Published in final edited form as:

Science. 2020 December 11; 370(6522): . doi:10.1126/science.aaz4910.

Genetic interaction mapping informs integrative structure determination of protein complexes

Hannes Braberg[1,2,#], Ignacia Echeverria[1,2,3,#], Stefan Bohn[1,2,4,+,#], Peter Cimermancic[3,++,#], Anthony Shiver[5,+++], Richard Alexander[1], Jiewei Xu[1,2,4], Michael Shales[1,2], Raghuvar Dronamraju[6], Shuangying Jiang[7], Gajendradhar Dwivedi[8,§], Derek Bogdanoff[9], Kaitlin K. Chaung[9], Ruth Hüttenhain[1,2,4], Shuyi Wang[1], David Mavor[2,3], Riccardo Pellarin[3,++++], Dina Schneidman[3], Joel S. Bader[10], James S. Fraser[2,3], John Morris[11], James E. Haber[8], Brian D. Strahl[6], Carol A. Gross[12], Junbiao Dai[7], Jef D. Boeke[13,14,15,16,*], Andrej Sali[2,3,11,*], Nevan J. Krogan[1,2,4,17,*]

[1]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA.

[2]Quantitative Biosciences Institute, University of California, San Francisco, San Francisco, CA 94158, USA.

[3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA.

[4]Gladstone Institutes, San Francisco, CA 94158, USA.

[*]**Correspondence to:** Nevan J. Krogan - nevan.krogan@ucsf.edu, Andrej Sali - sali@salilab.org, Jef D. Boeke - jef.boeke@nyulangone.org.

[+]Present address: Department of Molecular Machines and Signaling, and Department of Molecular Structural Biology, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

[++]Present address: Verily Life Sciences, 269 E Grand Ave, South San Francisco, CA 94080, USA

[+++]Present address: Department of Bioengineering, Stanford University, Stanford, CA 94305, USA

[++++]Present address: Institut Pasteur, Structural Bioinformatics Unit, Department of Structural Biology and Chemistry, CNRS UMR 3528, C3BI USR 3756 CNRS & IP, Paris, France

[§]Deceased

[#]These authors contributed equally to this work.

Author contributions: H.B., I.E., P.C., J.D.B., A.Sa and N.J.K. designed the research; H.B., S.B., A.Sh, R.A., J.X., R.D., S.J., G.D., D.B., K.K.C., R.H. and S.W. performed experimental assays; I.E., P.C., R.P. and D.S. developed and performed integrative modeling; H.B., S.B., M.S. and D.M. analyzed experimental data; J.M. developed software; J.S.B., J.S.F., J.E.H., B.D.S., C.A.G., J.D., J.D.B., A.Sa and N.J.K. supervised the research; H.B., I.E., S.B., A.Sa and N.J.K. wrote the paper with input from all authors.

Competing interests: J.S.B. is a Founder and Director of Neochromosome, Inc. J.D.B is a Founder and Director of CDI Labs, Inc., a Founder of Neochromosome, Inc, a Founder and SAB member of ReOpen Diagnostics, and serves or served on the Scientific Advisory Board of the following: Sangamo, Inc., Modern Meadow, Inc., Sample6, Inc. and the Wyss Institute.

Data and materials availability: IMP modelling scripts, input data, and output results are available at http://integrativemodeling.org/systems/pemap. IMP is an open-source program, available at http://integrativemodeling.org. The clustered pE-MAP and correlation maps are available in the supplementary data, and can be visualized using the Java Treeview app (119), available at http://jtreeview.sourceforge.net/. The stEMAP app source is publicly available at https://github.com/RBVI/StEMAPApp and stEMAP will be available through the Cytoscape App store. stEMAP requires installation of Cytoscape and ChimeraX. Cytoscape is available at https://cytoscape.org/download.html and ChimeraX at https://www.rbvi.ucsf.edu/chimera /download.html. stEMAP also requires installation of the Sets, RINalyzer, StructureVizX, clusterMaker2 and ctdReader apps, all available through the Cytoscape apps menu. The modified version of PDB 1ID3 (including the unstructured H3 and H4 tails for illustrative purposes) is available in the supplementary data. The yeast RNAPII pE-MAP is available in the supplementary materials of (15). The bacterial RNAP point mutation data is available in the supplementary materials of (20). The library of histone mutant strains is available from J. Dai under a material transfer agreement with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. RNA-seq and ATAC-seq data have been deposited in NCBI's Gene Expression Omnibus (140) and are accessible through GEO Series accession number GSE160350 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160350).

[5]Graduate Group in Biophysics, University of California San Francisco, San Francisco, CA 94158, USA.

[6]Department of Biochemistry and Biophysics, University of North Carolina School of Medicine, Chapel Hill, NC 27599, USA.

[7]CAS Key Laboratory of Quantitative Engineering Biology, Guangdong Provincial Key Laboratory of Synthetic Genomics and Shenzhen Key Laboratory of Synthetic Genomics, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China

[8]Department of Biology and Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454, USA.

[9]Center for Advanced Technology, Department of Biophysics and Biochemistry, University of California, San Francisco, San Francisco, CA 94158, USA.

[10]Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

[11]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA.

[12]Department of Microbiology and Immunology and Department of Cell and Tissue Biology, University of California, San Francisco, San Francisco, CA 94158, USA.

[13]NYU Langone Health, New York, NY 10016, USA.

[14]High Throughput Biology Center and Department of Molecular Biology & Genetics, Johns Hopkins University School of Medicine, Baltimore, MD, 21205 USA.

[15]Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, NYU Langone Health, New York, NY 10016

[16]Department of Biomedical Engineering, NYU Tandon School of Engineering, Brooklyn NY 11201

[17]Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Abstract

Determining structures of protein complexes is crucial for understanding cellular functions. Here, we describe an integrative structure determination approach that relies on *in vivo* measurements of genetic interactions. We construct phenotypic profiles for point mutations crossed against gene deletions or exposed to environmental perturbations, followed by converting similarities between two profiles into an upper bound on the distance between the mutated residues. We determine the structure of the yeast histone H3-H4 complex based on ~500,000 genetic interactions of 350 mutants. We then apply the method to subunits Rpb1-Rpb2 of yeast RNA polymerase II, and subunits RpoB-RpoC of bacterial RNA polymerase. The accuracy is comparable to that based on chemical cross-links; using restraints from both genetic interactions and cross-links further improves model accuracy and precision. The approach provides an efficient means to augment integrative structure determination with *in vivo* observations.

A mechanistic understanding of cellular functions requires structural characterization of the corresponding macromolecular assemblies (1). Traditional structural biology methods, such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy (EM), rely on purified samples and are generally not applicable to heterogeneous samples, such as those of large, membrane-bound, or transient assemblies (2). Moreover, these methods do not determine the structures in their native environments, therefore increasing the risk of producing structures in non-functional states or missing relevant functional states.

Integrative structure determination has emerged as a powerful approach for determining the structures of biological assemblies (3). The motivation is that any system can be described most accurately, precisely, completely, and efficiently by using all available information about it, including varied experimental data (*e.g.* chemical cross-links, protein interaction data, small-angle X-ray scattering profiles) and prior models (*e.g.* atomic structures of the subunits). Integrative methods can often tackle protein assemblies that are difficult to characterize using traditional structural biology methods alone (1, 4-10). Spatial data generated by *in vivo* methods is especially useful for integrative structure determination (11). Therefore, high-throughput *in vivo* methods are needed to supplement low-throughput *in vivo* methods, such as single molecule Förster resonance energy transfer (FRET) spectroscopy (12).

Here, we describe how integrative structure modeling can benefit from spatial restraints derived from *in vivo* quantitative measurements of genetic interactions. A genetic interaction between two mutations occurs when the effect of one mutation is altered by the presence of the second mutation (Fig. 1A) (13). Positive genetic interactions (epistasis/suppression) arise when the double mutant is healthier than expected, whereas negative interactions (synthetic sickness) arise in relationships where the double mutant is sicker than expected. Single genetic interactions can often be difficult to interpret in isolation. A phenotypic profile, defined as a set of genetic interactions between a given mutation (*e.g.* a point mutation) and a library of secondary mutations (*e.g.* gene deletions), can be more informative (Fig. 1B) (14). A point mutant epistatic miniarray profile (pE-MAP) is comprised of such phenotypic profiles for all mutations in the analysis (Fig. 1C) (15). We have previously found a statistical association of the distance between two mutated residues in the wt structure and the similarity between their phenotypic profiles (*i.e.* phenotypic similarity) (15, 16) (Fig. 1D). This observation is in agreement with the expectation that mutations within the same functional region (*e.g.* active, allosteric, and binding sites) are likely to share more similar phenotypes than those that are distant in space (17-19). Here, we explore how to use these associations for determining *in vivo* structures of macromolecular assemblies using integrative modeling (Fig. 1E). To enable this analysis, we generated a large pE-MAP, by designing a comprehensive set of 350 mutations in histones H3 and H4 and crossing these against 1,370 gene deletions (or hypomorphic alleles for essential genes). We describe this pE-MAP and illustrate integrative structure determination by its application to three complexes of known structure: (i) the yeast histones H3 and H4; (ii) subunits Rpb1 and Rpb2 of yeast RNA polymerase II (RNAPII), using a pE-MAP dataset of 53 point mutants crossed against a library of 1,200 deletions and hypomorphic alleles (15); and (iii) subunits RpoB and RpoC of bacterial RNA polymerase (RNAP), using a chemical genetics miniarray profile (CG-MAP), where 44 point mutants were subjected to 83 different environmental stresses (*e.g.* treatments with chemicals and temperature shocks) (20).

A comprehensive pE-MAP of histones H3 and H4

Histones are central to chromatin structure and dynamics, as they make up the core of the nucleosome, the fundamental repeating unit of chromatin. The state of the nucleosome is controlled by histone post-translational modifications (PTMs) (21), including acetylation, methylation, phosphorylation and ubiquitination, that help maintain and regulate chromatin structure and transcription. Our library of point mutations in the core histones H3 and H4 was designed to comprise a comprehensive alanine-scan, as well as context-specific mutations of modifiable residues (e.g. lysine and arginine), such as charge removal/reversal and substitutions mimicking PTMs (22, 23). Partial deletions of the N-terminal tails of H3 and H4 were also included, as these regions play important and sometimes redundant roles in chromatin biology (24, 25). In budding yeast, histones H3 and H4 are expressed from two loci each, *HHT1/HHT2* and *HHF1/HHF2*, respectively. To ensure preservation of the native expression levels, we engineered each strain to include identical point mutations in both relevant loci with separate selection markers (HYG^R and *URA3*) (Fig. 2A). In total, we designed 479 histone mutants, of which 350 were amenable to pE-MAP analysis (Fig. 2B-D, Table S1); the remaining 129 mutants either were lethal or exhibited very poor growth, rendering them inaccessible to genetic analysis (Fig. S1). The histone mutants were crossed against a library of 1,370 gene deletions and hypomorphic alleles (Table S1) using our triple mutant selection strategy (26, 27) involving three different selectable markers (HYG^R and *URA3* to select for both copies of the histone alleles and KAN^R for the knockout library strains) (Fig. 2A, Methods) (26). Genetic interactions were quantified using the S-score (28), which measures the deviation of the double mutant fitness from the expected combined effect of the individual mutations (Methods). The pE-MAP screen was carried out in 3 biological replicates (Methods), which exhibit a high reproducibility (Fig. 2E), and the final S-scores (as depicted in Fig. 1C) are the averages of these replicates.

It has been shown that a pE-MAP can be used to predict protein-protein interactions (PPIs), by comparing the genetic interaction patterns between pairs of deletion mutants across all the point mutants (15, 29). On a global level, this is only possible if the point mutant set affects a broad group of processes and exhibits genetic interactions with the many different deletion mutants that encode the PPI proteins. Since the histone mutant collection perturbs only two proteins (H3 and H4), we set out to investigate whether the resulting phenotypic profiles are sufficient to predict PPIs among the 1,370 deletion mutants. Using a receiver operating characteristic (ROC) curve, we find that the histone pE-MAP predicts PPIs similarly to previous E-MAPs that affect more genes (15, 29) (Fig. 2F, Methods). This finding indicates that the combined set of histone point mutants affects a broad set of cellular processes, reflecting the multifunctional nature of histones H3 and H4 and their central role in controlling the global genetic environment of cells.

To gain insight into the regulatory hierarchy that drives the widespread functional effects of histone perturbations, we set out to examine the relationship between genetic interactions and gene expression changes. To this end, we determined the genome-wide gene expression levels for 29 representative histone mutants using RNA-seq and found no correlation between the expression change of a gene resulting from a given histone mutation and the corresponding S-score (Fig. 2G, Table S2). This indicates that observed genetic interactions between histone mutations and deletion mutants are due to complex regulatory patterns, rather than the histone mutation directly modulating the expression of the interacting gene.

The pE-MAP was clustered hierarchically along both dimensions (Fig. 3A, Data S1) and effectively recapitulates known protein complex and pathway memberships. For example, the pE-MAP identified COMPASS (30, 31), Swr1-C (32), and the Set2/Eaf3 pathway (33-35), as well as clusters of genes linked to telomere maintenance and Golgi/ER traffic (Fig. 3B, Data S1). Furthermore, mutations of histone residues in close proximity to each other (e.g. mutants of the H3 or H4 N-terminal tails) tend to show similar phenotypic profiles (Fig. 3, Fig. S2A). Overall, we find that histone tail deletion mutants give rise to stronger phenotypic profiles than the point mutants (Fig. S2B), reflecting the multiple residue perturbations and the importance of functional histone tails for cell homeostasis.

Phenotypic profile similarities are correlated with structural proximity

Similarities between pairs of phenotypic profiles in the histone H3-H4 pE-MAP were quantified by the maximal information coefficient (MIC) (36, 37) (Fig. 1D, Fig. S3, Methods). The MIC values between pairs of phenotypic profiles do not linearly correlate with the distances between the mutated residues in the wt structure (Pearson correlation coefficient of −0.07, Cα–Cα distances), but are informative about an upper distance bound between the residues (Fig. 4A, Fig. S3C). The upper distance bound was obtained by binning the MIC values into 20 intervals and selecting the maximum distance spanned by any pair of residues in each bin, followed by fitting a logarithmic decay function to these maximum distances (Fig. 4A, Fig. S3C, Methods). The data show that a pair of proximal point mutations are more likely to have a high MIC value than a pair of distal point mutations. However, not all proximal mutations have a high MIC value: most pairs of phenotypic profiles, even those for residues that are less than 16 Å apart, are highly dissimilar (94% of all pairs exhibit a MIC value <0.3). These observations justify converting the pE-MAP data into a Bayesian data likelihood that provides an upper bound on the distance spanned by the mutated residues (Fig. 4B, Methods). This Bayesian term objectively interprets the noise in the experimental data and allows us to quantify the uncertainty of the resulting structural models. The complete scoring function for evaluating any structural model also includes simple terms accounting for excluded volume and sequence connectivity, in addition to the Bayesian terms for all pairs of profiles in the pE-MAP with a MIC value above 0.3 (Fig. 5).

Spatial restraints derived from pE-MAP data can be used for integrative structure determination

An ensemble of the H3-H4 dimer configurations that satisfy the input information (*i.e.* the model) was found by exhaustive Monte Carlo sampling guided by the scoring function, starting with random initial configurations of the rigid comparative models of the H3 and H4 subunits (Fig. 5, Methods). The resulting ensemble is accurate and precise, as demonstrated by the similarity between the X-ray structure (PDB: 1ID3, (38)) and model contact maps (Fig. 6A-B). Specifically, the mean accuracy is 3.8 Å (Fig. 6C); the accuracy is defined as the average Cα root-mean-square deviation (RMSD) between the X-ray structure and each of the structures in the ensemble. The precision is 1.0 Å (Fig. 6C), which is defined as the average RMSD between all solutions in the ensemble. As a control, we also computed a model from randomly shuffled MIC values. The resulting model (Fig. 6D) is incorrect (mean accuracy of 15.8 Å and incorrect contact map; Fig. 6C-D) and imprecise (7.6 Å; Fig. 6C). As another control, we computed a model by a state-of-the-art protein-protein docking method (39), resulting in a model with an inferior accuracy of 6.9 Å (Fig. S4). Finally, we also mapped the accuracy and precision of the model as a function of the fraction of the pE-MAP data used (Methods). As expected, the more pE-MAP data are used, the more accurate and precise is the model (Fig. 6C).

To compute the structure of a protein complex for which the structures of the components are known, we estimate that 35-40 mutations per component are necessary to generate a complex model with precision sufficient to map the positions and relative orientations of the components (Fig. S5, Methods). What is a useful model precision depends on the questions asked (40). Fortunately, many questions can often be answered by models as precise as those obtained based on pE-MAP data (RMSD range of 1-15 Å). Some examples include describing the architecture and evolution of protein assemblies (8, 41), designing interface mutations (42), characterizing structural heterogeneity of protein complexes (42, 43), and mapping binding-induced structural changes (44). Importantly, the estimate of 35-40 mutations is an upper bound, and, in many cases, the number might be reduced by specifically exploiting point mutations that target surface residues and/or residues known to be functionally important, and by choosing substitutions likely to give rise to functional perturbations. The outcome of these calculations indicates the utility of the pE-MAP data for integrative structure determination.

The pE-MAP connects individual histone residues and regions to other associated complexes and processes

To examine whether the pE-MAP can identify interactions with complexes that are not stably associated with histones, we investigated the relationships between modifiable histone residues and their cognate enzymes (modifier pairs). Interestingly, we observed a dramatic increase in S-scores within specific modifier pairs, as compared to the overall genetic interaction distribution (Fig. 7A, Table S4). The positive S-scores reflect that a modifier and its target residue often are epistatic/suppressive because they function in the same pathway. To test if this pattern extends to phenotypic profile similarities, we integrated the histone pE-MAP into a merged map of previously collected genetic interaction data for gene deletions and hypomorphic alleles (45). We computed Pearson correlation coefficients for each histone mutant phenotypic profile across the merged map, generating a correlation map of 350 histone mutants against 4,414 whole gene perturbations (Data S3, Methods). In agreement with the individual S-scores, the specific modifier pairs exhibit significantly higher phenotypic profile correlations than the overall map (Fig. 7A, Table S4). These findings show that the pE-MAP can be used to pair specific residues to their respective modifiers, even though these are not stably associated with the histones. For example, when components of COMPASS, which methylates histone H3K4 (30, 31, 46), are deleted (*swd1*Δ*, swd3*Δ*, sdc1*Δ*, bre2*Δ), we observe strong positive S-scores with both H3K4 mutants (K4R and K4Q) as well as high correlations of the phenotypic profiles between the H3K4 mutants and COMPASS deletions (Fig. 7B, Data S1, Data S3).

To explore these relationships in a structural context, we developed a Cytoscape (47) app named stE-MAP (structure E-MAP) that interactively maps the genetic interactions of pE-MAP gene clusters onto the point mutated protein structure. stE-MAP connects Cytoscape to ChimeraX (48) and displays connections between a pre-defined set of genes and all mutated residues for which the underlying interactions pass user-defined criteria (Fig. S6A). We mapped the genetic connections between COMPASS and all histone residues with which it exhibits >0.2 median correlation (Methods). Only 5 residues pass this threshold, and the strongest connection is displayed by H3K4. The other 4 residues (H3K1-H3K3 and H3K5) are proximal and are thus likely to interfere with the interaction between COMPASS and H3K4 (Fig. 7C). This finding is particularly notable since these residues reside in the most distal region of the unstructured H3 N-terminal tail. Given that we do not have COMPASS point mutations in our dataset, we did not attempt to model this interaction. However, analysis of the MIC values associated with the H3 and H4 tails, and their relationship with the core domains, indicates that distance restraints for the histone tails could be derived from the pE-MAP data. Specifically, the MIC value distributions for the tail-core and tail-tail pairs of mutations are similar to that of the core-core mutations (Fig. S5C). This similarity indicates that we can derive distance restraints for the histone tails, thus, in principle supporting the feasibility of integrative structure modeling of disordered regions. In such modeling, to avoid overinterpretation of the data and to account for a possibility that the pE-MAP data of the histone tails reflect interactions with neighboring nucleosomes, we would have to include multiple nucleosome copies in the model and allow for assignment ambiguity, just as we do for distances inferred from chemical cross-links and protein proximity inferred from affinity co-purification (49).

We observe similar trends for other histone modifiers. For example, members of the Set2 pathway (Set2, Eaf3, Rco1, Ctk1) (33-35) rank highly in the distributions of S-scores or correlations for mutations of their target residue, H3K36 (Fig. S6B-C). Interestingly, we also found instances where different mutations of a single residue identify connections to different modifiers. For example, the phenotypic profile of the deacetylation mimic H3K56R is similar to that of deletion of *RTT109*, which encodes the H3K56 acetylase (Pearson correlation coefficient of 0.4) (29), whereas the acetylation mimic H3K56Q instead correlates with the profile generated from the deletion of the corresponding deacetylase, *HST3* (Pearson correlation coefficient of 0.35) (50) (Fig. 7D). H3K56R further correlates with *asf1*Δ, *rtt101*Δ, *mms1*Δ and *mms22*Δ, whose corresponding proteins play key roles in the H3K56 acetylation pathway and downstream H3 ubiquitylation (51, 52) (Fig. 7D). Accordingly, the stE-MAP app identified strong links between Hst3-Hst4 and H3K56Q, as well as Rtt109-Asf1 and H3K56R (Fig. S6D). While we find that it is often informative to group different mutations of the same residue together, these examples highlight the potential of these maps for deeper mechanistic insights where required.

Expanding on these findings, we built a gene set enrichment map connecting the modifiable histone residues to nuclear processes (Fig. S7A, Table S5, Methods). We observe both known and novel connections. For example, "DNA recombination & repair" is connected to 4 residues, and two of these, H3K56 and H3K79, have been shown to play key roles in yeast's DNA repair (53-57). Interestingly, we find that mutations of the other two residues (H3R63K and H4R36K) result in increased spontaneous mutation frequencies at the *URA3* locus, indicating that these residues also function in DNA repair (Fig. S7B-C, Table S6, Methods).

The gene set enrichment analysis also identified 13 residues connected to cryptic transcription (Fig. S7A). The pE-MAP includes 24 different mutations of these residues, and we tested their involvement in cryptic transcription by quantifying the abundance of transcripts at the 5' and 3' end of the *STE11* gene, using qPCR (Fig. S7D, Methods). In total, 16 mutations, distributed among 10 residues, increase 3' transcript abundance by >50% compared to wt (Table S7), and 9 mutants among 5 residues increase 3' transcription over two-fold, without major changes in 5' transcription (Fig. S7E). As expected, H3K36A, H3K36R, H3K36Q and *set2*Δ increase 3' transcript abundance strongly, as do mutations of H4K44, which is a residue known to affect cryptic transcription (58). Interestingly, H3K122A increases 3'-transcript abundance >15-fold and, using ATAC-seq, we find that the mutation gives rise to nucleosome free regions in *STE11* and other genes known to produce cryptic transcripts (Fig. S7F-H). H3K122A exhibits positive genetic interactions with deletion of the histone chaperone *SPT2* (S-score = 2.4) and the nucleosome remodeling factor *CHD1* (S-score = 4.9), which are both involved in cryptic transcription (59, 60). Accordingly, we find that deletion of either *SPT2* or *CHD1* suppresses the cryptic transcription phenotype observed in H3K122A to wt levels, even though *spt2*Δ or *chd1*Δ alone has no effect (Fig. S7I-K).

The integrative structure determination approach is transferable to other complexes

To test whether genetic interaction mapping can be used to determine the structure of other complexes, we examined a pE-MAP of RNAPII in budding yeast (15). This pE-MAP consists of 53 point mutants crossed against a library of 1,200 gene deletions and hypomorphic alleles. Interestingly, the association between MIC values and the upper distance bound is also apparent in this dataset (Fig. S8A-B), even though the protein sizes and mutational coverage of the polymerase system (up to ~1,700 residues and 1-2%, respectively) are vastly different from those of the histones (<140 residues and 85-90%). These observations suggest that our parameterization of the pE-MAP spatial restraint based on the histone data may be generally applicable. To evaluate this expectation directly, we next modeled subunits Rpb1 and Rpb2 of RNAPII using the Bayesian likelihood parametrization based on the histone pE-MAP. To illustrate the modeling of higher order complexes, we divided Rpb1 into two domains, thereby representing the system with three rigid-bodies (Table S8, Methods). We obtained a model with a mean accuracy of 16.8 Å and precision of 9.8 Å (Fig. 8A-D, Table S8). This positive result illustrates the generality of the pE-MAP based spatial restraints.

To further assess the utility of pE-MAP data for structure determination, we compared the RNAPII model obtained using pE-MAP to a model using 22 previously published chemical cross-links (XLs) (61). Cross-linking is widely used for integrative structure determination of macromolecular assemblies (2, 8). Interestingly, a model of yeast RNAPII based on the pE-MAP data is as accurate as that based on the cross-links (16.8 Å and 16.7 Å, respectively; Fig. 8D). Moreover, the accuracy and precision of the model improves if both datasets are used simultaneously (10.2 Å and 3.7 Å, respectively; Fig. 8D), indicating complementarity between the two types of data and demonstrating a premise of integrative structure determination (Fig. 5). While a cross-link between two residues may provide more direct structural information than the corresponding pE-MAP pair, the number of possible cross-links is limited by the number of proximal reactive residue pairs, whereas the number of pE-MAP pairs grows quadratically with every additional point mutation introduced. Therefore, the larger number of less precise pE-MAP restraints can lead to a more accurate model than a smaller number of more precise cross-links.

The integrative structure determination approach is transferable to other types of phenotypic profiles

To examine the applicability of our approach to other types of phenotypic profiles, we turned to a CG-MAP of 44 bacterial RNAP point mutations exposed to 83 different environmental stresses (*e.g.* chemical perturbations, temperature stress, and pH change) (20). We observe an association between MIC values and the upper distance bound, similar to that of the pE-MAP datasets (Fig. S8, Methods). We modeled the structure of subunits RpoB and RpoC of the bacterial RNAP with a mean accuracy of 15.0 Å and precision of 6.6 Å (Fig. 8E-H, Table S9). This result suggests that maps with relatively small numbers of orthogonal phenotypes per point mutation can be used to accurately predict the architecture of macromolecular assemblies. Considering that constructing large gene deletion libraries and crossing them against point mutations can be laborious, environmental phenotypic profiles may be a more efficient alternative for generating spatial restraints for integrative structure determination than genetic interaction phenotypic profiles.

Spatial restraints derived from pE-MAP data are comparable to other commonly used data types

Co-evolution information can also be used to predict the structure of protein assemblies (18, 62, 63). However, the success of such modeling is heavily dependent on the number of sequences in the input sequence alignments and the ability to discriminate interacting from non-interacting homologs in genomes with multiple paralogs (64). Using the RaptorX protein complex contact prediction server (65, 66), we predicted the interfacial contacts between RpoB and RpoC of the bacterial RNAP; the numbers of homologous sequences were insufficient for the yeast histones and RNAPII (Methods). Importantly, RaptorX is based on a combination of co-evolution analysis and a deep-learning algorithm that reduces the requirement for sequence homologs and improves accuracy (67). Other commonly used co-evolution methods (18, 62) did not identify any interfacial contacts. Similar to the pE-MAP and CG-MAP datasets, we observe a negative statistical association between the residue pair coupling strengths and the upper distance bound (Fig. S9). To mimic the pE-MAP restraint, we converted the top coupling strengths into upper distance bound restraints (Methods). The model ensemble computed from coevolution derived restraints includes two different sets of configurations (mean accuracy of 22.6 Å; model precision of 9.0 Å; Fig. 8H). Only a fraction of the bacterial RNAP structures computed using co-evolution derived restraints are as accurate as those computed using the CG-MAP restraints. The model precision and accuracy of the model improve slightly if both types of restraints are combined (mean accuracy of 14.5 Å; model precision of 6.5 Å; Fig. 8H).

Discussion

In summary, we show that the architectures of macromolecular assemblies can be determined using quantitative genetic interaction data collected *in vivo*. The accuracy and precision of such models are comparable to those of models based on chemical cross-linking or co-evolution analysis. A key premise of integrative modeling is that using several different types of data improves the accuracy and precision of the model. Because the pE-MAPs and CG-MAPs contain purely phenotypic measurements, collected in living cells, these datasets generate spatial restraints that are orthogonal to other commonly used data for integrative modeling. Since this data reflects *in vivo* structures, and is thus unlikely to share artifacts of biophysical methods, it could be of particularly high value in the integrative modeling process. The genetic interaction data may also allow for the characterization of complexes that are difficult to isolate and purify or those that are only transiently stable. Importantly, the equipment required for generating these data is basic and in particular the CG-MAPs can be generated efficiently. Recent developments in CRISPR/Cas9 based approaches have paved the way for multiplexed precision genome editing in yeast (68), allowing for rapid generation of CG-MAPs. Together, these methods make feasible the proteome-wide modeling of protein complex structures, guided by global protein-protein interaction maps (69). In addition to proteins, the approach is also applicable to assemblies containing nucleic acids, thus further expanding the scope of integrative structural biology. pE-MAPs and CG-MAPs are complementary to other high-throughput functional assays. For example, two recent studies have shown that deep mutational scans can be used for determining the fold of a small protein domain (70, 71). If these methods prove useful for multi-domain proteins or disordered regions (72), the measured phenotype changes could be used to derive additional restraints for the integrative structure determination.

The relationship between phenotypic pE-MAP measurements and structure can be uncertain. The reasons for this include mutations in distant positions that are part of an allosteric network and could give rise to similar profiles, mutations that are functionally irrelevant, and mutations that perturb gene expression, mRNA stability or translation. Additionally, the approach relies on the introduction of point mutations into the proteins of interest, which may result in structural changes. However, proteins often adapt to mutations by small local changes in their structure, maintaining their overall fold and function (73). Mutations that cause major misfolding of essential proteins and/or assemblies are uncommon in pE-MAPs since the resulting fitness defects typically prevent successful screening. The method could be improved by specifically designing point mutants that do not alter the structure and/or lead to aggregation, by selecting commonly allowed mutations as determined by divergent protein sequence alignments (74).

The aim of integrative structure determination is to model the structures of macromolecular assemblies. This often requires the structures of the individual components (from X-ray crystallography, NMR, cryo-EM, comparative modeling, or increasingly, *ab initio* structure prediction (65, 67, 75, 76)). The quality of the structures of the individual components and input data are crucial for integrative (or indeed any other) structural approaches, and one cannot achieve a precise structure from low-quality starting structures or data. Even so, there are numerous examples of utility of structural models at lower resolution (3). For example, these models can be used to explain the architectural principles of large assemblies (8, 41, 77, 78), describe the structural dynamics of protein complexes (42, 43), or rationalize the impact of many mutations (41). A lower resolution structure is also often a useful starting point for higher resolution structure characterization.

CRISPR/Cas9 genome editing (79) has proven highly effective for high-throughput genetic interaction mapping in mammalian cells (80, 81). To date, these efforts have relied on whole-gene perturbations, but methods for systematic generation of point mutants using CRISPR/Cas9 have recently been developed (82, 83), paving the way for mammalian pE-MAP screening. This advance provides a means for integrative structure determination of assemblies in human cells, and also allows for identification and characterization of functionally relevant structural changes that take place in disease alleles. Expanding this analysis to host-pathogen complexes (84-86) will be feasible by introducing specific mutations into the pathogenic genome and studying the phenotypic consequences using genetic interaction profiling of relevant host genes (87). Furthermore, several efforts are underway to generate multiscale models of entire cells (88-92). In such instances, high-throughput genetic interaction mapping could provide global insights into cellular organization and dynamics of different components, while also informing on the structures of individual assemblies.

Methods

Histone mutant strain construction

The histone H3/H4 mutant strain library was constructed essentially as described (22, 23). Briefly, the mutants (tail deletions, complete alanine-scan, and context specific point mutations) were generated in the YMS196 background (MATα *his3*Δ *leu2*Δ *ura3*Δ *can1::STE2pr-spHIS5 lyp1::STE3pr-LEU2*) (Table S1). First, the base strains were created by replacing the *HHT2-HHF2* locus with a *URA3*-containing cassette carrying a mutated *HHT2-HHF2* locus with their endogenous promoters. We randomly picked a few base strains and the mutated *HHT2-HHF2* loci were PCR amplified and validated by sequencing. Then the *HHT1-HHF1* locus was replaced with a HYG^R-containing cassette carrying a mutated version of the *HHT1-HHF1* locus, resulting in pE-MAP-amenable strains (Matα *his3*Δ *leu2*Δ *ura3*Δ *hht1-hhf1*::HYG^R *hht2-hhf2::URA3 can1::STE2pr-spHIS5 lyp1::STE3pr-LEU2*).

Histone mutant library validation

Libraries were validated in three steps: 1) Each mutant was constructed, transformed into bacteria, and sequenced; 100% sequence identity was required to pass quality control. 2) After the correct integration of histone mutants in the *HHT2-HHF2* locus, 5-10 yeast base strains from each 96-well plate were randomly selected and corresponding histone fragments were amplified and sequenced to ensure the identity of each mutant in the well and no cross-contamination during plasmid preparation and yeast transformation; 100% of these were correct. 3) After obtaining the yeast library with the second (HYG^R) cassette integrated, 5–10 yeast strains from each 96-well plate were also randomly selected. Both copies of histone mutants in each strain were amplified and sequenced to confirm the identity of mutations. All of them were correct.

pE-MAP analysis

Each of the histone H3/H4 mutant strains was crossed with 1370 MATa KAN^R marked deletion (non-essential genes) or DAmP (Decreased Abundance by mRNA Perturbation; essential genes) strains by pinning on solid media as described (15). Sporulation was induced and MATa haploid spores were selected by replica plating onto media containing canavanine (selecting *can1*Δ haploids) and S-AEC (selecting *lyp1*Δ haploids) and lacking histidine (selecting MATa spores). Triple mutant haploids were isolated on media containing hygromycin (selecting *hht1-hhf1 mutant cassette*) and G418 (selecting KAN^R marked deletion/DAmP), and lacking uracil (selecting *hht2-hhf2 mutant cassette*). Finally, triple mutant colony sizes were extracted using imaging software. The screen was carried out in 3 biological replicates with 3 technical replicates in each. 4 mutants (H4E73Q, H4H18A, H4121A and H4K44Q) failed screening in one biological replicate and the results for these are based on the two successful replicates. Detailed E-MAP experimental procedures are described in (26, 29, 93). Genetic interactions were quantified using S-scores (28), which are closely related to t-values. The S-score quantifies the deviation of the double (or triple) mutants from the expected combined fitness effects of the individual mutants and incorporates the reproducibility between technical replicates. The published S-scores represent the average S-scores across biological replicates.

Design of the pE-MAP spatial restraints

The distance restraint based on pE-MAP data was designed using the 308 single point mutants from the histone pE-MAP and the structure from the PDB 1ID3, as follows: 1) post-processing of the genetic interaction phenotypic profiles, 2) devising a phenotypic similarity metric between the phenotypic profiles, and 3) designing spatial restraints for integrative structure modeling using the phenotypic similarity values and the known nucleosome X-ray structure. Next, we describe each of these three steps in turn.

1) Post-processing of the genetic interaction phenotypic profiles: All missing values in the pE-MAP were imputed as the mean of the S-scores between the corresponding deletion mutant and all histone point mutants. To increase the signal-to-noise ratio of the pE-MAP, gene deletion mutants that mostly exhibited weak genetic interactions with the histone mutants were filtered out. To this end, the gene deletion profiles were ranked in descending order based on the counts of their S-scores that fell in either the top 2.5% of positive S-scores or the bottom 5% of negative S-scores, from the complete point mutant pE-MAP (cutoffs calculated after imputation). The more stringent cutoff for positive S-scores was chosen to reflect the smaller dynamic range for positive genetic interactions compared to negative genetic interactions. Gene deletions with the same count were then ranked in descending order by the mean of the absolute values of their highest and lowest score (Fig. S3A). The top fraction of the deletions, determined in step 3 below, were retained for computing the histone point mutant phenotypic profile similarities (below, Fig. S3).

2) Devising a phenotypic similarity metric between the phenotypic profiles: We computed the similarity between all pairs of histone phenotypic profiles using the maximal information coefficient (MIC, Fig. S3B), with the MIC parameters *alpha* and *c* set to 0.6 and 15, respectively, as suggested (36, 37). Many positions in the histones were mutated to several different residue types, giving rise to several phenotypic profiles for each of these positions. As a result, more than one MIC value would often be computed for a single residue pair. In such cases, only the highest MIC value was retained.

3) Designing spatial restraints for integrative structure modeling: Using the histone X-ray structure (PDB: 1ID3; (38)), we measured the Cα-Cα distance between all pairs of residues for which we computed a MIC phenotypic similarity score. The percentage of the top scoring phenotypic profiles (ranked by the genetic interaction scores; step 1) retained for further analysis was determined as follows. We compared the statistical association of the distances between two mutated residues with their phenotypic similarity by selecting the top 10%, 25%, 50%, and 100% of the ranked deletions (Fig. S3C). Although MIC values between phenotypic profiles do not linearly correlate with the distances spanned by the mutated residues in the wt structure (Pearson correlation coefficient of −0.07 when using the top 25% or top 50% of deletions), the MIC values provide an upper distance bound between the residues. The upper distance bound was obtained by binning the MIC values into 20 intervals and selecting the maximum distance spanned by any pair of residues in each bin, followed by fitting a logarithmic decay function (d_U) to the upper distance bounds:

$$d_U(MIC) = \begin{cases} \dfrac{log(MIC) - n}{k} & \text{if MIC} \leq 0.6 \\ 6.84 & \text{if MIC} > 0.6 \end{cases} \quad (1)$$

where k and n are −0.0147 and −0.41, respectively (Fig. S3C). We find that selecting the top 25% or 50% of the deletions had a comparable association between the upper distance bounds and the computed MIC values. The association was determined by computing the R and p-values of the Pearson correlation coefficient and association significance, respectively, for the log-transformed MIC values. In this work, we retained the top 25% of the ranked phenotypic profiles for computing the phenotypic profile similarities.

To effectively handle the uncertain relationship between the data and model, we use Bayesian inference for scoring alternative models by formulating spatial restraints as Bayesian data likelihoods (94). Formally, the posterior probability of model M given data D

and prior information I is $p(M|D, I) \propto p(D|M, I) \cdot p(M|I)$. The model, M, consists of a structure X and unknown parameters Y, such as noise in the data. The prior $p(M|I)$ is the probability density of model M given I. The prior can in general reflect information such as statistical potentials or a molecular mechanics force field; here, we only used excluded volume and sequence connectivity. The likelihood function $p(D|M, I)$ is the probability density of observing data D given M and I. The pE-MAP data was used to compute phenotypic similarities (*i.e.* MIC values) that inform distances between mutated residues pairs i, j. The likelihood of the entire pE-MAP dataset is the product over the individual observations between residue pairs i, j: $p(D|M, I) = \Pi_{i,j} N[d_{i,j}|f_{i,j}(X), \sigma_{i,j}]$, where $f_{i,j}(X)$ is a forward model that predicts the data point $d_{i,j}$ in D that would have been observed for structure X in an experiment without noise; $N[d_{i,j}|f_{i,j}(X), \sigma_{i,j}]$ is a noise model that quantifies the deviation between the predicted and observed data points.

We defined the forward model by inverting the relation between the upper distance bound and observed MIC values ($d_U(MIC)$, Eq. 1):

$$\boldsymbol{f}_{i,j}(X) = MIC(d_{i,j}) = \begin{cases} \exp(k \cdot d_{i,j} + n) & \text{if } d_{i,j} \le d_0 \\ 0.6 & \text{if } d_{i,j} > d_0 \end{cases} \tag{2}$$

where $d_0 = d_U\,(0.6)$. Our choice of a noise model is a lognormal distribution with a flat plateau for MIC values below the upper bound on the experimentally observed MIC values ($\mathrm{MIC}^{\mathrm{obs}}$):

$$P(\mathrm{MIC}_{i,j}^{obs} \mid \mathrm{MIC}_{i,j}, X, \sigma_{i,j}) = \\ \begin{cases} \dfrac{1}{N} & \text{if } \mathrm{MIC}_{i,j}^{obs} \ge \mathrm{MIC}_{i,j} \\ \dfrac{1}{M}\dfrac{1}{\sqrt{2\pi\sigma_{i,j}^2}\mathrm{MIC}_{i,j}^{obs}} exp\left[-\dfrac{1}{2\sigma_{i,j}^2} log^2\left(\dfrac{\mathrm{MIC}_{i,j}^{obs}}{\mathrm{MIC}_{i,j}}\right)\right] & \text{if } \mathrm{MIC}_{i,j}^{obs} < \mathrm{MIC}_{i,j} \end{cases} \tag{3}$$

Here, $\sigma_{i,j}$ are the noise parameters that can optionally be determined as part of the model, and N and M are normalization factors necessary to make the likelihood continuous. Lognormal noise models have previously been used to describe errors of inherently positive quantities (95). For computational efficiency, we used a single σ value for all residue pairs. An uninformative Jeffrey's prior is applied to σ to represent a lack of information on the bounds and distribution of this parameter (96).

Finally, a Bayesian term in the scoring function is defined as the negative logarithm of the posterior probability density: $S(M) = -log\, p(M|D, I)$. In the Bayesian view, the output model is in fact best equated to the posterior model density that specifies a distribution of alternative single models M with varying probability density, not a single model, although single representative or average models can always be proposed based on the posterior model density.

Calculation of similarity metrics for yeast RNAPII and bacterial RNAP datasets

Steps 1) and 2) from “Design of the pE-MAP spatial restraints*”* were repeated for the yeast RNAPII and bacterial RNAP datasets to generate the similarity metrics (MIC values) for these two systems, with the following modifications:

For yeast RNAPII, prior to imputing missing values in the pE-MAP, any deletion mutants that exhibit missing values with more than 15% of the point mutants were filtered out. This step is part of our pipeline but had no effect on the histone pE-MAP (because this pE-MAP does not contain any deletion mutant with more than 15% values missing). The number of ranked deletion mutants retained at the end of pE-MAP post-processing was chosen to be 25% of the number of deletions in the original unfiltered pE-MAP (in accordance with the histone pE-MAP processing).

For bacterial RNAP, due to the very small number of perturbations in this dataset, all the perturbations (instead of the top 25%) were retained for computing point mutant phenotypic profile similarities. In addition, due to differences in the experimental design for generating the yeast pE-MAPs and the bacterial RNAP CG-MAP, the bacterial RNAP MIC distribution had a ~2-fold higher median and greater spread than the other datasets. Correspondingly, the bacterial RNAP MIC distribution was normalized using linear scaling, decreasing its median to match that of the histone MIC distribution. Importantly, this step was based solely on the MIC distributions, without reliance on any structural information.

Integrative structure determination

Integrative structure determination for each system proceeded through the standard four stages (3-5, 8, 41, 97) (Fig. 5, Table S3, Table S8, Table S9): 1) gathering data, 2) representing subunits and translating data into spatial restraints, 3) configurational sampling to produce an ensemble of structures that satisfies the restraints, and 4) analyzing and validating the ensemble structures and data. The integrative structure modeling protocol (*i.e.* stages 2, 3, and 4) was scripted using the *Python Modeling Interface* (PMI) package, a library for modeling macromolecular complexes based on our open-source *Integrative Modeling Platform* (IMP) package (5), version 2.8 (https://integrativemodeling.org). Files containing the input data, scripts, and output results are available at http://integrativemodeling.org/systems/pemap and the nascent integrative methods benchmarking section of the worldwide Protein Data Bank (wwPDB) PDB-Dev repository for integrative structures and corresponding data (pdb-dev.wwpdb.org) (98).

1) Gathering data—To mimic realistic integrative structure determination, we did not rely on the known atomic structures of the subunits in the actual modeled complex (correct docking of exact bound structures based on geometric complementarity is easy, (99)). Instead, we computed comparative models of histones H3 and H4 based on their alignments with structures of the 1TZY (100) (89% and 92% sequence identity, respectively), using MODELLER, version 9.21 (101). The Cα-atom RMSDs between the crystal structures and comparative models is 2.8 and 5.5 Å for H3 and H4, respectively, corresponding to medium and low accuracy comparative models. The second major input information source was a pE-MAP dataset of 308 point mutations in histones H3 and H4 crossed against an array of

~1,370 gene deletion alleles, resulting in 946 MIC values above 0.3. Of these, 170 MIC values were converted into distance restraints between H3 and H4 residues (Table S3, Fig. S8).

Comparative models of subunits Rpb1 and Rpb2 of yeast RNAPII were computed based on template structures 6GMH (102) (54% sequence identity) and 4AYB (103) (43% sequence identity), respectively. The Cα RMSD between the crystal structures of subunit Rpb1 and Rpb2 (2E2H (104)) and their comparative models are 7.3 and 5.2 Å, respectively. A pE-MAP dataset of 53 single point mutants in yeast RNAPII (44 of which reside in subunits Rpb1 and Rpb2) and a library of ~1,200 gene-deletions resulted in 195 MIC values above 0.3. Of these, 123 MIC values were converted into distance restraints (Table S8, Fig. S8). In addition, we compared the RNAPII model based on the pE-MAP to a model based on 22 previously published chemical cross-links (XLs) (61).

The structures of subunits RpoB and RpoC of bacterial RNAP were obtained from the X-ray structure of the entire complex (4YG2) (105). A CG-MAP of 44 single point mutants of the two subunits and a library of 83 conditions (*e.g.* treatments with chemicals and temperature shocks) resulted in 109 MIC values above 0.3. Of these, 63 MIC values were converted into distance restraints between the subunits (Table S9, Fig. S8). In addition, we compared the bacterial RNAP model based on the CG-MAP to a model computed based on distance restraints derived from the interfacial contacts predicted using the RaptorX protein complex contact prediction server (65, 66)

2) Representing subunits and translating data into spatial restraints—To maximize computational efficiency while avoiding using too coarse a representation, we represented each complex in a multi-scale fashion. In particular, the subunits/domains of each complex were coarse-grained using beads of varying sizes representing either a rigid body or a flexible string, based on the available comparative models, as follows (Table S3, Table S8, Table S9). The comparative models were coarse-grained into two representations at different resolutions. First, we identified loop regions of at least 8 residues using DSSP (106, 107) and represented them by flexible strings of beads of up to 10 residues each. Second, for the remaining residues each bead corresponded to an individual residue, centered at the position of its Cα atom. With this representation in hand, we next translated the input information into spatial restraints as follows.

The defining and most important restraint for our method is extracted from the pE-MAP/CG-MAP. The collected pE-MAP and CG-MAP MIC values were used to construct the Bayesian term in the scoring function that restrained the distances spanned by the mutated residues as described above. The pE-MAP restraint was applied to the one residue-per-bead representation for the comparative models as well as to the flexible beads. To improve computational efficiency, we only considered point mutation pairs with MIC values greater than 0.3. This restraint was applied to all three complexes (Table S3, Table S8, Table S9). In addition to the pE-MAP data, integrative modeling can benefit from many other types of input information. Here, we have supplemented the pE-MAP/CG-MAP data by additional simple terms accounting for excluded volume and sequence connectivity. First, the excluded volume restraints were applied to each bead in the one-residue (or the closest) bead representations, using the statistical relationship between the volume and the number of residues that it covered (4, 108). Second, we applied the sequence connectivity restraint, using a harmonic upper bound on the distance between consecutive beads in a subunit, with a threshold distance equal to four times the sum of the radii of the two connected beads. The bead radius was calculated from the excluded volume of the corresponding bead, assuming standard protein density (4, 108). Moreover, we evaluated the utility of pE-MAP/CG-MAP data by considering two additional types of restraints. First, the 22 previously determined BS3 RNAPII cross-links (61) were used to construct a Bayesian term that restrained the distances spanned by the cross-linked residues (30 Å) (109, 110). The cross-link restraints were applied to the one residue-per-bead representation for the comparative models as well as flexible beads, only for RNAPII (Table S8). Second, we applied the evolutionary coupling restraints to determine the structures of the RpoB and RpoC subunits of bacterial RNA polymerase. Coupling strengths between residue pairs were obtained using the RaptorX ComplexContact server (http://raptorx.uchicago.edu/ComplexContact/) (65, 66) with default parameters. The top L/50 coupling strengths (Fig. S9) with sequence separation of 3 or greater were converted into distance restraints using a harmonic upper bound on the distances between the residues. The threshold distance was set to 12 Å. This restraint was applied only to a subset of bacterial RNAP modeling instances (Table S9).

Configurational sampling to produce an ensemble of structures that satisfy the restraints—The initial positions and orientations of rigid bodies and flexible beads were randomized. The generation of structural models was performed using Replica Exchange Gibbs sampling, based on the Metropolis Monte Carlo (MC) algorithm (110, 111). Each MC step consisted of a series of random transformations (*i.e.* rotation and translation) of the positions of the flexible beads and rigid bodies. Details about the MC runs for each system are in Tables S3, S8, S9.

Analyzing and validating the ensemble structures and data—Model validation follows four major steps (3, 112): (i) selection of the models for validation; (ii) estimation of sampling precision; (iii) estimation of model precision; (iv) quantification of the degree to which a model satisfies the information used to compute it. These validations are based on the nascent wwPDB effort on archival, validation, and dissemination of integrative structures (98, 113). We now discuss each one of these validations in turn.

(i) Selection of models for validation: The first step is to objectively define the ensemble of models that will be further analyzed. For each trajectory, we automatically determined the MC step at which all data likelihoods and priors have equilibrated (run equilibration step), and all prior frames are discarded (114). Discarding the initial, non-equilibrated steps of each run is helpful because non-typical early configurations (*e.g.* a random configuration of beads, an extended configuration of beads, and beads far apart from each other) are removed from the statistical sample used for posterior model estimates.

With this ensemble of sampled structures and their corresponding scores in hand, we analyze the data likelihoods and priors. We used HDBSCAN clustering, a hierarchical density-based clustering algorithm, to identify all high-density regions in the likelihoods and priors (115). If a single cluster was identified, we consider all the models after discarding the initial steps;

otherwise, we consider all models in the clusters that satisfy the input information, within the uncertainty of the data, for further analysis (below).

(ii) Estimation of sampling precision: Next, we estimate the precision at which sampling sampled the selected structures (sampling precision) (112); the sampling precision must be at least as high as the precision of the structure ensemble consistent with the input data (model precision). As a proxy for testing the thoroughness of sampling, we performed four sampling convergence tests: 1) verify that the scores of refined structures do not continue to improve as more structures are computed, 2) confirm that the selected structures in independent sets of sampling runs (Sample A and Sample B) satisfy the data equally well, 3) cluster the structural models and determine the sampling precision at which the structural features can be interpreted (Fig. S10), and 4) compare the localization probability density maps for each protein obtained from independent sets of runs. Details about all the tests are described in (112). For each modeling instance, the results from the convergence tests are summarized in Tables S3, S8, S9.

(iii) Estimation of model precision: In the third step, the model uncertainty (precision) is estimated. The most explicit description of model uncertainty is provided by the set of all models that are sufficiently consistent with the input information (*i.e.* the ensemble). Model precision can be quantified by the variability among the models in the ensemble; in the end, the ensemble can be described by one or more representative models and their uncertainties. For example, if the structures in the ensemble are clustered into a single cluster, the model precision is defined as the RMSD between models in the cluster. Importantly, the uncertainty may not be distributed evenly across the ensemble, such that some regions are determined at a higher precision than others.

(iv) Quantification of the degree to which a model satisfies the data used to compute it: An accurate structure needs to satisfy the input information used to compute it; all structures at computed precision that are consistent with the data are provided in the ensemble. A pE-MAP derived restraint is satisfied by a cluster of structures if the corresponding Cα–Cα distance in any of the structures in the cluster is lower than the distance predicted by the MIC value (Eq. 1). A BS3 cross-link restraint is satisfied by a cluster of structures if the corresponding Cα–Cα distance in any of the structures in the cluster is less than 30 Å (116). The remainder of the restraints are harmonic, with a specified standard deviation. Therefore, a restraint is satisfied by a cluster of structures if the restrained distance in any structure in the cluster is violated by less than 3 standard deviations, specified for the restraint. Tables S3, S8, S9 show that all models satisfy the input information within its uncertainty.

Benchmark—To benchmark the four-stage protocol described above, we computed the distribution of the accuracy for each structure in the ensemble of solutions obtained by integrative modeling. The accuracy is defined as the mean of Cα RMSD between the X-ray structure and each of the structures in the ensemble. The PDB accession code and accuracies for each modeling instance are summarized in Tables S3, S8, S9.

To assess the information content of the histone pE-MAP, we computed the models of the H3-H4 complex based on random subsets of the data. To this end, from the dataset of computed MIC values for pairs of mutated residues, we performed three independent random selections of 80%, 60%, 40%, and 20% of the data each. As expected, the more pE-MAP data used, the more accurate and precise are the models (Fig. 6C).

Finally, as another test, we computed the model based on datasets with randomly shuffled MIC values for the same pE-MAP/CG-MAP residue pairs, for each of the complexes.

Estimation of the number of mutations per protein—To estimate the suggested number of mutated positions per protein for integrative structure determination, we computed the number of mutations that would result in 4 or more MIC values above a 0.4, 0.45, 0.5, or 0.55 threshold. Based on our scoring function, MIC values above these thresholds will result in distance restraints with an upper distance bound in the 12-34 Å range. These distances are comparable to the upper distance bounds used for chemical cross-links (*e.g.* DSS, DSSO, EDC). A previous systematic study established that at least 4 chemical cross-links are needed to determine the binding mode of protein dimers if the subunit structures are known (*e.g.* from X-ray, NMR, or comparative models) (110). In general, adding more chemical cross-links does not further improve the accuracy, although it increases the precision of the resulting ensemble. By analogy, we estimate that, for systems in which the structures of the components are known, a good number of mutations per protein is 35-40 (Fig. S5A). This data can be used as a guideline to decide on the number of mutations to use for generating a pE-MAP or CG-MAP. Importantly, this estimate is an upper bound on the number of mutations, and in many cases, the number might be smaller for the following two reasons. First, this estimation was done assuming protein-wide mutations of residues, often to alanine. In practice, the number of necessary mutations can be reduced by specifically designing point mutations that target surface residues and/or residues known to be functionally important, and by choosing substitutions likely to give rise to functional perturbations. In general, we did not find a correlation between the secondary structure of the residue pairs and their associated MIC value (Fig. S5B). Second, this estimation only relies on the residue pairs with high MIC values. In contrast to chemical cross-links, the upper distance bounds of pE-MAP derived restraints are obtained from the statistical association between the MIC values and distances between residues. Consequently, residue pairs with low MIC values still carry structural information, even if at low resolution. Consistent with these considerations, the RNAPII dataset contains only 31 and 9 mutated residues for Rpb1 and Rpb2, respectively. Similarly, the bacterial RNAP dataset contains 23 and 15 mutated residues for Rpob and Rpoc, respectively.

Docking

To assess the relative value of pE-MAP restraints for structure determination, we computed the structures of the H3-H4 and RNAPII complexes by molecular docking. Specifically, we followed an integrative docking protocol (117) using the rigid-body docking program PatchDock (39). In each case, we used the same comparative models and rigid body definitions used for integrative modeling and default parameter values (Fig. S4, Fig. S11).

Visualizations

The pE-MAP was hierarchically clustered in both histone mutant and gene deletion dimensions using Cluster 3.0 (118) and displayed using Java Treeview (119) (Fig. 3). Images highlighting histone residues in context of the nucleosome structure (PDB: 1ID3 or its modified version in Data S2) were created using ChimeraX (Figs. 3A, 7C, S1D, S6) (48).

Distance of clustered pE-MAP profiles

First, all histone alleles affecting residues not included in the structural reference (PDB: 1ID3, H3A1-H3K37 and H4S1-H4R17) were removed and the remaining data (n = 222) clustered hierarchically using Cluster 3.0 (118). For each node of the clustergram, the mean distance among member residues was calculated and plotted vs the normalized branch length (where the first node is set to branch length = 0 and the last node to branch length = 1) of the respective node (Fig. S2A, red dots, and random distribution plotted in black).

Generation of the correlation map

Pearson correlation coefficients were computed for each of the 350 H3/H4 mutants against all genes/alleles (rows) in a merged map of previously published genetic interaction data (Dataset S4 from (45)). If the overlap between a histone mutant and a S-score vector from the merged map was < 150 scores, the resulting correlation was not considered (*i.e.* replaced by "NaN") (Data S3). Pearson correlation coefficients were chosen over MIC for this analysis because we found Pearson correlation more robust than MIC when many missing values were present.

Structural mapping of genetic interactions – stEMAP app

The hierarchically clustered pE-MAP data was imported into Cytoscape (47), creating an initial network, and then linked to a modified version of the nucleosome structure 1ID3 (Data S2) using the stEMAP app, developed to facilitate interactive exploration of the pE-MAP. The original nucleosome structure was modified by adding the N-terminal disordered regions of histone H3 and H4 and manually positioning them for clarity. The linking proceeds as follows: First, the structure is opened in ChimeraX (48) by structureVizX (120) and positioned in response to commands from the stEMAP app. Then, a residue interaction network (RIN) is created by the structureVizX app where nodes are positioned to reflect the nucleosome structure through the help of the RINalyzer app (121). Finally, the RIN network and the network created by the original cluster files are merged, and edges are drawn between genes and residues with interactions passing a user defined threshold (Fig. S6A). All of the preceding steps happen automatically through the stEMAP app interface, which takes as input any given PDB file and a short user-defined JSON configuration file defining interaction thresholds (here: correlation > 0.2), colors of edges (here: color-gradient from white to red for positive correlations), and display style of the structure in ChimeraX.

Selection of individual genes triggers the interacting residues to be selected and, in the ChimeraX window, those residues are shown as space-filling atoms, which are colored according to the edge colors. When multiple genes are selected (e.g. genes belonging to the same complex), there might be multiple edges connecting an individual residue. In this case, the color reflects the significance and consistency of the interactions (see below). To assist in interpretation and interactive exploration of complex data sets (i) colors are quantized into 10 bins, 5 positive and 5 negative, (ii) a heatmap is presented that shows only the values for the selected genes and their interacting residues, (iii) sets of genes belonging to a complex can be selected using the setsApp (122) and (iv) a slider provides a filter to restrict the selection to only those mutations with a minimum number of interactions.

To determine if a gene set is connected to a given residue, the stEMAP app calculates the median Pearson correlation coefficient across all genes of the gene set against all different mutations at that residue. If this median correlation is above the threshold of 0.2 (defined in the JSON file), the respective residue is colored according to the median. To instead determine if a gene set is connected to an individual mutation, the same method is used, except the median correlation coefficient is now calculated across all genes of the gene set against the single given mutation (instead of all mutations of the residue).

ROC curves

Only library deletion mutants that exist in both this study and the two previously published E-MAP datasets (Braberg et al. (15) and Collins et al. (29)), were included (n = 389) in this analysis. Based on their pE-MAP profiles, Pearson correlation coefficients were calculated for all pairwise combinations of these 389 deletion mutants. In order to determine the power of these correlations to predict physical interactions between encoded proteins an ROC curve was computed, where a physical interaction between proteins was defined if their PE score is greater than 2 (69). From the Collins et al. E-MAP, query strain profiles with more missing data than the sparsest histone mutant were removed, as were query mutants that also existed in the library mutant set. Since the Braberg et al. pE-MAP only includes 53 query mutants (rows), we used subsets of 53 query mutants each for the histone and Collins et al. E-MAPs when generating their ROC curves, to make all three systems comparable. To this end, for the Collins et al. E-MAP and histone pE-MAP, 53 query mutant profiles were randomly selected 1,000 times, and a ROC curve was generated for each run. The median areas under the ROC curves (AROCs) and corresponding ROC curves are reported together with the ROC curve of the pE-MAP from Braberg et al. in Fig. 2F.

RNA-seq expression analysis

10ml of overnight cultures of 29 histone mutant strains (Table S2) were harvested in mid-log phase ($OD_{600} \approx 1.0$) and washed with DEPC-ddH_2O. RNA was extracted with hot acidic phenol as described previously (123). RNA-seq libraries were generated using the QuantSeq 3’ mRNA-Seq Library Prep Kit FWD for Illumina (Lexogen). Single-end, 50 base reads were sequenced using an Illumina HiSeq 4000 sequencer. Reads were filtered for quality and aligned to the yeast genome using tophat (124). Non-unique reads and reads mapping to ribosomal RNA were removed prior to analysis. Transcript counts were extracted using htseq-count (125) and differential expression was measured using the Dseq2 package in R (126).

Identification of functional links between H3/H4 mutants and biological processes

The correlation map (Data S3) was used as the basis for this analysis. First, a curated annotation of all genes in the correlation map relevant to nuclear function was devised.

Biological process definitions for genes in nuclear processes were assigned manually based on literature and annotations from previous genetic interaction maps (29, 127, 128) (Table S5). To identify links between H3/H4 residues and nuclear processes that were highly correlated, we used a one-sided Mann-Whitney U test to compare the correlation distribution between the mutants of each H3/H4 residue and the members of each process to (i) the correlations between the same H3/H4 mutants and all genes not in that process, and to (ii) the correlations between the same process and all other H3/H4 mutants. The highest p-value of the comparison to (i) or (ii) was recorded. False discovery rates (FDR) for the links were then computed using the method of Benjamini and Hochberg (129), and are reported in Table S5. The most significant links with FDR $<10^{-6}$ were used for follow-ups.

Spontaneous Mutation Frequency

Cells were grown to saturation and then plated on YEPD and 5-FOA supplemented media. Mutants growing on 5-FOA were counted only after confirming that colonies growing on YEPD for all the strains were of equal size. The assay was repeated three times independently (Table S6).

MS quantification of H3K56ac levels

Sample preparation—Histone mutant cultures (wt, H3R63K and H4R36K) were harvested in mid-log phase ($OD_{600} \approx 1.0$) using a 250 mm ceramic filter funnel and 30 μm nitrocellulose membranes connected to high continuous wall suction. Yeast were removed from the nitrocellulose membrane and flash frozen for storage or used immediately for protein extraction. Per gram of yeast pellet, 3 ml of Yeast-Protein Extract Reagent (Y-PER; ThermoFisher Scientific) with added protease inhibitors (cOmplete Sigma-Aldrich, 1 tablet/50 ml), phosphatase inhibitors (PhosSTOP™ Sigma-Aldrich; 1 tablet/50 ml), histone deacetylase inhibitors (sodium butyrate 100 mM and nicotinamide 100 mM), and beta-mercaptoethanol (15 mM) were added. The suspension was mixed on a gyrator at 4°C for 30 minutes and centrifuged. Pellets were resuspended in fresh Y-PER medium, and extraction was repeated two additional times for a total of three extractions. Pellets were sequentially washed twice with 3ml ddH_2O per gram of yeast. Histone extraction was performed in the presence of 2.5 ml of 8M urea/0.4N sulfuric acid per gram of yeast protein pellets, incubated for 1 hour, centrifuged, and supernatants collected. Proteins were precipitated using a methanol-chloroform precipitation as previously described (130). Extracted proteins were trypsin digested; desalted and acetylated peptides were enriched as previously described (131).

Generation of selected reaction monitoring (SRM) assays for acetylation sites —Peptide mixtures (obtained from ThermoFisher) were analyzed by LC-MS/MS on a Thermo Scientific Orbitrap Fusion mass spectrometry system equipped with a Proxeon Easy nLC 1200 ultra high-pressure liquid chromatography and autosampler system. Samples were injected onto a C18 column (25 cm × 75 μm I.D. packed with ReproSil Pur C18 AQ 1.9 μm particles) in 0.1% formic acid and then separated with a 60 min gradient from 5% to 40% Buffer B (90% ACN/10% water/0.1% formic acid) at a flow rate of 300 nl/min. The mass spectrometer collected data in a data-dependent fashion, collecting one full scan in the Orbitrap followed by collision-induced dissociation MS/MS scans in the dual linear ion trap for the 20 most intense peaks from the full scan. Dynamic exclusion was enabled for 30 seconds with a repeat count of 1. Charge state screening was employed to reject analysis of singly charged species or species for which a charge could not be assigned. The raw data was matched to protein sequences using the MaxQuant algorithm (version 1.5.2.8) (132). Data were searched against a database containing SwissProt Human protein sequences concatenated to a decoy database where each protein sequence was randomized in order to estimate the FDR. Variable modifications were allowed for methionine oxidation and protein N-terminus acetylation and lysine acetylation. A fixed modification was indicated for cysteine carbamidomethylation. Full trypsin specificity was required. The first search was performed with a mass accuracy of +/− 20 parts per million and the main search was performed with a mass accuracy of +/− 4.5 parts per million. A maximum of 5 modifications were allowed per peptide. A maximum of 2 missed cleavages were allowed. The maximum charge allowed was 7+. Individual peptide mass tolerances were allowed. For MS/MS matching, a mass tolerance of 0.8 Da was allowed and the top 8 peaks per 100 Da were analyzed. MS/MS matching was allowed for higher charge states and water and ammonia loss events. The data were filtered to obtain a peptide, protein, and site-level false discovery rate of 0.01. The minimum peptide length was 7 amino acids. Selected reaction monitoring (SRM) assays were generated for selected acetylation sites. SRM assay generation was performed using Skyline (133). For all targeted proteins, proteotypic peptides and optimal transitions for identification and quantification were selected based on a spectral library generated from the shotgun MS experiments. The Skyline spectral library was used to extract optimal coordinates for the SRM assays, e.g. peptide fragments and peptide retention times. For each peptide the 5 best SRM transitions were selected based on intensity and peak shape.

Acquisition and quantification of acetylation sites by SRM—Digested peptide mixtures were analyzed by LC-SRM on a Thermo Scientific TSQ Quantiva MS system equipped with a Proxeon Easy nLC 1200 ultra high-pressure liquid chromatography and autosampler system. Samples were injected onto a C18 column (25 cm × 75 μm I.D. packed with ReproSil Pur C18 AQ 1.9 μm particles) in 0.1% formic acid and then separated with a 60 min gradient from 5% to 40% Buffer B (90% ACN/10% water/0.1% formic acid) at a flow rate of 300 nl/min. SRM acquisition was performed operating Q1 and Q3 at 0.7 unit mass resolution. For each peptide the best 5 transitions were monitored in a scheduled fashion with a retention time window of 5 min and a cycle time fixed to 2 sec. Argon was used as the collision gas at a nominal pressure of 1.5 mTorr. Collision energies were calculated by, $CE = 0.0348 \times (m/z) + 0.4551$ and $CE = 0.0271 \times (m/z) + 1.5910$ (CE, collision energy and m/z, mass to charge ratio) for doubly and triply charged precursor ions, respectively. SRM data was processed using Skyline (133). Protein significance analysis was performed using MSstats (134). Normalization of the intensities across samples was performed using the acetylated peptides H3K9_H3K14 (peptide containing both acetylation sites), H3K23 and H3K14 as global standards, which did not show any change across the mutants. Log_2-fold changes were calculated from three independent runs and plotted (Fig. S7C, Table S6).

Cryptic transcription - qPCR

Total RNA was extracted from 10 OD_{600} units of mid-log phase cells (wt and respective mutant strains) using hot acid phenol-chloroform extraction method as described. 10 μg of total RNA was DNAse I treated (Promega) followed by purification using an RNeasy minikit (Qiagen). 1 μg of DNAse I treated total RNA was used to synthesize cDNA using SuperScript III first strand synthesis system (Life Technologies) and random hexamer primers. cDNA was diluted 1:25 prior to amplification by PCR using primers designed for the 5' and the 3' ends of the *STE11* gene. qPCR was performed using SYBR green (Biorad) as described previously (135). Relative change in the transcript levels were estimated using the $\Delta\Delta C_t$ method described in (136) and were normalized to *ACT1* transcript (Table S7). Primers sequences are available upon request.

Western Blotting

Whole yeast cell lysates were prepared using TCA lysis as described previously (137). Lysates were subjected to immunoblotting according to standard procedures and proteins were detected using ECL Prime (Amersham Biosciences). Membranes were probed with αH3K36me3-antibody purchased from (Abcam, catalog #9050). GAPDH was used for loading control and detected using an antibody purchased from Sigma (catalog #A9521).

ATAC-seq

Yeast cells (2.5×10^6) were grown to mid-log phase, pelleted, washed with SB-buffer (1.4 M Sorbitol, 40 mM HEPES-KOH pH 7.5, 0.5 mM MgCl2), resuspended in 200 μl SB buffer + 10 mM DTT with 10 μl of 10 mg/ml 100T zymolyase (MP Biomedicals) solution and incubated for 5 min at 30 °C. Spheroblasted cells were washed with SB-buffer and incubated for 15 min at 37 °C in 25 μl transposase solution (12.5 μl 2x TD buffer, 1.25 μl Nextera enzyme, 11.25 μl water). DNA was purified (Qiagen MinElute DNA Purification Kit), amplified and barcoded by PCR. Purified PCR-products were sequenced using an Illumina HiSeq 4000 sequencer. Sequence reads were trimmed, aligned to the genome of *S.cerevisiae* (version SacCer3 from hgdownload.cse.ucsc.edu/downloads.html), and reads with a length <100 bp removed. Replicates belonging to an allele (wt, H3K36A, H3K122A, *set2*Δ) were merged and normalized to the smallest read number. For visualization of *STE11* read coverage using the IGV genome browser (138) (Fig. S7G), each track was scaled linearly so that the largest peak in the displayed window is the same height for all tracks. Count files were generated with "featureCounts v1.5.3" (139).

Gene body plots (Fig. S7H) were generated as follows: First, counts from genes reported to be targets of cryptic transcription (n = 11; *FLO8*, *AVO1*, *LCB5*, *SMC3*, *SPB4*, *APM2*, *DDC1*, *SYF1*, *OMS1*, *PUS4*, *STE11*), as well as counts 400bp up- and downstream of the respective gene bodies, were extracted. Second, up- and downstream regions were split into 50 bins of equal size (8bp), whereas the gene body was split into 300 equal bins, resulting in 400 bins for each gene in each tested strain (wt, *set2*Δ, H3K36A and H3K122A). Third, for each of the 400 bins the average for the 11 target genes was calculated. Fourth, each mutant allele was then scaled linearly so that the first bin (i.e. 400bp upstream of the gene body start) was equal to that of wt. Finally, the wt counts were subtracted from the mutant counts for each bin.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

We thank M. Eckhardt, I. Chemmama, Alain Verreault, David Truong, Pedro Beltrao and Assen Roguev for helpful discussions.

Funding: We acknowledge funding through NIH grants R01 GM084448, R01 GM084279, P50 GM081879, P50 AI150476, P01 HL089707, U54 NS100717 and R01 GM098101 to N.J.K., R01 GM083960, P50 GM082250, U19 AI135990, S10 OD021596 and P41 GM109824 to A.Sa, U54 RR020839 and U54 GM103520 to J.D.B., R35 GM118061 to C.A.G., GM126900 to B.D.S., P01GM105473 to J.E.H., U54 RR020839 and R01 GM084448 to J.S.B., OD009180 and GM123159 to J.S.F., GM008284 to A.Sh, National Key Research and Development Program of China (2017YFA0505103), Shenzhen Science and Technology Program (KQTD20180413181837372), Guangdong Provincial Key Laboratory of Synthetic Genomics (2019B030301006) and Bureau of International Cooperation, Chinese Academy of Sciences (172644KYSB20180022) to J.D., and the National Natural Science Foundation of China (31800069) to S.J.

References and Notes:

1. Alber F, Forster F, Korkin D, Topf M, Sali A, Integrating diverse data for structure determination of macromolecular assemblies. Annu. Rev. Biochem 77, 443–477 (2008). [PubMed: 18318657]
2. Herzog F et al., Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. Science 337, 1348–1352 (2012). [PubMed: 22984071]
3. Rout MP, Sali A, Principles for Integrative Structural Biology Studies. Cell 177, 1384–1403 (2019). [PubMed: 31150619]
4. Alber F et al., Determining the architectures of macromolecular assemblies. Nature 450, 683–694 (2007). [PubMed: 18046405]
5. Russel D et al., Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol. 10, e1001244 (2012). [PubMed: 22272186]
6. Ward AB, Sali A, Wilson IA, Biochemistry. Integrative structural biology. Science 339, 913–915 (2013). [PubMed: 23430643]
7. Alber F et al., The molecular architecture of the nuclear pore complex. Nature 450, 695–701 (2007). [PubMed: 18046406]
8. Lasker K et al., Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proc Natl Acad Sci U S A 109, 1380–1387 (2012). [PubMed: 22307589]
9. Loquet A et al., Atomic model of the type III secretion system needle. Nature 486, 276–279 (2012). [PubMed: 22699623]
10. Duan Z et al., A three-dimensional model of the yeast genome. Nature 465, 363–367 (2010). [PubMed: 20436457]
11. Plitzko JM, Schuler B, Selenko P, Structural Biology outside the box—inside the cell. Curr. Opin. Struct. Biol 46, 110–121 (2017). [PubMed: 28735108]
12. Dimura M et al., Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems. Curr. Opin. Struct. Biol 40, 163–185 (2016). [PubMed: 27939973]
13. Collins SR, Roguev A, Krogan NJ, Quantitative genetic interaction mapping using the E-MAP approach. Methods Enzymol. 470, 205–231 (2010). [PubMed: 20946812]
14. Beltrao P, Cagney G, Krogan NJ, Quantitative genetic interactions reveal biological modularity. Cell 141, 739–745 (2010). [PubMed: 20510918]
15. Braberg H et al., From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. Cell 154, 775–788 (2013). [PubMed: 23932120]
16. Braberg H, Moehle EA, Shales M, Guthrie C, Krogan NJ, Genetic interaction analysis of point mutations enables interrogation of gene function at a residue-level resolution: exploring the

applications of high-resolution genetic interaction mapping of point mutations. Bioessays 36, 706–713 (2014). [PubMed: 24842270]

17. Halabi N, Rivoire O, Leibler S, Ranganathan R, Protein sectors: evolutionary units of three-dimensional structure. Cell 138, 774–786 (2009). [PubMed: 19703402]
18. Marks DS et al., Protein 3D structure computed from evolutionary sequence variation. PLoS One 6, e28766 (2011). [PubMed: 22163331]
19. Diss G, Lehner B, The genetic landscape of a physical interaction. Elife 7, e32472 (2018). [PubMed: 29638215]
20. Shiver AL et al., Chemical-genetic interrogation of RNA polymerase mutants reveals structure-function relationships and physiological tradeoffs. bioRxiv, 2020.2006.2016.155770 (2020).
21. Huang H, Lin S, Garcia BA, Zhao Y, Quantitative proteomic analysis of histone modifications. Chemical reviews 115, 2376–2418 (2015). [PubMed: 25688442]
22. Dai J et al., Probing nucleosome function: a highly versatile library of synthetic histone H3 and H4 mutants. Cell 134, 1066–1078 (2008). [PubMed: 18805098]
23. Jiang S et al., Construction of Comprehensive Dosage-Matching Core Histone Mutant Libraries for Saccharomyces cerevisiae. Genetics 207, 1263–1273 (2017). [PubMed: 29084817]
24. Brownell JE et al., Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell 84, 843–851 (1996). [PubMed: 8601308]
25. Durrin LK, Mann RK, Kayne PS, Grunstein M, Yeast histone H4 N-terminal sequence is required for promoter activation in vivo. Cell 65, 1023–1031 (1991). [PubMed: 2044150]
26. Braberg H et al., Quantitative analysis of triple-mutant genetic interactions. Nature protocols 9, 1867–1881 (2014). [PubMed: 25010907]
27. Haber JE et al., Systematic triple-mutant analysis uncovers functional connectivity between pathways involved in chromosome regulation. Cell Rep 3, 2168–2178 (2013). [PubMed: 23746449]
28. Collins SR, Schuldiner M, Krogan NJ, Weissman JS, A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome biology 7, R63 (2006). [PubMed: 16859555]
29. Collins SR et al., Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature 446, 806–810 (2007). [PubMed: 17314980]
30. Miller T et al., COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. Proc Natl Acad Sci U S A 98, 12902–12907 (2001). [PubMed: 11687631]
31. Roguev A et al., The Saccharomyces cerevisiae Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. EMBO J 20, 7137–7148 (2001). [PubMed: 11742990]
32. Mizuguchi G et al., ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. Science 303, 343–348 (2004). [PubMed: 14645854]
33. Carrozza MJ et al., Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell 123, 581–592 (2005). [PubMed: 16286007]
34. Venkatesh S et al., Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes. Nature 489, 452–455 (2012). [PubMed: 22914091]
35. Keogh MC et al., Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. Cell 123, 593–605 (2005). [PubMed: 16286008]
36. Reshef DN et al., Detecting novel associations in large data sets. Science 334, 1518–1524 (2011). [PubMed: 22174245]
37. Albanese D et al., minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics 29, 407–408 (2013). [PubMed: 23242262]
38. White CL, Suto RK, Luger K, Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. EMBO J. 20, 5207–5218 (2001). [PubMed: 11566884]
39. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ, PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 33, W363–367 (2005). [PubMed: 15980490]
40. Baker D, Sali A, Protein structure prediction and structural genomics. Science 294, 93–96 (2001). [PubMed: 11588250]

41. Kim SJ et al., Integrative structure and functional anatomy of a nuclear pore complex. Nature 555, 475–482 (2018). [PubMed: 29539637]

42. Gutierrez C et al., Structural dynamics of the human COP9 signalosome revealed by cross-linking mass spectrometry and integrative modeling. Proc Natl Acad Sci U S A 117, 4088–4098 (2020). [PubMed: 32034103]

43. Molnar KS et al., Cys-scanning disulfide crosslinking and bayesian modeling probe the transmembrane signaling mechanism of the histidine kinase, PhoQ. Structure 22, 1239–1251 (2014). [PubMed: 25087511]

44. Kwon Y et al., Structural basis of CD4 downregulation by HIV-1 Nef. Nat Struct Mol Biol 27, 822–828 (2020). [PubMed: 32719457]

45. Ryan CJ et al., Hierarchical modularity and the evolution of genetic interactomes across species. Mol Cell 46, 691–704 (2012). [PubMed: 22681890]

46. Wysocka J et al., A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature 442, 86–90 (2006). [PubMed: 16728976]

47. Shannon P et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13, 2498–2504 (2003). [PubMed: 14597658]

48. Goddard TD et al., UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci 27, 14–25 (2018). [PubMed: 28710774]

49. Schneidman-Duhovny D, Pellarin R, Sali A, Uncertainty in integrative structural modeling. Curr Opin Struct Biol 28, 96–104 (2014). [PubMed: 25173450]

50. Maas NL, Miller KM, DeFazio LG, Toczyski DP, Cell cycle and checkpoint regulation of histone H3 K56 acetylation by Hst3 and Hst4. Mol Cell 23, 109–119 (2006). [PubMed: 16818235]

51. Han J et al., A Cul4 E3 ubiquitin ligase regulates histone hand-off during nucleosome assembly. Cell 155, 817–829 (2013). [PubMed: 24209620]

52. Tsubota T et al., Histone H3-K56 acetylation is catalyzed by histone chaperone-dependent complexes. Mol Cell 25, 703–712 (2007). [PubMed: 17320445]

53. Hyland EM et al., Insights into the role of histone H3 and histone H4 core modifiable residues in Saccharomyces cerevisiae. Mol Cell Biol 25, 10060–10070 (2005). [PubMed: 16260619]

54. Masumoto H, Hawke D, Kobayashi R, Verreault A, A role for cell-cycle-regulated histone H3 lysine 56 acetylation in the DNA damage response. Nature 436, 294–298 (2005). [PubMed: 16015338]

55. Giannattasio M, Lazzaro F, Plevani P, Muzi-Falconi M, The DNA damage checkpoint response requires histone H2B ubiquitination by Rad6-Bre1 and H3 methylation by Dot1. J Biol Chem 280, 9879–9886 (2005). [PubMed: 15632126]

56. Celic I et al., The sirtuins hst3 and Hst4p preserve genome integrity by controlling histone h3 lysine 56 deacetylation. Curr Biol 16, 1280–1289 (2006). [PubMed: 16815704]

57. Celic I, Verreault A, Boeke JD, Histone H3 K56 hyperacetylation perturbs replisomes and causes DNA damage. Genetics 179, 1769–1784 (2008). [PubMed: 18579506]

58. Du HN, Briggs SD, A nucleosome surface formed by histone H4, H2A, and H3 residues is needed for proper histone H3 Lys36 methylation, histone acetylation, and repression of cryptic transcription. J Biol Chem 285, 11704–11713 (2010). [PubMed: 20139424]

59. Chen S et al., Structure-function studies of histone H3/H4 tetramer maintenance during transcription by chaperone Spt2. Genes Dev 29, 1326–1340 (2015). [PubMed: 26109053]

60. Tran HG, Steger DJ, Iyer VR, Johnson AD, The chromo domain protein chd1p from budding yeast is an ATP-dependent chromatin-modifying factor. EMBO J 19, 2323–2331 (2000). [PubMed: 10811623]

61. Chen ZA et al., Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. EMBO J. 29, 717–726 (2010). [PubMed: 20094031]

62. Ovchinnikov S, Kamisetty H, Baker D, Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife 3, e02030 (2014). [PubMed: 24842992]

63. Cong Q, Anishchenko I, Ovchinnikov S, Baker D, Protein interaction networks revealed by proteome coevolution. Science 365, 185–189 (2019). [PubMed: 31296772]

64. Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A, Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. Proc. Natl. Acad. Sci. U. S. A 113, 12186–12191 (2016). [PubMed: 27729520]

65. Wang S, Sun S, Li Z, Zhang R, Xu J, Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Comput Biol 13, e1005324 (2017). [PubMed: 28056090]

66. Zeng H et al., ComplexContact: a web server for inter-protein contact prediction using deep learning. Nucleic Acids Res. 46, W432–W437 (2018). [PubMed: 29790960]

67. Wang S, Sun S, Xu J, Analysis of deep learning methods for blind protein contact prediction in CASP12. Proteins 86 Suppl 1, 67–77 (2018). [PubMed: 28845538]

68. Roy KR et al., Multiplexed precision genome editing with trackable genomic barcodes in yeast. Nat Biotechnol 36, 512–520 (2018). [PubMed: 29734294]

69. Collins SR et al., Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics 6, 439–450 (2007). [PubMed: 17200106]

70. Schmiedel JM, Lehner B, Determining protein structures using deep mutagenesis. Nat. Genet 51, 1177–1186 (2019). [PubMed: 31209395]

71. Rollins NJ et al., Inferring protein 3D structure from deep mutation scans. Nat. Genet 51, 1170–1176 (2019). [PubMed: 31209393]

72. Newberry RW, Leong JT, Chow ED, Kampmann M, DeGrado WF, Deep mutational scanning reveals the structural basis for alpha-synuclein activity. Nat Chem Biol 16, 653–659 (2020). [PubMed: 32152544]

73. Eyal E, Najmanovich R, Edelman M, Sobolev V, Protein side-chain rearrangement in regions of point mutations. Proteins 50, 272–282 (2003). [PubMed: 12486721]

74. Sasidharan R, Chothia C, The selection of acceptable protein mutations. Proc. Natl. Acad. Sci. U. S. A 104, 10080–10085 (2007). [PubMed: 17540730]

75. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin A, Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins 86 Suppl 1, 51–66 (2018). [PubMed: 29071738]

76. Ovchinnikov S, Park H, Kim DE, DiMaio F, Baker D, Protein structure prediction using Rosetta in CASP12. Proteins 86 Suppl 1, 113–121 (2018). [PubMed: 28940798]

77. Robinson PJ et al., Molecular architecture of the yeast Mediator complex. Elife 4, e08719 (2015). [PubMed: 26402457]

78. Luo J et al., Architecture of the Human and Yeast General Transcription and DNA Repair Factor TFIIH. Mol Cell 59, 794–806 (2015). [PubMed: 26340423]

79. Jinek M et al., A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science 337, 816–821 (2012). [PubMed: 22745249]

80. Shen JP et al., Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. Nat. Methods 14, 573–576 (2017). [PubMed: 28319113]

81. Du D et al., Genetic interaction mapping in mammalian cells using CRISPR interference. Nat. Methods 14, 577–580 (2017). [PubMed: 28481362]

82. Ma L et al., CRISPR-Cas9–mediated saturated mutagenesis screen predicts clinical drug resistance with improved accuracy. Proc. Natl. Acad. Sci. U. S. A 114, 11751–11756 (2017). [PubMed: 29078326]

83. Anzalone AV et al., Search-and-replace genome editing without double-strand breaks or donor DNA. Nature 576, 149–157 (2019). [PubMed: 31634902]

84. Gordon DE et al., A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 583, 459–468 (2020). [PubMed: 32353859]

85. Eckhardt M, Hultquist JF, Kaake RM, Huttenhain R, Krogan NJ, A systems approach to infectious disease. Nature reviews. Genetics 21, 339–354 (2020).

86. Gordon DE et al., Comparative Host-Coronavirus Protein Interaction Networks Reveal Pan-Viral Disease Mechanisms. Science in press, (2020).

87. Gordon DE et al., A Quantitative Genetic Interaction Map of HIV Infection. Mol Cell 78, 197–209 e197 (2020). [PubMed: 32084337]

88. McGuffee SR, Elcock AH, Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. PLoS Comput. Biol 6, e1000694 (2010). [PubMed: 20221255]

89. Takamori S et al., Molecular anatomy of a trafficking organelle. Cell 127, 831–846 (2006). [PubMed: 17110340]

90. Wilhelm BG et al., Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins. Science 344, 1023–1028 (2014). [PubMed: 24876496]

91. Singla J et al., Opportunities and Challenges in Building a Spatiotemporal Multi-scale Model of the Human Pancreatic β Cell. Cell 173, 11–19 (2018). [PubMed: 29570991]

92. Thul PJ et al., A subcellular map of the human proteome. Science 356, (2017).

93. Schuldiner M, Collins SR, Weissman JS, Krogan NJ, Quantitative genetic analysis in Saccharomyces cerevisiae using epistatic miniarray profiles (E-MAPs) and its application to chromatin functions. Methods 40, 344–352 (2006). [PubMed: 17101447]

94. Rieping W, Habeck M, Nilges M, Inferential structure determination. Science 309, 303–306 (2005). [PubMed: 16002620]

95. Rieping W, Habeck M, Nilges M, Modeling errors in NOE data with a log-normal distribution improves the quality of NMR structures. J. Am. Chem. Soc 127, 16026–16027 (2005). [PubMed: 16287280]

96. Jeffreys H, An invariant form for the prior probability in estimation problems. Proc. R. Soc. Lond. A Math. Phys. Sci 186, 453–461 (1946). [PubMed: 20998741]

97. Sali A et al., Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. Structure 23, 1156–1167 (2015). [PubMed: 26095030]

98. Burley SK et al., PDB-Dev: a Prototype System for Depositing Integrative/Hybrid Structural Models. Structure 25, 1317–1318 (2017). [PubMed: 28877501]

99. Chen R, Mintseris J, Janin J, Weng Z, A protein--protein docking benchmark. Proteins: Struct. Funct. Bioinf 52, 88–91 (2003).

100. Wood CM et al., High-resolution structure of the native histone octamer. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun 61, 541–545 (2005).

101. Sali A, Blundell TL, Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol 234, 779–815 (1993). [PubMed: 8254673]

102. Vos SM et al., Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. Nature 560, 607–612 (2018). [PubMed: 30135578]

103. Wojtas MN, Mogni M, Millet O, Bell SD, Abrescia NGA, Structural and functional analyses of the interaction of archaeal RNA polymerase with DNA. Nucleic Acids Res. 40, 9941–9952 (2012). [PubMed: 22848102]

104. Wang D, Bushnell DA, Westover KD, Kaplan CD, Kornberg RD, Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. Cell 127, 941–954 (2006). [PubMed: 17129781]

105. Murakami KS, X-ray crystal structure of Escherichia coli RNA polymerase σ70 holoenzyme. J. Biol. Chem 288, 9126–9134 (2013). [PubMed: 23389035]

106. Joosten RP et al., A series of PDB related databases for everyday needs. Nucleic Acids Res. 39, D411–419 (2011). [PubMed: 21071423]

107. Kabsch W, Sander C, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983). [PubMed: 6667333]

108. Shen M-Y, Sali A, Statistical potential for assessment and prediction of protein structures. Protein Sci. 15, 2507–2524 (2006). [PubMed: 17075131]

109. Erzberger JP et al., Molecular architecture of the 40S⋅ eIF1⋅ eIF3 translation initiation complex. Cell 158, 1123–1135 (2014). [PubMed: 25171412]

110. Shi Y et al., Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. Mol. Cell. Proteomics 13, 2927–2943 (2014). [PubMed: 25161197]

111. Swendsen RH, Wang JS, Replica Monte Carlo simulation of spin glasses. Phys. Rev. Lett 57, 2607–2609 (1986). [PubMed: 10033814]

112. Viswanath S, Chemmama IE, Cimermancic P, Sali A, Assessing Exhaustiveness of Stochastic Sampling for Integrative Modeling of Macromolecular Structures. Biophys. J 113, 2344–2353 (2017). [PubMed: 29211988]

113. Vallat B, Webb B, Westbrook JD, Sali A, Berman HM, Development of a Prototype System for Archiving Integrative/Hybrid Structure Models of Biological Macromolecules. Structure 26, 894–904.e892 (2018). [PubMed: 29657133]

114. Chodera JD, A Simple Method for Automated Equilibration Detection in Molecular Simulations. J. Chem. Theory Comput 12, 1799–1805 (2016). [PubMed: 26771390]

115. McInnes L, Healy J, Astels S, hdbscan: Hierarchical density based clustering. The Journal of Open Source Software 2, 205 (2017).

116. Merkley ED et al., Distance restraints from crosslinking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine-lysine distances. Protein Sci. 23, 747–759 (2014). [PubMed: 24639379]

117. Schneidman-Duhovny D et al., A method for integrative structure determination of protein-protein complexes. Bioinformatics 28, 3282–3289 (2012). [PubMed: 23093611]

118. de Hoon MJ, Imoto S, Nolan J, Miyano S, Open source clustering software. Bioinformatics 20, 1453–1454 (2004). [PubMed: 14871861]

119. Saldanha AJ, Java Treeview--extensible visualization of microarray data. Bioinformatics 20, 3246–3248 (2004). [PubMed: 15180930]

120. Morris JH, Huang CC, Babbitt PC, Ferrin TE, structureViz: linking Cytoscape and UCSF Chimera. Bioinformatics 23, 2345–2347 (2007). [PubMed: 17623706]

121. Doncheva NT, Klein K, Domingues FS, Albrecht M, Analyzing and visualizing residue networks of protein structures. Trends Biochem Sci 36, 179–182 (2011). [PubMed: 21345680]

122. Morris JH et al., setsApp for Cytoscape: Set operations for Cytoscape Nodes and Edges. F1000Res 3, 149 (2014). [PubMed: 25352980]

123. Collart MA, Oliviero S, Preparation of yeast RNA. Curr Protoc Mol Biol Chapter 13, Unit13 12 (2001).

124. Kim D et al., TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14, R36 (2013). [PubMed: 23618408]

125. Anders S, Pyl PT, Huber W, HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015). [PubMed: 25260700]

126. Love MI, Huber W, Anders S, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 15, 550 (2014). [PubMed: 25516281]

127. Costanzo M et al., The genetic landscape of a cell. Science 327, 425–431 (2010). [PubMed: 20093466]

128. Wilmes GM et al., A genetic interaction map of RNA-processing factors reveals links between Sem1/Dss1-containing complexes and mRNA export and splicing. Mol Cell 32, 735–746 (2008). [PubMed: 19061648]

129. Benjamini Y, Hochberg Y, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B 57, 289–300 (1995).

130. Wessel D, Flugge UI, A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. Anal Biochem 138, 141–143 (1984). [PubMed: 6731838]

131. Downey M et al., Acetylome profiling reveals overlap in the regulation of diverse processes by sirtuins, gcn5, and esa1. Mol Cell Proteomics 14, 162–176 (2015). [PubMed: 25381059]

132. Cox J, Mann M, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26, 1367–1372 (2008). [PubMed: 19029910]

133. MacLean B et al., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26, 966–968 (2010). [PubMed: 20147306]

134. Choi M et al., MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics 30, 2524–2526 (2014). [PubMed: 24794931]

135. Dronamraju R et al., Spt6 Association with RNA Polymerase II Directs mRNA Turnover During Transcription. Mol Cell 70, 1054–1066 e1054 (2018). [PubMed: 29932900]

136. Livak KJ, Schmittgen TD, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25, 402–408 (2001). [PubMed: 11846609]

137. Dronamraju R, Strahl BD, A feed forward circuit comprising Spt6, Ctk1 and PAF regulates Pol II CTD phosphorylation and transcription elongation. Nucleic Acids Res 42, 870–881 (2014). [PubMed: 24163256]

138. Robinson JT et al., Integrative genomics viewer. Nat Biotechnol 29, 24–26 (2011). [PubMed: 21221095]

139. Liao Y, Smyth GK, Shi W, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014). [PubMed: 24227677]

140. Edgar R, Domrachev M, Lash AE, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30, 207–210 (2002). [PubMed: 11752295]

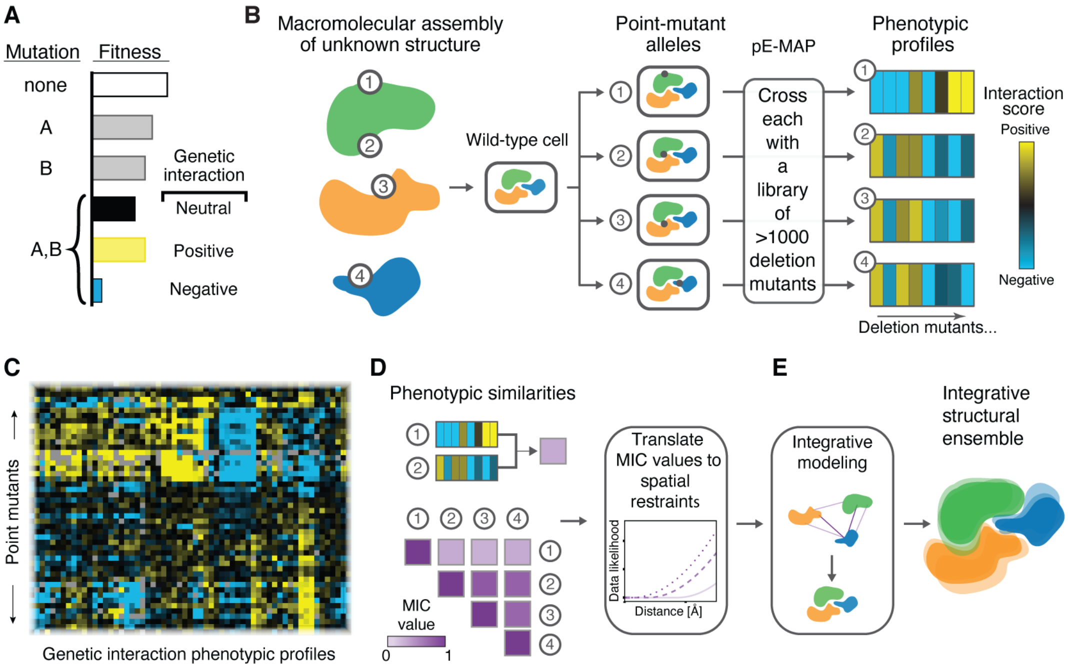
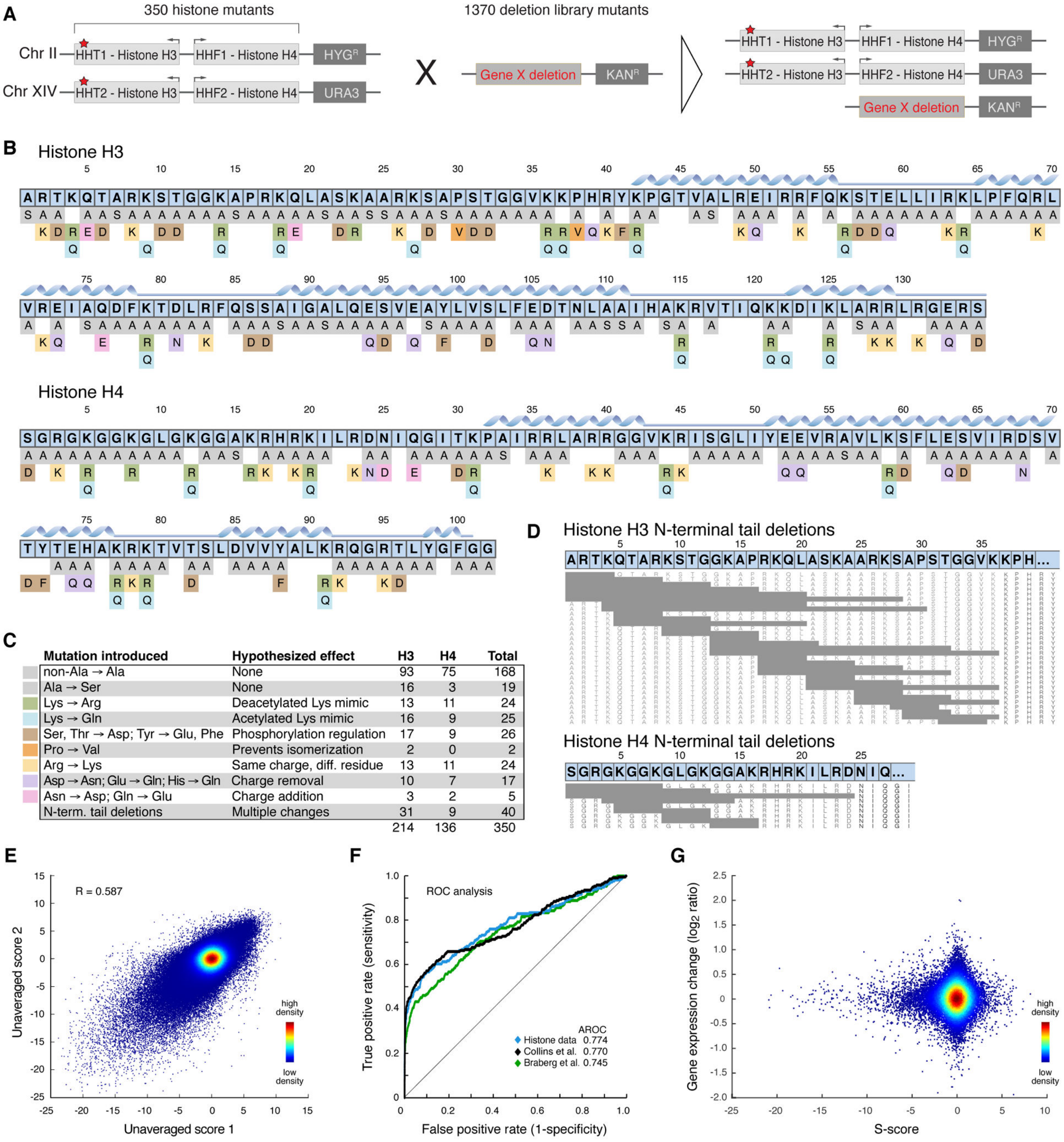


Fig. 1. Building spatial restraints from pairwise genetic perturbations.

(A) Genetic interactions arise when the combined fitness defect of a double mutant deviates from the expected multiplicative growth defect of the two single mutants. **(B)** The generation of a pE-MAP relies on a collection of point mutations, which is constructed by systematic mutagenesis of genes that encode the subunits of a macromolecular assembly (mutations labeled 1-4). The point mutant strains are then crossed against a library of gene deletions, followed by fitness measurement and subsequent calculation of genetic interaction scores to obtain the phenotypic profiles. **(C)** An example subset of a pE-MAP of point mutants crossed against a library of gene deletions. **(D)** Each pairwise combination of phenotypic profiles is transformed into a single MIC value that reflects the similarity between the two profiles. The MIC values are translated into spatial restrains for integrative modeling. **(E)** The MIC values and other input information are used for integrative structure modeling. An ensemble of structures that satisfy the input information is obtained.



Mutation introduced	Hypothesized effect	H3	H4	Total
non-Ala → Ala	None	93	75	168
Ala → Ser	None	16	3	19
Lys → Arg	Deacetylated Lys mimic	13	11	24
Lys → Gln	Acetylated Lys mimic	16	9	25
Ser, Thr → Asp; Tyr → Glu, Phe	Phosphorylation regulation	17	9	26
Pro → Val	Prevents isomerization	2	0	2
Arg → Lys	Same charge, diff. residue	13	11	24
Asp → Asn; Glu → Gln; His → Gln	Charge removal	10	7	17
Asn → Asp; Gln → Glu	Charge addition	3	2	5
N-term. tail deletions	Multiple changes	31	9	40
		214	136	350

Fig. 2. Genetic interrogation of histones H3 and H4 at a residue-level resolution.

(A) Each histone mutant strain was modified at both native loci (*HHT1* & *HHT2* for H3 or *HHF1* & *HHF2* for H4, red stars) and crossed against a library of 1370 different deletion mutants (or hypomorphic alleles for essential genes). **(B)** Schematic of the histone point mutants analyzed in this study (Table S1). Secondary structure elements are indicated as ribbons above the amino acid sequence. The mutations are color-coded according to the mutation introduced (**Fig. 2C**). Mutations resulting in inviable strains or strains too sick for genetic analysis are shown in Fig. S1. **(C)** Table of histone mutant categories and their hypothesized effects (color coding as in **Fig. 2B**). **(D)** Overview of viable H3 and H4 tail deletion mutants amenable to pE-MAP analysis. The amino acid sequences of the wt alleles are shown on top (residues 1-39 of histone H3 and 1-27 of histone H4). Grey bars signify the deleted residues in H3 and H4. **(E)** Reproducibility of histone pE-MAP S-scores between biological replicates. Plotted are all S-score pairs among the biological replicas, which include triplicate measurements for 346 histone alleles and duplicates of 4 alleles (H4E73Q, H4H18A, H4121A and H4K44Q). **(F)** ROC curves showing the power to predict physical interactions between pairs of proteins from this pE-MAP (blue) as well as previously published pE-MAP (green, (15)) and E-MAP (black, (29)) data. **(G)** Relationship between gene expression ($\log_2$ fold-change over wt) and S-scores of 29 H3 and H4 alleles (Table S2). Data from all 1,256 deletion library mutants that were measured in both RNAseq expression and pE-MAP analysis are plotted.

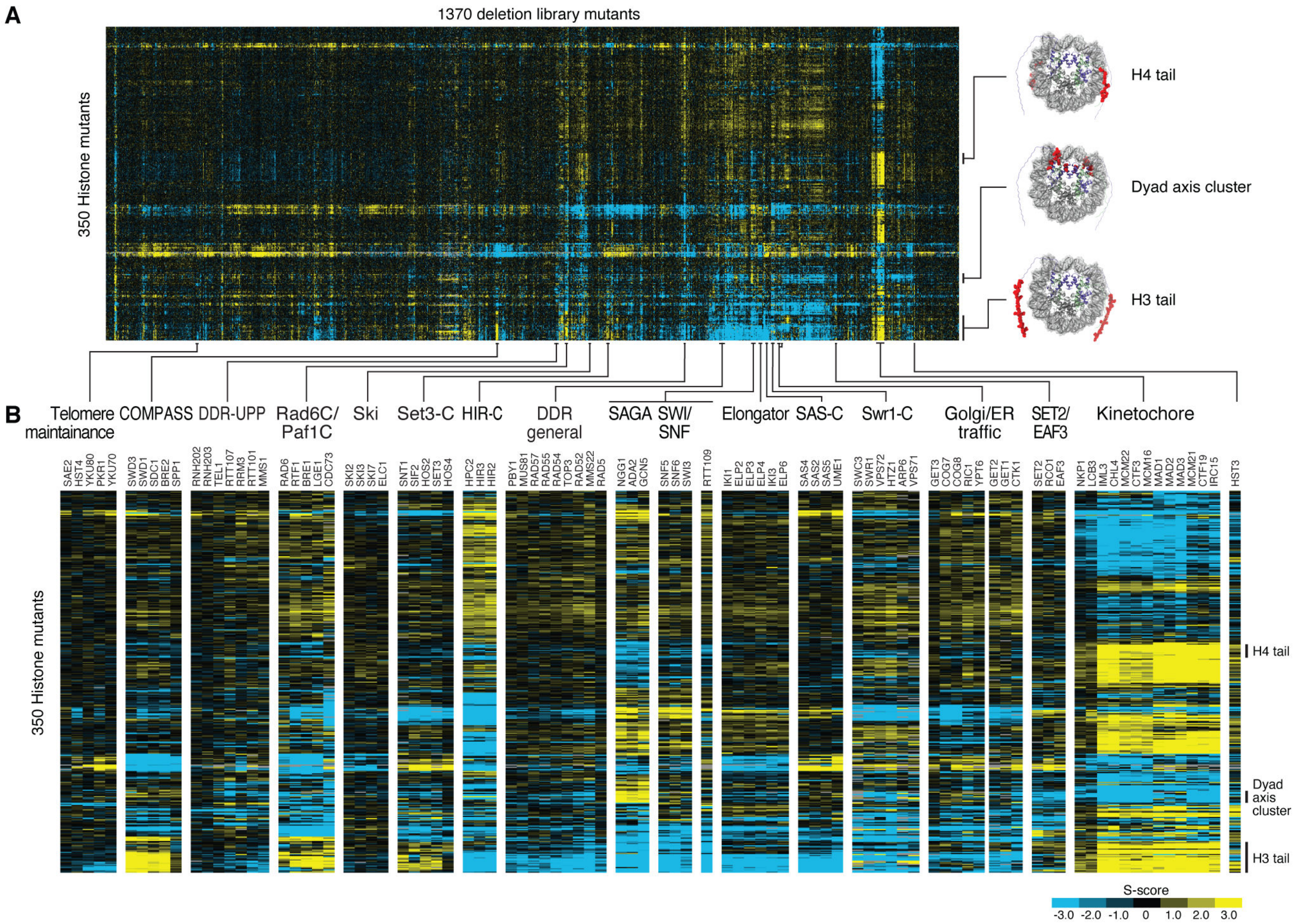


Fig. 3. The genetic interaction landscape of histones H3 and H4.

(A) Hierarchically clustered pE-MAP of 350 histone H3 and H4 alleles screened against a library of 1,370 deletion mutants or hypomorphic alleles. The pE-MAP consists of more than 479,000 genetic interactions. Positive- (suppressive/epistatic) and negative (synthetic sick) genetic interactions are colored in yellow or blue, respectively. Examples of histone alleles with similar genetic interaction profiles are highlighted on the right side in context of the nucleosome structure. The nucleosome structure is modified from PDB 1ID3 (Data S2), with H3 in purple, H4 in green, and mutated or deleted residues highlighted in red. N-terminal tail residues of H3 and H4 not included in 1ID3 are visualized as strings on the periphery. **(B)** Examples of genetic interaction profiles of gene clusters belonging to known protein complexes or biological pathways are highlighted and their genetic interaction profiles enlarged from **Fig. 3A**. DDR - DNA damage/repair, UPP - ubiquitin proteasome pathway.

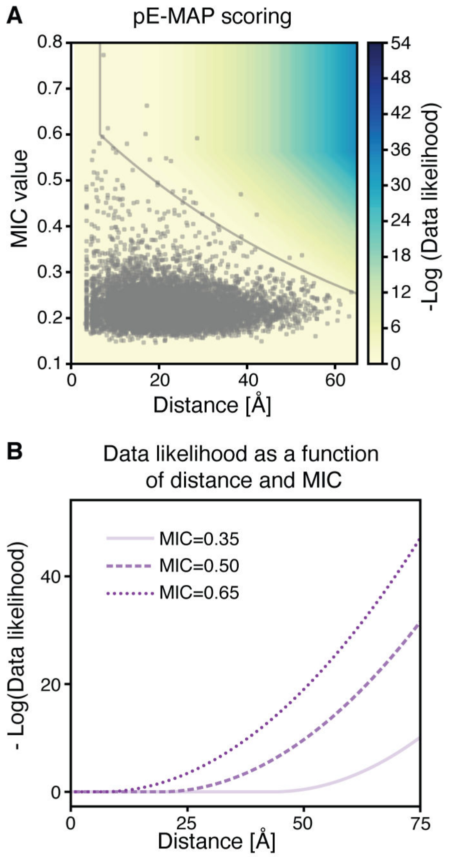


Fig. 4. Generation of the scoring function

(A) Relationship between pairwise distances and MIC values. The solid grey line represents the logarithmic decay fit to the upper distance bounds (Methods, Eq. 1). The background color gradient reflects how the data likelihood depends on MIC value and distance. (**B**) -Log of the data likelihood as a function of distance for different MIC values (Methods).

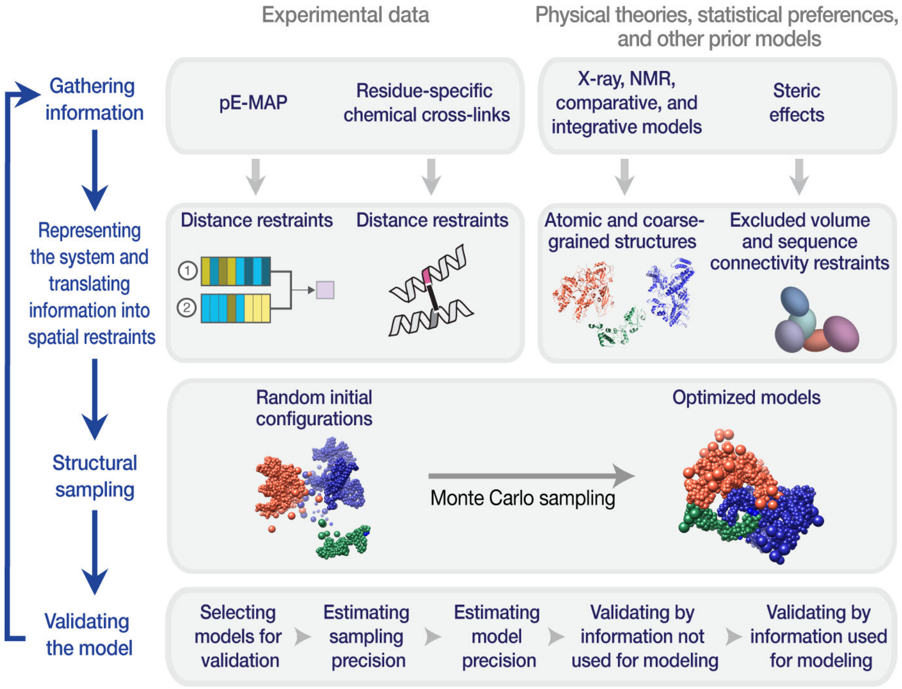

Experimental data
Physical theories, statistical preferences, and other prior models
Gathering information
pE-MAP
Residue-specific chemical cross-links
X-ray, NMR, comparative, and integrative models
Steric effects
Representing the system and translating information into spatial restraints
Distance restraints
Distance restraints
Atomic and coarse-grained structures
Excluded volume and sequence connectivity restraints
1
2
Structural sampling
Random initial configurations
Monte Carlo sampling
Optimized models
Validating the model
Selecting models for validation
Estimating sampling precision
Estimating model precision
Validating by information not used for modeling
Validating by information used for modeling

Fig. 5. Description of the integrative modeling workflow.

The four stages include: (1) gathering all available experimental data and prior information; (2) translating all information into a representation of the assembly components and a scoring function for ranking alternative assembly structures; (3) sampling structural models; and (4) validating the model. In this example, the representation of the components of a complex is based on comparative models of its components. The scoring function consists of spatial restraints that are obtained from pE-MAP and/or cross-linking experiments (evolutionary coupling analysis is not indicated in this scheme) as well as excluded volume and sequence connectivity restraints. The sampling explores the configurations of rigid components, searching for those assembly structures that satisfy the spatial restraints as well as possible. The goal is to obtain an ensemble of structures that satisfy the input data within the uncertainty of the data used to compute them. The sampling precision is estimated, and models are clustered and evaluated by the degree to which they satisfy the input information used to construct them as well as omitted information. The protocol can iterate through the four stages until the models are judged to be satisfactory, most often based on their precision and the degree to which they satisfy the data.

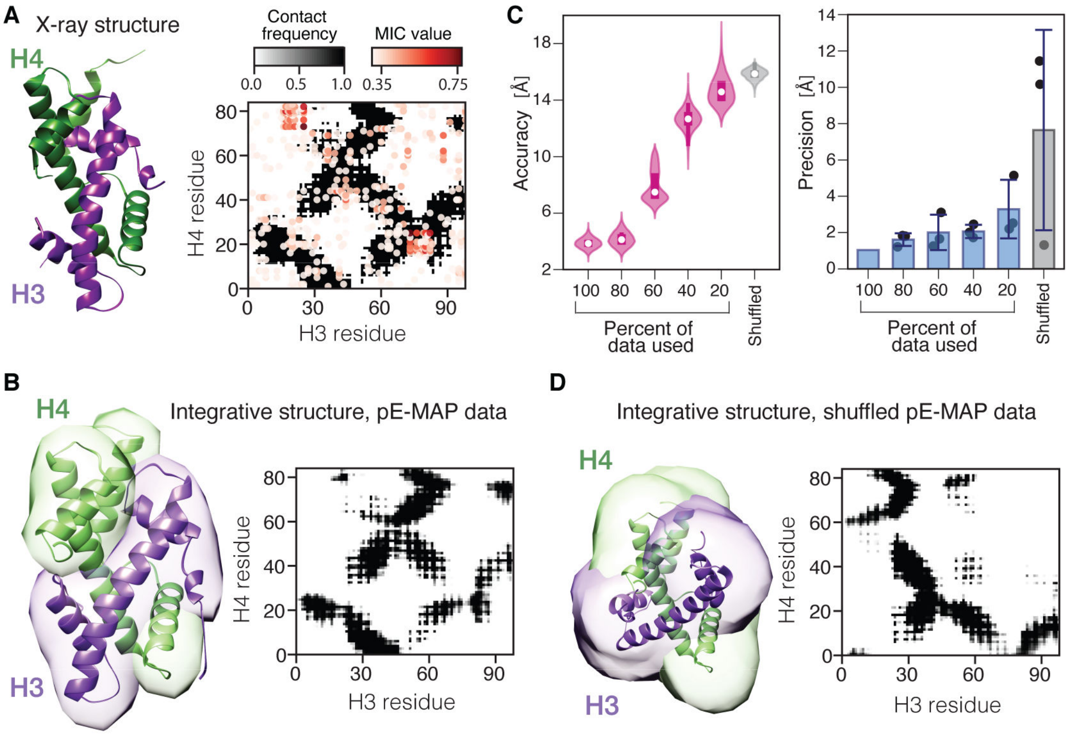
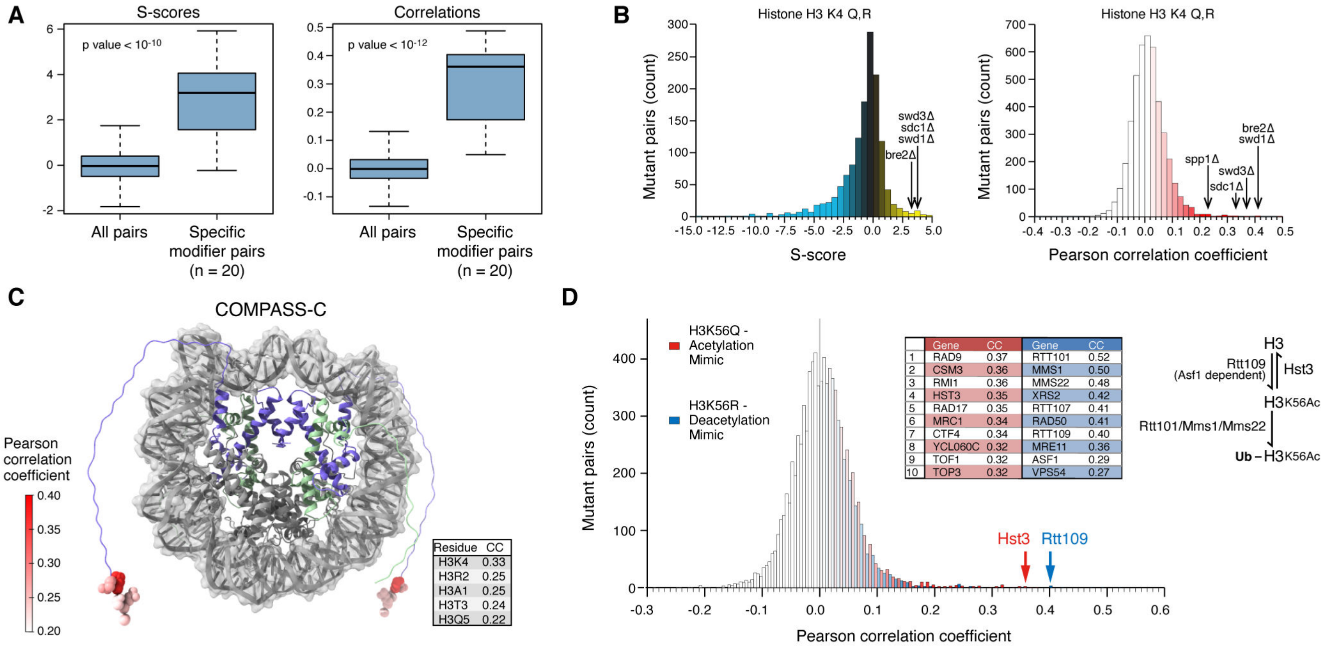


Fig. 6. Integrative structure determination of histones H3 and H4.

(A) The native structure of the histone H3-H4 dimer (PDB: 1ID3, left) and its contact map (right). In contact maps, the intensity of gray is proportional to the relative frequency of residue-residue contacts in the models (cutoff distance of 12 Å). For X-ray structures, the contact frequency is either 0 (white) or 1 (black). The circles correspond to the pairs of restrained residues, with the intensity of red proportional to the MIC value (MIC > 0.3), showing that the pairs of residues with high MIC values are distributed throughout the proteins. **(B)** The localization probability density of the ensemble of structures is shown with a representative (centroid) structure from the computed ensemble embedded within it (left) and the corresponding contact map (right). The localization probability density map represents the probability of any volume element being occupied by a given protein. **(C)** Distributions of accuracy (left) of structures in the ensembles and model precisions (right) based on the full pE-MAP dataset, resampled datasets that consider fractions of the data, and using shuffled MIC values. The white dots represent median accuracies and the error bars represent the standard deviations of model precision over three independent realizations (shown as dots). **(D)** Localization probability density and centroid structure (left), and contact map (right), computed with shuffled MIC values (Methods).



Residue	CC
H3K4	0.33
H3R2	0.25
H3A1	0.25
H3T3	0.24
H3Q5	0.22

	Gene	CC	Gene	CC
1	RAD9	0.37	RTT101	0.52
2	CSM3	0.36	MMS1	0.50
3	RMI1	0.36	MMS22	0.48
4	HST3	0.35	XRS2	0.42
5	RAD17	0.35	RTT107	0.41
6	MRC1	0.34	RAD50	0.41
7	CTF4	0.34	RTT109	0.40
8	YCL060C	0.32	MRE11	0.36
9	TOF1	0.32	ASF1	0.29
10	TOP3	0.32	VPS54	0.27

Fig. 7. Connecting individual histone residues and regions to other associated complexes

(A) Comparison of S-scores and Pearson correlation coefficients of phenotypic profiles of modifier-residue pairs to the overall data. Only residues with a single known modifier and modifiers with a single known target residue were included (Table S4). p-values were calculated using two-sided Wilcoxon rank sum tests. The whiskers of the boxplots extend to a maximum of 1.5 × IQR (interquartile range) and outliers are not plotted. **(B)** Average distributions of S-scores (left) and phenotypic profile correlations (right) of H3K4 mutants (mean of H3K4Q and H3K4R). Members of the COMPASS complex that exhibit a mean S-score >2.5 or a mean genetic interaction profile correlation >0.2 with H3K4 mutants are highlighted. The COMPASS complex is responsible for H3K4 methylation. **(C)** Mapping of genetic interaction profile correlations to COMPASS complex members on the structure of the nucleosome (modified PDB 1ID3, Data S2). N-terminal tail residues of H3 and H4 not included in 1ID3 are visualized as strings on the periphery. Only residues that exhibit a median genetic profile correlation >0.2 with the COMPASS subunits are highlighted (Methods). H3 is depicted in purple, H4 in light green, and H2A/H2B and DNA in grey. The red color gradient reflects the strength of the correlation between each residue and the COMPASS members, calculated as the median correlation between the residue's tested mutations and the COMPASS members. **(D)** Distributions of genetic interaction profile correlations of H3K56Q (acetylation mimic) and H3K56R (deacetylation mimic). Correlations of key H3K56ac-level regulators, Rtt109 (acetylating) and Hst3 (deacetylating), are highlighted. The cartoon outlines the H3K56 acetylation pathway and its role in H3 ubiquitylation. Rtt109 acetylates H3K56 via an Asf1-dependent mechanism, which promotes ubiquitylation of H3 by Rtt101-Mms1 and Mms22. These 5 gene deletions are all found among the top 10 most similar to the deacetylation mimic H3K56R, whereas deletion of the H3K56 deacetylase Hst3 instead gives rise to a profile similar to the acetylation mimic H3K56Q (table inset). CC, Pearson correlation coefficient.

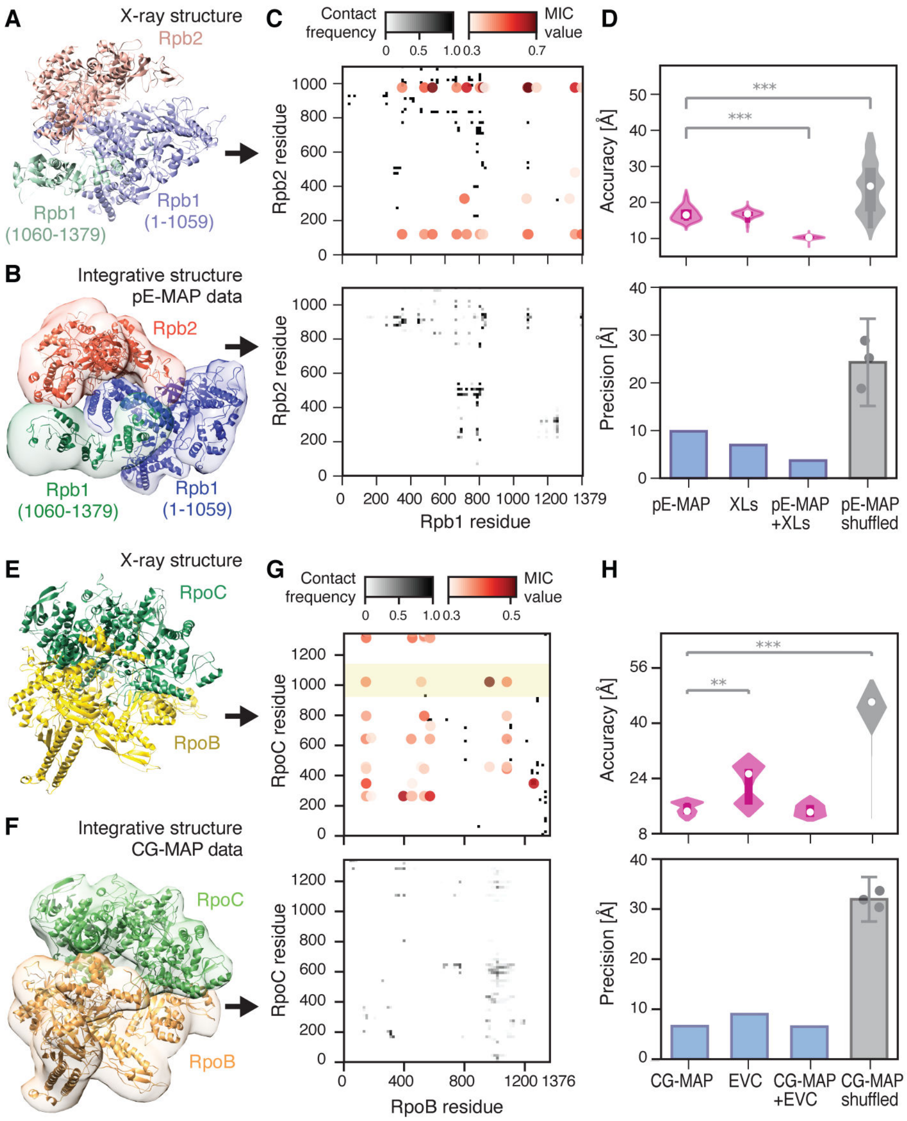


Fig. 8. Integrative structure determination of yeast RNAPII and bacterial RNAP.

(A) The native structure of Rpb1-Rpb2 (PDB: 2E2H) showing its three rigid-body components. Rpb1 was split into two domains, as shown. **(B)** The localization probability density of the ensemble of the three rigid-body structures is shown with a representative (centroid) structure from the computed ensemble embedded within it. **(C)** Contact maps computed for the X-ray structure (top) and model using the pE-MAP dataset (bottom). The circles correspond to the pairs of restrained residues, with the intensity of red proportional to the MIC value (MIC > 0.3). **(D)** Distributions of accuracy (top) for all structures in the ensemble and model precisions (bottom) for the computed ensembles based on pE-MAP and cross-link (XL) data. The white dots represent median accuracies. Error bars represent the standard deviations of model precisions over three independent realizations (shown as dots). ***p value < 10^{-12}. **(E)** Structure of subunits RpoB and RpoC from bacterial RNAP (PDB: 4YG2). **(F)** The localization probability density of the ensemble of the RpoB-RpoC structures with a representative (centroid) structure from the computed ensemble embedded within it. **(G)** Contact maps computed for the X-ray structure (top) and model using the CG-MAP dataset (bottom). The shaded yellow band represents a region missing in the X-ray structure. **(H)** Distributions of accuracy (top) for all structures in the ensemble and model precisions (bottom) for the ensembles based on CG-MAP and evolutionary coupling (EVC) data. The white dots represent median accuracies. The error bars represent the standard deviations of model precision over three independent realizations (shown as dots). **p value $< 10^{-6}$, ***p value $< 10^{-12}$.