Supplemental Information

Classification of Single Particles from Human Cell Extract Reveals Distinct Structures

Eric J. Verbeke, Anna L. Mallam, Kevin Drew, Edward M. Marcotte, David W. Taylor
## Key Resources Table

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deposited Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20S Proteasome Core</td>
<td>This paper</td>
<td>XXXX</td>
</tr>
<tr>
<td>26S single-capped Proteasome</td>
<td>This paper</td>
<td>YYYY</td>
</tr>
<tr>
<td><strong>Experimental Models: Cell Lines</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEK293T</td>
<td>ATCC</td>
<td>CRL3216</td>
</tr>
<tr>
<td><strong>Software and Algorithms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appion</td>
<td>(Lander et al., 2009)</td>
<td><a href="http://nramm.nysbc.org/software/">http://nramm.nysbc.org/software/</a></td>
</tr>
<tr>
<td>FindEM</td>
<td>(Roseman, 2004)</td>
<td></td>
</tr>
<tr>
<td>RELION</td>
<td>(Scheres, 2012)</td>
<td><a href="http://www2.mrc-lmb.cam.ac.uk/reli">http://www2.mrc-lmb.cam.ac.uk/reli</a> on/index.php/Main_Page</td>
</tr>
<tr>
<td>CTFFIND4</td>
<td>(Rohou and Grigorieff, 2015)</td>
<td><a href="http://grigoriefflab.janelia.org/ctf">http://grigoriefflab.janelia.org/ctf</a></td>
</tr>
<tr>
<td>cryoSPARC</td>
<td>(Punjani et al., 2017)</td>
<td><a href="https://cryosparc.com/">https://cryosparc.com/</a></td>
</tr>
<tr>
<td>Chimera</td>
<td>(Pettersen et al., 2004)</td>
<td><a href="https://www.cgl.ucsf.edu/chimera/">https://www.cgl.ucsf.edu/chimera/</a></td>
</tr>
<tr>
<td>SearchGUI</td>
<td>(Vaudel et al., 2011)</td>
<td><a href="http://compomics.github.io/projects/searchgui.html">http://compomics.github.io/projects/searchgui.html</a></td>
</tr>
<tr>
<td>Cytoscape</td>
<td>(Shannon et al., 2003)</td>
<td><a href="http://www.cytoscape.org/">http://www.cytoscape.org/</a></td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formvar/Carbon 400 mesh, Copper approx. grid hole size: 42µm</td>
<td>Ted Pella, inc.</td>
<td>01754-F</td>
</tr>
</tbody>
</table>
Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, David W. Taylor dtaylor@utexas.edu (D.W.T)

Experimental Model and Subject Details

HEK293T cells (ATCC CRL3216) cultured at 37°C in DMEM (Gibco) supplemented with 10% (v/v) FBS (Life Technologies) were continually split over 7 days to give four 10-cm dishes of adherent cells.

Methods Details

Cell Culture and Extract Preparation

HEK293T cells were harvested at 80-100% confluence without trypsin by washing in ice cold phosphate buffered saline (PBS) pH 7.2 (0.75 mL; Gibco) and placed on ice. Cells (approximately 10 mg) were lysed on ice (5 min) by resuspension in Pierce IP Lysis Buffer (0.8 mL; 25 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% NP-40 and 5% glycerol; Thermo Fisher) containing 1x protease inhibitor cocktail III (Calbiochem). The resulting lysate was clarified (17,000 g, 10 min, 4°C) and filtered (Ultrafree-MC filter unit (Millipore); 12,000 g, 2 min, 4°C).

Biochemical Fractionation Using Native Size-Exclusion Chromatography

Size-exclusion chromatography (SEC) was performed at 4°C on an AKTA FPLC (GE Healthcare). Approximately 6 mg of soluble protein was applied to a Superdex 200 10/300 GL analytical gel filtration column (GE Healthcare) equilibrated in PBS, pH 7.2 at a flow rate of 0.5 mL min⁻¹. Fractions were collected every 0.5 mL. The elution volume of molecular weight standards (Thyroglobulin, 670,000 Da; γ-globulin, 158,000 Da; Ovalbumin, 44,000 Da;
Myoglobin, 17,000 Da; Vitamin B$_{12}$, 1,350 Da; Biorad) were additionally measured to calibrate the column (Fig. 2A). Fraction 4 (concentration ~1 mg/mL) was deemed most likely to contain a high number of large complexes, as determined by $A_{280}$, and was subjected to further proteomic and structural analysis.

Mass Spectrometry

50 µL of Fraction 4 (Fig. 2A) was denatured and reduced in 50 % 2,2,2-trifluoroethanol (TFE) and 5 mM tris(2-carboxyethyl)phosphine (TCEP) at 55 °C for 45 minutes, followed by alkylation in the dark with iodoacetamide (55 mM, 30 min, RT). Samples were diluted to 5 % TFE in 50 mM Tris-HCl, pH 8.0, 2 mM CaCl$_2$, and digested with trypsin (1:50; proteomic grade; 5 hours: 37 °C). Digestion was quenched (1 % formic acid), and the sample volume reduced to ~100 uL by speed vacuum centrifugation. The sample was washed on a HyperSep C18 SpinTip (Thermo Fisher), eluted, reduced to near dryness by speed vacuum centrifugation, and resuspended in 5 % acetonitrile/ 0.1 % formic acid for analysis by LC-MS/MS. Peptides were separated on a 75 µM x 25 cm Acclaim PepMap100 C-18 column (Thermo) using a 3–45 % acetonitrile gradient over 60 min and analyzed on line by nanoelectrospray-ionization tandem mass spectrometry on an Orbitrap Fusion (Thermo Scientific). Data-dependent acquisition was activated, with parent ion (MS1) scans collected at high-resolution (120,000). Ions with charge 1 were selected for collision-induced dissociation fragmentation spectrum acquisition (MS2) in the ion trap, using a Top Speed acquisition time of 3-s. Dynamic exclusion was activated, with a 60-s exclusion time for ions selected more than once.
Proteomic and Bioinformatics Analyses

The mass spectrometry data was processed using searchGUI and PeptideShaker (Vaudel et al., 2011, 2015). Data was searched against a target-decoy human database downloaded from Universal Protein Resources Database (UniProtKB/Swiss-Prot comprising human proteins supplemented with common contaminants). Fixed modifications of carboxyamidomethylated cysteine and variable modifications of oxidized methionine and acetylation of protein N-terminus were permitted to allow for detection of modified peptides. Peptide spectral matches, peptides and proteins were considered positively identified if detected within a 1% false discovery rate cut off (based on empirical target-decoy database search results). This screening procedure resulted in 1,907 distinct human proteins. To facilitate mapping to a protein ID, we used UniProtKB accession numbers as a common identifier and the UniProt ID mapping tool to interconvert different gene and protein identifiers.

Relative abundance for each complex was determined using label-free quantification based on peptide spectral matches of each subunit. Protein length was used for normalizing the number of peptide spectral matches observed for each protein using the Normalized Spectral Abundance Factor (NSAF) (Zybailov et al., 2006) as calculated by PeptideShaker (Vaudel et al., 2015). The NSAF values for all proteins in a complex were then averaged to estimate the relative abundance of each complex. Proteins expected to participate in a complex as predicted by hu.MAP or CORUM, which were not identified by MS, were assigned a NSAF value of zero.

The hierarchical network of protein complexes in Figure S1 was created by determining the percent of shared subunits between all complexes. For a predicted protein complex A with subunits \( \{a_1, a_2, \ldots, a_n\} \) and B with subunits \( \{b_1, b_2, \ldots, b_m\} \), the similarity score (S) of A to B was calculated by finding the intersection of A and B divided by the size of set A as follows (equation 1).
\[ S = \frac{|A \cap B|}{|A|} \]  

(1)

If the similarity score between complexes was 90% or greater, it was considered a related complex. The network of related complexes was then visualized using Cytoscape with edges being attributed to the similarity score (Shannon et al., 2003).

**Negative Stain Electron Microscopy Sample Preparation**

4 µL of fractionated human cell lysate was applied to a glow-discharged 400-mesh continuous carbon grid. After a 1 min adsorption, the sample was negatively stained with five consecutive droplets of 2% (w/v) uranyl acetate solution, blotted to remove residual stain, and air-dried in a fume hood.

**Electron Microscopy**

Data was acquired using a JEOL 2010F transmission electron microscope operated at 200 keV with a nominal magnification of x60,000 (3.6 Å at the specimen level). Each image was acquired using a 1 s exposure time with a total dose of ~30-35 e Å⁻² and a defocus between −1 and −2 µm. A total of 1,250 micrographs were manually recorded on a Gatan OneView.

**3D Reconstruction and Analysis**

Two independent particle stacks were generated from the same 1,250 micrographs using either template or manual particle picking. The contrast transfer function (CTF) of each micrograph was estimated using CTFFIND4 (Rohou and Grigorieff, 2015). FindEM (Roseman, 2004) was used for template-based particle picking using a reference-free 2D class averages of our negatively stained 60S Ribosome from *Saccharomyces cerevisiae* (a gift from A. Johnson). ~97,000 and ~37,000 particles were selected by template picking and manually selecting particle images,
respectively. All image pre-processing was done in Appion (Lander et al., 2009). After removing junk particles, 31,731 particles were left from template picking and 35,381 particles from manual picking, respectively. The majority of junk classes from template picking can be attributed to the picking of particles within aggregates and two particles as one. Particle box size was set to 576 Å x 576 Å.

Reference-free 2D class averages were generated with 300 classes for both data sets using RELION (Scheres, 2012). Next, 3D classification was performed using RELION to create 30 classes of both data sets. The structure of DNA-dependent protein kinase catalytic subunit was chosen as an initial model using a negative stain structure low-pass filtered to 60 Å as a starting model (Sibanda et al., 2017) (Figure S3). Autorefine in RELION was used to refine the putative single-capped 26S proteasome structure from the manually-picked data set using the corresponding class reconstruction low-pass filtered to 60 Å as a starting model. The manual picked data set was used for subsequent analysis using cryoSPARC (Punjani et al., 2017). cryoSPARC was used to ab initio reconstruct 5, 10 and 15 3D models. The class corresponding to the 20S proteasome from the 10-model run, containing 3,150 particles, was then subjected to homogeneous refinement using cryoSPARC.

Random particle models were generated using RELION with the template picked particle data set. Each model was reconstructed using the mean number of particles from the 30 models in Figure S3, ~1000 particles. Particles were sampled without replacement. Model error (E) was calculated for each RELION generated model by taking the harmonic mean of their respective rotational accuracy (R) and translational accuracy (T) as determined using RELION (equation 2). Model error values were normalized between 1 and 2.

\[ E = \frac{2}{\left( \frac{1}{R} \right) + \left( \frac{1}{T} \right)} \] (2)
We then performed a two-sided Kolmogorov-Smirnov test between the distribution of model error from our models and the distribution of model error from the random particle models.

Several high-abundance complexes from our MS data with identifiable, previously solved structures were used to compare with our top 3 models generated using RELION. All models were first low-pass filtered to 30 Å before being aligned using Chimera’s Fit in Map function (Pettersen et al., 2004). The cross-correlation score was then calculated by using the model with a larger volume as the region of computation, essentially sliding the larger complex across the smaller complex.

**Data and Software Availability**

The EM reconstruction for both the 20S and 26S (presented in Figure 4B) were deposited in the EM Data Bank (EMDB) under accession codes EMD-XXXX, EMD-YYYY, respectively.

**Supplemental Information**

**Figure S1. Hierarchical network of related protein complexes**

Subset of the hierarchical network showing related complexes identified by MS in our sample. Each node represents a protein complex and is identified by name or by cluster number (Table S1). The size of each node depicts the molecular weight of the complete complex. Node fill color gradient represents the relative abundance of the complex determined by label-free quantification (see Methods). Node border color gradient represents the percent of subunits in a complex identified by MS. Arrows between nodes indicate at least 90% similarity in subunit composition between source and target node.

**Figure S2. Classification of particles using RELION**

(A) Reference-free 2D class averages of 31,731 template picked particles generated using
RELION. The size of each box is 576 Å x 576 Å. The 2D class averages are sorted by the number of particles belonging to each class. Highlighted boxes show examples of similar 2D classes from both particle selection methods.

(B) Reference-free 2D class averages of 35,381 manual picked particles generated using RELION. The size of each box is 518 Å x 518 Å. The 2D class averages are sorted by the number of particles belonging to each class.

(C) 30 classes generated using RELION from the template picked data set with the reference model shown in grey. Models are binned by their rotational-translational error score (a measure of the internal consistency of the model) as shown in Figure 3C.

**Figure S3. Cross-correlation comparison of top 3 RELION models to complexes identified by MS.**


**Figure S4. 3D models using cryoSPARC with k = 5,10,15 and related Fourier shell correlations curves**

(A) Reconstructed 3D models from 35,381 manually picked particles when sorted into 5, 10 and 15 classes by cryoSPARC. The 20S proteasome core is highlighted in gold.

(B) Comparison of 20S proteasome core models from 5, 10 and 15 classes.
(C) FSC curves for the single-capped 26S proteasome (red) and 20S core proteasome (gold) shown in Figure 4B. Nominal resolutions were estimated to be 31 Å and 20.4 Å using the 0.143 gold-standard FSC criterion for the single-capped 26S and 20S core proteasome, respectively.

Supplementary Tables and Figures:

Figure S1.
Figure S2.

A  Template Picked 2D Class Averages
B  Manual Picked 2D Class Averages
C  Rotational-Translational Error Score

Most Accurate  Least Accurate

Reference Model

Rotational-Translational Error Score

$[-2\sigma, -\sigma]$  $[-\sigma, \sigma]$  $[\sigma, +\sigma]$  $[+\sigma, +2\sigma]$  $[+2\sigma]$
Figure S3.
Figure S4.

A

cryoSPARC Models

K = 5

K = 10

K = 15

B

150 Å

200 Å

K = 5

K = 10

K = 15

C

1

0.9

0.8

0.7

0.6

0.5

0.4

0.3

0.2

0.1

0

100 50 33.3 25 20 16.7 14.3

Resolution (Å)

FSC at 0.143

31 Å

20.4 Å

cryoSPARC 20S Proteasome Core (K = 10)

RELION 26S Proteasome Single-Capped