Published in final edited form as:

Structure. 2022 April 07; 30(4): 498–509.e4. doi:10.1016/j.str.2021.12.005.

Quantitative Mining of Compositional Heterogeneity in Cryo-EM Datasets of Ribosome Assembly Intermediates

Jessica N. Rabuck-Gibbons[1], **Dmitry Lyumkis**[1,2], **James R. Williamson**[1,*]

[1]Department of Integrative Structural and Computational Biology, Department of Chemistry, and The Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

[2]Laboratory of Genetics and Helmsley Center for Genomic Medicine, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA

Summary

Single particle cryo-electron microscopy (cryo-EM) offers a unique opportunity to characterize macromolecular structural heterogeneity by virtue of its ability to place distinct particle populations into different groups through computational classification. However, there is a dearth of tools for surveying the heterogeneity landscape, quantitatively analyzing heterogeneous particle populations after classification, deciding how many unique classes are represented by the data, and accurately cross-comparing reconstructions. Here, we develop a workflow that contains discovery and analysis modules to quantitatively mine cryo-EM data for sets of structures with maximal diversity. This workflow was applied to a dataset of *E. coli* 50S ribosome assembly intermediates, which is characterized by significant structural heterogeneity. We identified more detailed branch points in the assembly process and characterized the interactions of an assembly factor with immature intermediates. While the tools described here were developed for ribosome assembly, they should be broadly applicable to the analysis of other heterogeneous cryo-EM datasets.

Graphical Abstract

*Corresponding author: jrwill@scripps.edu.

Author Contributions

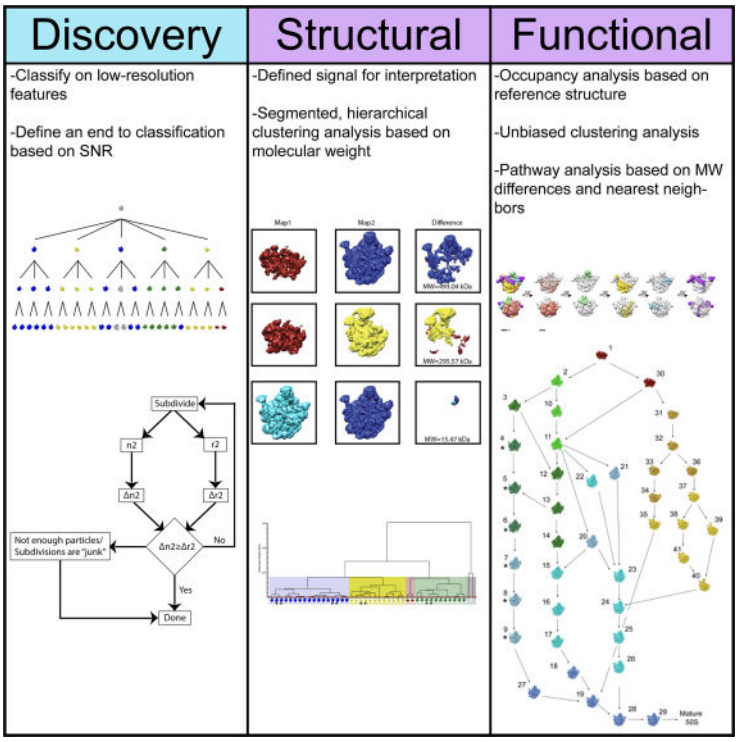
Jessica N. Rabuck-Gibbons: Conceptualization, Investigation, Methodology, Software, Formal Analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization.

Dmitry Lyumkis: Conceptualization, Investigation, Writing – Review & Editing, Resources.

James R. Williamson: Conceptualization, Methodology, Software, Data Curation, Writing – Review & Editing, Visualization, Supervision, Project Administration, Funding Acquisition.

Declaration of Interests

The authors declare no competing interests.



Rabuck-Gibbons et al. developed a workflow for the quantitative analysis of cryo-EM data starting with 3D classification. The workflow provides criteria for determining the end of classification and accurately cross-comparing reconstructions. This workflow was applied to *E. coli* 50S intermediates to identify structures and branchpoints in the assembly pathway.

Keywords

Cryo-electron microscopy; Single particle analysis; Ribosome biogenesis; Heterogeneity analysis; Classification

Introduction

Cryo-electron microscopy (cryo-EM) is a rapidly evolving, powerful technology for solving the structures of a wide variety of biological assemblies. The “resolution revolution” in cryo-EM (Kühlbrandt, 2014), caused in part by advances in direct electron detectors and improved data acquisition and analysis workflows, has led to high-resolution structural insights into a wide variety of biological processes performed by macromolecular assemblies (Fernandez-Leiro and Scheres, 2016). There have been steady, but consistent improvements to achievable resolution, and the collective tools are now enabling structure determination at true atomic resolution (Bartesaghi et al., 2015; Nakane et al., 2020; Tan et al., 2018; Yip et al., 2020; Zhang et al., 2020). There have also been numerous advances in workflows for analyzing structurally heterogeneous particle populations, and data processing software now routinely include strategies for handling distributions of structures that arise from compositional or conformational changes in the macromolecular species of interest

(Elmlund and Elmlund, 2012; Gao et al., 2004; Grant et al., 2018; Klaholz, 2015; Liao et al., 2015; Lyumkis et al., 2013; Nakane et al., 2018; Punjani and Fleet, 2021a, b; Scheres, 2016; Spahn and Penczek, 2009; Wang et al., 2013; White et al., 2017; Zhong et al., 2021). However, most current cryo-EM workflows still focus on achieving the maximum possible resolution, which requires selecting and averaging potentially heterogeneous subsets of the data in the interest of increasing the particle count for the homogeneous regions of a map. This strategy comes at the expense of either eliminating particle populations that do not conform to the predominant species or neglecting dynamic and labile regions of reconstructed maps, which are often of biological interest.

Another challenge in cryo-EM heterogeneity analysis is that there is no way to define the number of distinct structures in a given dataset *a priori*. It is up to the researcher to employ a classification strategy and to heuristically determine the number of distinct classes. Furthermore, there is no set procedure to determine the threshold for examining map features and differences between maps. Thresholds are often set in a subjective manner in order to best display the features of interest in the maps, although an approach was recently described where a voxel-based false discovery rate could be determined to establish a noise threshold for contouring (Beckers et al., 2019). Thus, determining the final number of classes in a dataset and quantitatively comparing a set of maps in order to tell a concise biological story with statistical significance remains a challenge.

The process of ribosome assembly provides a useful case study for mining and quantitatively assessing structural heterogeneity in cryo-EM data. The bacterial 70S ribosome is a complex macromolecular machine composed of three ribosomal RNAs (rRNAs) and ~50 ribosomal proteins (r-proteins) that form a large 50S subunit and a small 30S subunit. Ribosome assembly occurs within several minutes *in vivo*, and the process includes transcription and translation of the rRNAs and r-proteins, folding of the rRNA and r-proteins, and docking of the r-proteins on the rRNA scaffold. rRNA folding events and proper r-protein binding are facilitated by ~100 ribosome assembly factors. Given the efficiency and speed of the assembly process, structural intermediates are difficult to isolate and purify. However, perturbations in ribosome assembly lead to the accumulation of numerous structural intermediates, which collectively inform molecular mechanisms of ribosome assembly (Davis et al., 2016; Jomaa et al., 2014; Li et al., 2013; Ni et al., 2016; Rabuck-Gibbons et al., 2020; Sashital et al., 2014; Shajani et al., 2011; Stokes et al., 2014; Sykes et al., 2010). The major parts of the ribosome that are often present or missing in assembly intermediates are the central protuberance (CP), the L7/12 and L1 stalks, and the base (Figure 1A).

We previously developed a genetic approach by which the amount of a given r-protein, in our case bL17, could be titrated by the addition of the small molecule homoserine lactone (HSL) (Davis et al., 2016). Limiting the amount of bL17 induced a roadblock in ribosome assembly, causing intermediates to accumulate. In the first work using the bL17-lim strain (Davis et al., 2016), we identified thirteen distinct structures that fell into four main structural classes (Figure 1B). Here, we will continue to use the nomenclature for the main classes used by Davis and Tan, et al. These categories, ordered least to most mature, are the B class which is missing the base, CP, and both stalks, the C class in which the base is formed, but the central protuberance (CP) is either misdocked or altogether

missing, the D class in which the base of the 50S ribosome is missing, and the E class, which contains both the base and the CP, but has variability in the presence or absence of the stalks. Some of the "missing" regions (primarily rRNA, but they may also include r-proteins) described above are not present in the reconstructed maps but are in fact present in the sample and within individual particle images, meaning that they contribute to "biological noise". This becomes relevant for some of the decisions that need to be made in the data analysis workflow, as will be discussed below. In previous work, the four main initial classes belonging to the 50S assembly intermediates (B, C, D, E) were each further subdivided by an additional round of subclassification, resulting in thirteen distinct structures. While classes were identified belonging to the 30S and 70S (F class and A class in Davis *et al.,* 2016), they are not explicitly described in our previous work or in the work described here. Previously, several different subclassification schemes were attempted using heuristics to determine the number of subclasses, but no attempt was made to establish quantitative criteria by which the subclassification or coverage of relevant classes would be complete.

As our goal is to define broad trends in ribosome assembly through various perturbations, it is important to quantitatively assess differences between intermediates that accumulate under various specific conditions and to organize them into a ribosome assembly landscape (Bernstein et al., 2004; Davis et al., 2016; Harnpicharnchai et al., 2001; Jomaa et al., 2011; Loerke et al., 2010; Nikolay et al., 2018; Razi et al., 2017; Uicker et al., 2006). To this end, we developed a data processing framework to analyze cryo-EM datasets methodically and quantitatively in order to assess the number of distinct structures, the significant differences among them, and to place these structures into a biological context. When we apply our complete workflow to a dataset of ribosome assembly intermediates from bL17-lim, we discover a total of forty-one different structures that are identifiable based on a defined set of cutoff parameters. These structures include several intermediates that are, to our knowledge, novel, including an immature assembly intermediate, and an independent pathway contingent on the binding of a ribosome assembly factor, as well as late-stage assembly intermediates. Together, these are organized into a revised assembly landscape for the 50S ribosomal subunit under bL17-lim conditions.

Results and Discussion

An overview of the heterogeneity processing workflow

There are two main phases in the framework for systematic analysis of heterogeneous ensembles of macromolecular conformations (Figure 2). The first phase is a discovery phase, which begins with iterative rounds of hierarchical classification and sub-classification using a defined set of thresholding parameters. The goal of this first phase is to uncover the broad spectrum of distinct classes in a cryo-EM dataset, starting with traditional pre-processing and data cleaning steps (e.g. motion correction, particle picking, CTF estimation, and initial 2D and 3D classification). The initial data cleaning steps defined here are intended to be very lenient, such that the only particles removed from the dataset are clear artifacts or molecular species that are not of interest. For example, in the case of 50S ribosome assembly intermediate analysis, we remove particles that are clearly 30S ribosomal subunits, 70S ribosomes, and/or other large contaminating (clearly defined) macromolecular

assemblies from the stack, but we do not remove any classes that could possibly be 50S assembly intermediates. After the cleaning steps, an iterative subclassification strategy is used to parse out molecular heterogeneity. After an initial round of classification, each class (class X) is subjected to a n=2 subclassification, resulting in two potential subclasses, X1 and X2. Both subclasses are processed and binarized, and then difference maps X1-X2 and X2-X1 are calculated, to determine if there is more heterogeneity that can be mined from each class X. The differences are defined as changes in molecular weight, which directly correspond to the voxel-based difference densities and are calculated using 0.81 Da/Å^3 (see Methods). If the difference volumes don't reach a chosen molecular weight or resolution threshold, then the subclassification is rejected, and further subclassification is terminated. If neither of these two criteria are reached, the binary subclassification process is iteratively repeated until one of the convergence criteria are met.

The second phase is an analysis phase, which is intended to quantitatively define and distinguish structural features between maps, and further, to establish the number of structural states using a given set of quantitative cutoffs. During this hierarchical difference analysis, the full matrix of difference maps is calculated, and the molecular weight differences in the difference maps are used as a metric to cluster the classes, which can be visualized as a particle dendrogram. A line can be drawn through the dendrogram at a chosen molecular weight threshold, and the maps for any set of branches in the dendrogram below this threshold can be combined.

In order to distinguish between classes that have large amounts of compositional heterogeneity, the resulting set of maps are compared to a catalog of coarse-grained structural features that are calculated from a reference structure, in this case the bacterial 50S ribosomal subunit. It is convenient to use features such as rRNA helices and r-proteins, that may be present or absent in various classes. The presence of these coarse-grained reference features is quantitatively analyzed using hierarchical clustering to organize and visualize the patterns of variation among the final set of particle classes. For our dataset of bacterial ribosome assembly intermediates, these features are used to place the observed classes into a putative assembly pathway, based on a principle of parsimonious folding and unfolding.

A divisive resolution-limited subclassification approach facilitates identifying novel species

A major challenge in the analysis of heterogeneous datasets is the accurate identification of a broad diversity of structural states. To address this, we developed a classification strategy to mine an experimental cryo-EM dataset for distinct particle populations. Classification and refinement of particle classes can be undertaken using a variety of software packages, and we have adopted the latest version of FrealignX, whose code base is also implemented within *cis*TEM (Grant et al., 2018; Lyumkis et al., 2013). We note that most processing packages that are capable of classifying single-particle cryo-EM data can be employed for this purpose (Nakane et al., 2018; Punjani and Fleet, 2021a, b; Scheres, 2016; Zhong et al., 2021).

In typical cryo-EM workflows, 3D classification is performed several times, with different choices for the total number of classes (*n*). If *n* is too small, the resulting classes may have averaged properties leading to loss of structural diversity but potentially higher resolution in the homogeneous regions. If *n* is too large, the data is subdivided into nearly identical classes, but each class is characterized by lower resolution, because the particle count contributing to the class decreases. For the characterization of intrinsically heterogeneous datasets such as those encountered during ribosome assembly, the goal of 3D classification is to capture the full range of structural diversity, as opposed to a select few well-resolved classes. Therefore, we developed an iterative subclassification strategy to systematically mine the data and identify distinct structural intermediates, including species that are rare and underpopulated.

With the knowledge that our test dataset harbored at least thirteen intermediates (Davis et al., 2016), we started with *n*=10 in order to evaluate parameters for subclassification. The ten initial classes are shown in Figure 3A. While we expected that we would find the previous B, C, D and E classes in the dataset, the B-class was not present, and rather, multiple classes that are subtle variations of the E-class were present. This exemplifies one of the pitfalls of classification that we term “hiding”, where subclasses can be mixed, only to emerge at subsequent stages of subclassification. A survey of various classification parameters within FrealignX revealed that lowering the *res_high_class* parameter, which is the resolution of the data to be used for classification, ameliorated class hiding and had a strong effect on the classes that emerged. This parameter is typically set to just below the estimated resolution limit of the data. However, by setting *res_high_class* to 20Å, the gross class heterogeneity increased, and the expected B-class emerged (Figure 3B). The resolution threshold for classification is frequently defaulted and determined automatically during classification, but it may also be explicitly set by the user or limited to the resolution of the first Thon ring (Scheres, 2012, 2016; Scheres et al., 2008). With the well-defined ribosome assembly case study, we show that a lower resolution threshold during classification helps to identify particle subsets that are substantially distinct from the predominant species, as we were able to identify the B class in the ten classes once the *res_high_class* parameter was changed.

We also examined different iterative subclassification strategies to further ameliorate “hiding” with various numbers of classes used for each stage of subclassification. In order to test the success of these strategies, we selected a final *n* of ~30, which was chosen because it provided a convenient number to evaluate a variety of subclassification schemes, and because it was close to twice the final number of classes found in the original bL17-lim dataset (Davis et al., 2016). The five classification schemes (Figure 3C) tested were: (1) a simple 1-round classification with *n*=30, (2) a 2-round hierarchical classification of 6 initial classes, each subdivided into 5 (n_1=6,n_2=5; total n=30,), (3) a 3-round hierarchical subclassification of 2 initial classes each subdivided into 3, with a second subdivision into 5 (n_1=2,n_2=3,n_3=5; total n=30), (4) a 3-round hierarchical classification of 5 initial classes subdivided into 3, then subdivided into 2 n_1=5,n_2=3,n_3=2; total n=30,), and (5) a 5-round hierarchical binary subclassification strategy, where 2 initial classes were subdivided into 2 until n=32 was reached (n_1=2,n_2=2,n_3=2,n_4=2,n_5=2; total n=32,).

With the sole exception of the simple single-round classification with n=30, all of these divisive schemes yielded classes that, to our knowledge, have not been previously identified (Supplemental Figure 1, indicated by *). Furthermore, the iterative divisive approaches produced the greatest range of structural diversity and avoided grouping together dissimilar classes. This observation is perhaps not unexpected, as it is well known that a divisive classification approach avoids local minima within the search space and is more robust than attempting to produce a final number of classes directly (Gray, 1984; Sorzano et al., 2010). Qualitatively, a first round of classification where n_1 is on the order of the number of major classes works well, followed by smaller subdivisions. As an example, a three round subclassification scheme with [$n_1 = 5$, $n_2 = 3$, $n_3 = 2$], for a total of 30 final classes, identified the greatest number of unique structures, as shown in Supplemental Figure 1. For this reason, we proceeded with the $n_1 = 5$, $n_2 = 3$, $n_3 = 2$ approach for our work, although we note that the optimal classification scheme will likely vary with the distinct heterogeneity spectrum for each unique dataset.

The phenomenon of hiding is linked to both the chosen resolution threshold and the number of classes chosen for the initial classification. Hiding is ameliorated by *both* a lower resolution threshold and choosing an initial number of classes that spans the major classes. Our results indicate that there is a “goldilocks” phenomenon for the initial subclassification, where too few or too many initial classes is suboptimal. Given that the observed final classes are relatively independent of the details of the subclassification, we turned our attention to the criteria for termination of subclassification.

Defining an endpoint for subclassification

The determination of when subclassification is complete is a key question in cryo-EM analysis. Many times, classification is considered finished if a specific region of interest can be resolved to a satisfactory resolution, depending on what question(s) the user wishes to address. However, this subjective approach may be insufficient for the purpose of uncovering hidden features and discovering new structural states, especially if there are multiple datasets to be compared. To guide the analysis of our bL17-lim dataset, and to establish a protocol that can be used to analyze other data with statistical significance, our goal was to establish metrics by which we could confidently terminate the subclassification. We adopted a simple metric to determine the endpoint of subclassification. For any given class at any stage of subclassification, a test subclassification is performed with n=2. If the two resulting subclasses differ by less than a chosen noise threshold, or by less than a chosen molecular weight threshold, then subclassification is complete, and the subdivision is rejected. Conversely, if the thresholds are exceeded, the subclassification is retained, and the two resulting classes are iteratively subjected to additional subclassification until the termination thresholds are met (Figure 2).

There are at least two types of noise that need to be considered in the difference analysis that are used to conclude subclassification. First, there is the intrinsic noise floor in the map that arises from averaging noisy image data during the reconstruction process. Second, there is biological noise, which can be broadly attributed to conformational and compositional heterogeneity, resulting in density above the intrinsic noise floor that cannot be interpreted

in terms of a structure or slight shifts of well-defined elements that may or may not be significant (Supplemental Figure 2). For example, in the case of ribosome assembly, there are portions of rRNA that are present in the sample, but do not resolve to a reasonable structure (Davis et al., 2016). To characterize a diverse set of classes, the goal is to identify significant differences that exceed chosen thresholds for these noise components.

A three-step process was developed to remedy the above challenges, based on the estimation of the real space noise in a given map. First, a low-pass filter is used to reduce high-frequency information in the map (low-pass filter threshold, Table 1). Clearly, this is inadvisable if high resolution is the goal for the experiment, but for heterogeneity analysis, resolution is secondary to differentiating between broader conformational and compositional differences. For ribosome assembly, since the RNA helical regions are clearly visible at 10Å resolution, we chose this resolution as our low-pass filter threshold. Second, it is important that the soft spherical mask typically applied during classification is removed, and the standard deviation of the unmasked map (σ_{map}) is calculated using standard cryo-EM analysis programs. While the signal from the macromolecular object is included in this calculation, that contribution to the standard deviation is negligible if the box size is sufficiently large, so that voxels containing true signal represents 1–2% of the total map volume. Effectively, σ_{map} provides a crude estimate of the intrinsic map noise. There are several other ways to calculate a noise threshold, most recently as demonstrated with the program developed by Beckers et al. (Beckers et al., 2019), which uses a false discovery rate (FDR) to determine the threshold used for visualization and analysis; alternatively one can use the noise sampled from the periphery of the map. The values of $3\sigma_{\text{map}}$ are highly correlated to the contour levels based on FDR as shown in Figure S3 but the $3\sigma_{\text{map}}$ threshold generally exceeds the FDR threshold, and is thus more conservative. Due to the prevalence of unresolved features in the ribosome data, we have used $3\sigma_{\text{map}}$ as a convenient threshold to eliminate noise. Third, each map is then binarized using a $3\sigma_{\text{map}}$ threshold such that intensities greater than $3\sigma_{\text{map}}$ were set to 1, and intensities less than $3\sigma_{\text{map}}$ were set to 0 (binarization threshold, Table 1). Other thresholds could be devised and implemented, as long as they are applied consistently across classes. These thresholded, binarized maps are used for the remainder of the analysis. Using these maps is advantageous because the “noise” from flexible regions is removed from the map, and there is a clear boundary of which parts of a structure are analyzed. Further, binarization facilitates coarse-grained analysis and eliminates the need for scaling.

To define the endpoint to classification, the above filtering and binarization steps are applied after a test n=2 subclassification of a given class X into class X1 and X2. If either class X1 or class X2 do not have a resolvable map, as defined by the resolution limit (r-limit, Table 1), then the classification process is terminated. If the differences between class X1 and class X2 are less than the volume limit (v-limit, Table 1), the subclassification is also terminated. However, if the differences between class X1 and X2 are greater than the v-limit, then the subclassification is retained, and classes X1 and X2 are in turn further subdivided into 2 classes. This process then repeats on all classes until either the r-limit or the v-limit are reached. This set of limits provides a consistent and quantitative basis for iterative subclassification.

Segmented difference analysis between map identifies the number of structural states

Having discovered the structural variants in the data, we then asked how the different maps compare to one another and where/what are the major differences. To address this question, we developed a strategy to quantitatively assess similarities between the classes. While the classification approach in the discovery phase is designed to terminate once the structural features were no longer distinguishable using the r-limit or v-limit, this procedure does not guarantee that individual structures within the collective set of reconstructions are all distinct from one another. More specifically, a situation can arise where two similar classes emerge from hiding in different branches of the subclassification tree.

In the first step, difference maps are calculated between all of the binarized, thresholded maps. Such difference maps are useful to identify regions of density that are distinct between classes, and in our case, provide both qualitative and quantitative insight into structural relationships between distinct assembly intermediates. Two specific examples for distinct “D-classes” are shown in Figure 4A–B. The first two columns display two distinct maps arising from some point during classification. The raw difference maps are shown in the third column (map1-map2, red; map2-map1, blue). The approximate molecular weight of these differences is also indicated. These difference maps are then segmented to remove “dust” that may arise from minor conformational or compositional variations between maps. This dust cannot be interpreted in biological terms but may add up to a significant molecular weight (Table 1 segmentation threshold, Figure 4). Such difference maps can be computed for all pairwise combinations of reconstructions arising from the classification procedure.

The pairwise difference maps are useful for both qualitative and quantitative downstream analyses. To parse through structural differences, define an accurate final number of *unique* structural variants in the data, and combine particles contributing to similar maps, we employed a simple hierarchical clustering approach based on the positive/negative molecular weight differences between structures. Based on the clustering, it is possible to pare down the maps and combine particles from similar reconstructions, even if they arise from different starting points in the classification (Figure 5A). At this stage, two classes can be combined if the molecular weight differences between the two classes are less than a given threshold. Since the branchpoints of the dendrogram provide a measure of *molecular weight* differences between maps, they can serve as a guide for analyzing the similarity between classes overall based on the nodes of the dendrogram (Figure 5B). In the example in Figure 5B, the dendrogram reveals that the leftmost structure is distinct from the other two and needs to be treated independently, whereas the latter two can be combined into a single class. Thus, although there are forty-two distinct structures in Figure 5A, after hierarchical clustering analysis and the subsequent merging of similar maps, there are forty-one distinct structures that will go forward in the analysis pathway. Collectively, these procedures enable us to identify the number of structural states within the data, given the limitations associated with identifying novel classes in the discovery phase and according to the established criteria in the analysis phase, defined above.

Defining relationships between distinct structures

An important step in analyzing differences between classes discovered within the above procedures for heterogenous cryo-EM data analysis is to define *where* differences between two maps are located. If a model (e.g. an atomic model or a cryo-EM structure) exists as a reference, and if the reconstructed maps differ primarily by compositional variation, then it is straightforward to use the model for interpreting the collective set of maps (Davis et al., 2016) in an "occupancy analysis." In the case of bacterial ribosome assembly, we have a well-defined reference model (Figure 6A). This reference structure is broken into its individual r-RNA and r-protein parts, yielding theoretical cryo-EM densities for each component (Figure 6B). Such individual densities can then be directly compared to densities arising from experimental cryo-EM classification. It is important that the reference densities are generated at (approximately) the same resolution as the experimental densities arising from hierarchical clustering and difference analyses. The theoretical maps are then binarized, which enables comparing the theoretical maps to the binarized experimental maps arising from subclassification. Each binarized class (Figure 6C) is then compared to each theoretical feature map by counting overlapping voxels and normalizing to the theoretical volume, to define the fractional occupancy of the selected feature in the map that can be completely present, partially present due to partial flexibility or a misdocked r-protein or helix, or completely missing (Figure 6D). The complete set of fractional occupancies are given as an n by m matrix of values between 0 and 1, where n describes the set of classes and m defines the number of features.

The resulting fractional occupancies can be visualized as a heat map and subjected to hierarchical clustering to organize the classes and features (Figure 6E). Clustering along the feature (x-axis) groups elements (in this case, r-proteins and rRNAs), and clustering along the map (y-axis) groups the maps according to their occupancy. As expected, the B, C, D, and E, maps cluster well together. The occupancy matrix facilitates the visualization of large blocks of structural features that co-vary across the particle classes, providing cooperative folding blocks (Figure 6F) (Davis et al., 2016). This procedure enables a quantitative comparison of distinct sets of maps that differ by compositional variants. We note that this procedure is not currently compatible with conformational variability or density that is not represented in the reference. However, if there are multiple reference models that differ by discrete conformational changes, the current protocol can be extended to competitively compare occupancies against different reference models.

Ordering structures in a ribosome assembly pathway

In the final step that is relevant to defining an assembly process, we developed a module that uses molecular weight differences to place ribosome assembly intermediates into a pathway. In this analysis, a "folding" matrix is calculated from the molecular weight difference that would need to be added to a given map to create a second map, and the "unfolding" matrix is calculated from the molecular weight that would need to be subtracted from one map to create a second. Each element of the folding/unfolding matrix can be considered as the driving force/barrier for a structural transition between two classes. By postulating that folding proceeds by incremental assembly, with minimal unfolding, a parsimonious transition graph can be constructed with allowed passages between classes based on simple

criteria – there is a molecular weight cutoff unfolding transitions, and there is a limit set to the number of transitions emanating from each class. Large unfolding events are unlikely, given the large number of states that are close in molecular weight, but small unfolding events must be permitted to allow for structural rearrangements required to transition between classes. Finally, it is likely that structural transitions proceed from a finite manifold of close intermediates. The folding and unfolding matrices can be used to construct a directed graph of allowed transitions using these criteria, as shown in Figure 7.

Analysis of bL17-lim data using the quantitative heterogeneity mining protocol

Our quantitative mining protocol was developed using data collected from newly purified assembly intermediates from the previously characterized bL17-limitation strain (Davis et al., 2016). We collected new cryo-EM data (Supplementary Table 1) and subjected it to our workflow. In the discovery phase, we employed an updated high-resolution limit for refinement (*res_high_class*=20Å) and an initial n5>n3>n2 hierarchical classification scheme, followed by additional rounds of binary subdivision. All maps were binarized according to the $3\sigma_{map}$ threshold determined individually for each map. To determine if subclassification was complete, we selected a v-limit of 1.5 kDa. The rationale for this choice is that 1.5 kDa represents the size of the smallest RNA helix present in the bacterial ribosome and therefore corresponds to the smallest feature that we would like to capture in the data. For our purposes, smaller features can be assumed to be either biological and/or experimental noise. After iterative subclassification, the total number of classes is forty-two. The similarity between all of the maps was then analyzed by the hierarchical clustering analysis as described above, and at this stage, pairs of classes were combined with a 10 kDa difference threshold (Figure 5A, dotted red line). This cutoff was chosen because it is close to the average molecular weight of all proteins and rRNA features, and we wished to reduce the complexity of our data. We found one pair of structures that were similar to one another according to our established criteria for biological significance, and the particles belonging to these classes were accordingly combined (Figure 5A). Thus, using this protocol, a total of forty-one ribosome assembly intermediates were identified using quantitative metrics and similarity analysis, with minimal heuristic intervention.

The classes were then subjected to an occupancy analysis to view the sets of cooperative folding blocks across the different classes. For the bL17-lim dataset here, the maps are compared to the reference crystal structure (PDB 4ybb) (Figure 6A). The reference 4ybb structure is filtered to 10Å and segmented into volumes corresponding to individual r-proteins and rRNA helices, resulting in 139 theoretical map segments (Figure 6B). The fractional analysis revealed five major structural blocks (Figure 6F) The largest, block I (red) is composed of structural elements that are largely present in all of the classes. These elements are found on the back of the ribosome and represent the structural core that can form without bL17. Block II (green) represents the central protuberance, which is fully formed in the D and E classes but is either missing or mis-docked in the B and C classes. Block III (yellow) maps to the base of the ribosome and the L1 stalk. These features are mostly present in the C and E classes but are missing in the B and D classes. These two blocks represent parallel pathways in assembly (Davis et al., 2016), as it is unlikely that the base of the ribosome would be unfolded or disordered in order to form the central

protuberance, and vice versa. Block IV (blue) represents density that is specific to the base of the L7/12 stalk and is mostly present in the D and some of the E classes. Finally, block V (purple) represents density that is mostly missing in all maps, and is composed of h68, bL9, the L7/12 stalk, and the top of the L1 stalk. These represent features that are among the last of the ribosome to fold (like h68) or are flexible elements (the stalks). bL9 is a special case, as the conformation in the crystal structure is an artefact due to crystallization; in cryo-EM structures, bL9 wraps around to the interface between the 30S and 50S subunits and is often flexible. These central blocks are similar to the ones that we discovered previously (Davis et al., 2016), but this updated occupancy matrix will allow us to compare the blocks that arise from other depletion or deletion strains in order to explore the cooperative block-like behavior or ribosome assembly in future work.

The ordering module was used to calculate an initial pathway in the absence of bL17-lim, which was modified by hand, as elements like the mis-docked central protuberance and non-native structural elements can have large effects on molecular weight differences but may arise earlier in the order of assembly. We found the same initial super classes as previously reported (B, C, D and E classes). While the classes we found were similar to the initial bL17 data (Supplemental Figure 4), the new classes enabled refinement of our bL17-lim ribosome assembly pathway. First, we found a YjgA-dependent pathway through the assembly process (Figure 7, classes denoted by *). YjgA was only bound if the central protuberance and the L1 stalks were present. We also discovered three potential parallel processes in the C class where the earliest event could either be the completion of the L1 stalk, the partial docking of the central protuberance, or the formation of the base of the L7/12 stalk. We did not previously observe the formation of the base of the L7/12 structure in the assembly pathway for any class. We also found an immature B class (Figure 7, structure 1) and an immature D class where the base was missing, but the L7/12 stalk was absent or present (Figure 7, structures 2 and 3), which were not present in the original set of 13 structures (Davis et al., 2016). In particular, the immature B class represents the least mature pre-50S intermediate identified to date. We also identified several structures that seem to be transition points between the two classes (Figure 7, structures 6 and 7), and we observe formation of density at the base of the structures, which is lacking in other D classes and is present in other E classes. These discoveries inform a better understanding of ribosome assembly in the context of bL17 limitation, and the data analysis process will allow us to quantitatively assess cryo-EM data from other limitation strains and ribosome assembly defects.

Conclusions

Heterogeneity analysis in cryo-EM provides exciting opportunities to discover new biology, but current workflows suffer from numerous challenges. The work here addresses three challenges that researchers face in the analysis of cryo-EM data, as exemplified using a case study of ribosome assembly intermediates: establishing a divisive approach to classification with well-defined endpoints to discover novel structures, a comprehensive difference analysis between distinct structures, and the application of well-defined criteria (thresholds) for limiting classification. The application of specific thresholds and limits (Table 1) has been critical to the success of analyzing ribosome assembly intermediate data. The

implementation of this workflow has allowed us to identify forty-one ribosome assembly intermediates, twenty-eight of which are distinct from the original 13 intermediates previously identified. These intermediate include an independent pathway for the assembly factor YjgA and the earliest intermediate discovered to date in the ribosome assembly process. The initial bL17-lim pathway proposed in 2016 used a heuristic, user-selected number of classes, and there was no quantitative way to determine if subclassification was complete. This work was motivated, in part, by the need for a consistent way to systematically analyze heterogeneity in multiple datasets resulting from different perturbations to ribosome assembly. The discovery and analysis modules of this workflow provide a powerful analysis for quantitatively interrogating heterogeneous cryo-EM data for complex biological processes.

STAR Methods

Resource Availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. James Williamson (jrwill@scripps.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

All original code is available in this paper's supplemental information. The raw dataset containing 481 movies, as well as the raw particle stack is deposited into EMPIAR under accession code EMPIAR-10841. All cryo-EM maps are deposited into the EMDB, as noted in the Key Resources Table.

Experimental Model and Subject Details.

Cells were grown and ribosomal particles were isolated as in (Davis et al., 2016). Briefly, strain JD321 was grown in M9 media (48mM Na2HPO4, 22mM KH2PO4, 8.5mM NaCl, 10mM MgCl2, 10mM MgSO4, 5.6mM glucose, 50mM Na3*EDTA, 25mM CaCl2, 50mM FeCl3, 0.5mM ZnSO4, 0.5mM CuSO4, 0.5mM MnSO4, 0.5mM CoCl2, 0.04mM d-biotin, 0.02mM folic acid, 0.08mM vitamin B1, 0.11mM calcium pantothenate, 0.4nM vitamin B12, 0.2mM nicotinamide, 0.07mM riboflavin, and 7.6mM (14NH4)2SO4]) with tetracycline (10 mg/mL), chloramphenicol (35 mg/mL),and limiting conditions HSL (0.1 nM) and harvested at OD=0.5. Cells were lysed in Buffer A (20mM Tris-HCl, 100mM NH4Cl, 10mM MgCl2, 0.5mM EDTA, 6mM b-mercaptoethanol; pH 7.5) by a mini bead beater, and the clarified lysate was fractionated on a 10–40% w/v sucrose gradient (50mM Tris-HCl, 100mM NH4Cl, 10mM MgCl2, 0.5mM EDTA, 6mM b-mercaptoethanol; pH 7.5).

Method Details

Electron Microscopy Data Collection—Fractions containing the ribosomal intermediates were spin-concentrated with a 100 kDa MW filter (Amicon) and buffer exchanged into Buffer A. 3 μl aliquot of this sample was added to a plasma cleaned

(Gatan, Solarus) 1.2mm hole, 1.3mm spacing holey gold grids (Russo and Passmore, 2014). Grids were manually frozen in liquid ethane, and single-particle data was collected using Leginon on a Titan Krios microscope (FEI) with a K2 summit direct detector (Gatan) in super-resolution mode (pixel size of 0.66Å at 22,500 magnification). A dose rate of ~5.8e^-/pix/sec was collected across 50 frames with a fluence of 33–35e^-/Å^2 at a tilt of −20° to compensate for preferred orientation (Tan et al., 2017b). A total of 481 movies were collected.

Cryo-EM data pre-processing—Movies were motion corrected using MotionCor2 (Zheng et al., 2017) within the Appion package (Lander et al., 2009), and CTF parameters were estimated using gCTF (Zhang, 2016). Particle selection was performed using FindEM template-matching (Roseman, 2004) and were extracted with a boxsize of 160 pixels. The extracted particles were classified in 2D using Relion (Scheres, 2016), at which point we removed the first set of classes that were clearly resembling either 30S ribosomal subunits, 70S ribosomes, other well-defined non-ribosomal macromolecular assemblies, or particles represented by poorly defined 2D classes. Subsequently, we performed a single round of 3D classification in Relion, initiating with a C class and asking for 10, classes. Results of 3D classification allowed us to further remove particles using the criteria above, leaving us with a 123,804 particle stack that could be then subjected to our iterative subclassification analysis, described below.

FrealignX Classifications—After conversion from Relion to FrealignX parameters, global refinements were performed in FrealignX, and all occupancies were randomized across the parameter files. A final value of 20Å was selected for *res_high_class*, and after every 10 cycles of classification/refinement, all classes were aligned to a C class scaffold using custom scripts for a 3D alignment with Chimera (Pettersen et al., 2004) while running FrealignX. The alignment to a C class at the end of every 10 cycles effectively puts all maps in frame with respect to one another and minimizes the chance of map drift during the course of classification, which might lead to mis-classification. For each classification step, 50 refinement/classification cycles were performed. After initial classification, each class was selected in a parameter file for subsequent rounds of classification using the merge_classes.exe in cisTEM (Grant et al., 2018) and custom scripts. The occupancies were randomized across the parameter files, and the same cycle of 50 cycles of refinement/classification interspersed with 3D alignment with Chimera every 10 cycles. For the unbinned maps, particles were re-extracted in Relion with a box size of 320, and the final parameter files in FrealignX were used to initiate a final 10 rounds of classification to achieve higher resolution. FSC curves and Euler plots were generated by FrealignX and cisTEM (Grant et al., 2018), and 3DFSC plots were calculated by the 3DFSC server (Tan et al., 2017a). The SCF was calculated according to the process in (Baldwin and Lyumkis, 2020, 2021). The 3DFSCs and all maps shown were visualized in Chimera (Pettersen et al., 2004), and the details for each map are indicated in Data S1.

Quantification and Statistical Analysis

Calculation of σ values—For analysis, each map was first filtered to 10Å. To calculate σ which was used as a measure of noise, each map was unmasked by expanding the

outer_radius in FrealignX so that the spherical particle mask would be larger than the box size. The Fourier folding of signal along the edges of the box was negligible. Relion 2.1 was used to calculate the σ value using the *relion_image_handler* command. Relion 2.1 was then used to create binarized maps using the *relion_image_handler* command, and the binarization threshold was set to 3σ.

Hierarchical clustering analysis—Thresholded, binarized maps were given as input to a custom Mathematica script (Wolfram Research, 2020). The Mathematica script calculated the segmented difference maps between all maps and calculated the molecular weights of the differences maps (in kilodaltons) using Equation 1(Ludtke, 2016):

$$MW = n_{pixels} * pixelsize^3 * \rho/1000$$

Density (ρ) is 0.81 daltons/Å^3. The MW difference matrix was clustered using the Euclidean distance metric and Ward's linkage and displayed in a dendrogram. Similar maps were averaged together after hierarchical clustering analysis using EMAN2 ((Ludtke, 2016).

Occupancy Analysis—The thresholded and binarized maps were given as input, and the reference map from the *E. coli* 50S subunit crystal structure (PDB ID 4YBB) was segmented into 139 elements comprised of individual ribosomal proteins and rRNA helices according to the 23S secondary structure. Theoretical densities for each r-protein and rRNA helix were calculated for each element at 10Å using the pdb2mrc command from EMAN. Prior to binarization, voxels that had overlapping theoretical density from two structural elements, were assigned to the smaller of the two theoretical volumes so that each pair of volumes is nonoverlapping. Each voxel density was binarized to either 0 or 1 using a threshold of 0.016, which is the threshold that gave the approximately correct molecular weight for individual r-proteins and rRNAs helices. The relative volumes in the binarized experimental and reference maps were calculated, which gave a fractional occupancy between 0 and 1 for each element. The occupancy values were clustered across the rows (classes) and columns (rRNA/protein elements) using an unsupervised hierarchical clustering using the Euclidean distance metric and Ward's linkage method, as implemented in Mathematica (Method S1).

Parsimonious folding/unfolding matrices.—A pathway diagram was constructed by using the n x n molecular weight difference matrices, $\mathbf{M_f}$ and $\mathbf{M_u}$, from a set of n structures. Each difference map (M_i-M_j) has negative elements corresponding to folding that occurs in the transition from class i to class j, and positive elements that correspond to unfolding that occurs in the transition from class i to class j. The volume changes for folding and unfolding form the elements of $\mathbf{M_f}$ or $\mathbf{M_u}$, noting that $\mathbf{M_u} = \mathbf{M_f}^{\mathbf{T}}$. The matrices $\mathbf{M_f}$ and $\mathbf{M_u}$ are used to construct a directed graph **G**, comprised of the set of vertices v_i, and a set of directed edges, e_{ij}, representing the allowed transitions between classes, The set of edges is initialized as the set of e_{ij} where $M_{u,i,j} > M_{f,i,j}$, such that only net folding transitions are allowed. The set of edges is pruned using two global parameters: θ_{unf} as a maximum threshold for unfolding, and n_{branch}, as a limit on the number of transitions emanating from a single class. The unfolding threshold limits unreasonable structural rearrangements, while the branching threshold limits transitions to a small set of the closest transitions. Edges are eliminated if

the unfolding exceeds the threshold such that $M_{u,i,j} > \theta_{unf}$, unless elimination of the edge results in a disconnected graph **G**. Next, for each vertex v_i, the set of remaining edges e_{ik} emanating from v_i, are sorted into the order based on the $M_{f,i,k}$, retaining at most the n_{branch} edges, again, unless deleting the edge would result in a disconnected graph **G**. The resulting transition graph **G** should have one or more *source* vertices (classes) that are the earliest classes in the assembly pathway, and one or more *sink* vertices that are the most mature classes in the pathway. Tuning of the parameters θ_{unf} and n_{branch}, adjusts the connectivity and degree of branching of the resulting graph. The graph vertices are annotated with thumbnails of the map, followed by manual layout of the graph into a sensible order in Adobe Illustrator. The values of θ_{unf} and n_{branch} used to generate the graph in Figure 7 were 390 kDa and 3, respectively (Method S1).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Molecular graphics and analyses were performed with the USCF Chimera package (supported by NIH P41 GM103311). This work was supported by grants from the NIH DP5-OD021396 and U54 AI150472 (to DL) R35-GM136412 (to JRW).

References

Baldwin PR, and Lyumkis D (2020). Non-uniformity of projection distributions attenuates resolution in Cryo-EM. Progress in biophysics and molecular biology 150, 160–183. [PubMed: 31525386]

Baldwin PR, and Lyumkis D (2021). Tools for visualizing and analyzing Fourier space sampling in Cryo-EM. Progress in Biophysics and Molecular Biology 160, 53–65. [PubMed: 32645314]

Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne JL, and Subramaniam S (2015). 2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. Science 348, 1147–1151. [PubMed: 25953817]

Beckers M, Jakobi AJ, and Sachse C (2019). Thresholding of cryo-EM density maps by false discovery rate control. IUCrJ 6, 18–33.

Bernstein KA, Gallagher JE, Mitchell BM, Granneman S, and Baserga SJ (2004). The small-subunit processome is a ribosome assembly intermediate. Eukaryotic cell 3, 1619–1626. [PubMed: 15590835]

Davis JH, Tan YZ, Carragher B, Potter CS, Lyumkis D, and Williamson JR (2016). Modular assembly of the bacterial large ribosomal subunit. Cell 167, 1610–1622. e1615. [PubMed: 27912064]

Elmlund D, and Elmlund H (2012). SIMPLE: Software for ab initio reconstruction of heterogeneous single-particles. Journal of structural biology 180, 420–427. [PubMed: 22902564]

Fernandez-Leiro R, and Scheres SH (2016). Unravelling biological macromolecules with cryo-electron microscopy. Nature 537, 339–346. [PubMed: 27629640]

Gao H, Valle M, Ehrenberg M, and Frank J (2004). Dynamics of EF-G interaction with the ribosome explored by classification of a heterogeneous cryo-EM dataset. Journal of structural biology 147, 283–290. [PubMed: 15450297]

Grant T, Rohou A, and Grigorieff N (2018). cisTEM, user-friendly software for single-particle image processing. elife 7, e35383. [PubMed: 29513216]

Gray R (1984). Vector quantization. IEEE Assp Magazine 1, 4–29.

Harnpicharnchai P, Jakovljevic J, Horsey E, Miles T, Roman J, Rout M, Meagher D, Imai B, Guo Y, and Brame CJ (2001). Composition and functional characterization of yeast 66S ribosome assembly intermediates. Molecular cell 8, 505–515. [PubMed: 11583614]

Jomaa A, Jain N, Davis JH, Williamson JR, Britton RA, and Ortega J (2014). Functional domains of the 50S subunit mature late in the assembly process. Nucleic Acids Research 42, 3419–3435. [PubMed: 24335279]

Jomaa A, Stewart G, Martín-Benito J, Zielke R, Campbell TL, Maddock JR, Brown ED, and Ortega J (2011). Understanding ribosome assembly: the structure of in vivo assembled immature 30S subunits revealed by cryo-electron microscopy. Rna 17, 697–709. [PubMed: 21303937]

Klaholz BP (2015). Structure sorting of multiple macromolecular states in heterogeneous cryo-EM samples by 3D multivariate statistical analysis. Open Journal of Statistics 5, 820.

Kühlbrandt W (2014). The resolution revolution. Science 343, 1443–1444. [PubMed: 24675944]

Lander GC, Stagg SM, Voss NR, Cheng A, Fellmann D, Pulokas J, Yoshioka C, Irving C, Mulder A, and Lau P-W (2009). Appion: an integrated, database-driven pipeline to facilitate EM image processing. Journal of structural biology 166, 95–102. [PubMed: 19263523]

Li N, Chen Y, Guo Q, Zhang Y, Yuan Y, Ma C, Deng H, Lei J, and Gao N (2013). Cryo-EM structures of the late-stage assembly intermediates of the bacterial 50S ribosomal subunit. Nucleic acids research 41, 7073–7083. [PubMed: 23700310]

Liao HY, Hashem Y, and Frank J (2015). Efficient estimation of three-dimensional covariance and its application in the analysis of heterogeneous samples in cryo-electron microscopy. Structure 23, 1129–1137. [PubMed: 25982529]

Loerke J, Giesebrecht J, and Spahn CM (2010). Multiparticle cryo-EM of ribosomes. In Methods in enzymology (Elsevier), pp. 161–177.

Ludtke SJ (2016). Single-particle refinement and variability analysis in EMAN2. 1. In Methods in enzymology (Elsevier), pp. 159–189.

Lyumkis D, Brilot AF, Theobald DL, and Grigorieff N (2013). Likelihood-based classification of cryo-EM images using FREALIGN. Journal of structural biology 183, 377–388. [PubMed: 23872434]

Nakane T, Kimanius D, Lindahl E, and Scheres SH (2018). Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. Elife 7, e36861. [PubMed: 29856314]

Nakane T, Kotecha A, Sente A, McMullan G, Masiulis S, Brown PM, Grigoras IT, Malinauskaite L, Malinauskas T, and Miehling J (2020). Single-particle cryo-EM at atomic resolution. Nature 587, 152–156. [PubMed: 33087931]

Ni X, Davis JH, Jain N, Razi A, Benlekbir S, McArthur AG, Rubinstein JL, Britton RA, Williamson JR, and Ortega J (2016). YphC and YsxC GTPases assist the maturation of the central protuberance, GTPase associated region and functional core of the 50S ribosomal subunit. Nucleic acids research 44, 8442–8455. [PubMed: 27484475]

Nikolay R, Hilal T, Qin B, Mielke T, Bürger J, Loerke J, Textoris-Taube K, Nierhaus KH, and Spahn CM (2018). Structural visualization of the formation and activation of the 50S ribosomal subunit during in vitro reconstitution. Molecular cell 70, 881–893. e883. [PubMed: 29883607]

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, and Ferrin TE (2004). UCSF Chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry 25, 1605–1612. [PubMed: 15264254]

Punjani A, and Fleet DJ (2021a). 3D Flexible Refinement: Structure and Motion of Flexible Proteins from Cryo-EM. bioRxiv.

Punjani A, and Fleet DJ (2021b). 3D Variability Analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. Journal of Structural Biology 213, 107702. [PubMed: 33582281]

Rabuck-Gibbons JN, Popova AM, Greene EM, Cervantes CF, Lyumkis D, and Williamson JR (2020). SrmB rescues trapped ribosome assembly intermediates. Journal of molecular biology 432, 978–990. [PubMed: 31877323]

Razi A, Guarné A, and Ortega J (2017). The cryo-EM structure of YjeQ bound to the 30S subunit suggests a fidelity checkpoint function for this protein in ribosome assembly. Proceedings of the National Academy of Sciences 114, E3396–E3403.

Roseman A (2004). FindEM—a fast, efficient program for automatic selection of particles from electron micrographs. Journal of structural biology 145, 91–99. [PubMed: 15065677]

Russo CJ, and Passmore LA (2014). Ultrastable gold substrates for electron cryomicroscopy. Science 346, 1377–1380. [PubMed: 25504723]

Sashital DG, Greeman CA, Lyumkis D, Potter CS, Carragher B, and Williamson JR (2014). A combined quantitative mass spectrometry and electron microscopy analysis of ribosomal 30S subunit assembly in E. coli. Elife 3, e04491. [PubMed: 25313868]

Scheres SH (2012). RELION: implementation of a Bayesian approach to cryo-EM structure determination. Journal of structural biology 180, 519–530. [PubMed: 23000701]

Scheres SH (2016). Processing of structurally heterogeneous cryo-EM data in RELION. In Methods in enzymology (Elsevier), pp. 125–157.

Scheres SH, Núñez-Ramírez R, Sorzano CO, Carazo JM, and Marabini R (2008). Image processing for electron microscopy single-particle analysis using XMIPP. Nature protocols 3, 977–990. [PubMed: 18536645]

Shajani Z, Sykes MT, and Williamson JR (2011). Assembly of bacterial ribosomes. Annual review of biochemistry 80, 501–526.

Sorzano C, Bilbao-Castro J, Shkolnisky Y, Alcorlo M, Melero R, Caffarena-Fernández G, Li M, Xu G, Marabini R, and Carazo J (2010). A clustering approach to multireference alignment of single-particle projections in electron microscopy. Journal of structural biology 171, 197–206. [PubMed: 20362059]

Spahn CM, and Penczek PA (2009). Exploring conformational modes of macromolecular assemblies by multiparticle cryo-EM. Current opinion in structural biology 19, 623–631. [PubMed: 19767196]

Stokes JM, Davis JH, Mangat CS, Williamson JR, and Brown ED (2014). Discovery of a small molecule that inhibits bacterial ribosome biogenesis. Elife 3, e03574. [PubMed: 25233066]

Sykes MT, Shajani Z, Sperling E, Beck AH, and Williamson JR (2010). Quantitative proteomic analysis of ribosome assembly and turnover in vivo. Journal of molecular biology 403, 331–345. [PubMed: 20709079]

Tan YZ, Aiyer S, Mietzsch M, Hull JA, McKenna R, Grieger J, Samulski RJ, Baker TS, Agbandje-McKenna M, and Lyumkis D (2018). Sub-2 Å Ewald curvature corrected structure of an AAV2 capsid variant. Nature communications 9, 1–11.

Tan YZ, Baldwin PR, Davis JH, Williamson JR, Potter CS, Carragher B, and Lyumkis D (2017a). Addressing preferred specimen orientation in single-particle cryo-EM through tilting. Nature methods 14, 793–796. [PubMed: 28671674]

Tan YZ, Baldwin PR, Davis JH, Williamson JR, Potter CS, Carragher B, and Lyumkis D (2017b). Addressing preferred specimen orientation in single-particle cryo-EM through tilting. Nature methods 14, 793. [PubMed: 28671674]

Uicker WC, Schaefer L, and Britton RA (2006). The essential GTPase RbgA (YlqF) is required for 50S ribosome assembly in Bacillus subtilis. Molecular microbiology 59, 528–540. [PubMed: 16390447]

Wang Q, Matsui T, Domitrovic T, Zheng Y, Doerschuk PC, and Johnson JE (2013). Dynamics in cryo EM reconstructions visualized with maximum-likelihood derived variance maps. Journal of structural biology 181, 195–206. [PubMed: 23246781]

White H, Ignatiou A, Clare D, and Orlova E (2017). Structural study of heterogeneous biological samples by cryoelectron microscopy and image processing. BioMed research international 2017.

Wolfram Research, I. (2020). Mathematica, Version 12.2 edn (Champaign, Illinois: Wolfram Research, Inc.).

Yip KM, Fischer N, Paknia E, Chari A, and Stark H (2020). Atomic-resolution protein structure determination by cryo-EM. Nature 587, 157–161. [PubMed: 33087927]

Zhang K (2016). Gctf: Real-time CTF determination and correction. Journal of structural biology 193, 1–12. [PubMed: 26592709]

Zhang K, Pintilie GD, Li S, Schmid MF, and Chiu W (2020). Resolving individual atoms of protein complex by cryo-electron microscopy. Cell research 30, 1136–1139. [PubMed: 33139928]

Zheng SQ, Palovcak E, Armache J-P, Verba KA, Cheng Y, and Agard DA (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nature methods 14, 331. [PubMed: 28250466]

Zhong ED, Bepler T, Berger B, and Davis JH (2021). CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. Nature Methods 18, 176–185. [PubMed: 33542510]

Highlights

- Method to quantitatively mine and characterize significant structural heterogeneity
- Cryo-EM structure determination of 41 assembly intermediates
- Identification of branchpoints in the ribosome assembly process

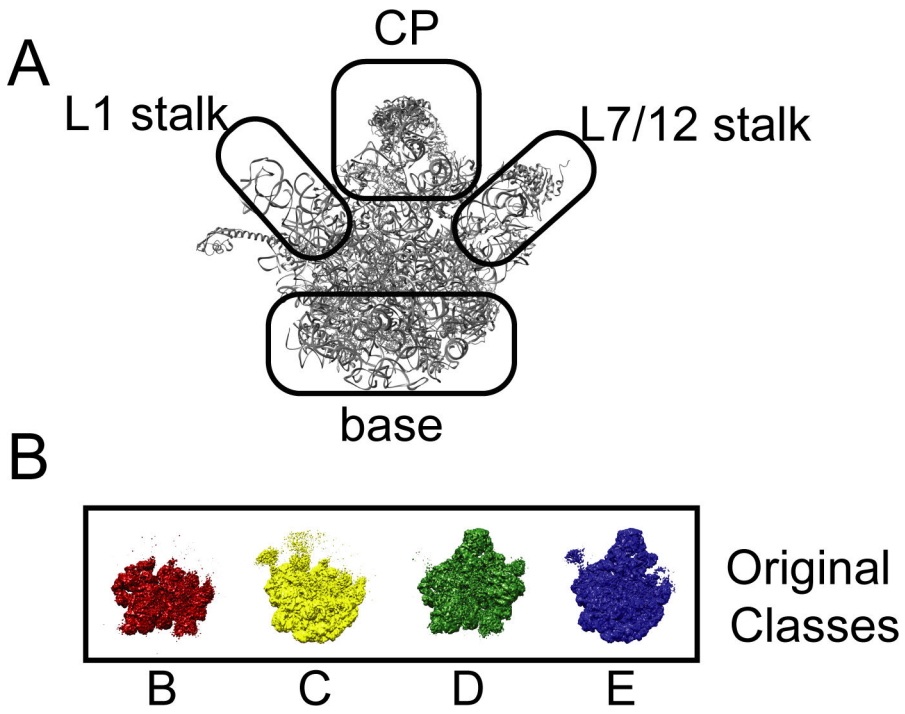


Figure 1. Description of the bacterial large ribosomal subunit and prior assembly intermediates identified by cryo-EM.

(A) PDB ID 4YBB labeled with prominent features identifiable on the large ribosomal subunit, including the central protuberance (CP), base, L1 stalk, and L7/12 stalks. These terms are used throughout the paper. (B) Primary classes identified within the original bL17-lim dataset (Davis *et al.*, 2016). From left to right: B class (red), C class (yellow), D class (green), and the E class (blue).

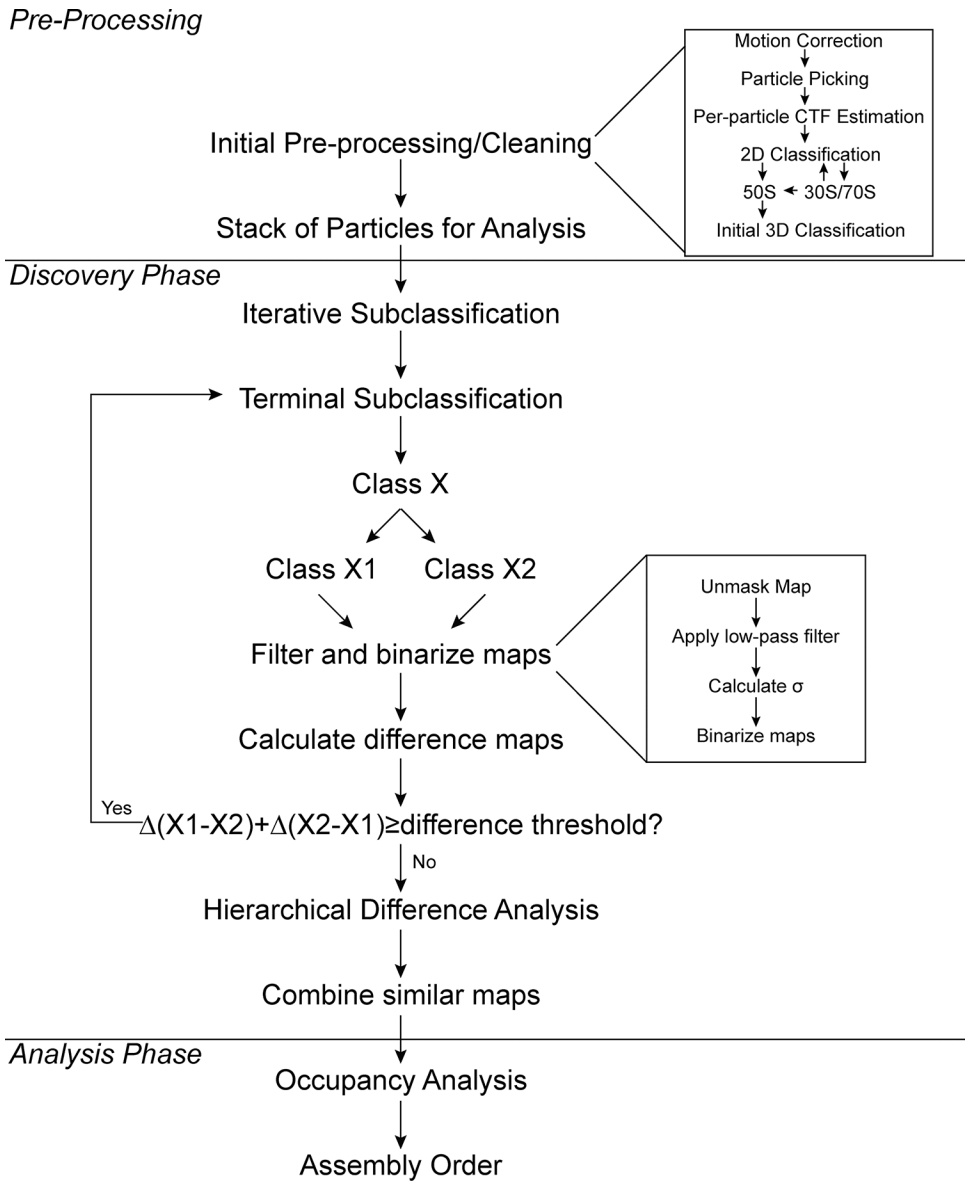


Figure 2. Workflow for cryo-EM heterogeneity analysis.

The workflow is divided into three major phases, including (i) a pre-processing phase, which is typical of all cryo-EM datasets, followed by (ii) a discovery and an (iii) analysis phase, which are described in detail here.

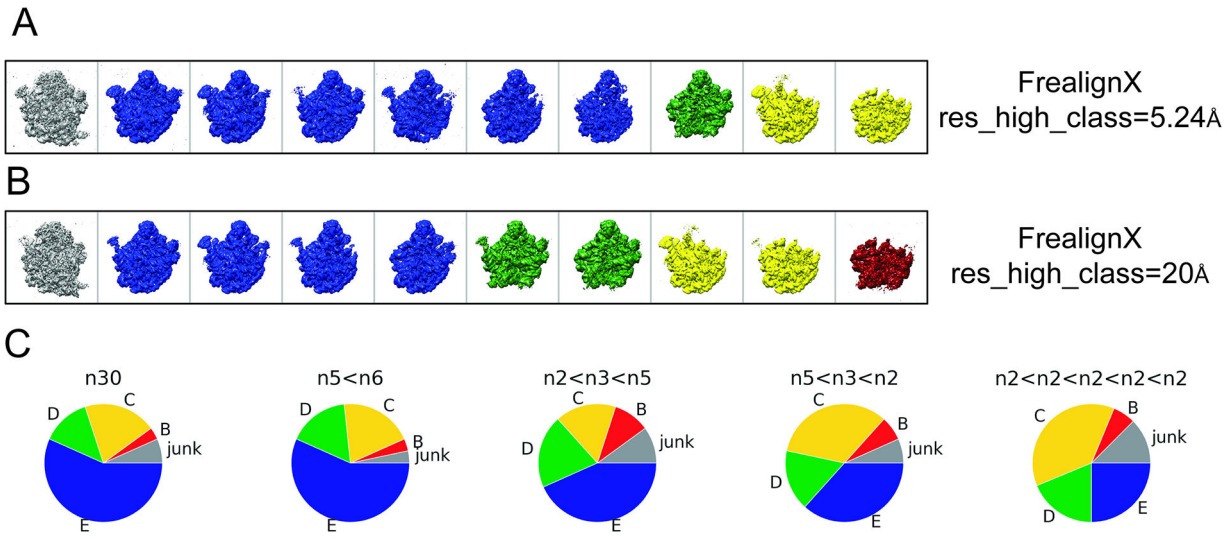


Figure 3. A divisive resolution-limited subclassification approach facilitates identifying rare structural variants.

Maps and sectors are color coded according to the major subclasses: B(red), C(yellow), D(green), E(blue). in all three panels. (A) FrealignX classification with *res_high_class* parameter set to Nyquist (5.24Å). (B) FrealignX classification with the *res_high_class* parameter set to 20Å. Using a lower resolution cutoff eliminates the “hiding” of the B-class, and generally leads to identification of a broader range of classes. (C) Comparison of the number of subclasses resulting from five different classification schemes.

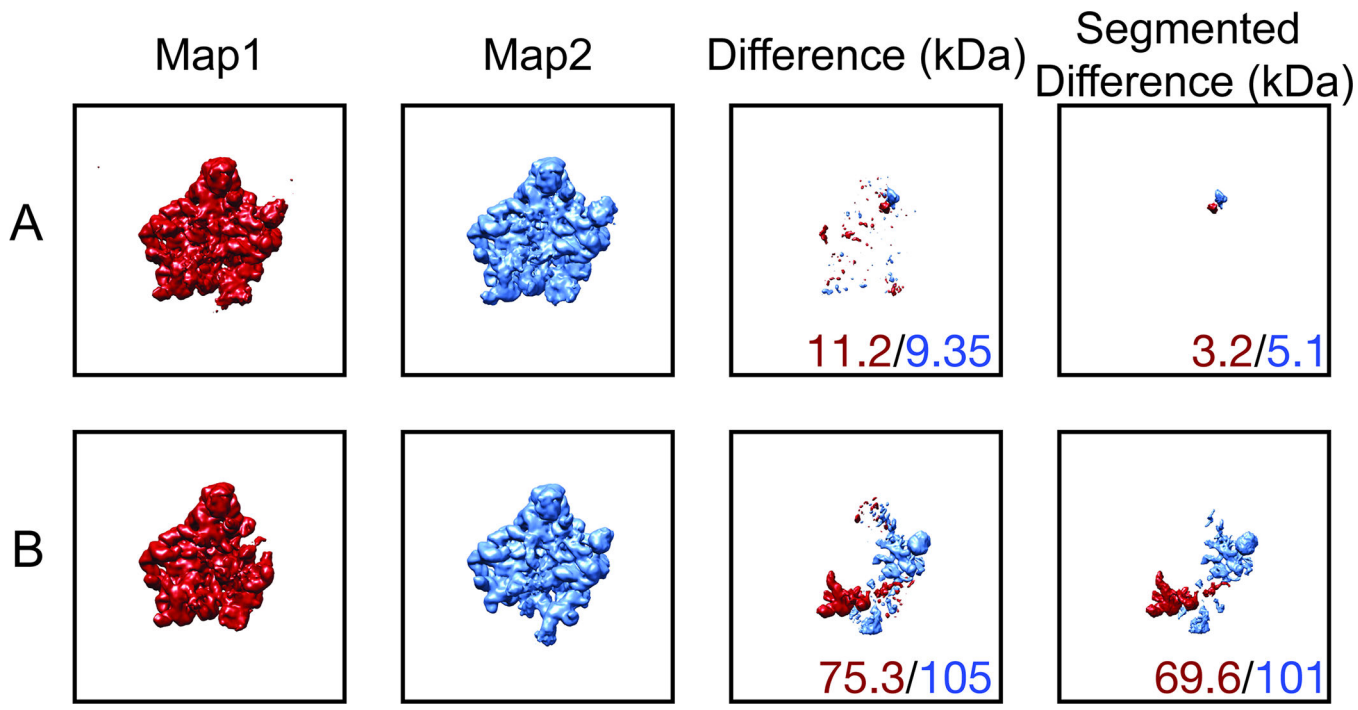


Figure 4. Segmented difference analysis helps to define molecular weight differences between map pairs.

Difference analyses are displayed for two examples. Numbers indicate positive (Map1-Map2) and negative (Map2-Map1) molecular weight differences. (A) Example where two maps would have been considered different before segmentation but are not different after segmentation. (B) Example where two maps are different both before and after segmentation.

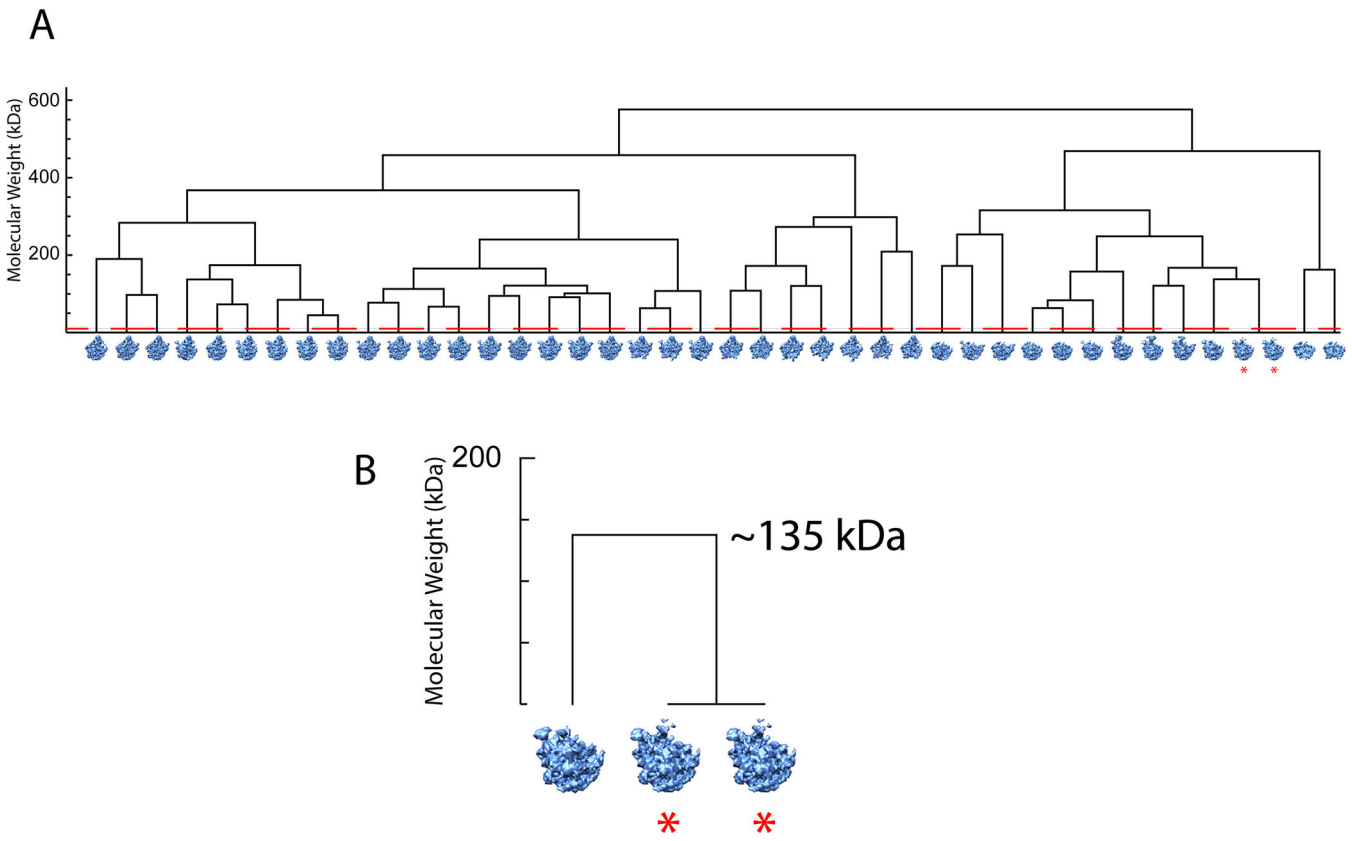


Figure 5. Hierarchical clustering is used to combine similar maps under a given threshold.

(A) Hierarchical clustering analysis of the maps that result after the terminal subclassification (total n=42). The red dashed line indicates the 10kDa MWCO used to combine similar maps at this step, and the red stars indicate maps that are combined after this analysis. After combining similar maps, the final number of classes is thus forty-one. (B) Close-up example of two combined maps in (A). The leftmost structure is distinct from the other two by ~135 kDa and needs to be treated independently, whereas the two rightmost structures can be combined into a single class.

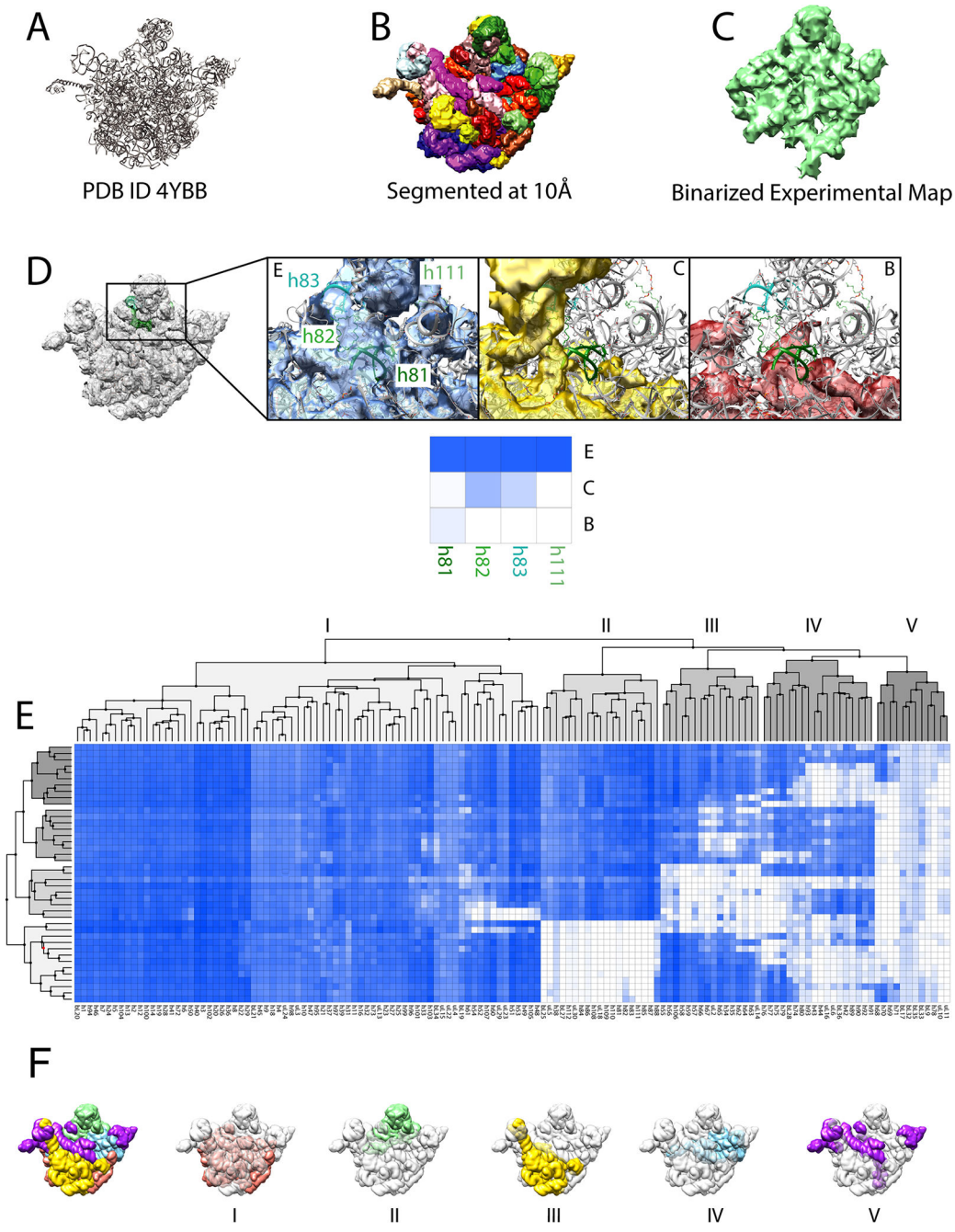


Figure 6. Results of occupancy analysis on the full dataset mapped onto the ribosomal scaffold.

(A) Reference crystal structure 4ybb. (B) Binarized maps of the individual proteins and rRNA helices created by segmenting the crystal structure into 139 individual helices and proteins, and calculating theoretical 10Å maps in Chimera. (C) An example of a binarized experimental map arising from sub-classification. The pixels from the binarized experimental map that are located in the theoretical binarized map are counted and normalized to an occupancy value of 0–1. (D) Example of an E, C, and B class with rRNA helix occupancies that are present, partially present, or fully missing. (E) Occupancy analysis plot, where the individual proteins and helices are shown on the x-axis, the

experimental maps are on the y-axis, and the normalized occupancy values are shown from white (0) to dark blue (1). Hierarchical clustering of both structure elements and experimental maps was performed on the occupancy matrix using a squared Euclidean distance metric and Ward's linkage. (F) Occupancy analysis blocks mapped back to the reference structure 4YBB, and the numbering system is the same as in (E).

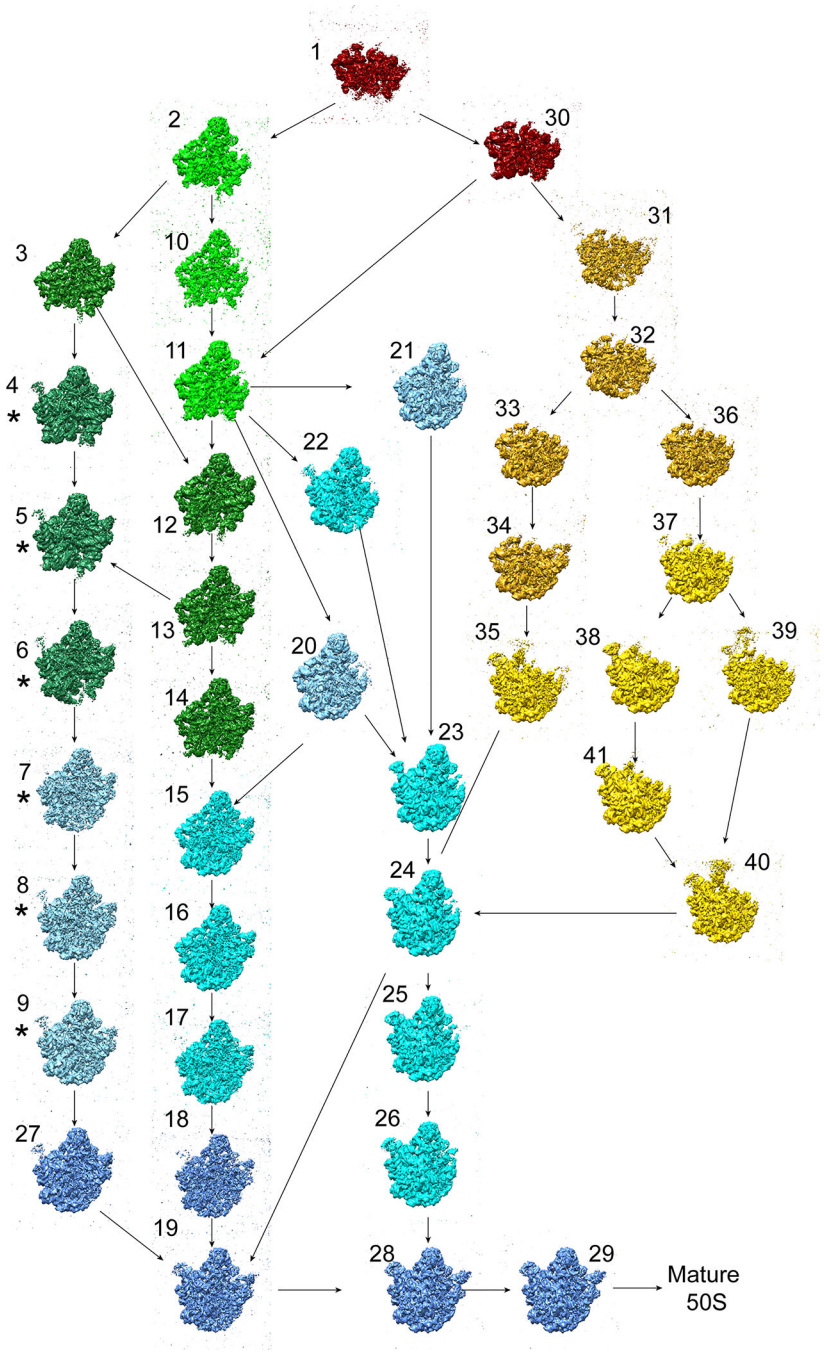


Figure 7. Revised ribosome assembly map from bL17-lim.

The assembly pathway is drawn by analyzing the folding and unfolding molecular weight matrices and revised by hand.

Table 1.

Description of limits and thresholds

	Threshold/limit	Description	Value used in this paper
Limits	r-limit	The minimum resolution necessary for a map	10Å
	v-limit	volume limit: molecular weight difference limit for terminal subdivision	1.5 kDa
Thresholds	Low pass filter threshold	used to normalize resolution between maps and to focus on lower-resolution differences between maps	10Å
	Binarization threshold	threshold at which maps are binarized; pixel values below this limit are set to 0, values above this limit are set to 1	3σmap
	Segmentation threshold	defines the volume of dust to be removed from difference maps	1.5 kDa
	Difference threshold	defines the lower limit for acceptable differences between maps	10 kDa

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
JD321: E. coli NCM3722 [rplQ::cat, pJD075]	Davis, Tan et al. 2016	N/A
Chemicals, Peptides, and Recombinant Proteins		
N-(β-Ketocaproyl)-L-homoserine lactone	Santa Cruz Biotech.	Cat#: sc-205396
Deposited Data		
Cryo-EM density map: Class 20	This paper	EMD-24491
Cryo-EM density map: Class 16	This paper	EMD-24492
Cryo-EM density map: Class 15	This paper	EMD-24499
Cryo-EM density map: Class 17	This paper	EMD-24508
Cryo-EM density map: Class 27	This paper	EMD-24509
Cryo-EM density map: Class 9	This paper	EMD-24510
Cryo-EM density map: Class 8	This paper	EMD-24515
Cryo-EM density map: Class 21	This paper	EMD-24517
Cryo-EM density map: Class 23	This paper	EMD-24520
Cryo-EM density map: Class 22	This paper	EMD-24521
Cryo-EM density map: Class 40	This paper	EMD-24527
Cryo-EM density map: Class 35	This paper	EMD-24538
Cryo-EM density map: Class 39	This paper	EMD-24529
Cryo-EM density map: Class 37	This paper	EMD-24543
Cryo-EM density map: Class 38	This paper	EMD-24546
Cryo-EM density map: Class 41	This paper	EMD-24550
Cryo-EM density map: Class 41	This paper	EMD-24555
Cryo-EM density map: Class 19	This paper	EMD-24558
Cryo-EM density map: Class 18	This paper	EMD-24559
Cryo-EM density map: Class 28	This paper	EMD-24561
Cryo-EM density map: Class 29	This paper	EMD-24562
Cryo-EM density map: Class 25	This paper	EMD-24563
Cryo-EM density map: Class 26	This paper	EMD-24564
Cryo-EM density map: Class 24	This paper	EMD-24565
Cryo-EM density map: Class 2	This paper	EMD-24567
Cryo-EM density map: Class 3	This paper	EMD-24568
Cryo-EM density map: Class 12	This paper	EMD-24571
Cryo-EM density map: Class 14	This paper	EMD-24573
Cryo-EM density map: Class 13	This paper	EMD-24574
Cryo-EM density map: Class 11	This paper	EMD-24626
Cryo-EM density map: Class 10	This paper	EMD-24627
Cryo-EM density map: Class 4	This paper	EMD-24631
Cryo-EM density map: Class 5	This paper	EMD-24633

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Cryo-EM density map: Class 6	This paper	EMD-24634
Cryo-EM density map: Class 7	This paper	EMD-24658
Cryo-EM density map: Class 33	This paper	EMD-24659
Cryo-EM density map: Class 36	This paper	EMD-24660
Cryo-EM density map: Class 34	This paper	EMD-24661
Cryo-EM density map: Class 32	This paper	EMD-24662
Cryo-EM density map: Class 31	This paper	EMD-24669
Cryo-EM density map: Class 1	This paper	EMD-24671
Cryo-EM density map: Class 30	This paper	EMD-24673
Raw movies and final particle stack for a dataset of bL17-limited E. coli ribosome assembly intermediates	This paper	EMPIAR-10841
Experimental Models: Organisms/Strains		
JD321: E. coli NCM3722 [rplQ::cat, pJD075]	Davis, Tan et al. 2016	N/A
Software and Algorithms		
Leginon	National Resource for Automated Molecular Microscopy	https://emg.nysbc.org/redmine/projects/leginon/wiki/Leginon_Homepage
Appion	National Resource for Automated Molecular Microscopy	https://emg.nysbc.org/redmine/projects/appion/wiki/Appion_Home
CTFFind4	The Grigorieff Lab	https://grigoriefflab.umassmed.edu/ctffind4
GCTF	Zhang, 2016	http://www.mrc-lmb.cam.ac.uk/kzhang/Gctf/
FindEM	Roseman, 2004	http://emg.nysbc.org/redmine/projects/software/wiki/FindEM
Relion 2.1	Scheres, 2012	https://www3.mrc-lmb.cam.ac.uk/relion/index.php?title=Main_Page
FrealignX	Grant, 2018,Lyumkis, 2013	https://cistem.org/
EMAN	Ludtke et al., 1999	http://blake.bcm.edu/emanwiki/EMAN2
Chimera	UCSF Resource for Biocomputing, Visualization, and Informatics	https://www.cgl.ucsf.edu/chimera/
Hierarchical Clustering	This paper	Method S1
Occupancy Analysis	This paper	Method S1
Pathway Maker	This paper	Method S1
3DFSC	Tan et al., 2017a	https://3dfsc.salk.edu/
SCF	Baldwin and Lyumkis, 2020, 2021	https://github.com/LyumkisLab/SamplingGui