doi:10.1017/S1431927620014221
Microsc. Microanal. 26 (Suppl 2), 2020

Data Management Pipeline at the National Cryo-EM Facility

Matt Hutchison[1], Thomas Edwards[1], Joseph Finney[1], Helen Wang[1], Adam Wier[1], Ulrich Baxa[1] and Sriram Subramaniam[2]

[1]Frederick National Laboratory, Frederick, Maryland, United States, [2]University of British Columbia, Frederick, Maryland, United States

The resolution revolution in cryo-EM over the last several years has been enabled in part by a series of technological improvements that have introduced new challenges to the ways that cryo-EM facilities collect and process data. These challenges include the rate of data acquisition, the quality control of the data, and the evaluation of the effective quality of the sample as frozen on the EM grid.

The National Cryo-EM Facility (NCEF) was established by the National Cancer Institute (NCI) as a resource for researchers to collect cryo-EM data on targets related to cancer. The NCEF currently operates two Titan Krios microscopes, each equipped with a Falcon 3EC detector and a K3 direct electron detector at the end of a Gatan imaging filter. Imaging is performed using either the Latitude S or SerialEM [1] software packages.

A typical imaging session involves approximately 40 hours of data collection at a rate of 150 to 225 movies per hour in either counting or super-resolution mode. The resulting raw data rate is up to 2 Gbps, with full datasets ranging between 5 and 20 TB uncompressed and 1 to 6 TB when LZW-compressed. We observed a fivefold increase in the last year in data volume (Fig. 1), driven by the introduction of multi-shot beam shift imaging templates and the upgrade from a K2 to K3 detector.

To support continuous operation, the collected data must be transferred off of the detector's data server and compressed for distribution to the researcher. Real-time quality control consisting of CTF estimation (CTFFind4 [2]) and motion correction (MotionCor2 [3]) is also performed on each image. The applications are run with QC-optimized settings, using fast CTF search and binning super-resolution data by a factor of 2. A workstation collocated with the detector equipment handles this and transfers the data into a larger storage array. To handle the real-time computational requirement, our workstation is equipped with dual 10-core CPUs, a single NVIDIA 1080 Ti GPU, an NVME flash disk used as scratch space, 128GB of RAM, and 10 Gbps networking connecting it to both the detector equipment and the second-tier storage equipment. With this configuration, we are able to process up to 200 super-resolution movies per hour, rate-limited by the available CPU cores. When we upgrade the CPU configuration in this machine, the next-nearest rate limiting step will be motion correction, where we are limited to approximately 500 movies per hour on a single GPU.

To make the CTF and motion estimates available to the researcher and the facility microscopists during data collection, we generate a webpage using Scipion [4] in its stream processing configuration. This output includes a preview of each image, its power spectrum, a plot of the motion in the movie, and the estimated astigmatism, defocus, and resolution values. This is useful for making adjustments during imaging and monitoring for performance issues.

To provide feedback on the effective quality of the sample, we run a fully automated set of processes to generate a rough set of 2D classes after the first 500 images of the dataset are acquired. Particle picking is done with crYOLO [5] using the pre-trained general model. 2D classification is performed in two rounds with RELION [6], using the 10 most populated classes from the first round as the input to the second

https://doi.org/10.1017/S1431927620014221 Published online by Cambridge University Press

round. These steps are done in the Frederick National Laboratory cluster environment attached to our larger storage array and run in about 2 hours [7].

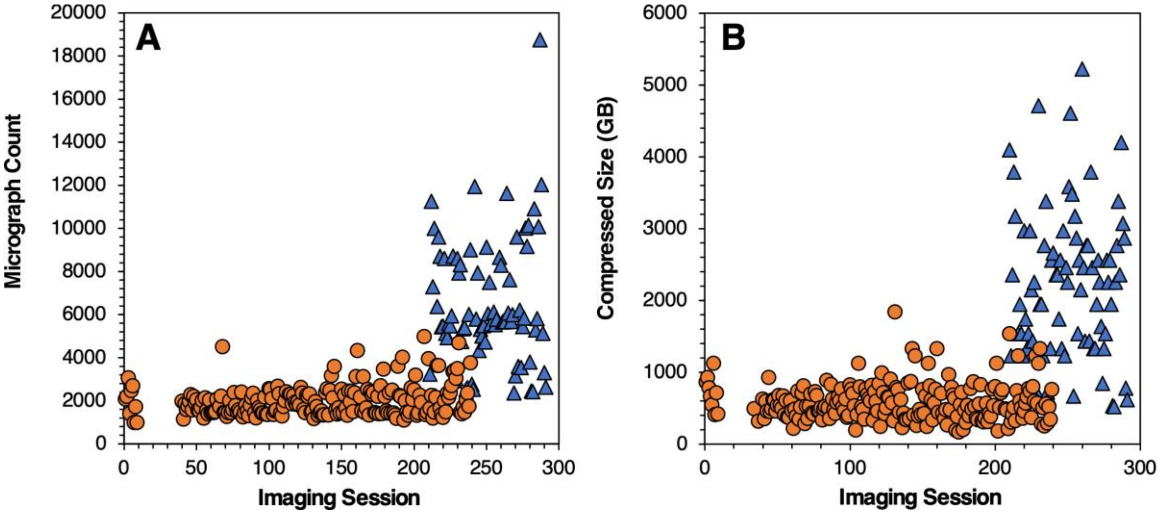

Figure 1. Data collection results from earlier data collection runs that utilized the K2 summit camera (orange circles) and later data collection runs that utilized the K3 camera (blue triangles). (A) shows the total number of movies collected in each imaging session. (B) shows the total compressed size for each imaging session.

References
[1] D Mastronarde, J Struct Biol. **152** (2005), p. 36-51.
[2] A Rohou and N Grigorieff, J Struct Biol. **192** (2015), p. 216-221.
[3] S Zheng et al., Nat Methods. **14** (2017), p. 331-332.
[4] JM de la Rosa-Trevin et al., J Struct Biol. **195** (2016), p. 93-99.
[5] T Wagner et al., Commun Biol. **2** (2019)
[6] S Scheres, J Struct Biol. **180** (2012), p. 519-530.
[7] This project has been funded with Federal funds from the Frederick National Laboratory for Cancer Research, National Institutes of Health, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

https://doi.org/10.1017/S1431927620014221 Published online by Cambridge University Press