C, triangles) and the SMC 5/6 complex (D, triangles).