Tools for near-atomic resolution in single-particle cryogenic electron microscopy

TOOLS FOR NEAR-ATOMIC RESOLUTION IN SINGLE-PARTICLE CRYOGENIC ELECTRON MICROSCOPY

Pavel Valerievich Afanasyev

The research in this thesis was performed in the group of prof. dr. Peter J. Peters (The Maastricht Multimodal Molecular Imaging Institute, Maastricht University, Maastricht, The Netherlands and The Netherlands Cancer Institute, Amsterdam, the Netherlands) and in the group of prof. dr. ir. Marin van Heel (Institute of Biology Leiden, NeCEN, Leiden University, Leiden, The Netherlands).

Printed by: Off Page, Amsterdam, The Netherlands
ISBN: 978-94-6182-670-1

TOOLS FOR NEAR-ATOMIC RESOLUTION IN SINGLE-PARTICLE CRYOGENIC ELECTRON MICROSCOPY

DISSERTATION

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus,
Prof. dr. L.L.G. Soete
in accordance with the decision of the Board of Deans,
to be defended in public
on Thursday 21 April 2016, at 14:00

by

Pavel Valerievich Afanasyev

born on 23 November 1988
in Leningrad, USSR

SUPERVISORS

Prof. dr. P. J. Peters

Prof. dr. ir. M. van Heel (Leiden University; Imperial College London, UK)

ASSESSMENT COMMITTEE

Prof. dr. F.F.C.S. Ramaekers (Chairman)

Prof. dr. R.M.A. Heeren

Prof. dr. K. Jalink (The Netherlands Cancer Institute)

Prof. dr. ir. A.J. Koster (Leiden University Medical Center)

Prof. dr. J. Neefjes (The Netherlands Cancer Institute)

Dr. N.S. Pannu (Leiden University)

Prof. dr. A. Sonnenberg (The Netherlands Cancer Institute)

Dr. N. van der Wel (Amsterdam Medical Center)

The research described in this thesis was accomplished with financial support by NanoNextNL, a micro and nanotechnology innovation consortium of the Government of the Netherlands and 130 partners from academia and industry (more information on www.nanonextnl.nl).
The printing of this thesis was partially financed by FEI Eindhoven Company and a grant of the *Stichting tot Bevordering van de Electronenmicroscopie in Nederland*.

Моим родителям

TABLE OF CONTENTS

CHAPTER 1

THE RAPIDLY EVOLVING FIELD OF SINGLE-PARTICLE CRYO-EM

Single-particle cryo-EM technique in structural biology

Almost a century ago, the advent of molecular biology opened up a new era of drug design by introducing the first drugs and vaccines based on an understanding of the underlying biochemistry. The discovery of novel drugs and vaccines can be boosted by a good understanding of the fundamental aspects of cell functioning. Modern structural biology studies also play an increasingly important role in rational drug design (Blundell et al. 2006; Scannell et al. 2012). In 2009 the Nobel Prize in Chemistry was awarded to Venkatraman Ramakrishnan, Thomas Steitz and Ada Yonath for revealing the atomic-resolution structure of the bacterial ribosome (Nobel Media 2014b). The availability of detailed structural information on the ribosome has, for example, facilitated the development of new improved antibiotics (Rodnina and Wintermeyer 2010). Structural information on the pathogens may suggest: (i) the use novel drugs; (ii) how to improve on existing drugs and vaccines; or (iii) how to suppress side-effects of the medications (Congreve et al. 2005).

More than thirty years of massive efforts have been invested in searching for vaccines preventing HIV transmission (Barre-Sinoussi et al. 2013), but this goal remains elusive. One of the most promising strategies is to design a vaccine, stimulating the production of broadly neutralizing antibodies against the virus (Earl et al. 2013). Such a vaccine could be based on the structure of the Env glycoprotein trimer. Solving the structure of this envelope protein could thus significantly contribute to the development of HIV-vaccines (Earl et al. 2013; Julien et al. 2013; Lyumkis et al. 2013; Bartesaghi et al. 2015).

Another worldwide major health challenge is tuberculosis, caused by *Mycobacterium tuberculosis*. In terms of mortality rate, tuberculosis is only second to HIV/AIDS. The efficacy of the commonly administrated BCG-vaccine is highly variable (Andersen and Woodworth 2014); new multiple-drug-resistant strains represent a severe threat to the human population (Fogel 2015). Knowledge of the structure of the secretion system of *Mycobacterium* species (Abdallah et al. 2007), could be essential for the development of new drugs/vaccines against tuberculosis.

In structural biology, NMR spectroscopy and especially X-ray crystallography are widely used to solve the 3D structures of various important protein complexes (Garman 2014; Wang et al. 2014). These methods, however, have their limitations. With NMR spectroscopy only small molecules (<40 kDa) can be easily studied. X-ray crystallography requires the proteins to first be crystallized, which can be a major challenge (Snyder et al. 2005). Moreover, both these methods normally require large amounts of purified protein. Single-particle cryogenic electron microscopy (cryo-EM) (Dubochet et al. 1988), in contrast, requires only small amounts of proteins in solution. A wide range of samples can be studied by cryo-EM where molecular weights can vary from 100 kDa (Maletta et al. 2014) to 70 MDa or even larger (Veesler et al. 2013).

In cryo-EM, the biological macromolecules are embedded in vitreous ice by plunge-freezing a thin layer of solution into a cryogen, typically ethane or propane, cooled at liquid nitrogen temperature. The sample is then imaged in the electron microscope at the liquid nitrogen temperature. The higher the electron exposure, the higher the signal-to-noise ratio (SNR) can be achieved in the resulting cryo-EM images. However, biological samples can only tolerate a relatively low electron dose without being severely damaged (Baker and Rubinstein 2010; Karuppasamy et al. 2011). The basic idea of single-particle cryo-EM is to increase the SNR by averaging many images of particles in the same orientation. This averaging allows one to study biological samples by cryo-EM in a close-to-native environment, under low-dose imaging conditions. Another important advantage of cryo-EM is the possibility to solve structures of heterogeneous samples: several conformational states of a complex present in the sample can be reconstructed simultaneously (Klaholz et al. 2004; Fischer et al. 2010).

Development of single-particle cryo-EM

The history of the single-particle electron microscopy starts with the design of the first electron microscope by Ernst Ruska and Max Knoll in 1931 (Knoll and Ruska 1932). In 1986 Ruska received the Nobel Prize for Physics for his invention (Nobel Media 2014c). In the electron microscope one uses a beam of accelerated electrons as a source of illumination. The wavelength of accelerated electrons is up to 5 orders of magnitude shorter than the wavelength of photons used in light microscopy. This allows the transmission electron microscope to achieve a much higher resolution level than the light microscope. In biological sciences electon microscopy allows observing different parts of the cellular and molecular ultrastructure.

The first structural studies of biomolecules (bacterial viruses) were already performed in the late 1940s (Luria et al. 1943; Harris 2015). Electron microscopy requires the sample to be imaged in vacuum, which is a far cry from the physiological cellular environment. The first techniques, which allowed EM imaging of biological samples, were the metal shadowing technique (Sharp et al. 1950) and the negative staining technique (Brenner and Horne 1959; van Bruggen et al. 1960). In the negative staining of the air–dried samples, grains of the negative stain (salts of heavy metals) often introduce artefacts, cause dissociation of the complex into smaller subunits, and limit the achievable resolution. Already in 1960 Humberto Fernández-Morán suggested to perform rapid freezing in liquid helium in order to keep biological samples hydrated (Fernandez-Moran 1960). Marc Adrian and Jacques Dubochet brought this technique to maturity by rapidly freezing samples in liquid ethane and propane, cooled au-bain-marie in liquid nitrogen (Adrian et al. 1984; Dubochet et al. 1988). This plunge-freezing maintains the solvent in a vitreous state, allowing the proteins to be observed in a close-to-native environment. A limitation of vitrified biological samples is the low contrast level in the resulting images.

As mentioned above, one of the fundamental limitations in cryo-EM is the radiation sensitivity of the sample. Frozen-hydrated biological samples can tolerate a maximum electron dose of ~20 e–/Å^2 without accumulating significant damage (Cheng et al. 2015). As a consequence of this low electron dose, the SNR of images is low, which affects subsequent data processing. The combination of vitrification and negative staining technique was developed (cryo-negative staining) and has been often used to increase contrast of the biological samples (Orlova et al. 1997; Adrian et al. 1998). However, this approach often suffers from artefacts and there may be resolution limits dependent on sample-preparation details. In cryo-negative staining, the stain grains (usually uranium salts or ammonium molybdate) are often not evenly distributed around the protein complicating the interpretation of the results (De Carlo and Stark 2010; De Carlo and Harris 2011).

The first groups involved in EM methodology developments for structural biology included the MRC Laboratory of Molecular Biology in Cambridge (Aaron Klug), and the Max-Planck-Institute of Biochemistry in Martinsried (Walter Hoppe; Hoppe 1974). A classical work of David De Rosier and Aaron Klug in 1968 (De Rosier and Klug 1968) formulated basic principles of three-dimensional reconstructions from EM images (tail of bacteriophage T4 in negative stain). The proposed Fourier-space techniques (De Rosier and Klug 1968; Crowther 1971) allowed obtaining reconstructions of highly symmetrical samples like helical oligomers (DeRosier and Moore 1970) and icosahedral virus capsids (Crowther et al. 1970), as well as two-dimensional crystals from bacterial membranes (Henderson and Unwin 1975; Unwin and Henderson 1975). This pioneering work led to awarding the Nobel Prize in Chemistry to Aaron Klug in 1982 "for his development of crystallographic electron microscopy and his structural elucidation of biologically important nucleic acid-protein complexes" (Nobel Media 2014a; Harris 2015).

In a parallel development, new methods and programs were developed for studying individual biological complexes. First experiments with aligning and averaging 2D images of crystal patches and individual molecules were performed in the late 1970s. Owen Saxton, Joachim Frank, Marin van Heel and various others contributed to these early developments for 2D-image analysis (Saxton and Frank 1977; Steinkilberg and Schramm 1980; van Heel and Hollenberg 1980). With these "single-particle processing" developments, the need emerged to elaborate general-purpose EM-oriented software packages supporting the repetitive processing of many images, for example. Hence, the first extensive software packages like: "Semper", "Spider", and "Imagic" emerged (Saxton et al. 1979; Frank et al. 1981; van Heel and Keegstra 1981). A significant infrastructural improvement was the introduction of large image stacks, facilitating the recursive processing of large numbers of individual molecular images (van Heel and Keegstra 1981).

The next level of processing reflects the three-dimensionality of single-particle data. Molecular

images are 2D projections of the underlying 3D structure, and one will thus typically see many different views of the 3D structure in a micrograph. Variance-based automatic particle picking (van Heel 1982), multi-reference alignment (van Heel and Stöffler-Meilicke 1985); multivariate statistical data compression (van Heel and Frank 1981); and automatic classification (van Heel 1984; 1989; Borland and van Heel 1990), allow obtaining good class averages from a mixed population of molecular images. Projection matching (Harauz and Ottensmeyer 1984; van Heel 1984; Penczek et al. 1994) and angular reconstitution (Vainstein and Goncharov 1986; van Heel 1987) allow one to assign Euler angles to the various projection images. Using these Euler-angle orientations one can then apply specialized filtered back-projection algorithms (Harauz and van Heel 1986; Radermacher 1988) to obtain the 3D structure of the investigated biological macromolecules.

Already by the end of 1990s it was possible to obtain reliable 3D-reconstructions of complexes with sub-nanometer resolution. For example, in 1997 the structure of the icosahedral capsid of the hepatitis B virus was solved at 7.4 Å and 9 Å (Bottcher et al. 1997; Conway et al. 1997). The entirely asymmetric *E. Coli* 50S large ribosomal subunit was already solved to a resolution of 7.5 Å in 1999 (Matadeen et al. 1999). However, for a long time even the highest attainable resolution in single-particle cryo-EM was insufficient to resolve secondary structure elements; and single-particle cryo-EM was often referred to as "blobology".

Challenges in obtaining high-resolution results by cryo-EM

Let us now consider fundamental factors, which limit the obtainable resolution in cryo-EM. In the first place, the resolution of the final 3D-reconstruction depends on the *nature of the sample*. The size of the complex is also essential: it is easier to obtain high-resolution results for larger complexes than for small ones (discussed in details below). Obtaining high-resolution structures for < 200 kDa proteins can still be a real challenge. It is easier to work with large symmetric proteins. Symmetry can significantly facilitate the 3D reconstruction process in various ways. First, symmetry improves the statistics of the dataset: all individual asymmetric units in a particle, contribute independently to the final 3D-reconstruction. Virus capsids with icosahedral symmetry are thus favorite test objects for cryo-EM, since they contain 60 (or more) asymmetric subunits. A high level of symmetry also facilitates the orientation search for a molecular view, which is the reason that virus capsids were among the first structures studied by cryo-EM.

The *quality of purification* of the protein is of great importance. If the sample is not purified well (contains fragments of other proteins, junk, etc.), that can be a problem in the analysis of the data. If the complex is present in different oligomeric states, it is recommended to perform a biochemical separation of those states. A heterogeneous dataset can also be split into homogeneous subsets computationally, but that may complicate the analysis. At the

same time, as was mentioned above, the ability to handle conformational heterogeneity of a complex is one of the most important innovations introduced in cryo-EM as "4D cryo-EM" (Klaholz et al. 2004).

The *sample preparation* for cryo-EM is also of primary importance. A poor vitrification of the sample may result in too thick vitreous-ice layers , contamination with crystalline ice or ethane. Thick vitreous-ice layers are a source of strong background noise in the images and are associated with significant defocus differences with in the ice layer. A good vitrification results in a thin vitreous-ice layer, containing the individual particles in all possible orientations. However, a too thin ice layer may also cause preferred orientation of the molecules with only one or just a few dominant characteristic views. Using different agents, treatment of the grids may help to reduce the problem of preferred orientations (for details see (Grassucci et al. 2007)).

The *spectral power distribution of cryo-EM images* is another important consideration in achieving high-resolution results. Any image (such as the common test "Lena" image (Figure 1a)) can be characterized in Fourier space in terms of its different special frequency components. The low frequency information describes the general features of the object in the image (Figure 1b), whereas the high-frequency information corresponds to the small fine image details (Figure 1c).

Figure 1. Simulations demonstrating contributions of low- and high-frequency information. **(a)** A test "Lena" image was **(b)** low-pass filtered (with 0.1 cut-off value) to demonstrate contribution of low-frequency information (general features, shades). **(c)** The difference image between (a) and (b) demonstrates contribution of the high-frequency information in the images (small details).

In cryo-EM, both the low and high frequency information are important for obtaining a reliable high-resolution 3D-reconstruction. One typically starts emphasizing the low-frequency components in the data (overall shape and characteristic views of the molecule) to later improve the reliability of the high-resolution details (secondary structure elements).

The electron microscope is a phase-contrast microscope. The image contrast is described by the phase contrast transfer function (CTF; (Wade 1992)), which starts at zero at the Fourier-

space origin and oscillates around zero as function of spatial frequency. The exact shape of the CTF is determined by the microscope and the imaging conditions used. For single-particle cryo-EM the CTF is primarily dependent on the defocus value, at which the image was collected. The highest instrumental resolution of an electron microscope is reached, when the sample is imaged very close to focus ("Scherzer focus"; corresponding to ~0.08 µm in Figure 2a). At those defocus levels, the image contrast is transferred very well in high frequency range. However, the low-frequency information transfer is very poor around Scherzer focus. The particles within the sample can thus hardly be seen at these imaging conditions. Therefore, particle picking, centering, initial alignments and angular assignment procedures can all fail. To increase the low-frequency contrast, data acquisition is performed at defocus values around 0.5 to 3 µm (Figure 2b). At large defocus, the presence of the particles can easily be detected, but at large defocus the CTF oscillates rapidly at high resolution, causing serious damping of the high-resolution information transfer. This damping is due to reasons including specimen thickness and limited temporal and spatial coherence of the illumination (van Heel 1978; van Heel et al. 2000).

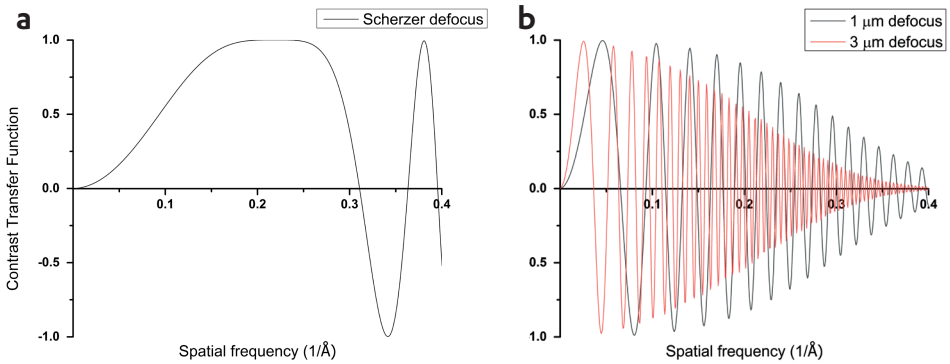

Figure 2. Simulated idealized phase contrast transfer functions for 200 keV microscope for three defocus levels: **(a)** Scherzer defocus (0.08 µm) is with a great transfer of high-frequency information. **(b)** 1 µm defocus (black) and 3 µm (red). The data from 3 µm defocus contains more low-frequency information, which is necessary for the particle detection. However, due to partial coherence and rapid oscillations, 1 µm defocus data has more high-frequency information, essential for high resolution. We used the following parameters in these simulations in the program TRANSFER of the IMAGIC-4D software (van Heel et al. 2012): 2.7 mm spherical aberration; 3.4 mm focal distance; 70 µm aperture; 0.5 µm source diameter; 500 Å object height.

Another important resolution-limiting issue is the stability of the microscope/sample during the data collection; the stage might be *drifting within the exposure time* (Kunath et al. 1984). Moreover, the particles may be moving within the ice layer during the electron exposure, a phenomenon known as "*beam-induce motion*" (Brilot et al. 2012; McMullan et al. 2015).

The "resolution revolution"

Recent developments of a new generation of direct electron detector devices EM cameras (DDD) (Faruqi and McMullan 2011; Li et al. 2013) triggered the "resolution revolution" in single-particle cryo-EM (Kühlbrandt 2014; Bai et al. 2015b; Cheng et al. 2015). Modern DDD (Direct Electron; FEI Falcon II; Gatan K2 Summit) have two major advances. First, their detective quantum efficiency (DQE, (Meyer and Kirkland 2000; McMullan et al. 2009)) is much higher, compared to common CCD (charge-coupled devices) and CMOS (complementary metal-oxide semiconductor) cryo-EM cameras. In DDD no scintillator is used – electrons are hitting a back-thinned layer of the sensor, reducing back-scattering and essentially the noise. This results in a higher camera performance, especially of the high frequency information, yielding a higher resolution. The second advantage is the readout speed. The new cameras acquire several frames (a "movie") during the exposure time. These frames can be aligned and the images can be corrected for the stage drifts and beam-induced particle motion (Brilot et al. 2012; Li et al. 2013).

Within the last years, the developments DDDs resulted in solving structures of ribosomes, multiple protein complexes and highly symmetrical virus structures by different groups with the resolution of around 3-4 Å or even below (Bai et al. 2015b; Cheng 2015; Cheng et al. 2015). Moreover, small molecules like the TRPV1 receptor (transient receptor potential cation channel subfamily V member 1) and the human γ-secretase were solved at 3.3 Å (Liao et al. 2013) and 3.4 Å (Bai et al. 2015a) respectively. The current absolute resolution record for cryo-EM specimens belongs to a β-galactosidase study, solved at 2.2 Å (Bartesaghi et al. 2015). Nevertheless, obtaining high resolution cryo-EM results is still not routine. Careful *image processing* is a paramount factor and is the main subject of this thesis. There can be many serious pitfalls during the data processing of very noisy complex data, leading to incorrect structures. Serious problems can be generated by reference bias and overfitting of the data. For example, in 2012 and 2013, Y. Mao and J. Sodroski published two papers on the structure of HIV-1 envelope glycoprotein trimer in NSMB (Mao et al. 2012) and PNAS (Mao et al. 2013) journals. The structures led to serious controversies in the cryo-EM field. In the raw data there were no particles seen in the micrographs, and the resulting 3D-reconstructions were suggested to be the result of reference-biased particle picking (Henderson 2013; Subramaniam 2013; van Heel 2013). More recently three independent studies, from three different groups, on the same HIV-1 structure were published (Bartesaghi et al. 2013; Julien et al. 2013; Lyumkis et al. 2013). These new structures were in good agreement with each other but not compatible with the results of Mao et al. (Mao et al. 2013).

Despite the recent success, image processing in single-particle cryo-EM is still not trivial. The main issues in the development of a solid unbiased methodology are: development of fast and effective algorithms for the movie alignments; unbiased approaches for the particle

picking; classification of heterogeneous datasets; automatic refinement procedures that are not prone to overfitting. Today, due to the developments of the data acquisition, the logistics of data handling and parallel processing has changed. Thousands of movies from a dataset, containing hundreds of thousands of particles, have to be processed quickly and with minimal user interference.

The aim of this thesis is to develop and test advanced methodologies for high-resolution single-particle cryo-EM and to apply those to real-life datasets.

Scope of the thesis

In *Chapter 2* we show how to take advantage of large datasets to characterize the image transducers used. We introduce a new approach for *a posteriori* camera normalization of large datasets from cryo-EM and suggest this approach for other fields of image processing to improve quality of the images.

In *Chapter 3* we assess characteristics of the cryo-EM detectors and introduce a new program to align frames from cryo-EM movies to correct for possible drifts and beam-induced motions of the molecules. We introduce the P-spectrum and the rotational averaged P-spectrum (RAP), as metrics to assess the success of the movie-alignment alignment algorithm.

In *Chapter 4* we cover the important issue of cross-correlation particle picking and demonstrate the danger of introducing reference bias.

Chapter 5 is dedicated to the description of the full single-particle cryo-EM image processing pipeline. Using our reference-free methodology we demonstrate how to obtain near-atomic resolution structure of the giant worm hemoglobin.

In *Chapter 6* we tackle a complicated problem of a heterogeneous dataset of a small protein EspB – a substrate of the type VII secretion system.

Finally, we provide a summarizing discussion of the work done in this thesis, and discuss the impact of the work for the single-particle cryo-EM field.

REFERENCES

Abdallah, A.M., Gey van Pittius, N.C., Champion, P.A., Cox, J., Luirink, J., Vandenbroucke-Grauls, C.M., Appelmelk, B.J. and Bitter, W. (2007) Type VII secretion-mycobacteria show the way. Nat Rev Microbiol 5(11): 883-891.

Adrian, M., Dubochet, J., Fuller, S.D. and Harris, J.R. (1998) Cryo-negative staining. Micron 29(2–3): 145-160.

Adrian, M., Dubochet, J., Lepault, J. and McDowall, A.W. (1984) Cryo-electron microscopy of viruses. Nature 308(5954): 32-36.

Andersen, P. and Woodworth, J.S. (2014) Tuberculosis vaccines – rethinking the current paradigm. Trends in Immunology 35(8): 387-395.

Bai, X.-c., Yan, C., Yang, G., Lu, P., Ma, D., Sun, L., Zhou, R., Scheres, S.H.W. and Shi, Y. (2015a) An atomic structure of human [ggr]-secretase. Nature 525(7568): 212-217.

Bai, X.C., McMullan, G. and Scheres, S.H. (2015b) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40(1): 49-57.

Baker, L.A. and Rubinstein, J.L. (2010) Chapter Fifteen - Radiation damage in electron cryomicroscopy. In Methods in enzymology, J.J. Grant, ed. (Academic Press), 371-388.

Barre-Sinoussi, F., Ross, A.L. and Delfraissy, J.F. (2013) Past, present and future: 30 years of HIV research. Nat Rev Microbiol 11(12): 877-883.

Bartesaghi, A., Merk, A., Banerjee, S., Matthies, D., Wu, X., Milne, J.L.S. and Subramaniam, S. (2015) Electron microscopy. 2.2 A resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor. Science 348(6239): 1147-1151.

Bartesaghi, A., Merk, A., Borgnia, M.J., Milne, J.L. and Subramaniam, S. (2013) Prefusion structure of trimeric HIV-1 envelope glycoprotein determined by cryo-electron microscopy. Nat Struct Mol Biol 20(12): 1352-1357.

Blundell, T.L., Sibanda, B.L., Montalvao, R.W., Brewerton, S., Chelliah, V., Worth, C.L., Harmer, N.J., Davies, O. and Burke, D. (2006) Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Philos Trans R Soc Lond B Biol Sci 361(1467): 413-423.

Bottcher, B., Wynne, S.A. and Crowther, R.A. (1997) Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. Nature 386(6620): 88-91.

Brenner, S. and Horne, R.W. (1959) A negative staining method for high resolution electron microscopy of viruses. Biochim Biophys Acta 34: 103-110.

Brilot, A.F., Chen, J.Z., Cheng, A., Pan, J., Harrison, S.C., Potter, C.S., Carragher, B., Henderson, R. and Grigorieff, N. (2012) Beam-induced motion of vitrified specimen on holey carbon film. J Struct Biol 177(3): 630-637.

Cheng, Y. (2015) Single-particle cryo-EM at crystallographic resolution. Cell 161(3): 450-457.

Cheng, Y., Grigorieff, N., Penczek, P.A. and Walz, T. (2015) A primer to single-particle cryo-electron microscopy. Cell 161(3): 438-449.

Congreve, M., Murray, C.W. and Blundell, T.L. (2005) Keynote review: Structural biology and drug discovery. Drug Discovery Today 10(13): 895-907.

Conway, J.F., Cheng, N., Zlotnick, A., Wingfield, P.T., Stahl, S.J. and Steven, A.C. (1997) Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy. Nature 386(6620): 91-94.

Crowther, R.A. (1971) Procedures for three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs. Philos Trans R Soc Lond B Biol Sci 261(837): 221-230.

Crowther, R.A., Amos, L.A., Finch, J.T., De Rosier, D.J. and Klug, A. (1970) Three dimensional reconstructions of spherical viruses by fourier synthesis from electron micrographs. Nature 226(5244): 421-425.

De Carlo, S. and Harris, J.R. (2011) Negative staining and cryo-negative staining of macromolecules and viruses for TEM. Micron 42(2): 117-131.

De Carlo, S. and Stark, H. (2010) Cryonegative staining of macromolecular assemblies. Methods in enzymology 481: 127-145.

De Rosier, D.J. and Klug, A. (1968) Reconstruction of three dimensional structures from electron micrographs. Nature 217(5124): 130-134.

DeRosier, D.J. and Moore, P.B. (1970) Reconstruction of three-dimensional images from electron micrographs of structures with helical symmetry. J Mol Biol 52(2): 355-369.

Dubochet, J., Adrian, M., Chang, J.J., Homo, J.C., Lepault, J., McDowall, A.W. and Schultz, P. (1988) Cryo-electron microscopy of vitrified specimens. Quarterly reviews of biophysics 21(2): 129-228.

Earl, L.A., Lifson, J.D. and Subramaniam, S. (2013) Catching HIV 'in the act' with 3D electron microscopy. Trends in microbiology 21(8): 397-404.

Faruqi, A.R. and McMullan, G. (2011) Electronic detectors for electron microscopy. Quarterly reviews of biophysics 44(3): 357-390.

Fernandez-Moran, H. (1960) Low-temperature preparation techniques for electron microscopy of biological specimens based on rapid freezing with liquid helium II. Ann N Y Acad Sci 85: 689-713.

Fischer, N., Konevega, A.L., Wintermeyer, W., Rodnina, M.V. and Stark, H. (2010) Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. Nature 466(7304): 29-333.

Fogel, N. (2015) Tuberculosis: A disease without boundaries. Tuberculosis (Edinb).

Frank, J., Shimkin, B. and Dowse, H. (1981) Spider—A modular software system for electron image processing. Ultramicroscopy 6(4): 343-357.

Garman, E.F. (2014) Developments in x-ray crystallographic structure determination of biological macromolecules. Science 343(6175): 1102-1108.

Grassucci, R.A., Taylor, D.J. and Frank, J. (2007) Preparation of macromolecular complexes for cryo-electron microscopy. Nat. Protocols 2(12): 3239-3246.

Harauz, G. and Ottensmeyer, F. (1984) Nucleosome reconstruction via phosphorus mapping. Science 226(4677): 936-940.

Harauz, G. and van Heel, M. (1986) Exact filters for general geometry three dimensional reconstruction. Optik 73: 146-156.

Harris, J.R. (2015) Transmission electron microscopy in molecular structural biology: A historical survey. Arch Biochem Biophys 581: 3-18.

Henderson, R. (2013) Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. Proc Natl Acad Sci U S A 110(45): 18037-18041.

Henderson, R. and Unwin, P.N. (1975) Three-dimensional model of purple membrane obtained by electron microscopy. Nature 257(5521): 28-32.

Hoppe, W. (1974) Towards three-dimensional "electron microscopy" at atomic resolution. Naturwissenschaften 61(6): 239-249.

Julien, J.-P., Cupo, A., Sok, D., Stanfield, R.L., Lyumkis, D., Deller, M.C., Klasse, P.-J., Burton, D.R., Sanders, R.W., Moore, J.P., Ward, A.B. and Wilson, I.A. (2013) Crystal structure of a soluble cleaved HIV-1 envelope trimer. Science 342(6165): 1477-1483.

Karuppasamy, M., Karimi Nejadasl, F., Vulovic, M., Koster, A.J. and Ravelli, R.B.G. (2011) Radiation damage in single-particle cryo-electron microscopy: effects of dose and dose rate. Journal of Synchrotron Radiation 18(Pt 3): 398-412.

Klaholz, B.P., Myasnikov, A.G. and Van Heel, M. (2004) Visualization of release factor 3 on the ribosome during termination of protein synthesis. Nature 427(6977): 862-865.

Knoll, M. and Ruska, E. (1932) Das elektronenmikroskop. Zeitschrift für Physik 78(5-6): 318-339.

Kühlbrandt, W. (2014) Biochemistry. The resolution revolution. Science 343(6178): 1443-1444.

Kunath, W., Weiss, K., Sackkongehl, H., Kessel, M. and Zeitler, E. (1984) Time-resolved low-dose microscopy of glutamine-synthetase molecules. Ultramicroscopy 13(3): 241-252.

Li, X., Mooney, P., Zheng, S., Booth, C.R., Braunfeld, M.B., Gubbens, S., Agard, D.A. and Cheng, Y. (2013) Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat Methods 10(6): 584-590.

Liao, M., Cao, E., Julius, D. and Cheng, Y. (2013) Structure of the TRPV1 ion channel determined by electron cryo-microscopy. Nature 504(7478): 107-112.

Luria, S.E., Delbruck, M. and Anderson, T.F. (1943) Electron microscope studies of bacterial viruses. J Bacteriol 46(1): 57-77.

Lyumkis, D., Julien, J.-P., de Val, N., Cupo, A., Potter, C.S., Klasse, P.-J., Burton, D.R., Sanders, R.W., Moore, J.P., Carragher, B., Wilson, I.A. and Ward, A.B. (2013) Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. Science 342(6165): 1484-1490.

Maletta, M., Orlov, I., Roblin, P., Beck, Y., Moras, D., Billas, I.M.L. and Klaholz, B.P. (2014) The palindromic DNA-bound USP/EcR nuclear receptor adopts an asymmetric organization with allosteric domain positioning. Nat Commun 5.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Desormeaux, A., Finzi, A., Xiang, S.H. and Sodroski, J.G. (2013) Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. Proc Natl Acad Sci U S A 110(30): 12438-12443.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Xiang, S.H., Haim, H., Yang, X. and Sodroski, J. (2012) Subunit organization of the membrane-bound HIV-1 envelope glycoprotein trimer. Nat Struct Mol Biol 19(9): 893-899.

Matadeen, R., Patwardhan, A., Gowen, B., Orlova, E.V., Pape, T., Cuff, M., Mueller, F., Brimacombe, R. and van Heel, M. (1999) The Escherichia coli large ribosomal subunit at 7.5 A resolution. Structure 7(12): 1575-1583.

McMullan, G., Clark, A.T., Turchetta, R. and Faruqi, A.R. (2009) Enhanced imaging in low dose electron microscopy using electron counting. Ultramicroscopy 109(12): 1411-1416.

McMullan, G., Vinothkumar, K.R. and Henderson, R. (2015) Thon rings from amorphous ice and implications of beam-induced Brownian motion in single particle electron cryo-microscopy. Ultramicroscopy 158: 26-32.

Meyer, R.R. and Kirkland, A.I. (2000) Characterisation of the signal and noise transfer of CCD cameras for electron detection. Microscopy Research and Technique 49(3): 269-280.

Nobel Media, A. (2014a). The Nobel Prize in Chemistry 1982. Retrieved 21 Aug, 2015, from http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1982.

Nobel Media, A. (2014b). The Nobel Prize in Chemistry 2009 Retrieved 20 Aug, 2015, from http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2009/.

Nobel Media, A. (2014c). The Nobel Prize in Physics 1986. Retrieved 21 Aug, 2015, from http://www.nobelprize.org/nobel_prizes/physics/laureates/1986/.

Orlova, E.V., Dube, P., Harris, J.R., Beckman, E., Zemlin, F., Markl, J. and van Heel, M. (1997) Structure of keyhole limpet hemocyanin type 1 (KLH1) at 15 A resolution by electron cryomicroscopy and angular reconstitution. J Mol Biol 271(3): 417-437.

Penczek, P.A., Grassucci, R.A. and Frank, J. (1994) The ribosome at improved resolution: New techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles. Ultramicroscopy 53(3): 251-270.

Radermacher, M. (1988) Three-dimensional reconstruction of single particles from random and nonrandom tilt series. J Electron Microsc Tech 9(4): 359-394.

Rodnina, M.V. and Wintermeyer, W. (2010) The ribosome goes Nobel. Trends Biochem Sci 35(1): 1-5.

Saxton, W.O. and Frank, J. (1977) Motif detection in quantum noise-limited electron micrographs by cross-correlation. Ultramicroscopy 2(2-3): 219-227.

Saxton, W.O., Pitt, T.J. and Horner, M. (1979) Digital image processing: The semper system. Ultramicroscopy 4(3): 343-353.

Scannell, J.W., Blanckley, A., Boldon, H. and Warrington, B. (2012) Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov 11(3): 191-200.

Sharp, D.G., Lanni, F. and Beard, J.W. (1950) The egg-white inhibitor of influenza virus hemagglutination: II. Electron microscopy of the inhibitor. Journal of Biological Chemistry

185(2): 681-688.

Snyder, D.A., Chen, Y., Denissova, N.G., Acton, T., Aramini, J.M., Ciano, M., Karlin, R., Liu, J., Manor, P., Rajan, P.A., Rossi, P., Swapna, G.V.T., Xiao, R., Rost, B., Hunt, J. and Montelione, G.T. (2005) Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination. Journal of the American Chemical Society 127(47): 16505-16511.

Steinkilberg, M. and Schramm, H.J. (1980) Eine verbesserte Drehkorrelationsmethode für die Strukturbestimmung biologischer Makromoleküle durch Mittelung elektronenmikroskopischer Bilder. Hoppe-Seyler´ s Zeitschrift für physiologische Chemie 361(2): 1363-1370.

Subramaniam, S. (2013) Structure of trimeric HIV-1 envelope glycoproteins. Proc Natl Acad Sci U S A 110(45): E4172-4174.

Unwin, P.N. and Henderson, R. (1975) Molecular structure determination by electron microscopy of unstained crystalline specimens. J Mol Biol 94(3): 425-440.

Vainstein, B.K. and Goncharov, A.B. (1986) Determination of the spatial orientation of arbitrarily arranged identical particles of unknown structure from their projections. Doklady Acad. Nauk SSSR 287: 1131-1134.

van Bruggen, E., Wiebenga, E. and Gruber, M. (1960) Negative-staining electron microscopy of proteins at pH values below their isoelectric points. Its application to hemocyanin. Biochim Biophys Acta 42: 171-172.

van Heel, M. (1982) Detection of objects in quantum-noise-limited images. Ultramicroscopy 7(4): 331-341.

van Heel, M. (1984) Three-dimensional reconstructions from projections with unknown angular relationship. 8th European Congress Electron Microscopy Budapest. 2: 1347-1348.

van Heel, M. (1987) Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy 21(2): 111-123.

van Heel, M. (2013) Finding trimeric HIV-1 envelope glycoproteins in random noise. Proc Natl Acad Sci U S A 110(45): E4175-4177.

van Heel, M. and Frank, J. (1981) Use of multivariates statistics in analysing the images of biological macromolecules. Ultramicroscopy 6(1): 187-194.

van Heel, M., Gowen, B., Matadeen, R., Orlova, E.V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M. and Patwardhan, A. (2000) Single-particle electron cryo-microscopy: towards atomic resolution. Quarterly reviews of biophysics 33(4): 307-369.

van Heel, M. and Hollenberg, J. (1980) On the stretching of distorted images of two-dimensional crystals. In Electron microscopy at molecular dimensions (Springer), 256-260.

van Heel, M. and Keegstra, W. (1981) IMAGIC: A fast, flexible and friendly image analysis software system. Ultramicroscopy 7(2): 113-129.

van Heel, M., Portugal, R., Rohou, A., Linnemayr, C., Bebeacua, C., Schmidt, R., Grant, T. and Schatz, M. (2012) Four-dimensional cryo electron microscopy at quasi atomic resolution: IMAGIC 4D. International Tables for Crystallography F: 624-628.

van Heel, M. and Stöffler-Meilicke, M. (1985) Characteristic views of E. coli and B. stearothermophilus 30S ribosomal subunits in the electron microscope. EMBO 4(9): 2389.

van Heel, M.G. (1978) Imaging of relatively strong objects in partially coherent illumination in optics and electron optics Optik 49(4): 389-408.

Veesler, D., Ng, T.-S., Sendamarai, A.K., Eilers, B.J., Lawrence, C.M., Lok, S.-M., Young, M.J., Johnson, J.E. and Fu, C. (2013) Atomic structure of the 75 MDa extremophile Sulfolobus turreted icosahedral virus determined by CryoEM and X-ray crystallography. Proceedings of the National Academy of Sciences 110(14): 5504-5509.

Wade, R.H. (1992) A brief look at imaging and contrast transfer. Ultramicroscopy 46(1): 145-156.

Wang, G., Zhang, Z.T., Jiang, B., Zhang, X., Li, C. and Liu, M. (2014) Recent advances in protein NMR spectroscopy and their implications in protein therapeutics research. Anal Bioanal Chem 406(9-10): 2279-2288.

CHAPTER 2

A POSTERIORI CORRECTION OF CAMERA CHARACTERISTICS FROM LARGE IMAGE DATA SETS

Pavel Afanasyev[1,2†], Raimond B.G. Ravelli[2†], Rishi Matadeen[3], Sacha De Carlo[3,4], Gijs van Duinen[4], Bart Alewijnse[1,2], Peter J.Peters[2], Jan-Pieter Abrahams[1], Rodrigo V. Portugal[5], Michael Schatz[6], Marin van Heel[1,5,7*]

Scientific Reports 2015 (5), 10317

[1]Leiden Institute of Chemistry, Leiden University, 2333 CC Leiden, The Netherlands

[2]The Institute of Nanoscopy, Maastricht University, 6211 LK Maastricht, The Netherlands

[3]Netherlands Centre for Electron Nanoscopy (NeCEN), 2333 CC Leiden, The Netherlands

[4]FEI Company, 5651 GG Eindhoven, The Netherlands

[5]Brazilian Nanotechnology National Laboratory – LNNano, CNPEM, C.P. 6192, 13083-970 Campinas SP, Brasil

[6]Image Science Software GmbH, Gillweg 3, D-14193 Berlin, Germany

[7]Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, UK

[†]These authors contributed equally to this work

*Corresponding author

ABSTRACT

Large datasets are emerging in many fields of image processing including: electron microscopy, light microscopy, medical X-ray imaging, astronomy, etc. Novel computer-controlled instrumentation facilitates the collection of very large datasets containing thousands of individual digital images. In single-particle cryogenic electron microscopy, for example, large datasets are required for achieving quasi-atomic resolution structures of biological complexes. Based on the collected data alone, large datasets allow us to precisely determine the statistical properties of the imaging sensor on a pixel-by-pixel basis, independent of any *a priori* normalization routinely applied to the raw image data during collection ("flat field correction"). Our straightforward *a posteriori* correction yields clean linear images as can be verified by Fourier ring correlation (FRC), illustrating the statistical independence of the corrected images over all spatial frequencies. The image sensor characteristics can also be measured continuously and used for correcting upcoming images.

INTRODUCTION

In recent years, digital image data acquisition with image transducers like CCD/CMOS chips (Boyle and Smith 1970) has become the standard, superseding the earlier analogue imaging technologies. Digital image transducers contain millions of pixels, each having (slightly) different characteristics in response to identical input signals. Indeed, "dead" pixels (or rows, columns, spots of pixels) may not give any response at all. In contrast, "hot" pixels always may produce a strong output independent of the input signal. The image presented to the user is typically a corrected image in which such pixels have been replaced by averages of neighbouring pixels, rows, or columns. The sensitivity differences between pixels will also be affected by flat-field correction (Aikens et al. 1989). We will call all such pre-processing the "*a priori*" correction associated with the image transducer/sensor.

The precision required for the correction of a set of images depends on their intended use. For example, in standard digital photography each image is appreciated individually, and the image sensor flaws after the *a priori* correction are normally not discernible. This means that the image errors are small compared to the 6-12 bits dynamic range of the red, green, and blue channel image information. The situation changes when the image has only very small contrast variations (such as an image of a homogeneous white wall) especially when dirt has accumulated on the sensor surface, effectively changing its properties compared to when the *a priori* correction parameters were determined. In such cases flaws in an individual image can be directly visible.

When thousands or tens-of-thousands of images are to be studied extensively by advanced averaging algorithms, a simple *a priori* correction can prove insufficient. In the total average of a 10,000-images dataset, for example, the contrast of a fixed-pattern background image

increases 10,000-fold since this small but fixed pattern adds up coherently. At the same time, the contrast due to actual image information – uncorrelated from frame to frame – increases by only a factor 100 (= √10,000). The ratio of the residual fixed-pattern variance, over the variance of the summed image information, thus increases 10,000 fold in the averaging process. A typical example is shown in Figure 1a-c where ~10,000 cryo-EM images are averaged. The correct representation of the data in quantitative scientific image acquisition can thus be critical, depending on intended use.

The development of advanced digital cameras has been instrumental in pushing the resolution attainable by single-particle cryo-EM towards near-atomic levels (Yu et al. 2008; Campbell et al. 2012; Bai et al. 2013; Li et al. 2013; Amunts et al. 2014; Kühlbrandt 2014). This approach requires the use of large datasets to bring the noisy, low-contrast image information in the micrographs to statistical significance through extensive averaging procedures. In cryo-EM, the calibration of the *a priori* correction must be repeated regularly and performed under approximately the same conditions used in the subsequent data collection, since the pixel properties may change with the average exposure level (Li et al. 2013). Movie-mode data collection procedures require alignments based on correlation functions of images collected with the same sensor (Kunath et al. 1984; Li et al. 2013). Insufficiently corrected images can lead to alignments with respect to the spurious zero-image of the chip rather than to the information content of the individual image frames.

When a large digital image dataset is available, collected with the same image transducer, then that dataset itself can be used for the statistical characterization of every pixel in the sensor. Images from different parts of the sample are in principle uncorrelated; thus, when summing all images from such a large dataset, the image information averages out. What prevails is the systematic different response of the individual pixels to the same average exposure. Different pixels also exhibit a different sensitivity or "gain": very sensitive pixels will exhibit a larger standard deviation from the average exposure than do less sensitive pixels. For characterizing each pixel in the transducer, we study the statistics of associated pixel vector: the collection of all density measurements from that pixel throughput the dataset (Borland and van Heel 1990). We characterize each pixel in terms of the average density and standard deviation of its pixel vector and exploit that information to normalize its output.

RESULTS

A posteriori image dataset correction

We assume that a reasonable *a priori* correction has taken place which includes the masking out of dead (or "hot") pixels, column, rows etc. in the physical transducers thus avoiding "division-by-zero" problems. Our *a posteriori* correction can then have the simple form:

$$I_c(r) = \frac{I_m(r) - \overline{I(r)}}{\sigma(r)} \quad (1),$$

where the corrected image intensity $I_c(r)$ is derived from the measured raw image $I_m(r)$ by subtracting $\overline{I(r)}$, the average image over the full large dataset, normalized by the standard deviation image of that dataset $\sigma(r)$ (dimensionless units are used throughout). The thus corrected images are normalized to zero mean and a unity standard deviation per pixel vector throughout the dataset. The assumption behind this approach is that the input images used for the calculation of the average image $\overline{I(r)}$ are uniform and indiscriminate in terms of the position of objects in the image.

Examples

After the *a posteriori* data normalization, all pixel vectors will show the same statistical behaviour in that each pixel vector will have the same average density and the same standard deviation. As mentioned, camera manufacturers will typically replace dead (and hot) pixels by an average of surrounding pixels in their *a priori* correction. Thus, such flaws are normally not obvious in the average image $\overline{I(r)}$. However, since the density provided for those poorly performing pixels is some average of their surroundings, the variance (or standard deviation) of the values found for that specific pixel vector will typically be lower than that of its fully functional peers. One can thus see these manipulated pixels/row/patches to behave differently $\sigma(r)$ in the image derived from the full dataset.

In the cryo-EM example detailed in Figure 1, bad image areas can be discerned directly in the raw images (Figure 1a,b) but especially in the average image (Figure 1c) and the standard deviation image (σ-image, Figure 1d). Whereas a set of vertical lines and some horizontal lines (marked by two arrows) are not very prominent in the overall average image (Figure 1c) their "hiding" by averaging over an environment is clearly visible in the σ-image of the same sensor area (Figure 1d). The *a posteriori* corrected images contain significantly less artefacts (Figure 1e,f), however, the truly "dead" or "hot" pixel areas (like the columns and rows marked in Figure 1e) cannot be fully corrected since they only repeat the information from neighbouring pixels and contain no new information. Spurious frequency-dependent correlations between different images are discussed below (Figure 2).

As another example of the sensor correction, we use 1064 raw images of the Mars surface, collected by the Curiosity rover. These images of 1344x1200 pixels each (Figure 3a,b) are from the Mastcam-right camera of the rover and are/were available from (NASA). Their average (detail) is shown in Figure 3c, and their σ-image image (detail) in Figure 3d. The main purpose of the exercise was to correct for the different black-&-white sensitivities of the red-green-blue pixels of the "Bayer pattern" of the sensor in the raw images (before correction: Figure 3a,b; after correction: Figure 3e,f). Further anomalies of the sensor emerged (see legend of Figure 3). Further examples of the procedure applied to data from different fields of science

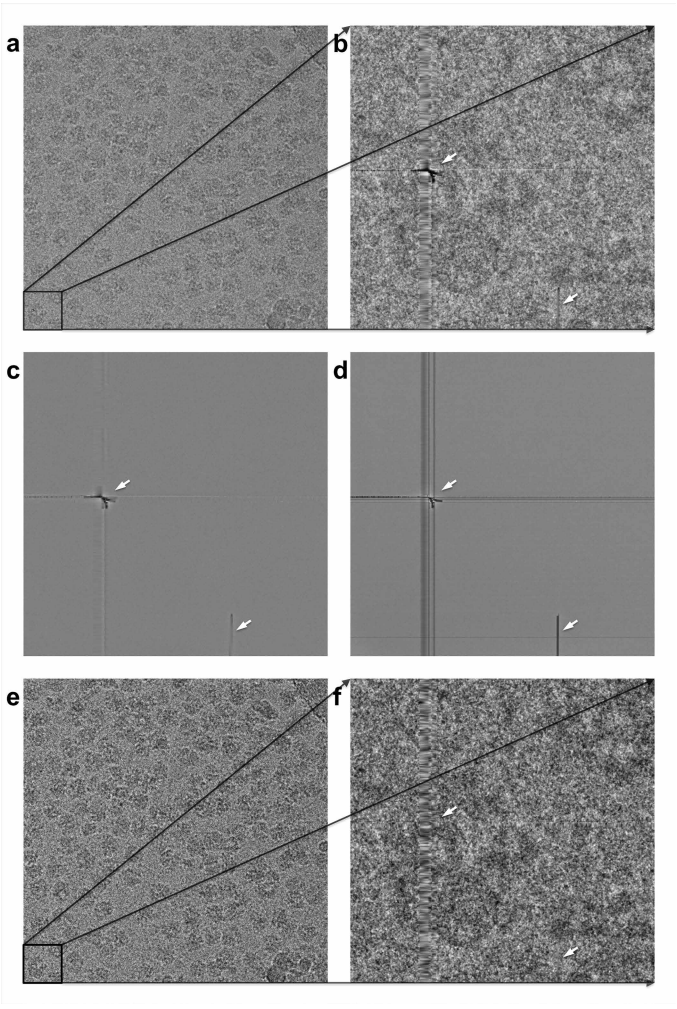

Figure 1. One full raw 4096x4096 ribosome image **(a)**, and one specific zoomed-in 512x512 patch (lower-left corner) extracted from that image **(b)**. This patch was the worst patch we could find in this experimental 4096x4096 back-thinned CMOS direct electron detection camera. The chip errors are such that they can be visually pinpointed in the image in spite of the *a priori* correction to the data. The average of all corresponding patch images from the full dataset – a total of 10821 images - is shown in frame **(c)**. Spurious vertical and horizontal lines and other serious 'fixed pattern' defects become clearly visible, while the ribosome images disappear altogether due to the averaging. The corresponding patch in the σ-image **(d)** reveals a strong bundle of about 16 vertical lines (marked by the left arrow) that were well suppressed by the standard *a priori* correction. While this suppression apparently included the averaging of pixel information in the immediate vicinity of these "dead" pixels and lines, the sensitivity of these chip areas collapses as is revealed by the dark areas in the σ-image. Moreover, the σ-image also reveals a thin horizontal line at the bottom of the patch (lower arrow) that had been corrected out in the average image (c), but again without compensating for the gain anomalies generated by the defect. The *a posteriori* corrected images are shown in panels **(e)** and **(f)** derived from the images shown in panels (a) and (b), respectively. Interestingly the *a posteriori* correction managed to improve on the dataset even by visual criteria although the more relevant metric is the FRC (see Figure 2). The amplitude spectra of the average and standard-deviation images are shown in the Supplementary Figure S1.

are given in the Supplementary Figures S1-S8.

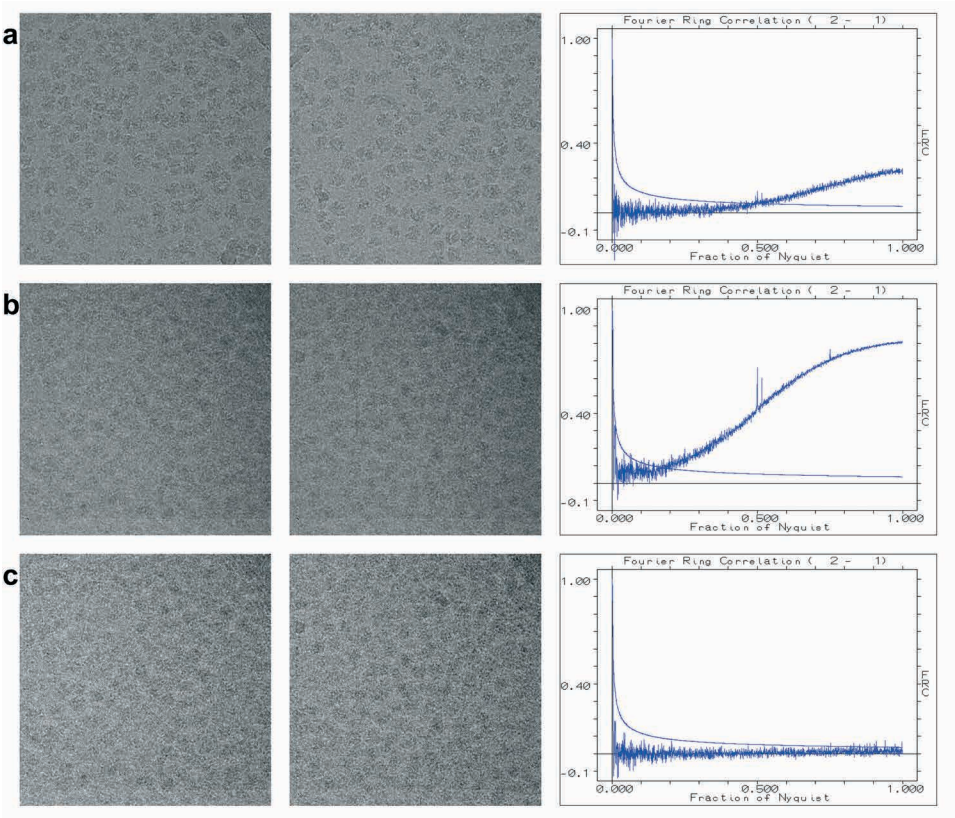

Figure 2. (a) Two different images collected on the same sensor (dataset of Figure 1) can show a strong correlation of the fine image details (high frequency Fourier space components) due to a common background pattern in the image transducer. The FRC curve shows that the high-frequency information fully exceeds (by more than 3σ) the level of expected random-noise correlations. **(b)** This effect can be exaggerated if we first average a number of raw input images to give two image average (12 different images per average are used here) and only then perform the FRC calculations between the two average images. The extra peaks at 0.5 and 0.75 of the Nyquist frequency in the FRC curve, are associated with fixed sensor readout patterns of the on-chip electronics. **(c)** The FRC of the same averaged image-sets illustrates that after the *a posteriori* correction the systematic background pattern is virtually removed from the data. Note that, in this 12-fold exaggerated critical test, the correction of the residual sensor pattern is close-to perfect. .

Validation by fourier ring correlation

To assess the quality of the *a posteriori* correction, we use the Fourier Ring Correlation (FRC), in which the normalized correlation coefficient is calculated between two different images over concentric rings in Fourier space (Saxton and Baumeister 1982; van Heel et al. 1982; Harauz and van Heel 1986; van Heel 1987; van Heel and Schatz 2005). This "gold standard" for assessing the reproducible resolution in two or three dimensions, has recently also become popular outside the field of electron microscopy (Vila-Comamala et al. 2011; Karplus

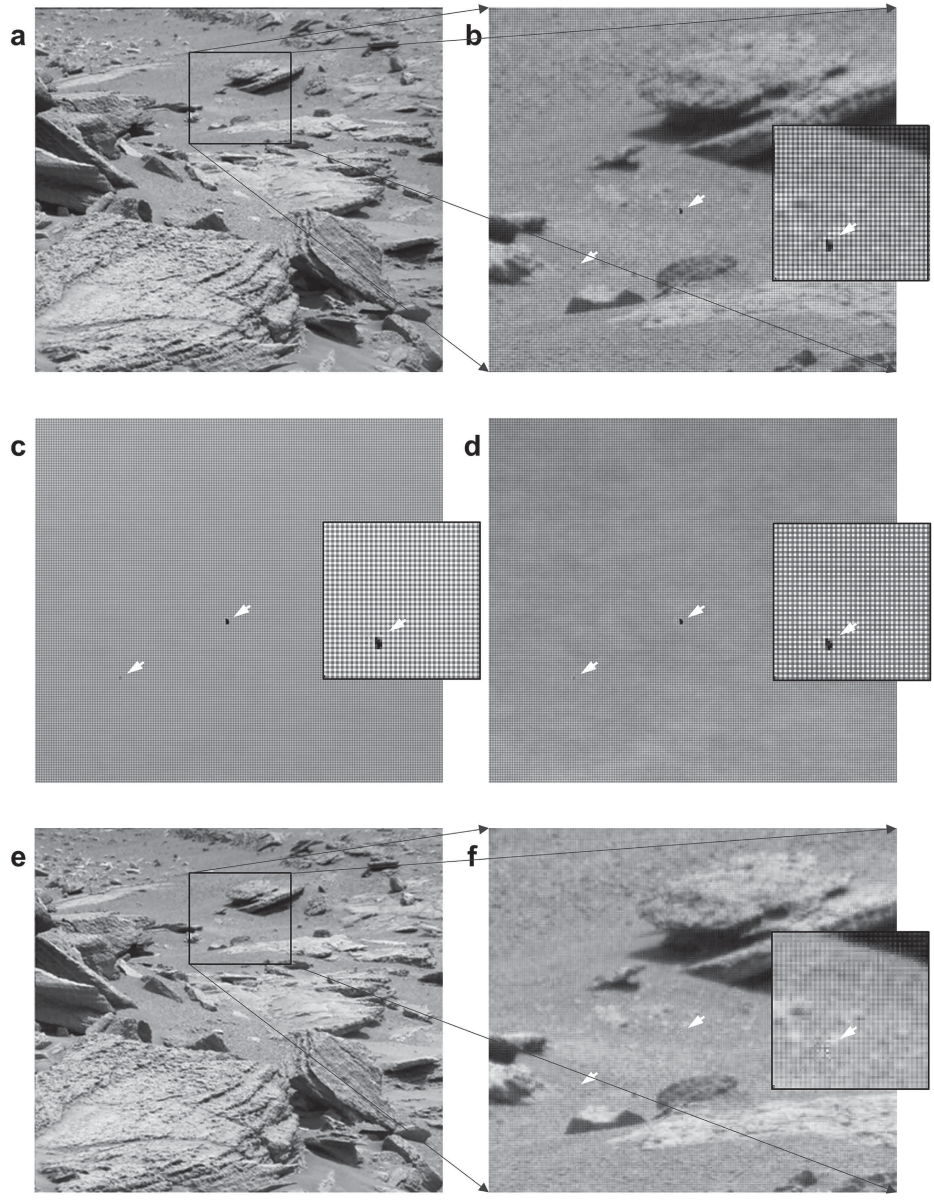

Figure 3. (a) A single 1344x1200 pixel image, from "Mastcam right" camera (MSSS-MALIN) of the NASA Mars rover "Curiosity", is shown together with a 300x336 pixel detail **(b)**. We used 1064 raw images of 1344x1200 pixels from this camera to find the average image (see 300x336 detail **(c)**; the central part of that is shown as an extra inset) and the σ-image image (300x336 detail: **(d)**). Apart from the strong visibility of the Bayer pattern in the average image of this camera, a block of 3x5 pixels with a very poor response is marked by a white arrow in the various "detail" images. A smaller anomaly visible in both the average - and standard deviation image is marked by another white arrow. The *a posteriori* corrected image is shown in panel **(e)** and in detail in **(f)**. The Bayer pattern is now largely invisible as are the other marked anomalies. The improvement of Fourier Ring Correlation between different images of this dataset by the *a posteriori* correction is discussed in Figure 4.

and Diederichs 2012; Banterle et al. 2013; Nieuwenhuizen et al. 2013). Here, however, we use the full FRC curve to assess the *independence* of two images as function of spatial frequency (Figure 2,4) rather than their cross resolution. The 3σ threshold curve indicates the maximum correlation levels expected between two independent random images. What is thus expected with a successful correction is an FRC curve that oscillates around zero and essentially never touches the positive (or negative) 3σ threshold curves. The behaviour of the FRC close to the origin is not very relevant since that reflects just a few pixels in Fourier space and may suffer from fluctuations between the extreme values "-1" and "+1".

When different object areas are imaged on the same part of the sensor, the FRC can show spurious correlations due to an insufficient *a priori* flat-field correction. In Figure 2, examples are given for the FRC curve between two such cryo-EM images of ribosomes before and after the *a posteriori* correction. The FRC curve for this comparison exceeds the random-fluctuations threshold of 3σ expected for uncorrelated images. To better visualize this effect we also averaged two groups of twelve images of different object areas collected on the same area of the chip, to emphasize the influence of the fixed background pattern (Figure 2b). The FRC peaks at 0.5 and 0.75 times the Nyquist frequency are due to systematic errors in the readout electronics of the camera. The "*a posteriori*" normalization corrects for the artificial correlations at high frequency as well as for the peaks at 0.5 and 0.75 times the Nyquist frequency (Figure 2c). Note that by averaging twelve frames we emphasize the influence of a failing *a priori* correction; the *a posteriori* correction is nevertheless capable of suppressing the undesired background-pattern correlations in this critical test. Figure 4 illustrates the spurious correlations existing between two arbitrary images chosen from the Mars dataset before and after the *a posteriori* correction.

DISCUSSION

In the processing of digital images, *a priori* flat-field corrections are routinely applied to every collected raw image. Apart from measures to smoothen errors like dead and hot pixels, the typical procedures applied to normalize the behaviour of all pixels in a camera includes the subtraction of a dark-image, and a pixel-by-pixel gain correction (Kunath et al. 1984; Aikens et al. 1989). More elaborate corrections may include refinements to compensate for non-linearity of the pixel sensitivity (Li et al. 2013). As we have seen in most cases tested (see also the various sensors discussed in the Supplementary Information), however, the routine *a priori* dataset corrections are insufficient for many advanced data processing needs. Indeed, especially the σ-image images derived from *a priori* corrected image datasets clearly reveal significant gain differences over the surface of the chip. Moreover, strong spurious Fourier-space correlations are revealed by FRC.

The proposed *a posteriori* correction aims at optimizing a full image dataset collected under

similar conditions, based on the idea that all pixels should behave equal in a statistical sense. Small departures from an average signal on an individual sensor will always give a linear response on output. The *a posteriori* corrections will generally give clean linear results, irrespective of, for example, the average exposure level of the raw images. With a reasonable *a priori* correction in place, the linearity of the output data from the well behaved pixels, after *a posteriori* correction, can extend over the full dynamic range of the image transducer.

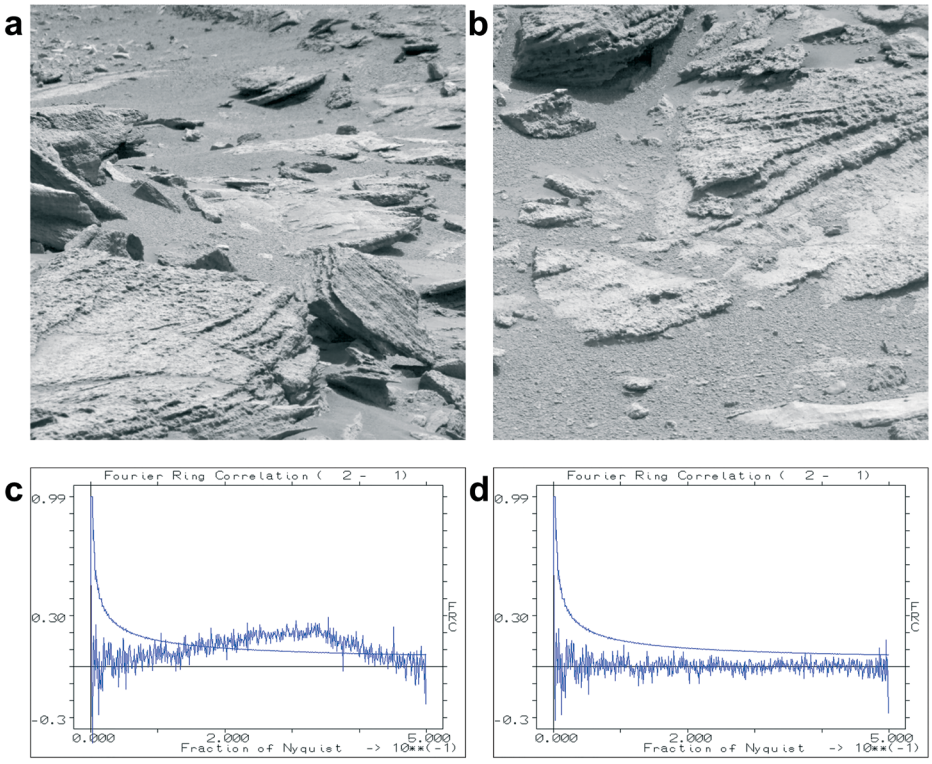

Figure 4. FRC of two Mars-rover images. Two typical images **(a, b)** taken from the Mastcam-right camera dataset (Figure 3) are compared to each other by cross-correlation as function of spatial frequency (the central 1200x1200 pixels part of the images were used). The FRC of the raw uncorrected images **(c)** shows significant correlations (above 3σ of the theoretically expected for random noise correlations) at around the 0.5 Nyquist frequency range, associated with the repeat of the 2x2 pixel Bayer patters of the sensor. After *a posteriori* correction the FRC oscillates around the zero value up to the Nyquist frequency **(d)**.

An important development in single-particle cryo-EM is the introduction of movie-mode data collection where, rather than taking a single individual image of each area of the sample, a full sequence of individual "movie" frames is collected. The movie frames are then aligned relative to each other and only then to be summed into a single overall average image (Kunath et al. 1984; Nejadasl et al. 2011; Campbell et al. 2012; Li et al. 2013). Although the rationale is 30 years old, three decades of instrumental and methodological developments have now brought

the cryo-EM approach into the realm of atomic resolution (Kühlbrandt 2014). Movie-mode data collection on direct electron detectors has contributed significantly to this development (Campbell et al. 2012; Li et al. 2013). Critical is the aligning of the individual low-dose frames constituting a "movie" (Kunath et al. 1984; Brilot et al. 2012; Campbell et al. 2012; Li et al. 2013) in producing the best possible image of the sample. Avoiding alignments of the movie frames to a fixed background patterns rather than to the actual image information is critical. Our *a posteriori* correction provides a new perspective for minimizing this problem.

The boundary between the *a priori* and *a posteriori* background correction is not always clearly defined. The individual pixel sensitivities can change over time or as a function of the ambient temperature (Vulovic et al. 2010). One may thus want to apply the *a posteriori* parameters extracted from a similar recent data collection as an *a priori* correction to a new dataset while it is being collected. A pragmatic approach for the flat field correction of a new dataset could thus be to use the average and sigma images of the last few thousand images collected of similar samples with any given camera. A further *a posteriori* correction may then also serve as a diagnostic tool to highlight new flaws arising in the image transducer or in the *a priori* data normalization.

In one example the average and sigma images from two different datasets, collected on the same camera but almost a year apart, revealed that some old dust particles remained in the same position on the chip, that new "dust" particles were deposited, and that some old dust particles had moved over to a different location (Supplementary Figure S7). Also in conventional photography, dust collection on the sensor of cameras with interchangeable lenses is a recurring problem. Software dust removal is often implemented in professional cameras or in post-processing software. In these cases, an input image is typically required of a homogeneous featureless area like a white wall for the dust removal. Whereas such a correction will visibly reduce the influence of new dirt on the sensor it cannot correct for the sensitivity loss suffered by pixels (partly) covered by the dirt on the sensor as does our proposed *a posteriori* correction. Finally, the normalization images derived from large datasets, i.e. the average and sigma images, characterize the state of health of the sensor at the time of data collection. This emphasizes the importance of long-term storage of raw datasets for quality control and validation purposes.

For many tasks in scientific image processing, the routine *a priori* correction of the image transducer properties is insufficient. This can be due to flaws of the image transducer or of the flat-field correction applied, but can also be a consequence of the limited number of grey values available in the corrected image and/or of the limited time allocated for measuring the transducer characteristics. We have shown that we can determine the characteristics of the digital imaging sensor directly from large datasets, allowing us to perform *a posteriori* data corrections that are matched to the experimental conditions. The statistical independence of

corrected images can readily be verified by FRC. The reduction of the fixed-pattern noise in the datasets leads to an overall improvement of the signal-to-noise ratio implying that more can be achieved with less data. The approach is simple to integrate in data-collection routines and leads to consistent datasets with a significantly reduced level of artefacts.

METHODS

The *a posteriori* determination of sensor characteristics and subsequent normalization of the data are based on the availability of a large dataset collected with the same camera under the same circumstances. For simplicity, we assume that basic errors in the image transducer such as missing pixels or row/columns ("dead" and "hot" pixels) have been corrected by a reasonable *a priori* correction procedure that a user does not normally have access to. This correction procedure will ensure that none of the pixels of the sensor will always produce the same numerical output value, leading to division-by-zero problems in our *a posteriori* correction procedures (equation (1)).

A further assumption is that the large dataset is homogeneous in a statistical sense (unimodal distributions of intensities and their standard deviations) and is not an agglomeration of diverse types of data collection on the same sensor. For example, an automatically collected cryo-EM dataset will typically also contain a number of entirely blank images or other deviant images which disturb the overall statistics. Such images are easily discarded after studying histograms of the average densities and/or the standard deviations of the individual images.

We also assume a homogeneous distribution of objects over the area of the images such that the chance of having a certain contrast at a certain position is isotropically distributed. The images should, for example, preferably not all have the same basic motif - such as a horizon running through the middle of the image with a bright sky above and a dark earth below the horizon. This issue was relevant in the processing of the differential interference contrast microscopy images (see Supplementary Figure S5). Another implicit assumption is that the illumination is uniform over the field of view of the imaging sensor. When that is not the case, say a microscope with a misaligned illumination, then the resulting density ramp will be interpreted as a sensitivity ramp of the sensor and the images corrected accordingly (see Supplementary Figure S5).

The *a posteriori* correction for the more general case than as formulated above (equation (1)), now including a target standard deviation σ_0 and a target average dataset density I_0, has the following form:

$$I_c(r) = \frac{I_m(r) - \overline{I(r)}}{\sigma(r)}\ \sigma_0 + I_0 \qquad (2),$$

where the corrected image intensity $I_c(r)$ is derived from the measured $I_m(r)$raw

image, by subtracting $\overline{I(r)}$, the average image over the full large dataset, normalized by the standard deviation image of that dataset $\sigma(r)$. Here the original contrast of the data is restored by multiplying the results by average standard deviation of all pixel vectors σ_0, and adding the original average density I_0 to all output images. The equation (2) can, but need not, be interpreted as dimensionless.

As discussed above, the FRC is used to identify spurious correlations between two images in the dataset. However, when those two images have also been used to calculate the average $\overline{I(r)}$ and sigma $\sigma(r)$ images that were used for the correction (as part of the full dataset), that procedure can introduce unintended correlations in the FRC curve. For smaller datasets in particular, to avoid such artificial correlations, it is thus best to correct the images to be used for FRC tests with an average and a sigma image to which they have not contributed.

A dedicated program (CAMERA-NORM) for *a posteriori* correction was developed in the context of the IMAGIC-4D software system (van Heel et al. 1996; van Heel et al. 2012). Complementing the examples given in the main paper, further examples from different fields are presented in the Supplementary Information.

Acknowledgements

We thank: Ralf Schmidt of Image Science GmbH, Berlin, for programming support; Linda Clijsters and Lenny Brocks of the Netherlands Cancer Institute (NKI), for providing the DIC dataset; Alexandre Cassago for collection of TVIPS camera data; Eric van Genderen for collecting the Medipix data; Mihoko Tame (NKI) for providing the Sony camera dataset. The Curiosity rover images are courtesy of NASA/JPL-Caltech. We thank the EMDB and Sjors Scheres for publicizing the EMPIAR-10002 dataset. Our research was financed in part by grants from: from the Dutch ministry of economic affairs Cyttron II FES-0908; HTS&M Initiative: FES-0901; by NanoNextNL of the Government of the Netherlands and 130 partners; from the BBSRC (Grant: BB/G015236/1)); from the Netherlands Organization for Scientific Research (NWO grant: 016.072.321); the Brazilian science foundations: CNPq (Grants CNPq-152746/2012-9 and CNPq-400796/2012-0), and the Instituto Nacional de C,T&I em Materiais Complexos Funcionais (INOMAT). We acknowledge the use of NeCEN electron microscopes (Leiden University) funded by NWO and the European Regional Development Fund of the European Commission.

REFERENCES

Aikens, R.S., Agard, D.A. and Sedat, J.W. (1989) Solid-state imagers for microscopy. Methods in Cell Biology 29: 291-313.

Amunts, A., Brown, A., Bai, X.C., Llacer, J.L., Hussain, T., Emsley, P., Long, F., Murshudov, G., Scheres, S.H. and Ramakrishnan, V. (2014) Structure of the yeast mitochondrial large ribosomal subunit. Science 343(6178): 1485-1489.

Bai, X.C., Fernandez, I.S., McMullan, G. and Scheres, S.H. (2013) Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. Elife 2: e00461.

Banterle, N., Bui, K.H., Lemke, E.A. and Beck, M. (2013) Fourier ring correlation as a resolution criterion for super-resolution microscopy. J Struct Biol 183(3): 363-367.

Borland, L. and van Heel, M. (1990) Classification of image data in conjugate representation spaces. Journal of the Optical Society of America and Optics Image Science and Vision 7(4): 601-610.

Boyle, W.S. and Smith, G.E. (1970) Charge coupled semiconductor devices. Bell System Technical

Journal 49(4): 587.

Brilot, A.F., Chen, J.Z., Cheng, A., Pan, J., Harrison, S.C., Potter, C.S., Carragher, B., Henderson, R. and Grigorieff, N. (2012) Beam-induced motion of vitrified specimen on holey carbon film. J Struct Biol 177(3): 630-637.

Campbell, M.G., Cheng, A., Brilot, A.F., Moeller, A., Lyumkis, D., Veesler, D., Pan, J., Harrison, S.C., Potter, C.S., Carragher, B. and Grigorieff, N. (2012) Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. Structure 20(11): 1823-1828.

Harauz, G. and van Heel, M. (1986) Exact filters for general geometry three dimensional reconstruction. Optik 73: 146-156.

Karplus, P.A. and Diederichs, K. (2012) Linking crystallographic model and data quality. Science 336(6084): 1030-1033.

Kühlbrandt, W. (2014) Biochemistry. The resolution revolution. Science 343(6178): 1443-1444.

Kunath, W., Weiss, K., Sackkongehl, H., Kessel, M. and Zeitler, E. (1984) Time-resolved low-dose microscopy of glutamine-synthetase molecules. Ultramicroscopy 13(3): 241-252.

Li, X., Mooney, P., Zheng, S., Booth, C.R., Braunfeld, M.B., Gubbens, S., Agard, D.A. and Cheng, Y. (2013) Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat Methods 10(6): 584-590.

NASA. Mars Science Laboratory; Curiosity Rover; Raw images. 2016, from http://mars.jpl.nasa.gov/msl/multimedia/raw/.

Nejadasl, F.K., Karuppasamy, M., Koster, A.J. and Ravelli, R.B.G. (2011) Defocus estimation from stroboscopic cryo-electron microscopy data. Ultramicroscopy 111(11): 1592-1598.

Nieuwenhuizen, R.P., Lidke, K.A., Bates, M., Puig, D.L., Grunwald, D., Stallinga, S. and Rieger, B. (2013) Measuring image resolution in optical nanoscopy. Nat Methods 10(6): 557-562.

Saxton, W.O. and Baumeister, W. (1982) The correlation averaging of a regularly arranged bacterial-cell envelope protein. J Microsc 127(Aug): 127-138.

van Heel, M. (1987) Similarity measures between Images. Ultramicroscopy 21(1): 95-99.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R. and Schatz, M. (1996) A new generation of the IMAGIC image processing system. J Struct Biol 116(1): 17-24.

van Heel, M., Keegstra, W., Schutter, W. and van Bruggen, E. (1982) Arthropod hemocyanin structures studied by image analysis. Life Chemistry Reports, Suppl. 1: 69-73.

van Heel, M., Portugal, R., Rohou, A., Linnemayr, C., Bebeacua, C., Schmidt, R., Grant, T. and Schatz, M. (2012) Four-dimensional cryo electron microscopy at quasi atomic resolution: IMAGIC 4D. International Tables for Crystallography F: 624-628.

van Heel, M. and Schatz, M. (2005) Fourier shell correlation threshold criteria. J Struct Biol 151(3): 250-262.

Vila-Comamala, J., Diaz, A., Guizar-Sicairos, M., Mantion, A., Kewish, C.M., Menzel, A., Bunk, O. and David, C. (2011) Characterization of high-resolution diffractive X-ray optics by ptychographic coherent diffractive imaging. Optics Express 19(22): 21333-21344.

Vulovic, M., Rieger, B., van Vliet, L.J., Koster, A.J. and Ravelli, R.B. (2010) A toolkit for the characterization of CCD cameras for transmission electron microscopy. Acta Crystallogr D Biol Crystallogr 66(Pt 1): 97-109.

Yu, X., Jin, L. and Zhou, Z.H. (2008) 3.88 A structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. Nature 453(7193): 415-419.

SUPPLEMENTARY INFORMATION

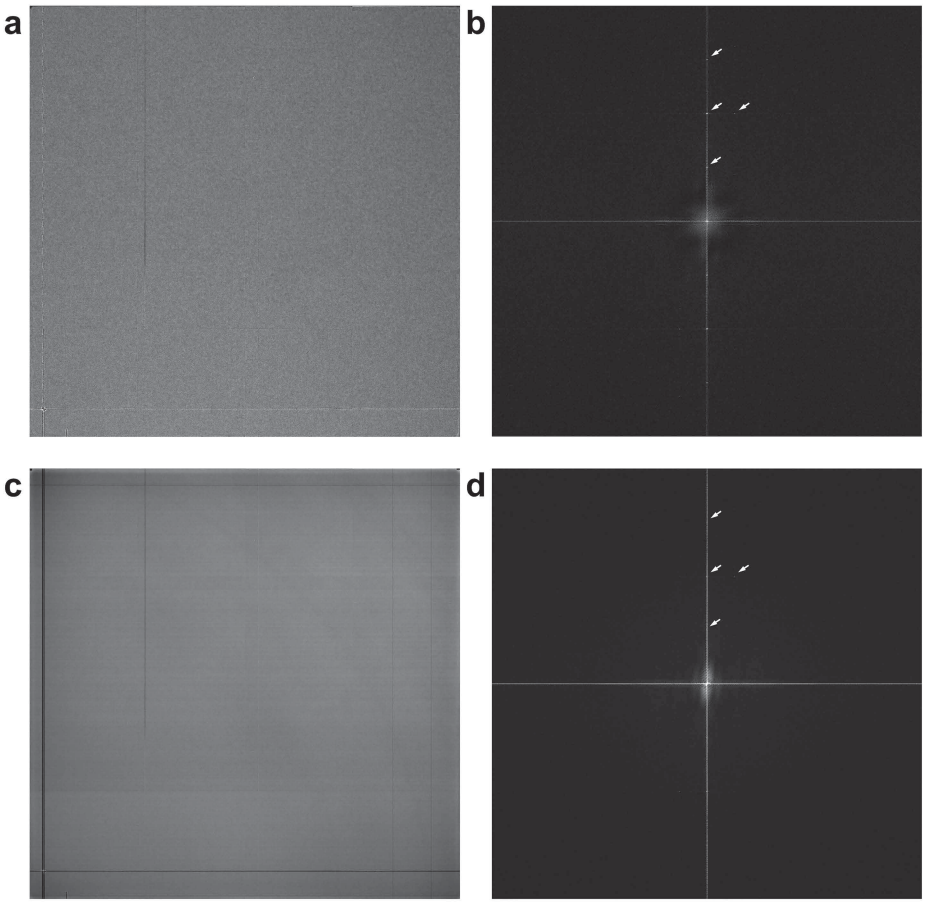

Figure S1. Fourier Spectra. **(a)** The average of the full dataset of 70S ribosome (Figure 1 of the main paper). This average image is of full size (4096x4096 pixels); note that the lower left corner of (a) is shown in detail in Figure 1c of the main paper. The amplitude spectrum of the full-scale average is shown in **(b)**. Highlighted by arrows in this spectrum are some peaks associated with the electronics of the camera leading to spurious peaks at ¼, ½ and ¾ of the Nyquist frequency. A further spurious peak is visible at ½ Nyquist in the vertical direction yet slightly offset to the right. These systematic peaks in the 2D amplitude spectra lead to the peaks revealed in the FRC calculation (main paper Figure 2). The sigma image of the full dataset **(c)** reveals the flawed areas of this experimental sensor. Its amplitude spectrum **(d)** shows similar features as does the total-average amplitude spectrum (b). Note that after the *a posteriori* correction of a large dataset their new average image and sigma image both become constants.

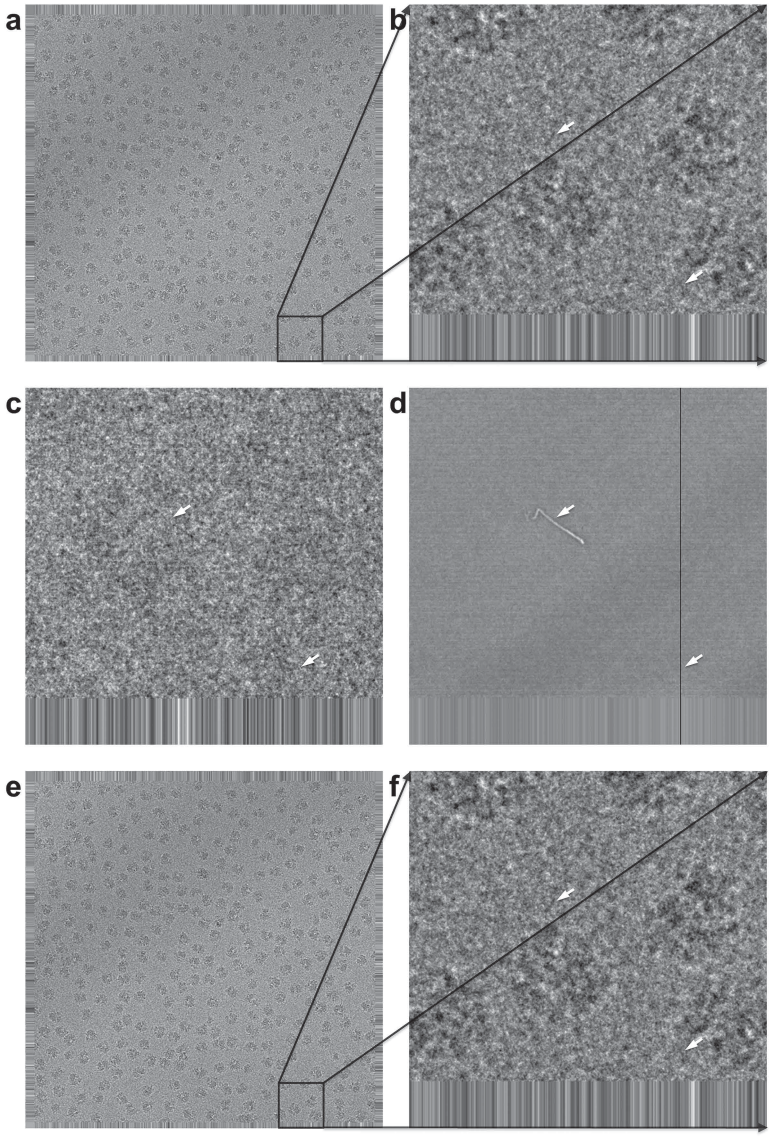

Figure S2. *A posteriori* correction of cryo-EM dataset of *S. cereviseae* 80S ribosomes. **(a)** A typical 4096x4096 pixel back-thinned FEI Falcon direct electron detector image from the publically available Electron Microscopy Pilot Image ARchive (EMPIAR) (http://pdbe.org/empiar; dataset EMPIAR-10002). The edges of the images each have a fixed repeating pattern ("apodization") which pattern differs from image to image in the 4,160 frames of this entry (4,160 =260x16 movie frames). For the details of this dataset see (Bai et al. 2013). In **(b)**, a 512x512 part of the lower edge of the first image (a) is shown in detail. Whereas the apodization of the edge is clearly visible, other anomalies such as dust particle and a missing vertical column remain invisible. In **(c)**, the overall average of these 4160 512x512 sub-images is shown; two arrows mark the positions of two anomalies that have apparently removed by *a priori* correction but are clearly visible in the standard deviation (sub-) image **(d)**. The *a posteriori* corrected images are shown in **(e)** and the detail area in **(f)**.

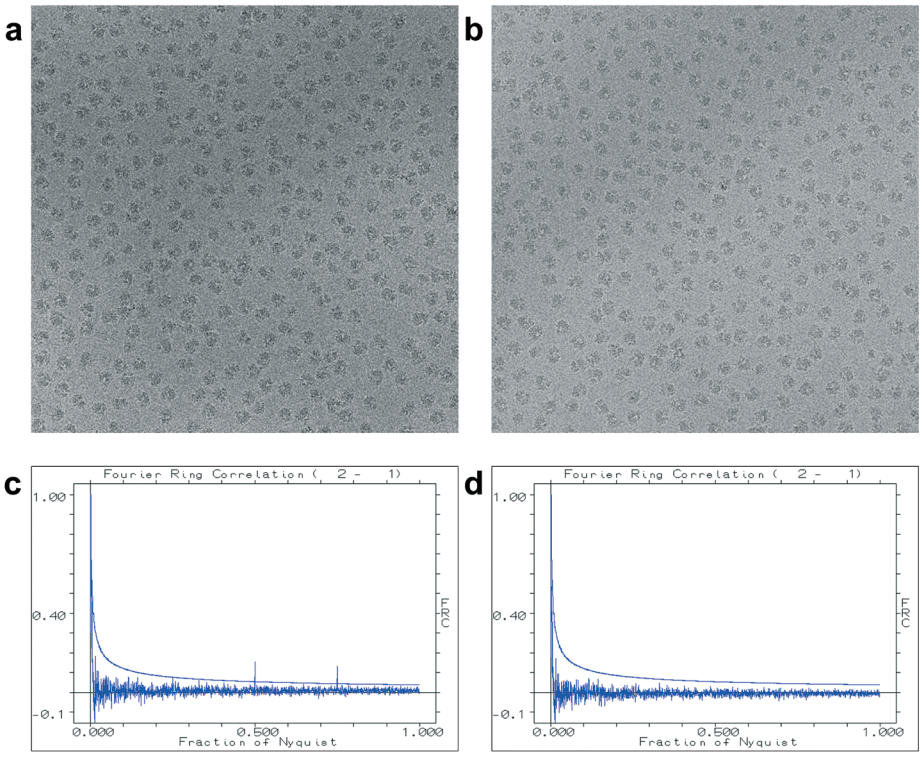

Figure S3. FRC of two ribosome micrographs. Two typical 3888 x 3888 sub-images of the 4096x4096 pixel images from the dataset shown/discussed in Figure S2. The *a priori* correction that had been applied to these images (Figure S2) was already quite good in providing uncorrelated images as exemplified in **(c)** which shows the FRC between images **(a)** and **(b)**. The *a posteriori* correction **(d)** succeeded in removing the remaining spurious correlation peaks at 0.5 and 0.75 Nyquist frequency that are correlated to the read-out electronics of the sensor. The sensor in this case was an experimental sensor back-thinned direct electron detector, comparable to the sensor used for generating the Figure 1 dataset.

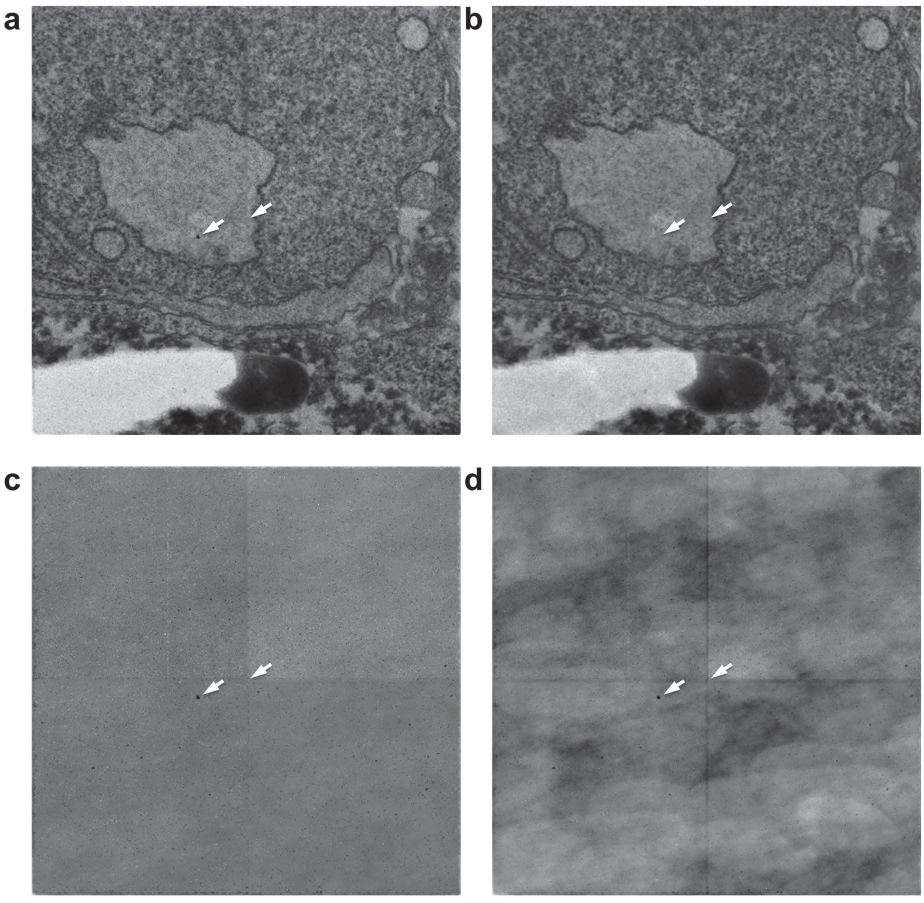

Figure S4. *A posteriori* correction of a Medipix camera dataset. A Medipix2 camera (Llopart et al. 2002; Nederlof et al. 2013) was used to acquire a dataset of stained cell sections consisting of 684 transmission electron microscopy images of each 512x512 pixels. The detector of this Medipix2 camera consists of 4 tiled sensors of 256x256 pixels each. The gaps between the four sensors cause a visible cross in the middle of the 512x512 image. The gaps were corrected for by an *a priori* correction just after data acquisition but remain clearly visible as a blurred band in the raw images **(a)**. In contrast to the other tests included in this publication we here had full access to the raw data since the *a priori* correction software was written and applied by one of us (JPA). For consistency, however, we fully ignored our specific knowledge of the details of this pre-processing algorithms. The *a priori* corrected images were thus *a posteriori* corrected **(b)**. The artificial cross in the middle of the images and the many dead - and hot pixel artefacts as well an anomalous (dotted-line) column disappeared after that correction. These anomalies were all clearly visible in the average image **(c)** and in the standard deviation image **(d)**. (Information on the Medipix2 camera can be found at the following website: http://medipix.web.cern.ch/medipix/).

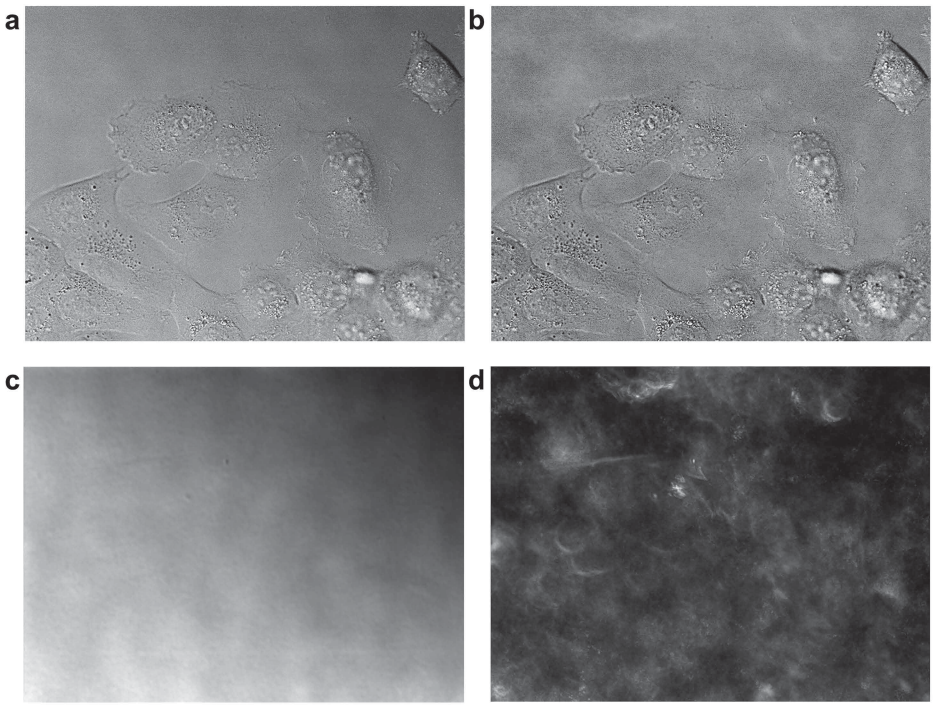

Figure S5. Differential interference contrast microscopy (DIC). This dataset was collected during an overnight live cell imaging experiment, and consists of 2,501 images **(a)** of 1344 x 1024 pixels each, acquired on a Hamamatsu ORCA-R2 CCD camera. The average of the dataset shows an uneven illumination with a ~25% density gradient in a diagonal direction over the full field of view **(c)**. This suggests that the illumination of the microscope was misaligned. The standard deviation image of the dataset **(d)** has residual patterns from the original image information in the dataset. In spite of the relatively large size of the dataset (2,501 images, remaining after excluding more than 1,000 images with a strongly differing standard deviation), the movement of the cells within the image frame were slow leading to specific areas exerting more contrast than others. The bright spots in the sigma images can be related to specific high-contrast objects being imaged which persist in specific places over prolonged periods of time. The removal of the background ramp in the corrected image **(b)** allows one to use the full available density-range for visualizing the object details in full contrast.

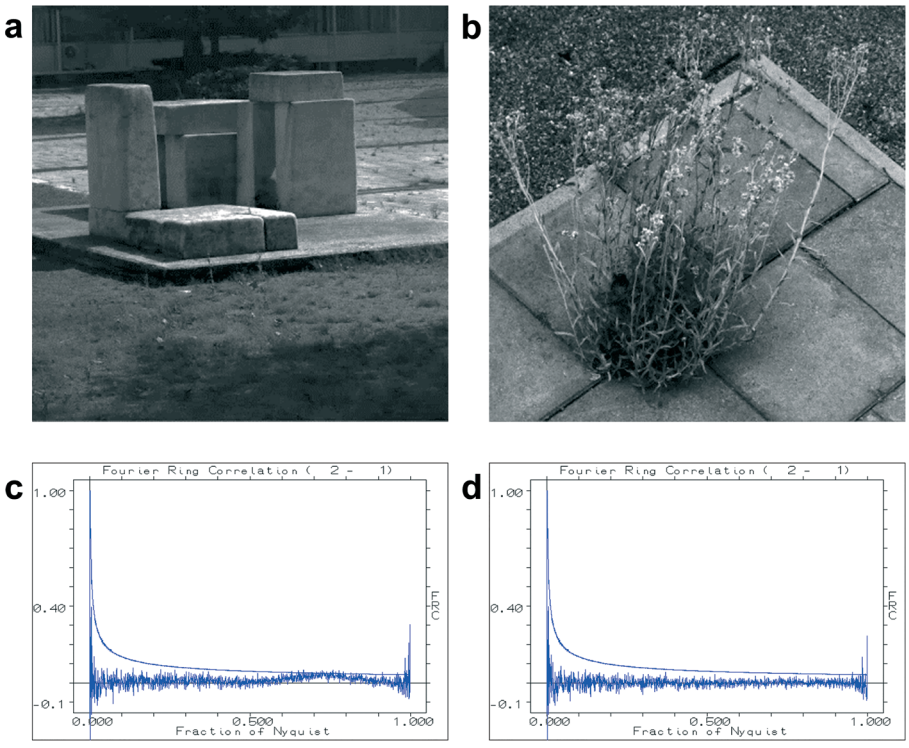

Figure S6. Standard Digital Photography. A Sony Alpha A350 consumer DSLR-camera was used to acquire a dataset of 666 RGB images 3072x4600 in ARW format. These images were converted to an uncompressed TIFF format treating each pixel as a separate grey-level measurement using LibRaw decoder (unprocessed_raw utility http://www.libraw.org/) followed by conversion into IMAGIC format (https://www.imagescience.de/formats) using the em2em converter (https://www.imagescience.de/em2em). The results are therefore fully comparable to the images produced by the Mastcam-right camera of the Mars rover (Figure 3). Two typical images **(a)** and **(b)** taken from the Sony camera dataset are compared to each other by FRC before **(c)** and after **(d)** *a posteriori* correction. These FRC curves indicate that the images correlate less at high resolution (~0.5 Nyquist and beyond) than do the Curiosity mars-rover images (Figures 3,4). This is probably due to the physical low-pass filter placed in front of the sensor in the A350 camera which removes high-resolution information prior to the image acquisition. Such a filter is not present in scientific grade sensors as used in the Mars rover camera.

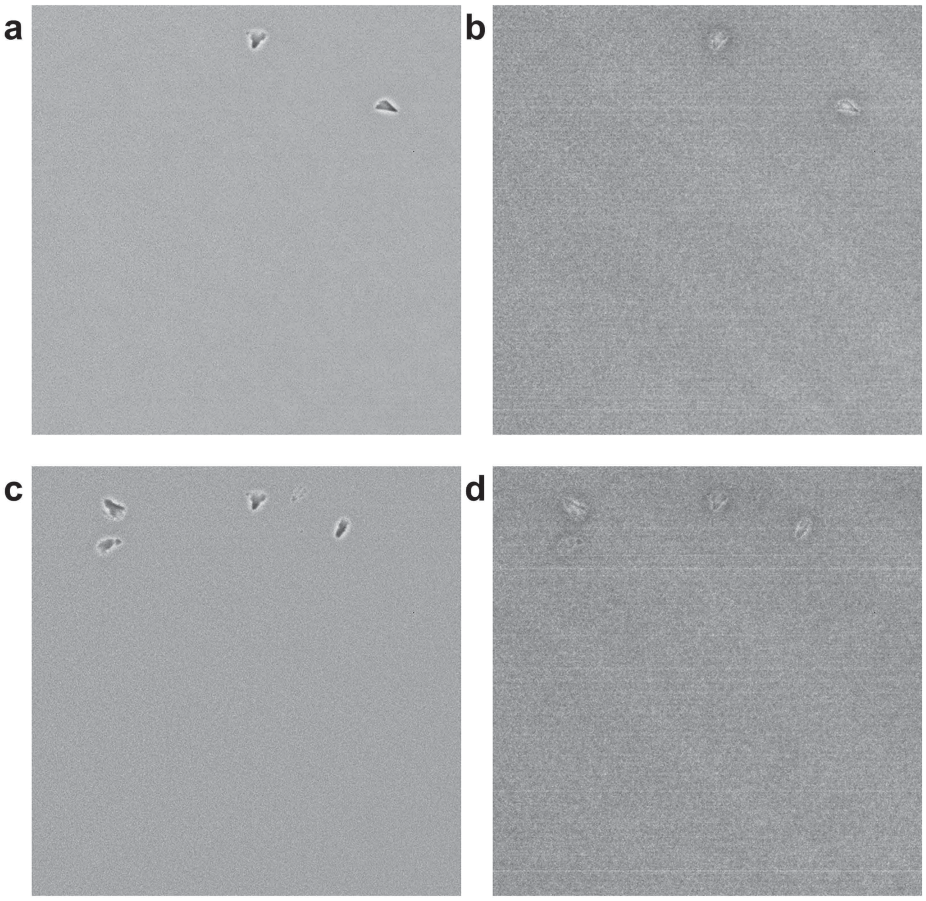

Figure S7. Dust changes over time. Two different cryo-EM datasets, collected on the same sensor but about a year apart, illustrate changes in "dust accumulation". **(a)** and **(b)** show the average and the standard deviation images of dataset "one" collected in May 2012 on an FEI Falcon-1 direct electron detector. Similarly, **(c)** and **(d)** show the average and the standard deviation images of dataset "two" collected in March 2013 on the same microscope and sensor. (All images are of the same 512x512 detailed area extracted from the full 4096x4096 images). The same "dust" particles can often be detected at the same position in the later dataset but some dust particles have changed position and/or orientation. New dust particles have appeared on the sensor after a year of use that were not seen in the earlier images.

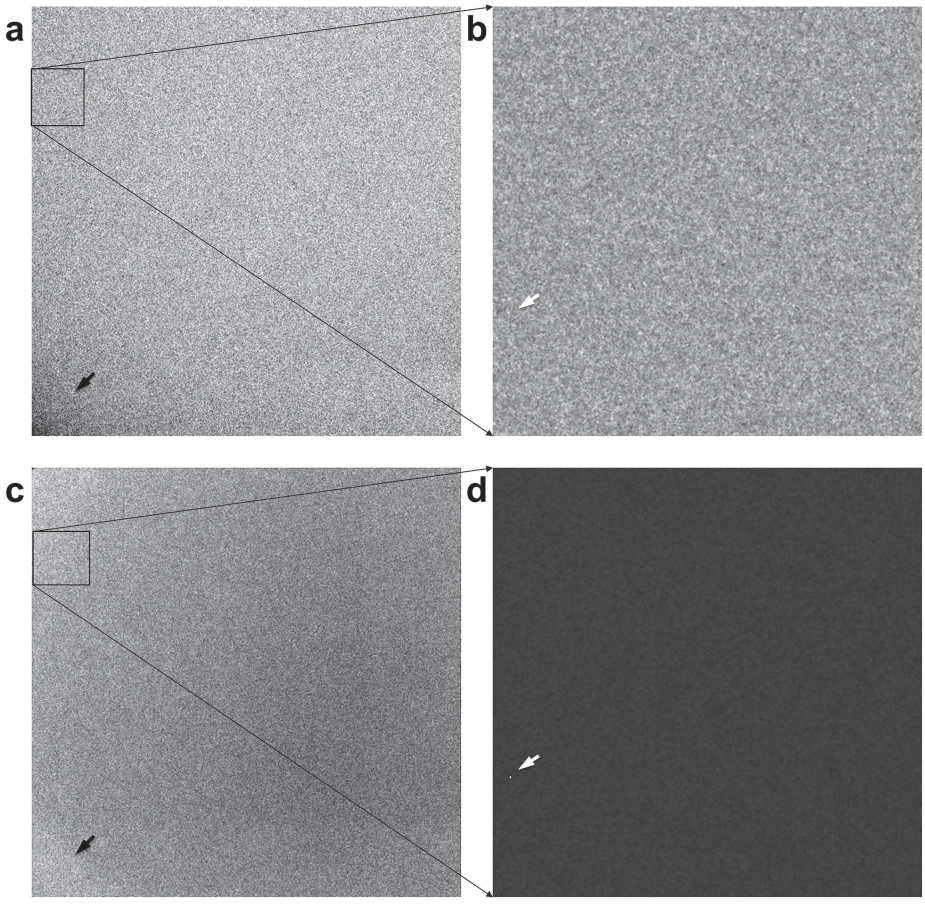

Figure S8. Dataset, collected manually over three consecutive days on a TVIPS F416 CMOS camera, contains 1,164 images. A significant number of images 174 were discarded based on the histograms of their individual statistics (average densities and standard deviations). This was due to the dataset being mixed with overview images and due to known instabilities in the camera electronics rendering some images unusable (the camera was already scheduled for replacement). The overall average image **(a)** has a slightly lower density in the lower–left corner (marked by a black arrow) yet that area also has a higher gain that the rest of the image **(c)**. A 512x512 detail from the average image marked with a black square in (a) is shown in **(b)**, revealing a very homogeneous density response in that patch. The corresponding σ-image 512x512 detail **(d)**, in contrast, reveals the presence of a deviant pixel with an extreme standard deviation behaviour.

REFERENCES

Bai, X.C., Fernandez, I.S., McMullan, G. and Scheres, S.H. (2013) Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. Elife 2: e00461.

Llopart, X., Campbell, M., Dinapoli, R., SanSegundo, D. and Pernigotti, E. (2002) MEDIPIX2: A 64-k pixel readout chip with 55'm2 elements working in single photon counting mode. Trans. Nucl. Sci. 49(5): 2279-2283.

Nederlof, I., van Genderen, E., Li, Y.W. and Abrahams, J.P. (2013) A Medipix quantum area detector allows rotation electron diffraction data collection from submicrometre three-dimensional protein crystals. Acta Crystallogr D Biol Crystallogr 69(Pt 7): 1223-1230.

CHAPTER 3

ASSESSING MOVIE-DATA IN CRYOGENIC ELECTRON MICROSCOPY

Marin van Heel[1,2,3†*], Pavel Afanasyev[1,4†], Michael Schatz[5*]

In preparation

[1]Leiden Institute of Biology, Leiden University, 2333 CC Leiden, The Netherlands

[2]Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, UK

[3]Brazilian Nanotechnology National Laboratory – LNNano, CNPEM, C.P. 6192, 13083-970 Campinas SP, Brasil

[4]The Maastricht Multimodal Molecular Imaging Institute, Maastricht University, 6229 ER Maastricht, The Netherlands

[5]Image Science Software GmbH, D-14193 Berlin, Germany

[†]These authors contributed equally to this work

*Corresponding authors

ABSTRACT

The recent boost in resolution, achieved in single-particle cryo-EM, was largely due to the introduction of direct electron detectors with greatly improved electron detection efficiency. These detectors can operate in "movie mode" allowing us to compensate for drift and beam-induced movements during data collection. We here present different spectrum-analysis approaches specifically designed for assessing movie-mode cryo-EM data. The "P-spectrum" automatically eliminates background ramps and can be used directly for contrast transfer function determination instead of traditional power spectra. We present a new iterative over-relaxation algorithm for the alignment of full-frame or sub-frame movie-mode data. We use the P-spectrum and the rotational averaged P-spectrum (RAP), as metrics to assess the success of the movie-alignment alignment algorithm. Finally, we discuss some fundamental aspects of spectrum calculations from movie-data.

INTRODUCTION

An important breakthrough in single-particle cryo-EM was the coming to maturity of movie data acquisition and processing, in which the overall electron exposure is fractionated over a number of "movie frames" that can later be processed separately. Already some 30 years ago it was realised that electron micrographs could be collected by registering the arrival of individual electrons on a sensor (Kunath et al. 1984). These electron arrivals were then integrated into "time-resolved low-dose images", nicknamed "movies", which were then used to study the decay of the biological samples as function of overall exposure. These early experiments, which already included "flat-field correction" and "movie alignments", however, were limited by the available technology: only ~256x256 pixel images could be recorded. At a time when data collection on film resolved up to ~9000x12000 pixels on a single photographic sheet, this was a serious disadvantage.

The recent progress of the movie-mode technology – now in combination with very efficient digital sensors typically recording ~4096x4096 pixels per image – has had a significant impact in terms of the resolution levels attainable in single-particle cryo-EM (Brilot et al. 2012; Bai et al. 2013; Grigorieff 2013; Kühlbrandt 2014). The development of advanced digital cameras has been instrumental in pushing the achievable resolution to near-atomic levels (Veesler et al. 2013b; Allegretti et al. 2014; Amunts et al. 2014; Bartesaghi et al. 2015; Campbell et al. 2015; Grant and Grigorieff 2015). The new back-thinned CMOS cameras are exposed directly to the incoming electrons ("direct electron detection" cameras). These cameras are a few times more efficient in electron detection than the earlier CCD cameras which first converted electrons into photons (Herrmann and Krahl 1982; Meyer and Kirkland 2000). For a recent review see (Faruqi and McMullan 2011; McMullan et al. 2014).

The movie-alignment problem in single-particle cryo-EM is not a simple optimisation problem

that has a unique solution as will be obvious to anyone who watched a cryo-EM sample disintegrate under electron exposure. Alignment issues (such as continuous drift during the collection of a movie) are relatively simple to solve (Kunath et al. 1984), yet, local movements relative to the overall movement of the frame already require more elaborate procedures. When the molecules move/rotate with respect to the local ice environment, the alignment may need to become particle-oriented (Brilot et al. 2012). Catastrophic events occurring during data collection such as the tearing of the vitreous ice or "boiling" of the ice in the grid-hole may also happen during the recording of a movie. For such events the simplest solution is to remove that movie from the dataset. Balanced decision making is thus required in order to effectively harvest a maximum of information from a cryo-EM sample in the available collection time (see Discussion).

The ultimate goal the (full-frame) movie alignment is to achieve a higher-resolution 3D structure from the improved processing (Campbell et al. 2012; Li et al. 2013; McMullan et al. 2014). The resolution typically measured by the Fourier Shell Correlation ("FSC" (Harauz and van Heel 1986). However, a good high-resolution molecular structure is many processing steps removed from the actual data collection. Therefore, such an indirect and sample-dependent metric cannot provide the necessary rapid feedback on the data-collection experiment.

The quality of movie alignments is typically illustrated by the power (or amplitude) spectrum of the total sum of the movie frames before and after movie alignment (Campbell et al. 2012; Li et al. 2013; Veesler et al. 2013a). We here follow a different strategy, namely that of the statistical evaluation during summing of the individual frames in Fourier space. This metric thus falls in the family of measures such as: the Q-factor (van Heel and Hollenberg 1980) used in peak-evaluation; the spectral signal-to-noise ratio (SSNR) (Unser et al. 1987); the S-image (Saß et al. 1989) used for assessing correlation-averaging results; and the Q-factor (Nejadasl et al. 2011) used for CTF evaluation.

THEORY

P-spectrum and Q-spectrum

In order to better evaluate images collected as a short burst of individual frames covering the same area of the sample ("movie-mode" data collection: collecting N images $I_{i\in N}(r)$ of the same sample area) we considered various options. The first, the "Q-spectrum" is a direct variant of the Q-factor (Nejadasl et al. 2011) from which it only differs by the subtraction by the random-noise expectation value ($1/\sqrt{N}$). We thus define the "Q-spectrum" as:

$$Q_S(f) = \frac{\left| \mathbb{F} \sum_N \left(I_{i \in N}(r) \right) \right|}{\sum_N \left| \mathbb{F} \left(I_{i \in N}(r) \right) \right|} - \frac{1}{\sqrt{N}} \quad (1),$$

We here calculate the amplitude-spectrum of the total sum of the movie frames and divide that by the sum of the amplitude spectra of the individual movie frames. In all earlier accounts these calculations were performed on complex structure factors in Fourier space. Due to the linearity of the Fourier transform, however, these summing operations are equivalent to the summing of the (real) amplitude spectra as formulated here. The subtraction of the random noise expectation value makes the Q-spectrum oscillate around zero at spatial frequencies where the spectra consist only of random noise. For practical reasons (discussed below) we found it mostly advantageous to use the "P-spectrum", which we define as follows:

$$P_S(f) = \frac{\left(\left| \mathbb{F} \sum_N \left(I_{i \in N}(r) \right) \right| \right)^2}{\sum_N \left(\left| \mathbb{F} \left(I_{i \in N}(r) \right) \right| \right)^2} - \frac{1}{N} \quad (2),$$

This P-spectrum differs from the "S-image" (Saß et al. 1989) only by the subtraction of the noise expectation ($1/N$). Due to the linearity of the Fourier transform, the statistical core of the P-spectrum calculation is simple one: it is a "square of the sum" divided by the "sum of the squares". Due to the subtraction of the expected average background fluctuations, the P-spectrum is not a classical power spectrum in the sense that it is not a positive-definite function. The advantage is again that the P-spectrum oscillates around zero when there is no signal in the (accumulated) movie frames. A metric derived from the P-spectrum is the Rotationally Averaged P-spectrum (RAP) which indicates the level of significant data collection as function of spatial frequency (see below).

Movie-alignments

Movie alignments come in different flavours. In the original (Kunath et al. 1984) publication, the alignment of the full 256x256 frames with respect to each other are addressed as "drift compensation". The alignment of individual molecular image frames was performed iteratively using running frame averages. This distinction between the movement of the image environment and of a specific molecule within its local environment exists until today. The larger the molecules and the more contrast they exert, the more a local refinement can be applied to the molecular images directly (Campbell et al. 2012; Bai et al. 2013; Veesler et al. 2013a). Here we focus on the alignment of the full image (or substantial subsection of the full image) irrespective of the individual molecular images in the field of view. We apply this normally after correcting for the camera characteristics (Afanasyev et al. 2015). The movie alignment algorithm we implemented as part of our standard alignment procedures (van Heel

et al. 1996; van Heel et al. 2012) has many options (like skipping the first frame), but the typical procedure starts with the total sum of the input image frames:

$$I_{Tot}(r) = \sum_{N} \left(I_{i \in N}(r) \right) \quad (3),$$

A standard cross-correlation ($\otimes$) search is then performed for each of the individual i frames leading to a shift-vector δ with respect to the original position of frame i.

$$Peak\left(I_{Tot}(r) \otimes I_{i \in N}(r) \right) \Rightarrow \delta \quad (4),$$

The original frames are then summed after shifting them to the position of maximum correlation, yet using an over-correction factor of κ to reinforce the good direction into which the frame is moving:

$$I'_{Tot}(r) = \sum_{N} \left(I_{i \in N}(r - (1+\kappa)\delta) \right) \quad (5),$$

The new improved total average of the aligned movies then serves as the new reference for the next iteration of movie-frame alignments as per equation (4). Typical κ values that optimally speed up the iterative movie alignment are in the range from ~0.8 to ~0.95. The procedure is halted when the average absolute shift of the movie frames falls below a pre-set minimum.

RESULTS

For testing our methodology, we used a dataset of earthworm (*Lumbricus terrestris*) hemoglobin, collected on the FEI Titan Krios microscope, equipped with: a 300 keV XFEG electron gun, a Cs-corrector, and a Falcon II direct-electron camera (Chapter 5). The movie-images were collected over a ~1 second period using 7-frames/movie (overall exposure ~40 e–/Å^2). Data was collected at 59000x magnification, corresponding to a (calibrated) pixel size of 1.12 Å. The sample has rather a high density of molecules but the selected images contain no carbon-foil edges from the supporting "Quantifoil" grid. Camera correction to reduce the influence of fixed pattern noise (Afanasyev et al. 2015), was applied to the full data set of all collected frames.

The P-spectra from the full movie data set (containing ~5235 movies with 7 frames each) shows a wide variety of cases: movies of empty holes (automatically sorted out), movies with a strong continuous drift (Figure 1a), and movies that were relatively stable over the 1 sec data collection time (Figure 2a). All movies were submitted to an alignment using the "ALIGN-MOVIE" command in IMAGIC-4D (van Heel et al. 2012), all using the same parameters (CCF alignment, max-shift 40 pixels, 22 iterations max, convergence criterion < 0.02 average pixels shift in one iteration, overcorrection factor 0.9 band-pass filter parameters 0.01-0.05 Nyquist).

Examples of P-spectra of the aligned movies are shown in (Figure 1b,2b). The rotational averages P-spectra, are shown on Figure 1c,d and Figure 2c,d. The height of the peak in these

curves is a function of: the signal and noise level in the movie frames, the mutual alignment of the movie frames, and of the number of (good) movie frames averaged. Note that since the drift is already substantial within each of the 7 movie frames of Figure 1a, the final resolution achieved in the high-drift direction, after movie alignment Figure 1b, is limited to ~7 times maximum resolution in the original unaligned average Figure 1a. Since the drift in movie images leading to Figure 2a is much smaller than in the case of Figure 1a, the final resolution achieved in the high-drift direction, after movie alignment Figure 2b, is significantly higher than in Figure 1b. The RAP at half the Nyquist frequency in Figure 2d is about twice as high as was the RAP in Figure 2c. This RAP indicates that the CTF information is retrievable from this image alone up to ~3.2 Å resolution (3.2 Å = 256/175*2.24 Å).

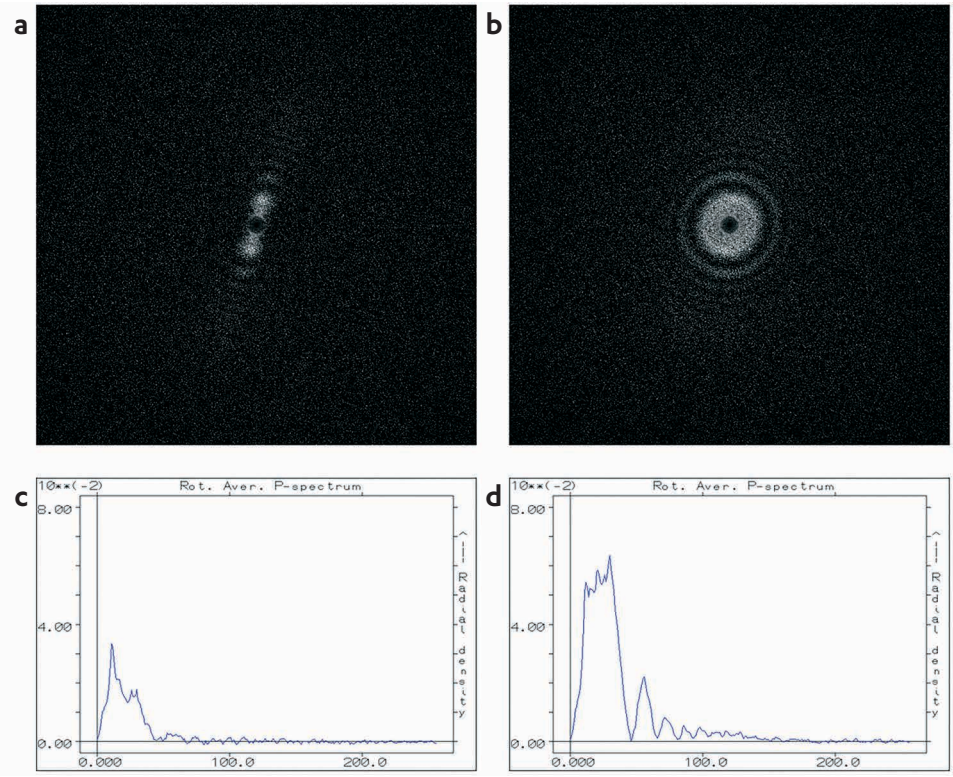

Figure 1. P-spectrum of movie-mode collected image of worm hemoglobin. The sample has a high density of molecules but contains no carbon-foil edges of the "Quantifoil" grid. **(a)** P-spectrum of the 7 movie-frames prior to their mutual alignment. **(b)** P-spectrum of the 7 movie-frames after their mutual alignment. **(c)** Rotational average of P-spectrum (a). **(d)** Rotational average of P-spectrum (b). The height of the peak in these curve is a function of: the signal and noise level in the movie frames, the mutual alignment of the movie frames, and of the number of (good) movie frames averaged. Note that since the drift is already substantial within each of the 7 movie frames, the final resolution achieved in the high-drift direction, after movie alignment (b), is limited to ~7 times maximum resolution in the original unaligned average (a). The success of a movie alignment can often be directly seen in the movie average at full resolution. For example, 512x512 fragments of these movie averages, before and after alignment, are presented in Figure S1.

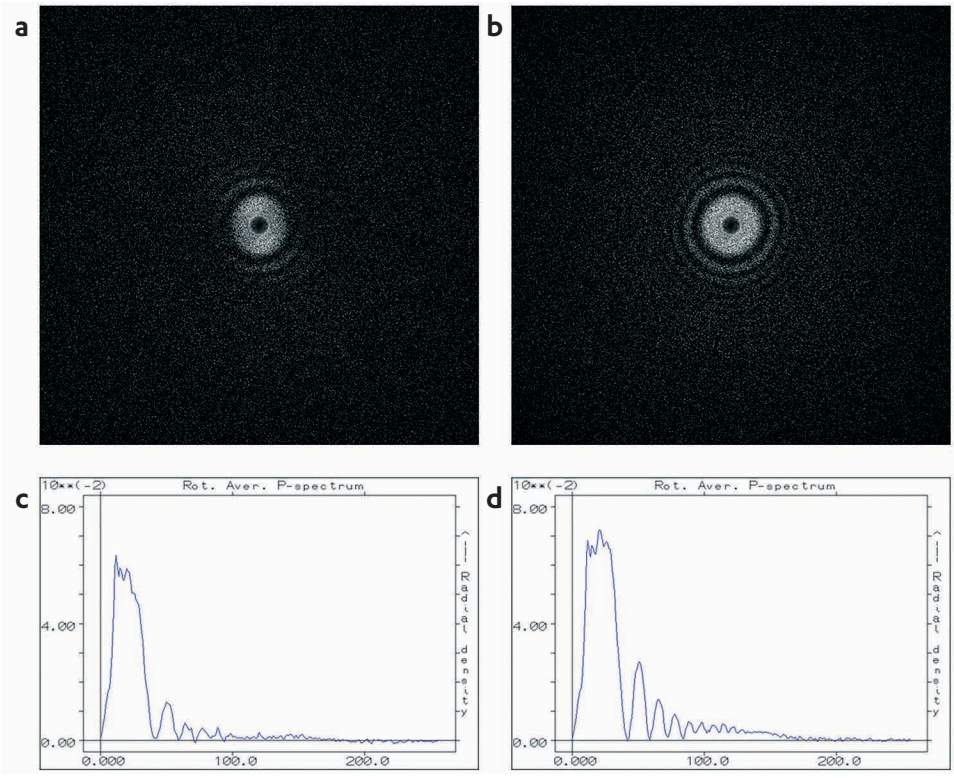

Figure 2. P-spectrum of movie-mode-collected micrograph of worm hemoglobin with moderate drift. **(a)** P-spectrum of the 7 movie-frames prior to their mutual alignment. **(b)** P-spectrum of the 7 movie-frames after their mutual alignment. **(c)** Rotational average of P-spectrum (a). **(d)** Rotational average of P-spectrum (b). Since the drift in this 7-frame movie is much less than those on (Figure 1), the final resolution achieved in the high-drift direction, after movie alignment (b), is significantly higher than in Figure 1. The RAP at half Nyquist (d) is about twice as high as was the RAP in Figure 1d. The RAP significantly departs from zero at spatial frequencies below 0.7 Nyquist, here corresponding to ~1/3.2 Å.

The quality of the P-spectrum is dependent on the amount of noise (dose-related) in the data and the contrast generated by the sample. To illustrate this behaviour, we added (even more) random white noise to the aligned experimental movie of Figure 2b,d at a standard deviation level of 1, 2 and 4 respectively (the original standard deviation of the movie frames was ~1). The effects are nicely illustrated by the rotationally averaged P-spectra of the same aligned movie before (Figure 3a) and after noise addition (Figure 3b-d). Note that the height of the peaks and the RAP in general is lower, the noisier the data is. Moreover, this experiment demonstrates that the low-frequency information from the sample is stronger that the high-frequency information. Thus, the addition of white noise affects the P-spectrum in the high-frequency regime more than in the low-frequency regime where the SNR is higher.

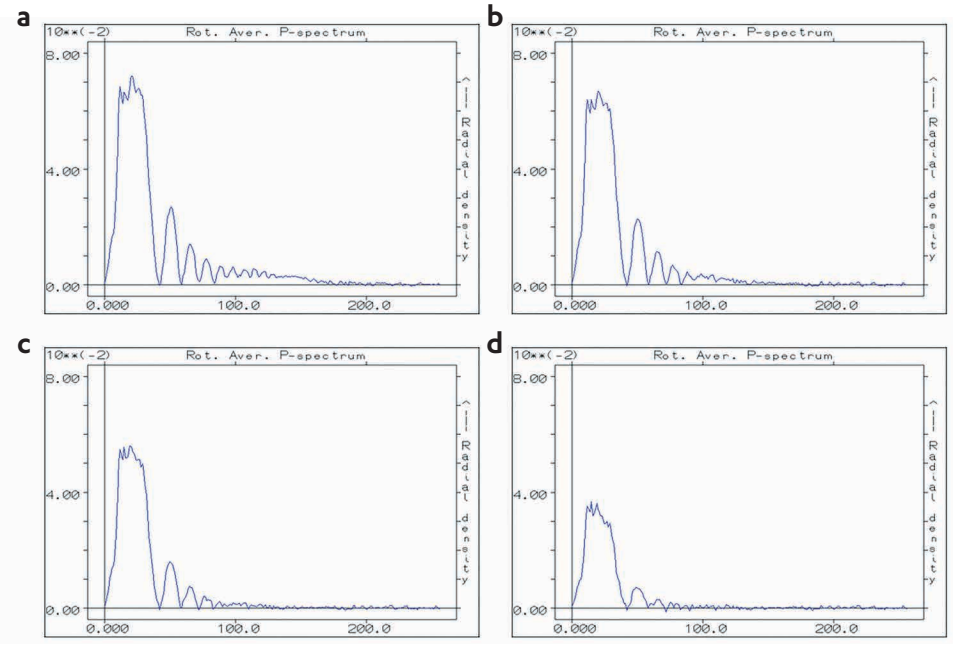

Figure 3. P-spectrum as function of SNR in the data. **(a)** Rotationally Averaged P-spectrum of the 7 movie-frames after mutual alignment (same as in Figure 2c). **(b)** RAP-spectrum of same aligned movie-sequence after adding random white noise with a sigma of unity (the raw movie frames (signal + noise) had also been normalised to a sigma of unity. **(c)** Same as (b) but with 2-sigma added noise. **(d)** Same as (b) but with 4-sigma added noise.

DISCUSSION

The P-spectrum removes the background ramp typically present in conventional power spectra (Mindell and Grigorieff 2003). What remains is the general decay of the CTF function caused by various physical effects – including the beam-induced motion – which is typically accounted for by a generic Gaussian low-pass filter (B-factor). Good movie alignments reduce this blurring contribution and can thus significantly improve the quality of the high-resolution data collection; the P-spectrum is a tool that may help monitoring this. The P-spectrum has a classical statistical basis: it is primarily a "square of the sum" of a set of measurements, divided by the "sum of the squares" of those measurements. The expectation value for this division when there is no consistent signal present (i.e. pure noise) is "1/N" where N is the number of movie-frames summed. We have included the subtraction of 1/N in the definition of the P-spectrum, which has the advantage that the P-spectrum thus has a zero expectation value in the absence of any signal. Note that a "signal" in this context means anything that is systematically present in all the frames we average. That means this is not just the signal associated with the molecules we are really interested in.

This zero expectation has the advantage that we can further integrate the P-spectrum over rings in Fourier space (RAP). If the number of independent pixels in each ring is "n" (with n

$\sim 2\pi|f|$), then the expected (normalized) standard deviation of the ring sum is $\sim\sqrt{n}$. We can therefore introduce a "3σ-threshold" for the RAP to better indicate to which resolution level the data collection is yielding significant information. (This is comparable to the FRC/FSC 3σ-threshold curve which indicates at which resolution level one is collecting significant information in a single-particle experiment (van Heel and Schatz 2005)). Note that with astigmatism present, the RAP may not perfectly represent the CTF minima and maxima but it would still reflect the level of information collected at that resolution. A certain amount of astigmatism in the data is advantageous to for achieving a good coverage of information over all spatial frequencies (van Heel et al. 2000).

The SSNR is defined as the power of signal (S^2) over the power of the noise (N^2) averaged over a ring in Fourier space (Unser et al. 1987) and this metric has very recently also been used in the context of movie alignments (Wang et al. 2014; Abrishami et al. 2015). One problem with the SSNR, however, is that the signal and the power are inseparably intertwined and thus approximations are required to reach realistic SSNR curves. Moreover, during the experiment part of the signal turns to noise due to radiation damage. Thus neither the signal nor the noise are constants but they change in the course of the experiment. The P-spectrum is obviously related to earlier metrics like the SSNR but one of its charms is that it represents a straightforward experimental result ("it is what is") rather than an estimate of some virtual entity.

The reason we prefer the P-spectrum over the Q-spectrum is that often, beyond the level of 2/3 of the Nyquist frequency, systematic interpolation/rounding-off errors make these measures start deviating from their expected zero-average. Because the P-spectrum is a squared metric (van Heel et al. 1992), whereas the Q-spectrum is not, the latter sometimes very visibly departs from the zero base-line whereas the P-spectrum remains well behaved (results not shown). This property facilitates automatic CTF-determination using P-spectra (Figure S2).

The processing time of a movie alignment depends on the size of the raw images and depending on the type of data the movie frames may be binned to say a 1/4, 1/16, or even 1/64 times the original size to speed the procedures. With clearly defined data, with a good SNR level, some 3-5 iterations suffice to achieve convergence whereas more iterations are required with poorer data. Our iterative algorithm is implemented in a coarse-grain MPI-parallelization environment (where each "core" performs the full alignment of a single movie), such that many movies for a large dataset may be aligned in parallel. For a direct feedback to the microscopy experiment we are considering a finer grain parallelization where the alignment of each movie frame is performed by an individual core.

When the goal of a data-collection session on an advanced cryo-EM instrument is achieving atomic-resolution structures, the data collection efficiency is of paramount importance, given its high cost. Long movies take more time to collect than individual integrated images and

that thus reduces the total number of particles imaged. Collecting many individual integrated images, increases the chance of harvesting good integrated particle images that do not require movie alignment (see for example: (Fischer et al. 2015)). This again depends on the stability of the sample the type of grids used, on the presence of thin carbon foil covering the holes, etc.

Conclusions

Our straightforward "P-spectrum" approach represents a convenient and direct way of assessing movie-mode cryo-EM data. The rotationally averaged P-spectrum can serve as an indicator of the level of information being collected per movie-micrograph. Our movie-alignment algorithm is aimed at the routine alignment of large, automatically collected datasets. We suggest that the P-spectrum and it rotationally averaged version will be a useful routine tools for optimising cryo-EM data acquisition. We note that often one can already obtain good spectra from movies by summing the spectra of individual unaligned movie frames (or of sums of consecutive movie frames).

Acknowledgements

We thank: Ralf Schmidt (Image Science GmbH, Berlin) for programming support; Bart Alewijnse (NeCEN, Leiden) for computing support, Peter Peters and Raimond Ravelli (Maastricht University) and Rodrigo Portugal (LNNano, Campinas) for discussions. We furthermore thank Rishi Matadeen and Sacha de Carlo for movie-mode data collection on our worm hemoglobin test sample. Our research was financed in part by: the Dutch ministry of economic affairs Cyttron II FES-0908; HTS&M Initiative: FES-0901; NanoNextNL of the Government of the Netherlands and 130 partners; from the BBSRC (Grant: BB/G015236/1); from the Netherlands Organization for Scientific Research (NWO grant: 016.072.321); the Brazilian science foundation: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-152746/2012-9 and CNPq-400796/2012-0). We acknowledge the use of NeCEN electron microscopes (Leiden University) funded by NWO and the European Regional Development Fund of the European Commission.

Author contributions statement

MvH: methodology; program development; data processing; paper writing. PA: data processing; paper writing; methodology. MS: program development, methodology.

REFERENCES

Abrishami, V., Vargas, J., Li, X., Cheng, Y., Marabini, R., Sorzano, C.Ó.S. and Carazo, J.M. (2015) Alignment of direct detection device micrographs using a robust Optical Flow approach. J Struct Biol 189(3): 163-176.

Afanasyev, P., Ravelli, R.B., Matadeen, R., De Carlo, S., van Duinen, G., Alewijnse, B., Peters, P.J., Abrahams, J.P., Portugal, R.V., Schatz, M. and van Heel, M. (2015) A posteriori correction of camera characteristics from large image data sets. Sci Rep 5: 10317.

Allegretti, M., Mills, D.J., McMullan, G., Kühlbrandt, W. and Vonck, J. (2014) Atomic model of the F420-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector. Elife 3: e01963.

Amunts, A., Brown, A., Bai, X.C., Llacer, J.L., Hussain, T., Emsley, P., Long, F., Murshudov, G., assalski,

Scheres, S.H. and Ramakrishnan, V. (2014) Structure of the yeast mitochondrial large ribosomal subunit. Science 343(6178): 1485-1489.

Bai, X.C., Fernandez, I.S., McMullan, G. and Scheres, S.H. (2013) Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. Elife 2: e00461.

Bartesaghi, A., Merk, A., Banerjee, S., Matthies, D., Wu, X., Milne, J.L.S. and Subramaniam, S. (2015) Electron microscopy. 2.2 A resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor. Science 348(6239): 1147-1151.

Brilot, A.F., Chen, J.Z., Cheng, A., Pan, J., Harrison, S.C., Potter, C.S., Carragher, B., Henderson, R. and Grigorieff, N. (2012) Beam-induced motion of vitrified specimen on holey carbon film. J Struct Biol 177(3): 630-637.

Campbell, M.G., Cheng, A., Brilot, A.F., Moeller, A., Lyumkis, D., Veesler, D., Pan, J., Harrison, S.C., Potter, C.S., Carragher, B. and Grigorieff, N. (2012) Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. Structure 20(11): 1823-1828.

Campbell, M.G., Veesler, D., Cheng, A., Potter, C.S. and Carragher, B. (2015) 2.8 Å resolution reconstruction of the Thermoplasma acidophilum 20 S proteasome using cryo-electron microscopy.

Faruqi, A.R. and McMullan, G. (2011) Electronic detectors for electron microscopy. Quarterly reviews of biophysics 44(3): 357-390.

Fischer, N., Neumann, P., Konevega, A.L., Bock, L.V., Ficner, R., Rodnina, M.V. and Stark, H. (2015) Structure of the E. coli ribosome-EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. Nature 520(7548): 567-570.

Grant, T. and Grigorieff, N. (2015) Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. Elife.

Grigorieff, N. (2013) Direct detection pays off for electron cryo-microscopy. Elife 2: e00573.

Harauz, G. and van Heel, M. (1986) Exact filters for general geometry three dimensional reconstruction. Optik 73: 146-156.

Herrmann, K.H. and Krahl, D. (1982) The detection quantum efficiency of electronic image recording systems. J Microsc 127(1): 17-28.

Kühlbrandt, W. (2014) Biochemistry. The resolution revolution. Science 343(6178): 1443-1444.

Kunath, W., Weiss, K., Sackkongehl, H., Kessel, M. and Zeitler, E. (1984) Time-resolved low-dose microscopy of glutamine-synthetase molecules. Ultramicroscopy 13(3): 241-252.

Li, X., Mooney, P., Zheng, S., Booth, C.R., Braunfeld, M.B., Gubbens, S., Agard, D.A. and Cheng, Y. (2013) Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat Methods 10(6): 584-590.

McMullan, G., Faruqi, A.R., Clare, D. and Henderson, R. (2014) Comparison of optimal performance at 300keV of three direct electron detectors for use in low dose electron microscopy. Ultramicroscopy 147: 156-163.

Meyer, R.R. and Kirkland, A.I. (2000) Characterisation of the signal and noise transfer of CCD cameras for electron detection. Microscopy Research and Technique 49(3): 269-280.

Mindell, J.A. and Grigorieff, N. (2003) Accurate determination of local defocus and specimen tilt in electron microscopy. J Struct Biol 142(3): 334-347.

Nejadasl, F.K., Karuppasamy, M., Koster, A.J. and Ravelli, R.B.G. (2011) Defocus estimation from stroboscopic cryo-electron microscopy data. Ultramicroscopy 111(11): 1592-1598.

Saß, H., Büldt, G., Beckmann, E., Zemlin, F., Van Heel, M., Zeitler, E., Rosenbusch, J., Dorset, D. and Massalski, A. (1989) Densely packed β-structure at the protein-lipid interface of porin is revealed by high-resolution cryo-electron microscopy. J Mol Biol 209(1): 171-175.

Unser, M., Trus, B.L. and Steven, A.C. (1987) A new resolution criterion based on spectral signal-to-noise ratios. Ultramicroscopy 23(1): 39-51.

van Heel, M., Gowen, B., Matadeen, R., Orlova, E.V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M. and Patwardhan, A. (2000) Single-particle electron cryo-microscopy: towards atomic resolution. Quarterly reviews of biophysics 33(4): 307-369.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R. and Schatz, M. (1996) A new generation of the IMAGIC image processing system. J Struct Biol 116(1): 17-24.

van Heel, M. and Hollenberg, J. (1980) On the stretching of distorted images of two-dimensional crystals. In Electron microscopy at molecular dimensions (Springer), 256-260.

van Heel, M., Portugal, R., Rohou, A., Linnemayr, C., Bebeacua, C., Schmidt, R., Grant, T. and Schatz, M. (2012) Four-dimensional cryo electron microscopy at quasi atomic resolution: IMAGIC 4D. International Tables for Crystallography F: 624-628.

van Heel, M. and Schatz, M. (2005) Fourier shell correlation threshold criteria. J Struct Biol 151(3): 250-262.

van Heel, M., Schatz, M. and Orlova, E. (1992) Correlation functions revisited. Ultramicroscopy 46(1): 307-316.

Veesler, D., Campbell, M.G., Cheng, A., Fu, C.Y., Murez, Z., Johnson, J.E., Potter, C.S. and Carragher, B. (2013a) Maximizing the potential of electron cryomicroscopy data collected using direct detectors. J Struct Biol 184(2): 193-202.

Veesler, D., Ng, T.-S., Sendamarai, A.K., Eilers, B.J., Lawrence, C.M., Lok, S.-M., Young, M.J., Johnson, J.E. and Fu, C. (2013b) Atomic structure of the 75 MDa extremophile Sulfolobus turreted icosahedral virus determined by CryoEM and X-ray crystallography. Proceedings of the National Academy of Sciences 110(14): 5504-5509.

Wang, Z., Hryc, C.F., Bammes, B., Afonine, P.V., Jakana, J., Chen, D.H., Liu, X., Baker, M.L., Kao, C., Ludtke, S.J., Schmid, M.F., Adams, P.D. and Chiu, W. (2014) An atomic model of brome mosaic virus using direct electron detection and real-space optimization. Nat Commun 5: 4808.

SUPPLEMENTARY FIGURES

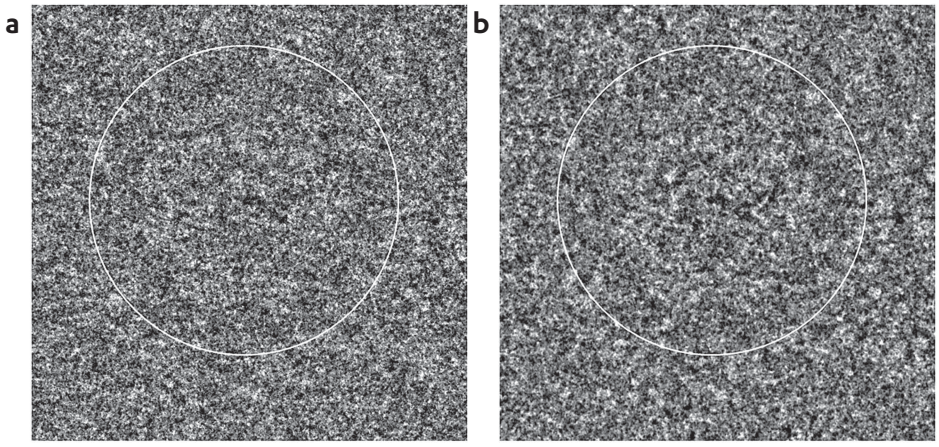

Figure S1. Effect of the movie alignments on the image and on CTF-determination based on the P-spectrum. **(a)** 512x512 patch from the unaligned movie sum (used in Figure 1a). **(b)** Same patch from the aligned movie sum. The top view in the center of this patch is marked by a white circle.

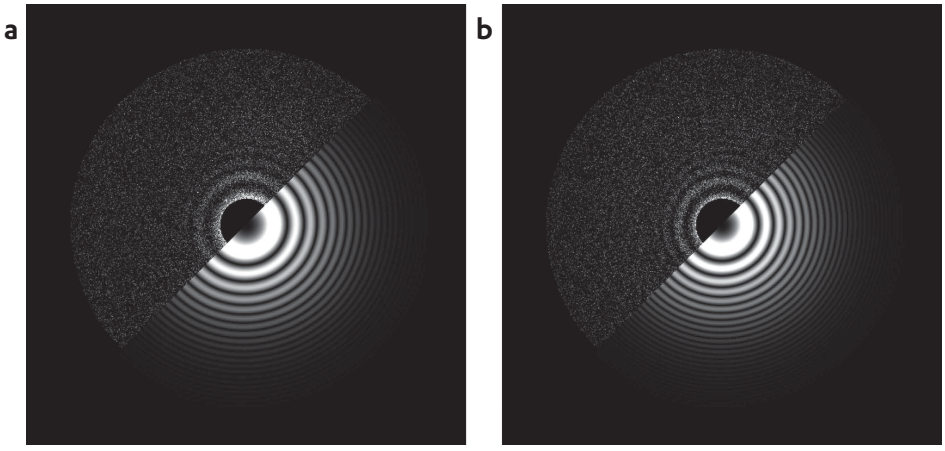

Figure S2. CTF-determination based on the P-spectrum. **(a)** Half-half image of the CTF-fit (top-left: experimental P-spectrum; bottom right fitted theoretical spectrum), based on P-spectrum of the movie used in Figure 1b (defocus: 0.790/0.736 µm). **(b)** Half-half image of the CTF-fit, based on P-spectrum of the movie used in Figure 2b (defocus: 0.974/0.936 µm). The Nyquist frequency (edge of the spectra) is at 1/2.22 Å.

CHAPTER 4

CAN 670,000 PARTICLES BE EXTRACTED FROM THE EMPIAR 10003 FULL DATA SET?

Pavel Afanasyev[1,2] and Marin van Heel[1,3,4]*

In preparation; partially in PNAS. 2013 (110), E4175-4177

[1]Leiden Institute of Biology, Leiden University, 2333 CC Leiden, The Netherlands

[2]The Maastricht Multimodal Molecular Imaging Institute, Maastricht University, 6229 ER Maastricht, The Netherlands

[3]Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, UK

[4]Brazilian Nanotechnology National Laboratory – LNNano, CNPEM, C.P. 6192, 13083-970 Campinas SP, Brazil

*Corresponding author

ABSTRACT

Single-particle cryogenic electron microscopy is an increasingly important technique in structural biology. However, various pitfalls exist during the processing of cryo-EM data towards high-resolution results. One such possible pitfall is the selection of individual particle images from the raw data with predetermined reference images ("particle picking"). This type of particle picking is equivalent to the alignment of individual particle images with respect to reference images and this process can suffer from reference bias. A controversy emerged relating to 3D-reconstructions of the HIV-1 Env trimer presented by the group of Sodroski (Mao et al. 2012; Mao et al. 2013b). These results have been questioned by three different authors (Henderson 2013; Subramaniam 2013; van Heel 2013). Very recently, in July 2015, the group deposited the full dataset they used for solving the HIV-1 Env trimer in the publically accessible EMPIAR data base. We here assess the properties of this newly deposited dataset. We have, unfortunately, not been able to find any recognizable trimer particles in the dataset. Even if we had found usable particles, the limited usable vitreous-ice areas in the full data set could - under ideal circumstances - not have yielded more than 50,000 particles whilst 670,000 had been deposited previously in EMPIAR.

INTRODUCTION

Over the last decade, cryogenic electron microscopy has become a new front-end technique in structural biology studies thanks to new instrumentation developments (Kühlbrandt 2014; Bai et al. 2015; Cheng 2015). Cryo-EM has recently been applied to many "hot issues" of molecular biology and medicine, where the structural data is essential for the fundamental understanding of the cellular processes at the molecular level (Amunts et al. 2014; Liao et al. 2014; Bai et al. 2015; Campbell et al. 2015). In particular, recent studies of the envelope glycoprotein (Env) trimer of human immunodeficiency virus HIV-1 (Liu et al. 2008; Tran et al. 2012; Bartesaghi et al. 2013; Julien et al. 2013; Lyumkis et al. 2013) have brought new insights on its architecture. HIV infects more than 2 million people a year (Barre-Sinoussi et al. 2013) and remains one of the major challenges for rational drug design and vaccine developments. One of the most promising strategies is to design an HIV vaccine, would be to stimulate the production of broadly neutralizing antibodies against HIV (Nabel et al. 2011; Earl et al. 2013). This can be done based on the knowledge of the structure of the Env trimer, composed of gp120/gp41 heterodimers. Therefore, the structural results of Env trimer, obtained by cryo-EM, are of great importance for the whole HIV field.

The first low-resolution structure of the Env trimer by cryo-EM was obtained already in 2008 (Liu et al. 2008) based on an electron tomographic study. The controversy emerged with the publications of cryo-EM structures of the HIV-1 Env trimer at 11 Å (Mao et al. 2012; Mao et al. 2013b), followed by a structure at 6 Å resolution (Mao et al. 2012; Mao et al. 2013b), which have

become the subject of serious criticism (Henderson 2013; Subramaniam 2013; van Heel 2013). The reported structures, according to the authors, were the result of two 3D-reconstructions from the same dataset, collected in 2009, according to their EMDB deposition EMD-5447 (EMDB 2015)). The experts in the single-particle cryo-EM, questioning the work, state that for the particle picking, Mao et al might have used reference-images of the desired structure, which were not resulting from the data (Henderson 2013; Subramaniam 2013; van Heel 2013). Moreover, in those two published papers (Mao et al. 2012; Mao et al. 2013b), there was no direct visual evidence of the presence any Env trimer particles in the presented micrographs (Henderson 2013; Subramaniam 2013; van Heel 2013).

Mao & Sodroski have deposited the 670,000 selected particles they used in the EMPIAR data base (as EMPIAR 10007) in March 2014. That dataset (deposited with no further details) we analyzed separately (results not shown). We concluded that these particles had been selected from the micrographs using masked reference images. From the classification of those images we were able to directly re-create the reference images used for particle picking. Interestingly the 670,000 particles dataset was remarkably clean in the sense that it contained no junk particles (like ice crystals) which normally are picked with preference due to their high contrast. In a separate report we had concluded that that dataset was the result of particle picking using the 2D projections of a pre-existing 3D model. No particle coordinates and no raw micrographs were deposited at the time (in March 2014) as had been requested by their critics (in October 2013).

More recently (in July 2015) the authors deposited the full dataset of raw micrographs which they allegedly used as input for the processing of both papers (Mao et al. 2012; Mao et al. 2013b). The full data set was again deposited without any additional explanation, and still without the requested list of coordinates describing from which images and which positions the 670,000 individual particles had been extracted. The earlier deposition of 10 defocus pairs by Mao & Sodrosky (EMPIAR-10003; deposited 23-08-2013) contained less than ~0.1% of the micrographs used (only 10 micrographs had been deposited from an expected number of 20,000 based on their published "typical" 4096x4096 micrographs). Here we analyze this newly deposited full dataset and draw conclusions on the viability of generating high-resolution reconstructions from that data.

RESULTS

We downloaded the "full" EMPIAR-10003 dataset, consisting of 4683 focal-pairs, the full dataset allegedly used for picking 670,000 particles resulting in the two publications (Mao et al. 2012; Mao et al. 2013b). Note that it had been reported earlier, that a total of 5991 micrographs were used to obtain their 3D-reconstructions (EMDB 2015). We used the 4683 images large-defocus dataset for a first visual assessment of the data set. We found the dataset to contain a

large number of unusable micrographs with contamination of different sources (Figures 1,2). We also found that the micrographs in the dataset were not unprocessed as was suggested. The images had been collected on a Gatan CCD camera which would produce images in the Gatan proprietary "DM" data format, but the data were deposited in the "Spider" format (Shaikh et al. 2008). The intensities were apparently inverted (the "particles" are dark against a light background). Whereas a majority had positive densities, a significant number of images had very negative values. The inversion of the densities was apparently performed with respect to the average density of each image sometimes causing negative densities in the images and thus complicating their interpretation. A significant number of images from the close-to-focus dataset had negative average densities whereas the far-from-focus images all had positive averages. The headers of the micrographs indicated, that the pairs (of Spider images) were created in July 2015 (a few days prior to their deposition) whereas the original dataset was collected in September 2009 (EMDB 2015). This new deposition is still lacking the requested particle coordinates.

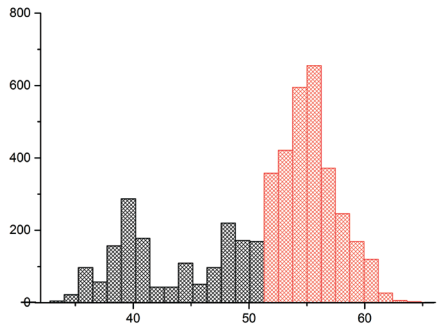

Figure 1. Histogram of the standard deviations of the large defocus images (4683). The area marked in bold corresponds primarily to carbon-foil images or otherwise heavily contaminated images that are not usable for processing.

Statistical analysis of the deposited micrographs showed that the "semi-automatically" acquired micrographs are not of uniform quality. Of the micrographs that were not contaminated, some were collected largely or fully on the carbon support (Figures 1,2). Particles picked from thick carbon support areas should not be used in image processing due to the large noise contribution from the supporting carbon foil. Having some carbon-foil visible in a corner of a micrograph in micrographs actually helps the visualization of Thon rings and thus helps the CTF-determination. However, in most micrographs the carbon support fully dominated the vitreous-ice area (if even present). Based on the analysis of the power spectrum (Figure 3) and visual inspection of the micrographs, we found that less than 1700 micrographs (~35% of the dataset), have enough vitreous ice to be usable for the data processing, in principle. We mention "in principle" because this assumes that usable trimer particles are indeed embedded in the vitreous ice. About 3000 micrographs out of the 4683 are not usable.

It can be difficult to distinguish whether a micrograph is taken from a vitreous ice area or completely on carbon support (Figure 3). A good indicator of the data acquisition on the carbon support is a large number of visible Thon rings on the amplitude spectrum of the

micrograph. Figure 3 contains such a comparison of the micrograph with small amount of carbon (Figure 3a) and the one, fully taken on carbon (Figure 3b). Their amplitude spectra, presented below on (Figure 3c,d), differ in the visibility of Thon rings. Micrograph at the Figure 3a contains a bit of carbon-foil at the bottom (which area has the same texture as all of Figure 3b).

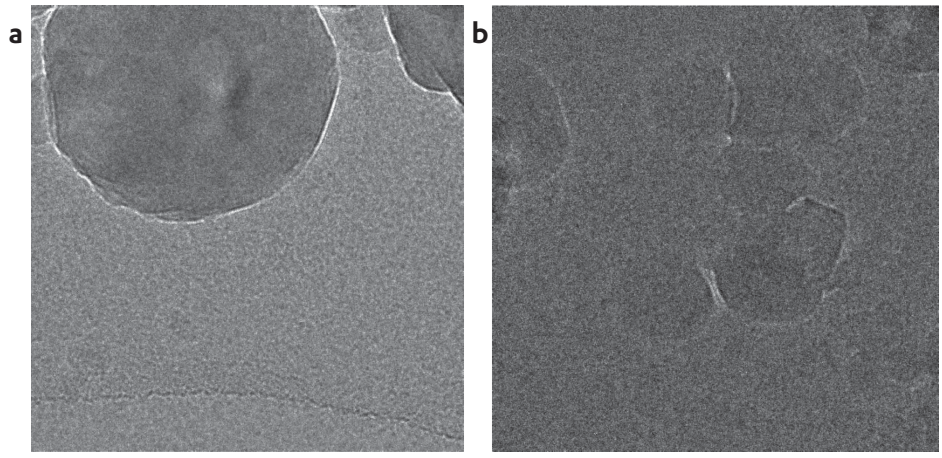

Figure 2. Examples of typical unusable micrographs from the full EMPIAR-10003 dataset. **(a)** Micrograph, contains a huge ice particle. Most of the micrograph area is on carbon foil; its edge is seen as a continuous line at the bottom of the image. This image is thus not usable for processing **(b)** The whole micrograph is taken of a thick ice layer; with contaminants and cannot be used for further processing. These micrographs are displayed with the same grayscale scales, illustrating that the sample in (b) is too thick.

With only 1700 usable micrographs, the particle density per micrograph must then be: 670,000/1700 ≈ 400 particles (of 256x256 pixels) per micrograph. A micrograph of 4096x4096 pixels cannot accommodate this amount of particles even without overlap. In a space-filling checkerboard arrangement one could fit 16x16 = 256 frames of each 256x256 in a 4096x4096 image. We have only been able to pick a maximum of ~145 randomly distributed particles per micrograph (Figure 4).

In the EMPIAR-10003 full dataset, we identified the original micrographs used in the Figure S1 of the Mao et al., 2013 paper (Mao et al. 2013b). In those two "typical" micrographs only 31 and 32 detected particles per micrograph were highlighted. Those two and other such "good" micrographs (for example, Figure 3a) did not seem to contain any particles of the size of the HIV trimer. Those particles are relatively big and should be visible at these high defocus levels (Henderson 2013; Subramaniam 2013). With ~31 particles/micrograph throughout the dataset one would have been able to pick 31*4683≈145,000 particles, assuming all the micrographs were usable. To achieve a total of 670,000 particles, the concentration of oligomers on a typical micrograph would have to be ~4.5 times higher (Figure 4). At such a high particles concentration one would in practice expect to see some clumps. The deposited data does not show any visible particles nor does it show any particle clusters.

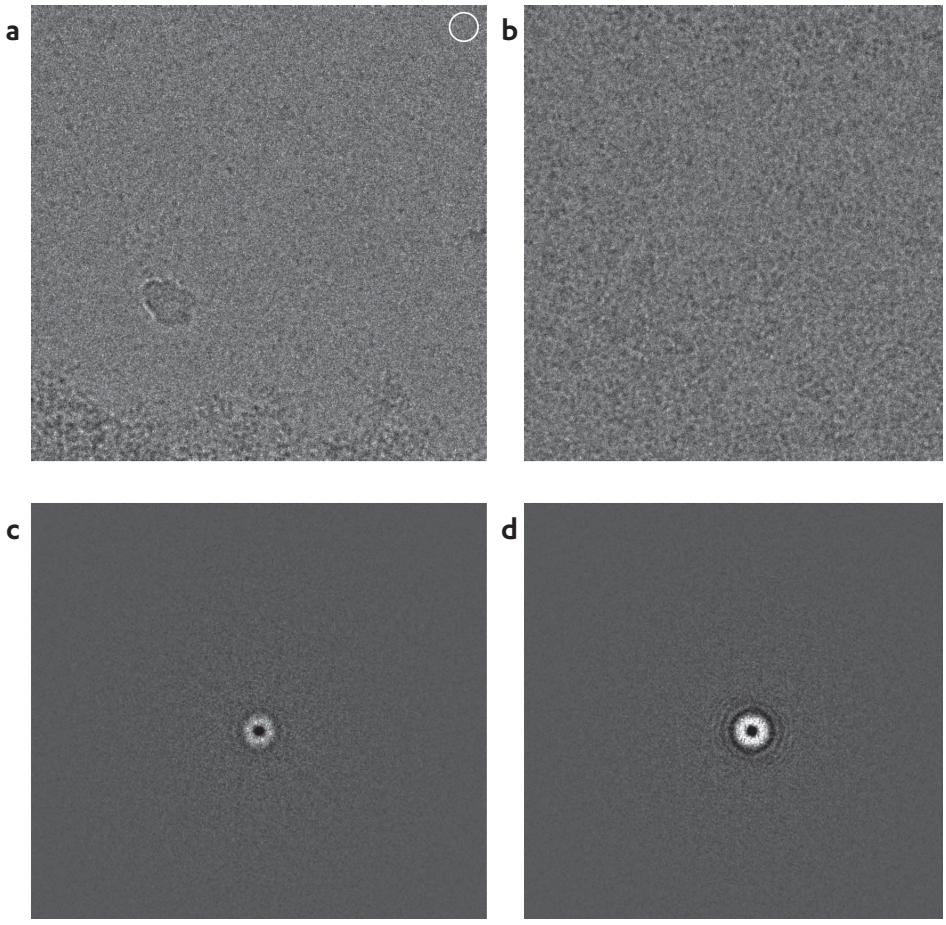

Figure 3. Example of far-from-focus micrographs, taken from a vitreous-ice area **(a)** and from a carbon-foil area **(b)**. The size of the ring at the top right corner in (a) corresponds to the expected size of the Env trimer. One can clearly see the edge of the carbon support at the bottom of the micrograph in (a). Note, that the same texture is seen everywhere in (b). The central part (2048x2048) of the amplitude spectra of images (a) and (b) are shown in **(c)** and **(d)** respectively. Carbon support film yields more Thon rings because its average contrast is higher than that of vitreous ice. The first CTF zero corresponds to ~1/20Å in these images.

DISCUSSION AND CONCLUSIONS

The EMPIAR archive (established in 2013) of the Electron Microscopy Databank (EBI-EMDB) allows free deposition, storage and download off full EM datasets. Thanks to this service, it is now possible to share large raw data sets in cryo-EM. The public debate around the Mao & Sodrosky structures started with the publication of three critical papers (Henderson 2013; Subramaniam 2013; van Heel 2013), together with the rebuttal of Mao and co-workers (Mao et al. 2013a). The main criticisms addressed the image processing, and especially the reference-biased particle picking. All critics explicitly asked for the raw data of Mao & Sodroski to be deposited in a publically accessible database (the new EMPIAR archive is tailored for that task).

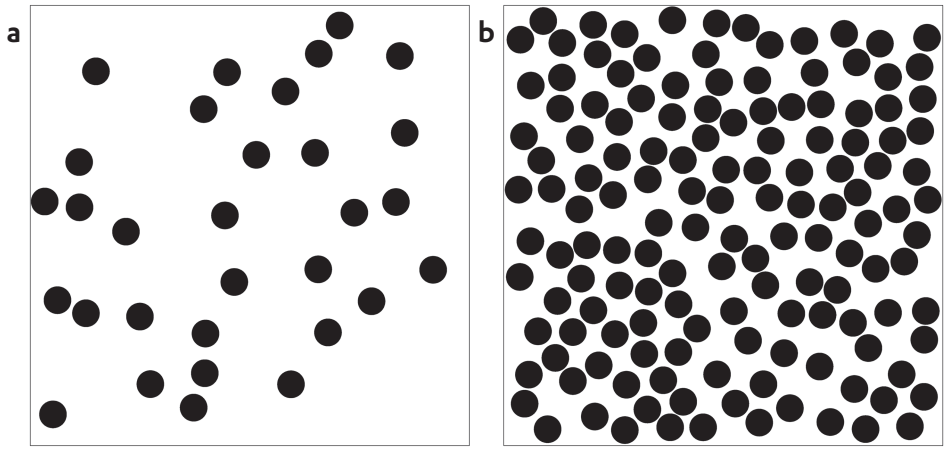

Figure 4. Simulation of different concentrations of the HIV-1 trimer particles per micrograph. **(a)** 31 particles of the size of the HIV trimer (diameter of 256 pixels) per a 4096x4096 micrograph like in the Figure S1 in Mao et al., 2013 (Mao et al. 2013b). **(b)** Concentration of 143 non-overlapping particles per micrograph, necessary to pick 670,000 particles from 4683 micrographs. Circles represent non-overlapping particle areas like in Mao et al., 2013 (Mao et al. 2013b). It proved impossible to place the required ~400 non-overlapping particles in a single 4096x4096 micrograph.

We have since seen the uncommented deposition of 670,000 particles in March 2014, and recently (July 2015), the also uncommented deposition of the full raw dataset which we focus on here. Independent of the continuing "Mao & Sodrosky" controversy, research efforts by three other independent groups have since resulted in high-resolution structures of the HIV trimer (Bartesaghi et al. 2013; Julien et al. 2013; Lyumkis et al. 2013). These three newly published structures are all in good agreement with each other but not with the results of Mao and co-workers.

Summarizing our findings:

1. The full dataset EMPIAR 10003 contains no visually discernable HIV Env trimers.
2. The data set does not contain a sufficiently large vitreous ice area to allow picking of more than 50,000 particles of the size of the HIV Env trimer.
3. The huge majority of the 670,000 particles deposited in EMPIAR 10007 (March 2014) could *not* have been selected from the EMPIAR 10003 (July 2015) full dataset by cryo-EM standards.The EMPIAR 10007 (March 2014) dataset was generated by explicitly searching for 2D projections of a preconceived 3D structure as suggested by their critics. For results generated by "reference bias" the nature of the input micrographs is not relevant (van Heel 2013).
4. We note that Mao & Sodrosky have still not deposited the full dataset including the co-ordinates of the picked as requested by their critics.
5. We note that their (uncommented) deposition of the selected particles in March 2014 as well as their current (uncommented) deposition of the full raw dataset (July 2015) have

caused significant delays in clarifying the scientific issues at hand.

6. We note that our research over the course of more than two years was performed without any specific funding to study these HIV/AIDS–related issues.
7. The strongly-worded rebuttal to their critics by Mao & Sodrosky (Mao et al. 2013a) has yet not been challenged publically. Our findings render any further discussion on the matter superfluous.

REFERENCES

Amunts, A., Brown, A., Bai, X.C., Llacer, J.L., Hussain, T., Emsley, P., Long, F., Murshudov, G., Scheres, S.H. and Ramakrishnan, V. (2014) Structure of the yeast mitochondrial large ribosomal subunit. Science 343(6178): 1485-1489.

Bai, X.C., McMullan, G. and Scheres, S.H. (2015) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40(1): 49-57.

Barre-Sinoussi, F., Ross, A.L. and Delfraissy, J.F. (2013) Past, present and future: 30 years of HIV research. Nat Rev Microbiol 11(12): 877-883.

Bartesaghi, A., Merk, A., Borgnia, M.J., Milne, J.L. and Subramaniam, S. (2013) Prefusion structure of trimeric HIV-1 envelope glycoprotein determined by cryo-electron microscopy. Nat Struct Mol Biol 20(12): 1352-1357.

Campbell, M.G., Veesler, D., Cheng, A., Potter, C.S. and Carragher, B. (2015) 2.8 Å resolution reconstruction of the Thermoplasma acidophilum 20 S proteasome using cryo-electron microscopy.

Cheng, Y. (2015) Single-particle cryo-EM at crystallographic resolution. Cell 161(3): 450-457.

Earl, L.A., Lifson, J.D. and Subramaniam, S. (2013) Catching HIV 'in the act' with 3D electron microscopy. Trends in microbiology 21(8): 397-404.

EMDB. (2015). Experimental metadata (xml) of EMD-5447. Retrieved 2 Sept, 2015, from ftp://ftp.ebi.ac.uk/pub/databases/emdb/structures/EMD-5447/header/emd-5447.xml.

Henderson, R. (2013) Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. Proc Natl Acad Sci U S A 110(45): 18037-18041.

Julien, J.-P., Cupo, A., Sok, D., Stanfield, R.L., Lyumkis, D., Deller, M.C., Klasse, P.-J., Burton, D.R., Sanders, R.W., Moore, J.P., Ward, A.B. and Wilson, I.A. (2013) Crystal structure of a soluble cleaved HIV-1 envelope trimer. Science 342(6165): 1477-1483.

Kühlbrandt, W. (2014) Biochemistry. The resolution revolution. Science 343(6178): 1443-1444.

Liao, M., Cao, E., Julius, D. and Cheng, Y. (2014) Single particle electron cryo-microscopy of a mammalian ion channel. Current opinion in structural biology 27: 1-7.

Liu, J., Bartesaghi, A., Borgnia, M.J., Sapiro, G. and Subramaniam, S. (2008) Molecular architecture of native HIV-1 gp120 trimers. Nature 455(7209): 109-113.

Lyumkis, D., Julien, J.-P., de Val, N., Cupo, A., Potter, C.S., Klasse, P.-J., Burton, D.R., Sanders, R.W., Moore, J.P., Carragher, B., Wilson, I.A. and Ward, A.B. (2013) Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. Science 342(6165): 1484-1490.

Mao, Y., Castillo-Menendez, L.R. and Sodroski, J.G. (2013a) Reply to Subramaniam, van Heel, and Henderson: Validity of the cryo-electron microscopy structures of the HIV-1 envelope glycoprotein complex. Proceedings of the National Academy of Sciences 110(45): E4178-E4182.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Desormeaux, A., Finzi, A., Xiang, S.H. and Sodroski, J.G. (2013b) Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. Proc Natl

Acad Sci U S A 110(30): 12438-12443.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Xiang, S.H., Haim, H., Yang, X. and Sodroski, J. (2012) Subunit organization of the membrane-bound HIV-1 envelope glycoprotein trimer. Nat Struct Mol Biol 19(9): 893-899.

Nabel, G.J., Kwong, P.D. and Mascola, J.R. (2011) Progress in the rational design of an AIDS vaccine. Philos Trans R Soc Lond B Biol Sci 366(1579): 2759-2765.

Shaikh, T.R., Gao, H., Baxter, W.T., Asturias, F.J., Boisset, N., Leith, A. and Frank, J. (2008) SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nat Protoc 3(12): 1941-1974.

Subramaniam, S. (2013) Structure of trimeric HIV-1 envelope glycoproteins. Proc Natl Acad Sci U S A 110(45): E4172-4174.

Tran, E.E.H., Borgnia, M.J., Kuybeda, O., Schauder, D.M., Bartesaghi, A., Frank, G.A., Sapiro, G., Milne, J.L.S. and Subramaniam, S. (2012) Structural mechanism of trimeric HIV-1 envelope glycoprotein activation. PLoS Pathog 8(7): e1002797.

van Heel, M. (2013) Finding trimeric HIV-1 envelope glycoproteins in random noise. Proc Natl Acad Sci U S A 110(45): E4175-4177.

SUPPLEMENTARY INFORMATION

Env trimer structure out of noise

Reference bias introduced at the particle picking stage can easily lead to a desirable 3D-reconstruction, even independent on the original imaged data. To illustrate the possibility of obtaining a 3D-reconstruction of the HIV-1 Env trimer even from the "particles" picked out of pure random noise, we performed an experiment, encompassing Mao's pipeline of image processing. We picked 670,000 particles from artificially generated random white noise micrographs, using 50 forward projections from the 11 Å resolution HIV-1 gp160 protein structure (EMD-5418) as the references for correlation particle picking. The particle picking was performed in the PICK-M-ALL program in IMAGIC-4D software (van Heel et al. 2012). Particles were classified without any alignments in 10,000 classes, 100 of which were randomly extracted and used for angular reconstitution. The unrefined 3D reconstruction, obtained from those classes without any alignments (Figure S1), yielded a 13-Å cross-resolution (van Heel and Schatz 2005), which is even better than the ~18 Å found between the two published structures by Mao et al. (Mao et al. 2012; Mao et al. 2013), although the latter is supposed to be a refined version of the former (van Heel 2013). Those two deposited maps are entirely uncorrelated at ~11-Å resolution (Fourier shell correlation ~ 0). This shows, that the particles for the 6-Å map could be picked/aligned with other reference images and not with those derived from the 11-Å map, as stated.

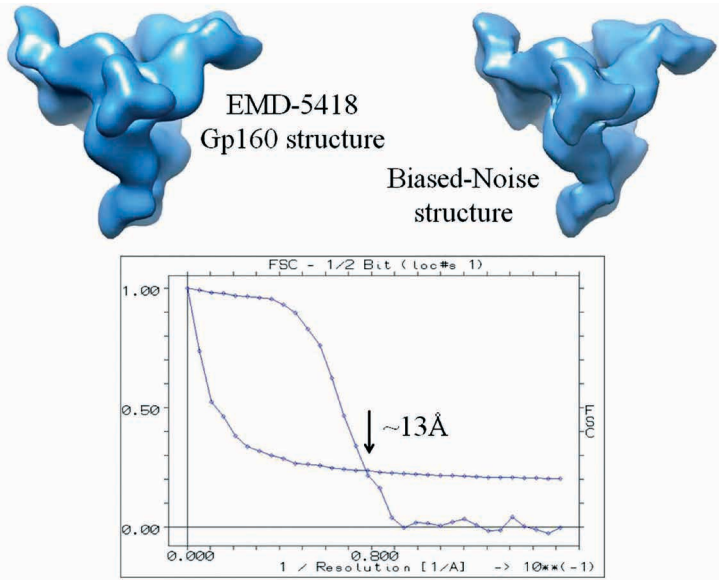

Figure S1. Generating an HIV-1 Env trimer from random noise. EMD-5418 map compared to the 3D-reconstruction, based on the "particles", picked out of noise. The cross-resolution between the biased noise map and the gp160 search structure is 13 Å.

Reference-free particle picking

To avoid any reference bias in particle picking, we present here an approach for the automatic reference-free particle picking, implemented in our IMAGIC-4D software package (van Heel et al. 2012). Purely reference-free automatic particle picking approach has already been suggested in 1982 (van Heel 1982), where the local variance particle picking was used. In fact, in variance particle picking everything significant (has a higher variance than the background) is being picked without any a priori knowledge about the shape of the particle. Its algorithm is described in detail in (van Heel 1982) and it is implemented in IMAGIC-4D PICK-M-ALL program. The local variance in an image, calculated over a convolution area $A(\vec{r})$ ("variance area"), is determined by:

$$Var_A(\vec{r}) = I^2(\vec{r}) \otimes \frac{A(\vec{r})}{N} - \left(I(\vec{r}) \otimes \frac{A(\vec{r})}{N} \right)^2 \quad (1),$$

where N is a number of pixels inside $A(\vec{r})$. As a new option in this program, modulation particle picking can be used, where instead of the local variance, local modulation is calculated:

$$Mod_A(\vec{r}) = \left| I(\vec{r}) \right| \otimes \frac{A(\vec{r})}{N} \quad (2),$$

which avoids squaring amplitudes in calculations.

This approach assumes picking everything what has a high local variance, therefore a lot of "false-positives" are to be expected: carbon edges, ice particles or any other possible junk. Those are quite easy to get rid of based on the histograms of modulation image densities and of the peak height. If the sample does not contain much high-contrast junk and, one could perform eigenimage analysis and classification of the picked particles.

The class averages will represent different low-resolution views of the particles. If the statistics of the dataset and the quality of the sample preparation are satisfactory, already at this stage one could create an initial 3D-reconstruction using angular reconstitution. Note, that this way, without applying any alignments, one could create purely reference-free 3D-reconstruction. If the quality of the class averages does not allow one to create a reliable 3D-reconstruction, the best rotationally averaged classes might be used as templates for the next round of particle picking. Those templates will only come from the data itself and nowhere else. The false-positive particles picked by using this approach can be easily sorted out at the later stage of the eigenimage analysis (for details see Chapter 5).

REFERENCES

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Desormeaux, A., Finzi, A., Xiang, S.H. and Sodroski, J.G. (2013) Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. Proc Natl Acad Sci U S A 110(30): 12438-12443.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Xiang, S.H., Haim, H., Yang, X. and Sodroski, J. (2012) Subunit organization of the membrane-bound HIV-1 envelope glycoprotein trimer. Nat Struct

Mol Biol 19(9): 893-899.

van Heel, M. (1982) Detection of objects in quantum-noise-limited images. Ultramicroscopy 7(4): 331-341.

van Heel, M. (2013) Finding trimeric HIV-1 envelope glycoproteins in random noise. Proc Natl Acad Sci U S A 110(45): E4175-4177.

van Heel, M., Portugal, R., Rohou, A., Linnemayr, C., Bebeacua, C., Schmidt, R., Grant, T. and Schatz, M. (2012) Four-dimensional cryo electron microscopy at quasi atomic resolution: IMAGIC 4D. International Tables for Crystallography F: 624-628.

van Heel, M. and Schatz, M. (2005) Fourier shell correlation threshold criteria. J Struct Biol 151(3): 250-262.

CHAPTER 5

SINGLE-PARTICLE CRYO-EM BASED ON ALIGNMENT BY CLASSIFICATION (ABC): *LUMBRICUS TERRESTRIS* HEMOGLOBIN AT NEAR-ATOMIC RESOLUTION

Pavel Afanasyev[1,2†], Charlotte Linnemayr-Seer[3,4†], Raimond B.G. Ravelli[2,5], Rishi Matadeen[5], Sacha De Carlo[5,6], Bart Alewijnse[5], Rodrigo V. Portugal[7], Navraj S. Pannu[8], Michael Schatz[9], Marin van Heel[1,4,7*]

In preparation

[1]Institute of Biology Leiden, Leiden University, 2333 CC Leiden, The Netherlands.

[2]The Maastricht Multimodal Molecular Imaging Institute, Maastricht University, 6229 ER Maastricht, The Netherlands

[3]Division of Internal Medicine, Inflammation Research, University Hospital Zürich, Switzerland

[4]Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, UK.

[5]Netherlands Centre for Electron Nanoscopy (NeCEN), 2333 CC Leiden, The Netherlands.

[6]FEI Company, 5651 GG Eindhoven, The Netherlands.

[7]Brazilian Nanotechnology National Laboratory – LNNano, CNPEM, C.P. 6192, 13083-970 Campinas SP, Brasil.

[8]Biophysical Structural Chemistry, Leiden University, 2300 RA Leiden, The Netherlands,

[9]Image Science Software GmbH, D-14193 Berlin, Germany.

[†]These authors contributed equally to this work

[*]Corresponding author

ABSTRACT

Single-particle cryogenic electron microscopy (cryo-EM) can now yield near-atomic resolution structures of biological complexes. The very low signal-to-noise ratio raw data in cryo-EM demands a robust, unbiased approach for elucidating the structural information contained in the micrographs. We here present a detailed reference-free pipeline for obtaining multiple three-dimensional (3D) reconstructions from heterogeneous populations ("4D"), with near-atomic resolution. The methodologies used in this pipeline include: *a posteriori* camera correction; improved movie-alignment algorithms; novel spectra calculations for movie quality control; movie-based full-dataset CTF-determination; Fourier-space multivariate statistical classification, and 4D structural refinements. All alignments are performed using the "alignment by classification" (ABC) approach avoiding the perils of reference bias. All Euler angle orientations are assigned based on angular reconstitution rather than on projection matching. As a practical example of our comprehensive "ABC-4D" approach, we present the structure of the giant hemoglobin of *Lumbricus terrestris* at an average resolution of ~3.8 Å.

INTRODUCTION

Since the beginning of single-particle electron microscopy more than three decades ago, we have seen a continuous improvement of the microscope instrumentation; the specimen preparation techniques; the image detectors; the advent of automatic data collection, and of the image processing methodology on ever more powerful parallel computers (van Heel and Frank 1981; Adrian et al. 1984; Henderson 1995; van Heel et al. 2000; Suloway et al. 2005). These developments, together with the impressive recent advances in electron detectors (Faruqi et al. 2005; Milazzo et al. 2005; McMullan et al. 2009) have led to the solving of biological structures with near-atomic resolution; for recent reviews see: (Kühlbrandt 2014; Cheng 2015; Cheng et al. 2015; Nogales 2016). The new direct electron detectors (FEI Falcon II, Gatan K2 Summit and Direct Electron DE-20 camera) represent a significant increase in quantum efficiency compared to the previous generation of image transducers, resulting in a greatly improved high-resolution data-collection performance (McMullan et al. 2014).

Fast movie-mode data acquisition, now conveniently possible with these new cameras, allows one to routinely compensate for specimen drift during the exposure, thus reviving an important old idea (Kunath et al. 1984). One may now also correct for beam-induced movements of the individual particles locally *within* the full acquired images (Brilot et al. 2012; Campbell et al. 2012; Li et al. 2013). Automatic data collection schemes (Suloway et al. 2005), now allow taking thousands of high-quality micrographs over period of days without human interaction. This changes the requirements of data handling and the logistics of the necessary image processing. The improved statistics of modern, large datasets greatly facilitate obtaining high-resolution results. Subtle effects can be brought to statistical significance. For example,

the raw large dataset contains inherent information about the properties of the detector which, in turn, can be used to retrospectively correct the dataset for camera imperfections (Afanasyev et al. 2015).

In spite of all these new developments, obtaining near-atomic resolution structures is not yet a routine operation, and many pitfalls must be avoided. Single-particle cryo-EM requires a thorough, critical analysis at all stages of the processing. High-resolution 3D-reconstructions result from the classification and averaging of a large number of particles in all possible orientations and, possibly, in multiple conformations (4D). Even under ideal circumstances, the particle images will have a very low signal-to-noise ratio (SNR) to as to minimize radiation damage.

A true pitfall of single-particle cryo-EM approaches therefore is "reference bias" resulting from correlation alignment of these very noisy molecular images with respect to a given reference image (Boekema et al. 1986; Dube et al. 1993; Stewart and Grigorieff 2004). It is essential to avoid reference-bias in detecting the particles in the collected micrographs or in determining their spatial orientations (Henderson 2013; Subramaniam 2013; van Heel 2013). Our philosophy is to exclusively exploit the information emerging from the dataset itself: all the information we need must be present in the raw data and, for example, external "starting models" are to be avoided.

The primary computational tool to achieve our goal of extracting the inherent information present in a dataset is multivariate statistical analysis ("MSA") eigenvector data compression and classification (for review see (van Heel et al. 2009)). The MSA approach is used especially in the context of the "alignment-by-classification" (ABC) as a method for avoiding reference bias (Dube et al. 1993). We here present a complete pipeline for a robust, reference-free analysis of heterogeneous macromolecular structures by single particle cryo-EM.

Oxygen-carrying hemoglobins are present in the red blood cells of most vertebrates, typically as a 67 kDa tetramer containing a total of four myoglobin folds (Perutz 1978). Myoglobin and hemoglobin were among the very first biomolecular structures elucidated by X-ray crystallography and the hemoglobin family remains one of the best studied families of structures represented in the PDB data base (RCSB Protein Data Bank 2015). Nevertheless, that wealth of structural information has exclusively been the result of X-ray crystallography experiments, where good crystal bonds may dictate the conformational state of the structure and not necessarily the ligand bound to its active site. Thus, the mechanism of oxygen binding and the associated conformational-state changes, are still not fully understood. Single-particle cryo-EM technique may be able to address these open questions.

In some invertebrates, like the common earthworm *Lumbricus terrestris*, hemoglobin is a large extracellular oligomer with D6 pointgroup symmetry, containing a total of 144 heme-containing subunits. Each giant hemoglobin is arranged as 12 protomers (1/12th units) , each

containing 12 heme groups, assembled around a central scaffold of 36 linker chains (Royer et al. 2000). Each protomer possesses a local 3-fold symmetry axis, first elucidated by cryo-EM (Schatz et al. 1995), which relates three *abcd* tetramers to each other (Royer et al. 2000). The *abcd* tetramers have a quasi-C2 symmetry axis, whereby the *a*-*d* and the *b*-c dimers each share a further local diad axis not perpendicular to that tetramers quasi-C2 axis. This *abcd* tetramer of globin folds thus does not resemble the classical quasi-D2 structure of the tetrameric mammalian hemoglobins (Royer et al. 2000).

The acellular giant worm hemoglobin is a potential candidate for use in oxygen-carrying blood substitutes (artificial blood) (Hirsch et al. 1997; Zal et al. 2001; Elmer et al. 2012; Roche et al. 2015). A number of crystallographic studies of the giant annelid hemoglobins are currently available: *L. terrestris* at 5.5 Å (Royer et al. 2000) and at 3.5 Å resolution (Royer et al. 2006); and *Glossoscolex paulistusa* at 3.2 Å (Ruggiero Bachega et al. 2015).

The giant hemoglobin of *L. terrestris* has been a favorite sample for methodological studies in electron microscopy because of its large size (3.6 MDa), its high symmetry, and ease of specimen preparation (van Bruggen and Weber 1974; Crewe 1983; Boekema and van Heel 1988; Schatz et al. 1995; de Haas et al. 1996; Mouche et al. 2001). An earlier worm hemoglobin dataset, acquired on a FEI Falcon I camera, resulted in a 6 Å-resolution map deposited in the EMDB database (EMD-2825). In our current study we use a dataset acquired in movie-mode on a FEI Falcon II camera.

The quality of 3D cryo-EM results depends on many factors preceding the actual data processing, and especially on the quality and stability of the sample (Kastner et al. 2008). Much then depends on the specifics of specimen preparation for cryo-EM: the thickness of the vitreous water layer, the concentration and the distribution of the particles within that layer, the stability of the microscope during data collection, the quality of the camera, etc. In our data processing pipeline we have incorporated various recent developments in the IMAGIC-4D software system (van Heel et al. 1996; van Heel et al. 2012), detailed in the step-by-step procedures described below. The specific computer programs and important commands used in this pipeline are provided below.

RESULTS

Step 1. Sample preparation

The hemoglobin samples were collected as described in the Materials & Methods section. The fresh hemolymph was diluted for cryo-EM sample preparation, leading to a relatively dirty sample with extraneous material appearing in the images, and a high particle density (Figure 1). Our procedures include an effective "*in silico* purification" as will be detailed below.

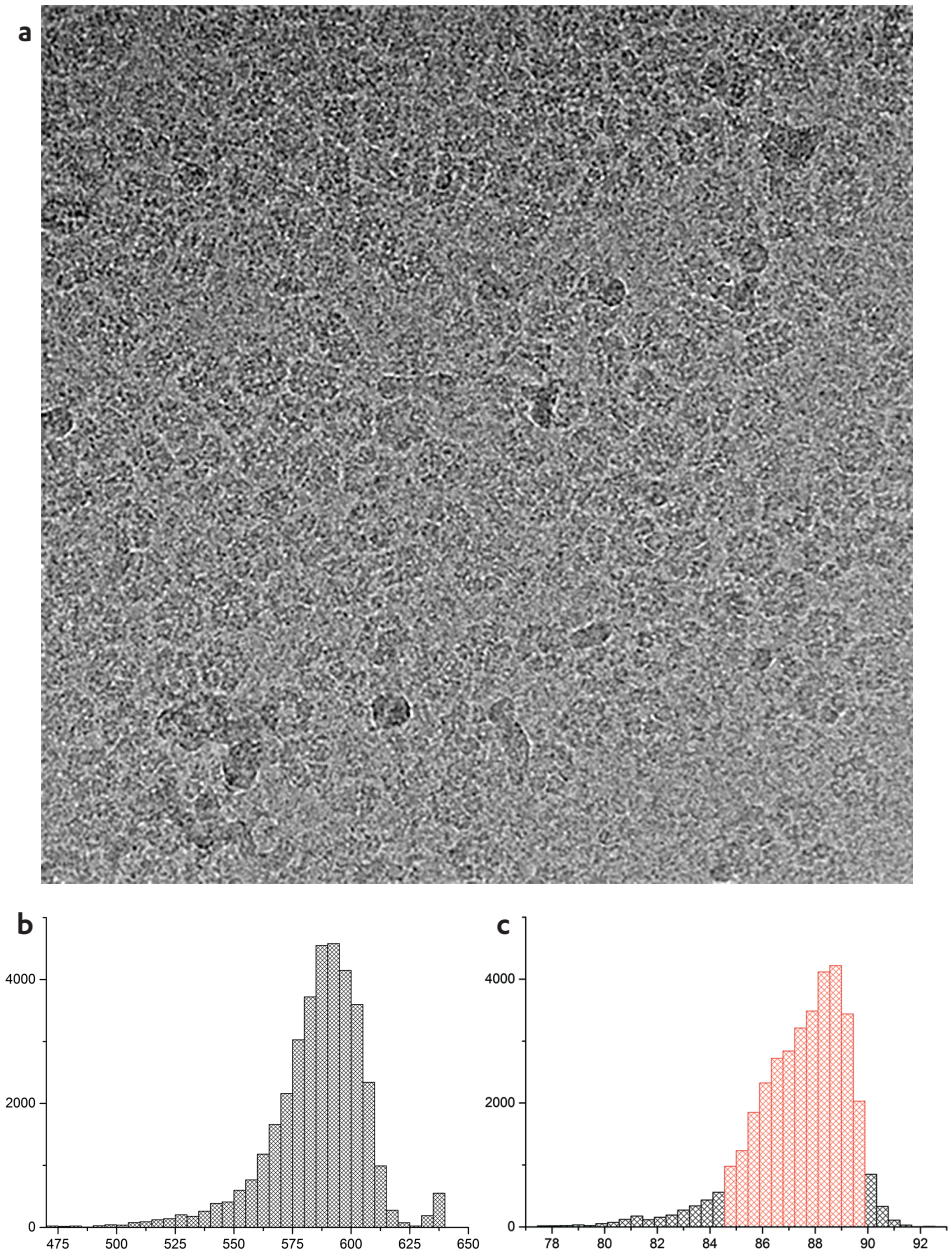

Figure 1. Full dataset statistics. **(a)** Typical micrograph (7-frames average, coarsened/binned to 512x512 size) of the highly concentrated worm hemoglobin sample. Histograms of the average density **(b)** and the standard deviation **(c)** of all the frames in the dataset (36645 frames or a total of 5,235 movies). The histogram range marked in red was used for the calculation of the statistical images used for the camera normalization procedure. Note, that the image (a) contains many high-contrast particles as well as some "junk".

Step 2. Data collection

The images were collected on an FEI Titan Krios microscope as detailed in the Materials & Methods. In cryo-EM, the complexes can often hardly be seen in the micrographs when the microscope is close to focus (less than ~0.5µm defocus) due to the lack of phase contrast in the low-resolution regime. Far from focus, more than at say: ~2 µm, the high-resolution information becomes compromised (van Heel 1978; van Heel et al. 2000). Given the good contrast/visibility of the particles, we targeted the data collection at defocus values of 1 µm and 1.2 µm, sufficient to discern the individual particles in the images. The experimental defocus spread and the presence of some astigmatism then yielded a good homogeneous coverage of the image information over all spatial frequencies (see below).

Step 3. Initial pre-processing and assessment of the micrographs

The images, as stored by the EPU supervisor in an FEI MRC format variant, were converted to IMAGIC stacks using the EM2EM conversion program (Image Science: Michael Schatz 2015). Within the total number of 5,235 movies collected from our sample there was a significant number of images which contained only junk (not vitrified specimen, grid bars edges, strong ice contamination etc.). Such images were excluded from the dataset based primarily on standard statistical metrics like averages and standard deviations (sigma) (Figures 1b,c). The commands SURVEY and HEADERS were used for deactivating those images. Finally, all micrographs de-activated using such criteria were physically removed from the dataset stack (EXCLUSIVE-COPY) and our final dataset thus contained only the 4062 selected good movies (28434 frames; 78% of the full raw dataset).

Step 4. Camera Correction

An *a posteriori* camera correction was then applied to avoid spurious correlations due to the various camera imperfections (Afanasyev et al. 2015). The camera correction is performed after removing extreme outlier images from the dataset but prior to the movie alignments. Figure 1c shows the histogram of the standard deviations of all movie frames. Marked in red, are all frames used for calculating the pixel-vector statistics for normalization. A large dataset consisting of many images collected with the same camera reflects the characteristics of each individual pixel. The normalization procedure was performed on the full dataset (CAMERA-NORMALIZATION). The frames were then prepared for further analysis (PREPARE-IMAGES) by: high-pass filtering to remove gradual background fluctuations, and contrast inversion, to make proteins "white" against a dark background.

Step 5. Correction of Anisotropic Magnification

We noticed that the NeCEN KRIOS 2 instrument (equipped with a CEOS Cs corrector (Corrected Electron Optical Systems GmbH 2015)), yielded images with an anisotropic magnification difference of 2.6% (see: chapter 3) with the maximum magnification in a 36.2° diagonal direction. For the 300 Å-diameter worm hemoglobin structure (sampled at ~1.12 Å per pixel)

this implies an orientation-dependent magnification fluctuation of ~4 pixels at the outer edge, seriously limiting the achievable resolution (to ~5.5 Å). The anisotropic magnification was detected from the dataset itself as an ellipticity of the ~3.6 Å and ~2.2 Å water rings in the MSA-based spectrum analysis of the dataset (Figure 2). This figure shows the second eigenimage of all spectra of the full data set, determined from the full data set itself during the CTF determination (see below). A correction program (ANISOTROPIC-MAGNIFICATION) was developed to re-interpolate the raw dataset. This anisotropic magnification compensation changed the average linear pixel size from ~1.12 Å to ~1.11 Å.

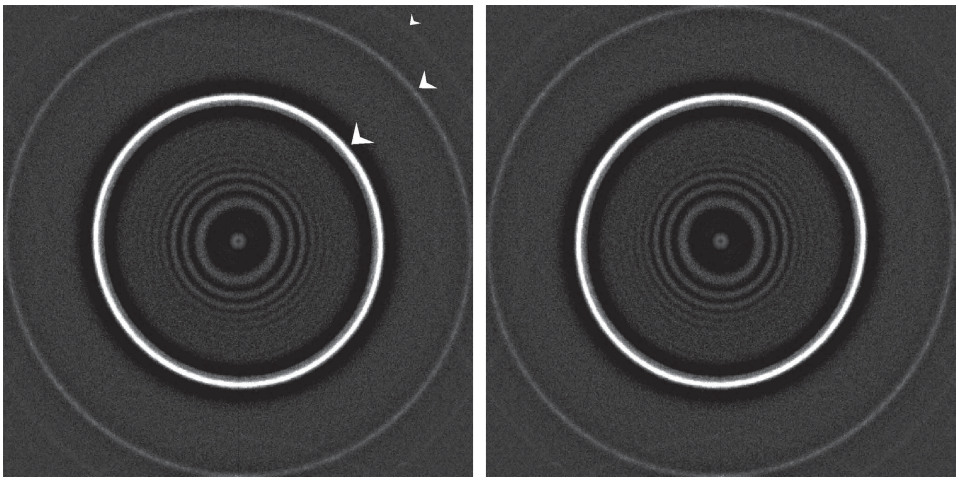

Figure 2. Anisotropic magnification of the Titan Krios 2 instrument became evident from an eigenvector analysis of all amplitude spectra of the worm hemoglobin dataset (see Step 7). The second eigenimage in the eigenvector analysis revealed a 2.6% ellipticity of the ~3.6 Å water ring (marked with large arrow in the left spectrum) and the 2.2 Å water ring (medium-sized arrow). Note that a further water/ice ring can be seen at 1.8 Å (small arrow) which is only visible in the corners of the image (beyond the Nyquist frequency) but reflects back into the image everywhere else due to aliasing. The right spectrum is the same as the left one but now mirrored horizontally therewith changing the orientation of the water-ring ellipticity for illustration purposes. The main ellipticity direction is in a close-to-diagonal (~36°) direction such that the vertical and the horizontal magnifications in the dataset are actually almost identical. The Thon rings visible in the center of the spectrum are elliptical due to an astigmatism of ~1000 Å. This ellipticity is unrelated to the ellipticity of the water rings due to anisotropic magnification. The two effects may, however, become entangled if the anisotropic magnification is not corrected prior to the CTF determination.

Step 6. Determining the CTF parameters

The electron microscope is a phase-contrast microscope and the linear transfer of the image information in the instrument is described by the phase contrast transfer function (CTF). The CTF starts at zero at the Fourier-space origin and oscillates around zero as function of spatial frequency (Scherzer 1949; van Heel 1978; van Heel et al. 2000). Our automatic defocus and astigmatism determination is based on the best correlation between the theoretical CTF-function and the experimentally measured spectra (van Heel et al. 2000). We here used the average of the power spectra of individual movie-frames (which is invariant to drift) as

the basic measurement for the CTF determination (MOVIE-SPECTRA; detailed in Chapter 3). This experimentally measured result is subsequently filtered to further suppress the strong influence of background ramps (van Heel et al. 2000; Mindell and Grigorieff 2003). In line with the "full-data-set CTF correction" philosophy (van Heel et al. 2012), the large set of spectra was then submitted to eigenvector analysis and automatic classification (van Heel et al. 2009). The defocus parameters of all class averages of amplitude spectra was then determined (Figure 3).

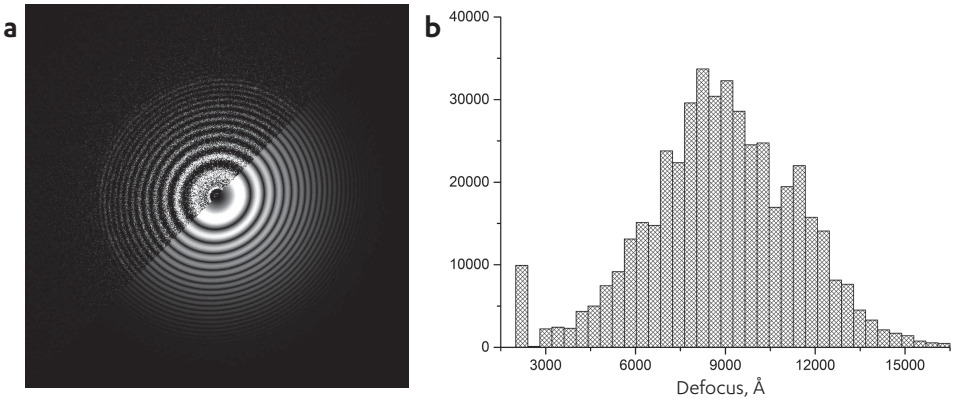

Figure 3. (a) A selected class-average with the matching theoretical CTF-spectrum fitted. Top-left corner: class average of the movie power spectra (1280x1280 patches). Bottom-right corner: theoretical CTF, corresponding to a defocus of 7913 Å/9047 Å; astigmatism: 1134 Å; astigmatism angle: 133°. Thon rings in good class-average spectra are visible beyond ~2/3 of the Nyquist frequency, corresponding to a resolution of ~3.1 Å. **(b)** Histogram of the defocus values (in Å) of ~450,000 movie frames patches of 1280x1280 pixels. The peak at the low-defocus edge of this histogram corresponds to images with a defocus less than 0.22 microns which we classified as "junk".

Step 7. Movie alignments

During the exposure time of a micrograph (1-2 sec), the stage or more locally, the sample, may suffer drifts limiting the attainable resolution (see Chapter 1 and (Kunath et al. 1984). Such movements can be corrected by the movie-alignment procedure described in Chapter 3. The *a posteriori* camera correction described above, is an important step preceding the movie-alignment procedures which may otherwise be trapped by the fixed background pattern in each frame. We first performed full-frame translational movie alignments on the dataset using the command ALIGN-MOVIES. The quality of the full-frame movie alignments was assessed by comparison of spectra before and after alignment (as discussed in Chapter 3). Figures 4a,b demonstrate typical images of a P-spectra before and after alignment, showing that even seriously drifted movies may be restored. The information was here lost in the drift direction, but the aligned images are still well suited for processing. When these in-frame drifts become too large, the aligned image frames will not contain much high resolution information in the drift direction Figure 4. We have discarded such large-drift movies from further processing. The final movie alignment was performed on sixteen checkerboard 1280x1280 overlapping

patches (4x4 cut from the full 4096x4096 frames), after CTF correction of the individual 1280x1280 movie frames (see below).

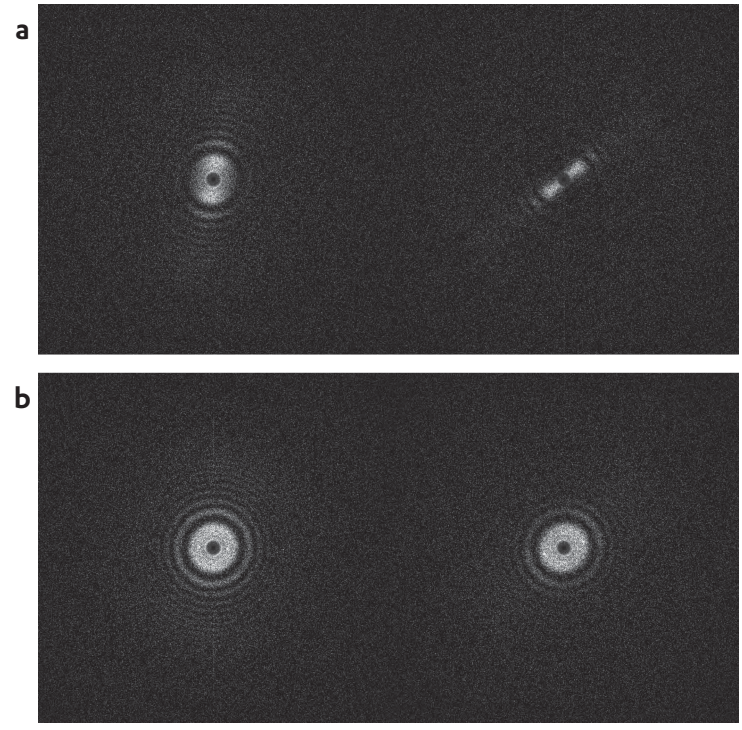

Figure 4. Two typical P-spectra of heavily drifted movies before **(a)** and after **(b)** movie alignments. The second movie experiences a heavy drift, which does not allow to restore enough of the high-resolution information (only two "zeros" in the direction of the drift were restored). Movies with such drifts should be discarded for obtaining high resolution 3D-reconstruction.

Step 8. CTF correction

All individual members of each class where assigned the defocus parameters of that class average. Some class averages were found to be of "poor" quality and the members of those classes were excluded from further processing. The CTF-correction by phase flipping was then performed on the individual movie frames (CTF-FLIP).

Step 9. Initial particle picking

For the initial particle picking we coarsened (binned) the CTF-corrected movie averages by a factor of eight, followed by a band-pass filter operation. To avoid any form of reference bias during particle picking, we prefer using variance or modulation for a first particle picking (van Heel 1982; van Heel 2013) and (Chapter 4). However, because of the high particle density of often overlapping particles, the variance images become continuous; we thus visually selected several typical particles (Figure 1a). These particles where then averaged rotationally (AVERAGE-ROTATIONAL) to create first templates for particle picking (Figure 5). Particles were first picked

(PICK-M-ALL) from the 500 movie averages with the largest defocus, using the six templates simultaneously in a competitive manner. For the subsequent processing the newly extracted (unbinned) particle images were binned by a factor of four.

he first round of automatic particle picking with such indiscriminate references, typically yields "dirty" results, depending on the quality/purity of the input images. Real-life samples are full of high-contrast artefacts like: crystalline ice, carbon foil edges, fragments of molecules, contamination, etc. These artefacts lead to a large number of "false-positive" picks. We excluded such outliers primarily based on standard deviation values (the same approach as used for the micrographs selection). False-positive picks are also associated with random background fluctuations on the low end of the standard-deviation histograms, and were also excluded.

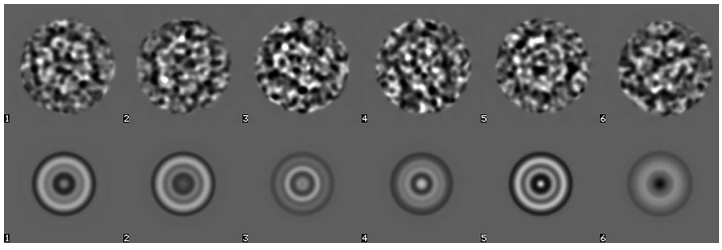

Figure 5. Top row: six manually selected particles (and their rotational averages) used as templates for an initial particle picking (for details see text).

Step 10. Eigenvector data compression and unsupervised classification

We use eigenvector data compression and classification methods to extract systematic information from the dataset (van Heel et al. 2009; van Heel et al. 2012). An initial MSA was performed (using modulation distances: (Borland and van Heel 1990)) of the band-pass filtered picked-particle stack (Figure 6) (MSA-RUN). The first 32 eigenimages (36 were used) are shown in Figure 6. Eigenimages #2 and #3 represent the main six-fold symmetry component of the dataset (Dube et al. 1993).

Class averages were obtained by hierarchical ascendant classification (HAC; MSA-CLASSIFY and MSA-SUM) with a new HYBRID option with increased processing speed, as proposed in (van Heel 1989). For this initial classification we aimed at ~20 members per class. Classes were sorted based on the number of members per class (SORT-IMAGES) and by an overall information-based quality criterion (MSA-SUM-CLASSES). A total of 15 class averages were selected at random from the best 500 class averages for Euler angle assignment by angular reconstitution. These selected class served as input for a first "random startup" 3D reconstruction (Figure 7).

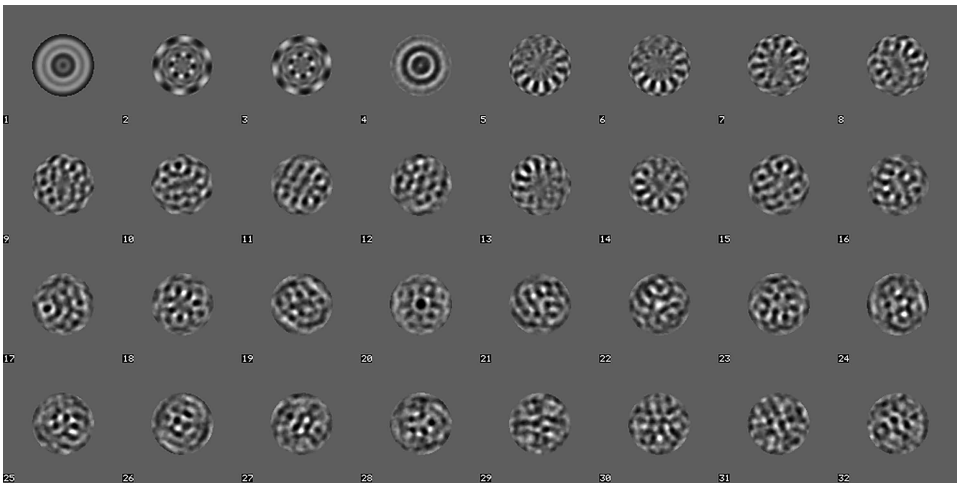

Figure 6. The first 32 eigenimages of the 20,000 picked particles using the rotationally-symmetric reference images depicted in Figure 5. The first eigenimage is rotationally symmetric reflecting that property of the first particle-picking references. Eigenimages #2 and #3 reflect the predominant six-fold symmetry of the worm hemoglobin "top views".

Step 11. Initial 3D-reconstruction by "random startup"

Angular reconstitution, based on the "common projections lines" theorem (van Heel 1987), allows determining the Euler angles of the 2D class averages and obtaining the first 3D-reconstruction without relying on any external information. We demonstrate here, that based on the 15 classes from Figure 7, we could obtain an initial 3D-reconstruction without using explicit alignments. We used the random startup option for the angular reconstitution (EULER program; option RANDOM) with imposed D6 symmetry. In this procedure random Euler angles are assigned to the class averages followed by a standard procedure (option

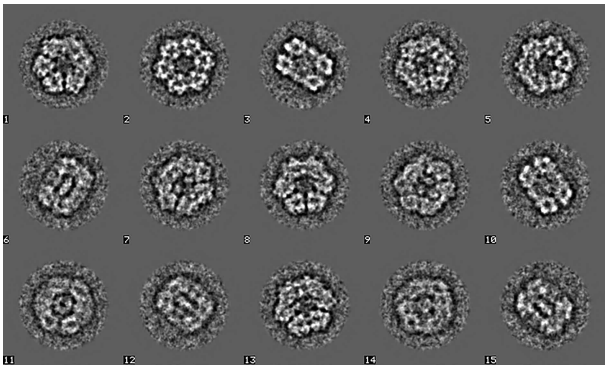

Figure 7. Class-averages of unaligned particles (i.e. the result of an "alignment by classification": ABC) used for the initial "random startup" 3D-reconstruction. Class-average #2 corresponds to a top view; class averages #3 and #10 to side views. The other class averages represent various intermediate projection images of the worm hemoglobin. Since at this stage of the processing the input images are coarsened (binned) by a factor of four, no information beyond the Nyquist frequency of 1/8.8 Å can be contained in these first class averages.

REFINE) for their iterative orientational refinement (Schatz et al. 1995; van Heel et al. 2000). The class averages were then used to create a 3D reconstruction (TRUE-THREED). To validate the results, the re-projections from the 3D reconstruction were compared to their corresponding class averages (Figure 8). A 3D automatic masking program (threed-auto-mask) was used to reduce the number of small artefacts surrounding the reconstructed 3D object.

An anchor-set (Schatz et al. 1995) was created from the masked 3D-reconstruction (THREED-FORWARD). These anchor-set re-projections are used to determine the Euler angle orientation of all other class-averages (EULER; option ANCHOR-SET). A new 3D-reconstruction is then obtained based on all good class averages. This process (3D-masking – new anchor set – angular reconstitution – 3D reconstruction) is iterated until convergence. We call this converged map our "initial" 3D-reconstruction.

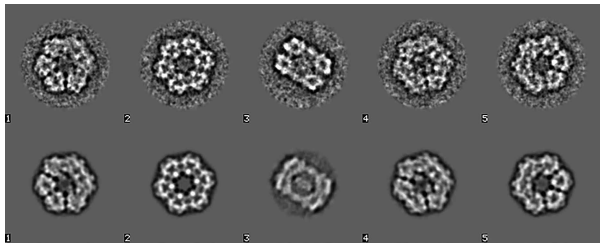

Figure 8. Five class-averages used in the random startup procedure (top row of the Figure 6) and the corresponding re-projections from the obtained 3D-reconstruction. A good correspondence of the class averages to their re-projections is a necessary condition for a valid reconstruction procedure. The input images were coarsened (binned) by a factor of four, yielding a Nyquist frequency of 1/8.8 Å. Effectively this implies that no information beyond ~1/15 Å can be present in our "initial" 3D reconstruction.

Step 12. Refined particle picking using the first 3D-reconstruction as a template

We used projections from the initial 3D-reconstruction, in all possible directions, as templates for a competitive correlation-function based particle picking. The newly picked 319,746 particles, were extracted from the CTF-corrected patches (CUT-IMAGE) at full sampling. Again, we can discard extreme images as described above. However, since we are now looking for specific views of the complex in all possible orientations, the MSA eigenvector analysis (as per "step 10") has become more specific and may be fully sufficient to remove all remaining "false-positive" particles. The eigenvectors of the newly picked particles dataset are now dominated by the properties of the giant hemoglobins. All remaining false positive picks lack these characteristics and their variances are thus not well covered the eigenvectors of the dataset. These not-characteristic images have a poor "representation quality" (the fraction of their variance described by the main eigenvectors). The poor representation quality (van Heel 1989) can be used directly to eliminate poor images from the dataset (Figure 9a). The poor variance of the corresponding class averages is obvious from the bimodal distribution of the standard deviations (Figure 9b).

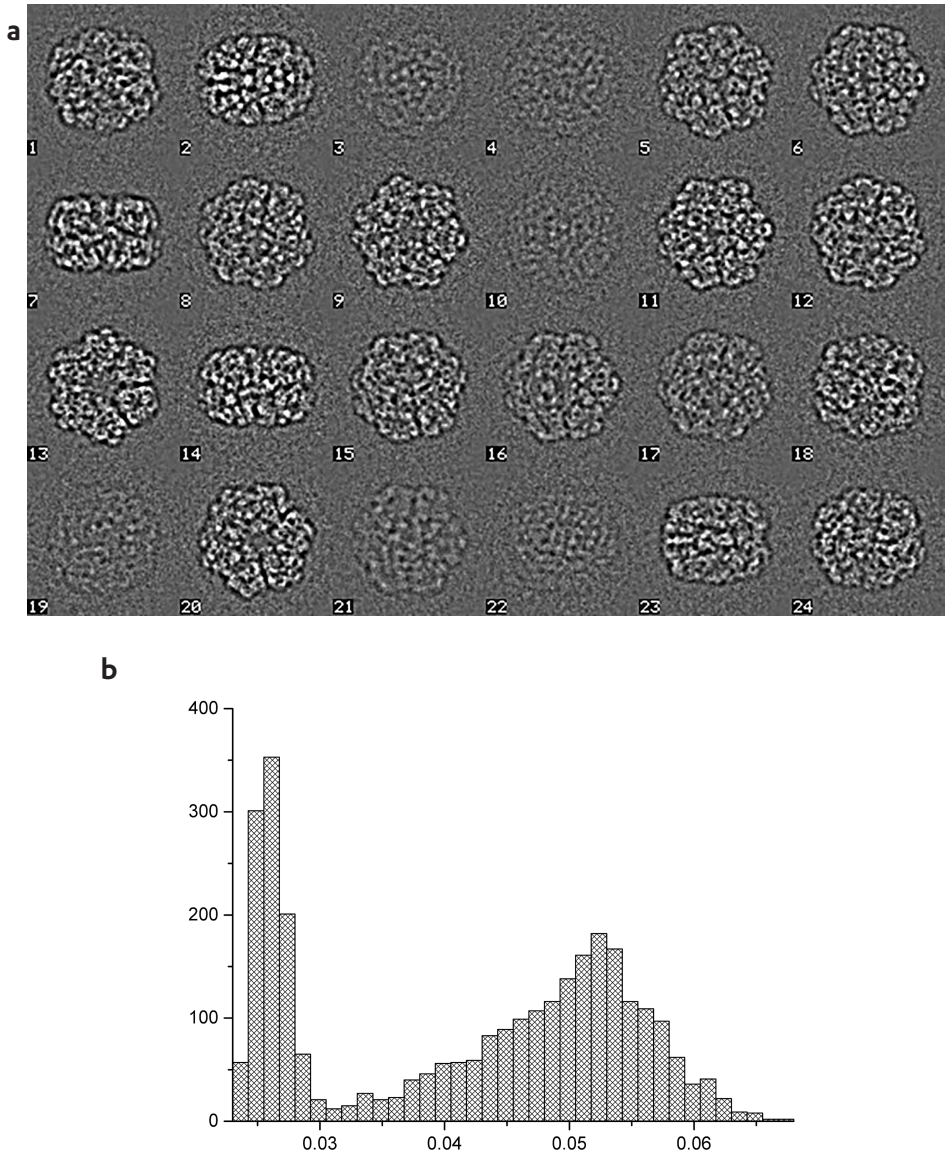

Figure 9. (a) Some class averages randomly extracted from the total 3000 class averages: the high-contrast class averages are associated with good views of the worm hemoglobin, whereas the low-contrast class averages are associated by the atypical "false-positive" particle detections. The false positives have a low representation quality with respect to the main eigenvectors of the dataset which now are dominated by the properties of the true molecular images. **(b)** The histogram of standard deviations of the 3000 class averages (derived from a total of 205,764 particle images) clearly shows a bimodal distribution in which all false-positive picks land into low-contrast class averages. Here we use these properties for a final automatic purification of the full dataset whilst the data are coarsened (binned) by a factor of 4, that is, while still using Nyquist frequency of ~1/9 Å.

Step 13. Improving overall alignment of the dataset: focus on class-average parameters

With the improved particle picking round, we now have a large dataset available with a sufficient number of images in all possible orientations. (We may, at this level of the processing, actually already be working with multiple 3D-reconstructions simultaneously). The goal of our iterative ABC procedures is to find the best overall alignment for the full dataset. The best alignment maximizes the variance of the dataset (i.e., eigenvalues) into the lower eigenimages of the system (van Heel et al. 2009; van Heel et al. 2012). The logistics of the approach is to first find the relative orientations of the 2D class averages (such as shown in Figures 7-8) with respect to the initial 3D reconstruction. These orientations are determined by finding the Euler angles of the class averages with respect to an anchor set (Schatz et al. 1995) derived from the initial 3D reconstruction(s), and the residual in-plane shifts with respect to that 3D reconstruction. The alignment and shift parameters found for the class-averages, are then applied to rotate/shift all original images contributing to that class (MOVE-BY-ALIGNED-CLASSUMS). The procedure is then iterated (MSA data compression; classification; Euler angle assignment; and 3D reconstruction).

These iterations gradually bring the full dataset to a common 6D co-ordinate system. All images gradually become better aligned within the full dataset and in subsequent classification rounds, similar images are thus more likely thus to end up in the same class (Figure 10).

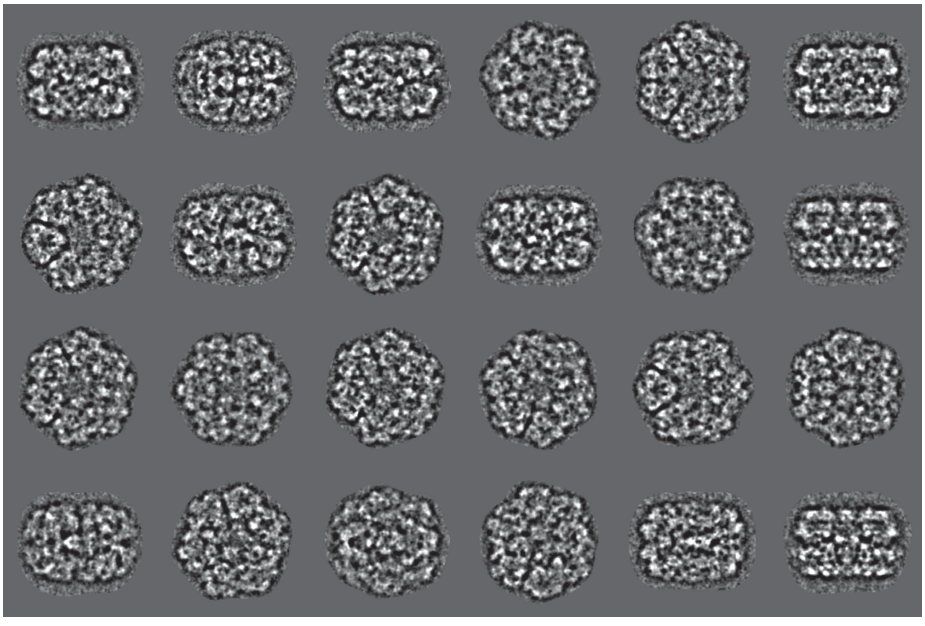

Figure 10. After the ABC alignment of the full dataset, all images are rotated and shifted to a common origin as dictated by the Euler orientations of the class averages to which they belong. Now operating at a coarsening (binning) level of just 2, the Nyquist frequency is at 1/4.4 Å.

Step 14. Refinements at full sampling level; Fourier-space MSA.

At each round of the refinements, the current resolution state is estimated by Fourier Shell Correlation (FSC) (Harauz and van Heel 1986; van Heel and Schatz 2005). The class averages associated with each of current 3D reconstructions (in the context of 4D processing, see below), are randomly split into two groups which are then compared by FSC. With the FSC resolution thresholds now exceeding half of the current Nyquist frequency (here at 1/4.44 Å) further refinements need to be pursued at full sampling (Nyquist: 1/2.22 Å). Refinement iterations continue following the same procedures: MSA data compression; automatic classification; Euler angles assignments and shifts w.r.t. the class averages and transferring those class parameters to the individual class members (MOVE-BY-ALIGNED-CLASSUMS).

At this level, the MSA data compression and classification was performed on the Fourier transforms of the images rather than on the images themselves (HERMITIAN-FOURIER-TRANSFORM). As the active MSA mask (van Heel et al. 2009), we use a Fourier-space ring mask starting at ~0.08 Nyquist, and ending at a high-frequency outer edge. This outer edge of the MSA ring mask is gradually increased from a value of ~0.4 Nyquist to a value of ~0.8 Nyquist. This mask thus gradually includes more high spatial frequencies during subsequent refinement iterations.

Step 15. Refinements in 4D

To deal with heterogeneities present in the data set, we use a 4D-refinement scheme where the fourth dimension stands for variability of 3D conformations of the structure (or any other systematic departure from a single 3D structure) (van Heel et al. 2012). The class averages, with assigned Euler angles, are randomly split into multiple groups, generating a number of separate 3D-reconstructions. Those 3D-reconstructions are then used to create their individual anchor-sets for the next round of Euler-angle assignments. Each class is then assigned to the 3D reconstruction associated with the anchor-set it fits best to (EACH-TO-BEST option);3D reconstructions which do not accumulate enough class averages during this 4D refinement are removed from the 3D-reconstructions pool in a "survival of the fittest" approach (van Heel et al. 2012).

The splitting into multiple 3D structures can be introduced at earlier stages of processing if appropriate (see Discussion). In our case, the 4D analysis of the dataset converged towards two different 3D reconstructions of which one of good quality and contained 56% of the molecular images (discussed below) and the other was of poorer quality and was discarded. We thus focused on the best global structure emerging. For the final Euler angle assignments, the number of members per class was reduced to N=1, that is, the particles were assigned Euler angles individually.

Step 16. Quality assessment and anisotropic resolution

The 3D/4D refinement iterations are stopped once convergence is achieved at a satisfactory level by different metrics. The standard half-maps FSC_{tot} crosses the ½ bit threshold curve at ~3.8 Å (Figure 11) indicating that enough information has been collected for interpretation at that resolution (FOURIER-SHELL-CORRELATION). The 3-sigma threshold of this reconstruction is at 3.5 Å, indicating that here we have collected significant information above the noise level (but not enough for structural interpretation). This 3-sigma threshold is below 2/3 of the Nyquist frequency, confirming that the data has been collected at a sufficiently high pixel sampling. Moreover, the FSC curves then oscillate around the zero value up to the Nyquist frequency and remains below the 3-sigma curve as is to be expected for well-behaved random correlations (van Heel and Schatz 2005). Thus no artifacts were introduced by the automatic masking of the various 3D reconstructions or by any other procedures.

For comparison, we have included the popular 0.143 threshold (Figure 11: dotted line) (Rosenthal and Henderson 2003). This criterion, however, is only valid in an asymptotical sense (i.e. for very large volumes, at high resolution) since all radius- and symmetry-dependency effects have been neglected (discussed in (van Heel and Schatz 2005)). Often more important than a single numerical resolution value is the shape of the FSC curve. The heights of "knee" area in FSC curve, for example, marked by a circle in Figure 10, indicates the area that typically has a strong influence of the appearance of the 3D reconstructions.

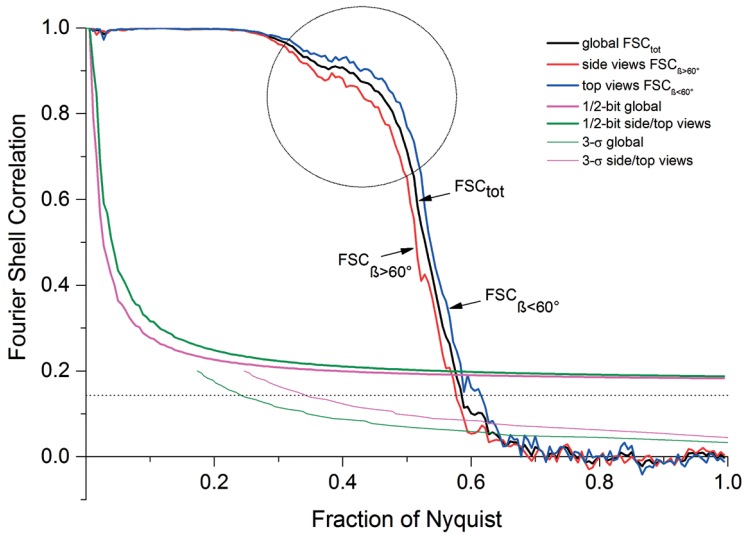

Figure 11. Fourier Shell Correlation (FSC) curves for evaluating the average reproducible resolution of the full 3D reconstruction. (The Nyquist frequency is at 1/2.22 Å, and thus 0.6 Nyquist corresponds to 1/3.7 Å spatial frequency). The $FSC_{\beta<60°}$ based on the contributions of class averages in the preferred top view orientations 0° < β < 60°, and the $FSC_{\beta>60°}$ assesses the contributions of the molecules in side view orientation 60°< β < 90°. The circle indicates the areas where a high FSC value has a strong influence on the quality of the reconstruction. Details in the main text.

We noticed that the dataset exhibited anisotropic resolution due to preferential particle orientations. In the structure presented below, a total of 85000 particle images were used, of which ~2/3 were top views and ~1/3 side views. The twice larger number of top views, makes that the statistical significance of the top views is better than that of the side views. Hence, the FSC curve in the direction of the top views should be better overall than that in the directions of the side-views. This effect, however, disappears in the "global" nature of the FSC: the integration is uniformly over the full spherical shells in Fourier space (Harauz and van Heel 1986). We have now implemented new options in the program to calculate the $FSC_{\beta<60°}$ or the $FSC_{\beta>60°}$ separately (ANISOTROPIC-RESOLUTION), whereby β is the particle's Euler angle orientation from the "North-pole", the major symmetry axis (6-fold here).The 60° choice can be replaced

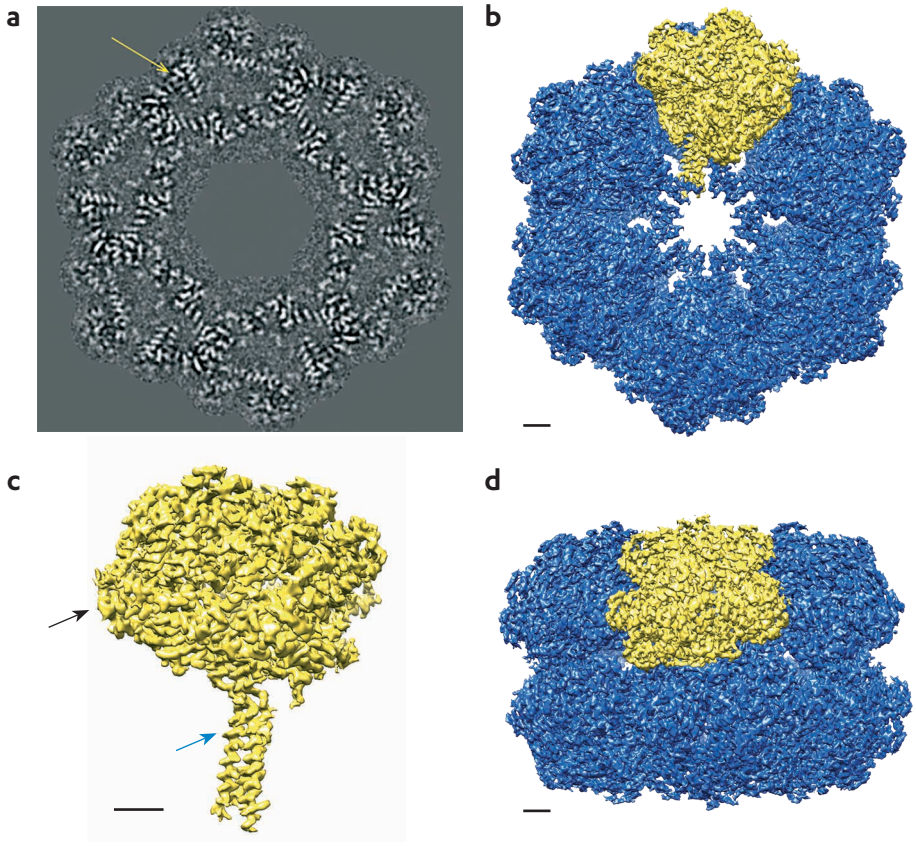

Figure 12. Final map of the worm hemoglobin. **(a)** Slice through the 3D cryo-EM reconstruction perpendicular to the main 6-fold symmetry axis, with a yellow arrow pointing at a heme group sandwiched between two alpha-helices; the proximal histidine contacting the iron in the heme group is also visible. **(b)** Top views (along the 6-fold axis) of the dodecameric hemoglobin with the asymmetric unit (1/12th of the overall D6 structure) highlighted in yellow. **(c)** The asymmetric unit ("protomer or "1/12th" unit) with a black arrow pointing at the globin fold shown in Figure 13b; and a blue arrow pointing at the helix detailed in Figure 13c. **(d)** Side view of the worm hemoglobin with one of the 1/12th subunit ("protomer") highlighted in yellow. (Scale bars are 20 Å).

by any other value, but the number of voxels within a Fourier-space sphere between 0°≤ β ≤60° equals the number of voxels between 60°≤ β ≤ 90°, making $FSC_{\beta<60°}$ and the $FSC_{\beta>60°}$ directly comparable in terms of the applicable thresholds (Figure 11). The global resolution of the main 3D reconstruction here is at ~3.8 Å; the top-view resolution is better at ~3.6 Å, whereas the side-view resolution is limited to only ~4.0 Å (Figure 11).

Step 17. Data interpretation, manifold separation, and fitting of atomic coordinates

The giant hemoglobin, with its D6 pointgroup symmetry, is arranged as 12 protomers (1/12th units), each containing 12 heme groups (Figure 12a,b). Since the 1/12th unit is the asymmetric unit of this oligomer (Figures 12,13), the 12 globin folds (Figure 13b) within the protomer are all determined independently. The "stem" of the mushroom-shaped 1/12th unit, is a hetero-trimer of linker chains, forming a triple coiled-coil structure.

With a global resolution of 3.8 Å the best areas of map locally have a significantly better resolution. We can thus clearly resolve the heme groups such as the inner heme groups (Figure 12c; Figure 13b) and distinguish various side chains in the coiled coil regions (Figure 12d; Figure 13c).

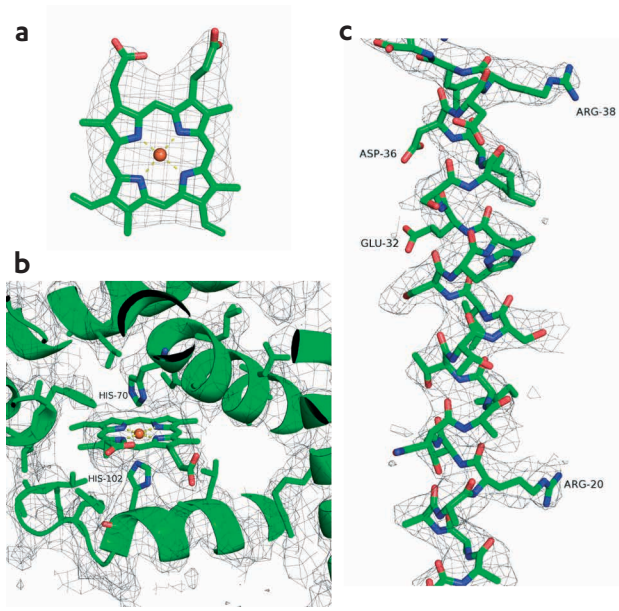

Figure 13. Detail of final cryo-EM map with a fitted atomic model. **(a)** Lateral view of one of the heme groups. **(b)** View of one the better resolved heme groups: both the proximal and the distal histidine (K-chain: His-102 and His-70) are well embedded in the cryo-EM density. **(c)** View of one alpha-helix in the triple coiled-coil region: the side chains fit nicely into the 3D-reconstruction, moreover, the acidic side chains (O-chain: Asp-36 and Glu-32 respectively) have lost some density due to radiation-damage effects.

DISCUSSION

We here have proposed a pipeline for obtaining 3D-reconstruction(s) by single-particle cryo-EM focused on the principle of MSA eigenvector data compression, alignment by classification (ABC), and angular reconstitution. Many old principles and recent ideas come together to form the comprehensive pipeline used here to reconstruct a large biological complex to quasi-atomic resolution. The philosophy of our processing pipeline is that all structural information must emerge from the dataset itself and not from external sources. Some *"a priori"* information external to the dataset is necessary such as: the nature of the sample, the microscope parameters used, etc. At the same time, there is no need for any external low-resolution "starting model" since the data itself is assumed to yield quasi-atomic 3D resolution information on the desired structure(s). We here discuss some of the main aspects of our processing pipeline.

The main tool for finding inherent information present in the micrographs is the eigenvector-eigenvalue data compression of the datasets which seeks to best describe the variances in the dataset as effectively as possible (van Heel et al. 2009). Our MSA eigenvector algorithms have been parallelized - also in terms of I/O operations - allowing for the efficient analysis of terabytes of cryo-EM data. A recent speed improvement is a "hybrid" option for unsupervised hierarchical ascendant classification (HAC) (van Heel 1989). These procedures can now group in the order of 10^6 images into several thousand class averages in a matter of hours on a standard notebook computer, while maintaining the principles of the minimum added intra-class variance known as the "Ward criterion".

All (multivariate) cryo-EM image processing is very sensitive to the power-spectrum distribution of the data. Thus any filtering (high-pass, band-pass, etc.) will strongly influence the results of the procedures. The new Fourier-space MSA eigenvector analysis was found to help significantly in focusing the analysis on the relevant spatial frequencies. Our favorite metric for performing the eigenvector analysis is the modulation metric (Borland and van Heel 1990). This metric is especially useful in Fourier-space eigenvector analysis since it balances out the over-representation of strong amplitudes that are a characteristic of the conventional principal component (PCA) analysis or, more generally speaking, characteristic of all cryo EM approaches that are based on squared correlation function (van Heel et al. 1992).

A somewhat different form of multivariate statistical analysis has been introduced in the form of maximum likelihood (Sigworth 1998; Scheres et al. 2005), which is an "explanatory factor analysis" technique (Fabrigar and Wegener 2011). This approach had a significant number of successes recently in refining the structure of large complexes (Amunts et al. 2014; Kühlbrandt 2014). The two multivariate approaches are considered largely similar in the statistical literature, but the PCA family of approaches (to which the modulation analysis method belongs) are "simply variable reduction techniques". In contrast, in the "exploratory factor

analysis" family (to which maximum likelihood belongs) are based on the assumption that an "underlying causal model" exists that has been formulated correctly. A further complication in comparing these two multivariate approaches is that in maximum-likelihood procedures in use in cryo-EM, multi-reference based alignments of 2D images or 3D volumes have intimately been intertwined in a single optimization procedure. In contrast, in our ABC-4D approaches the overall alignment of the data set is separated from the multivariate statistical data compression and automatic classification.

A good example of the impact of the large-scale MSA approach is the full-dataset CTF-correction based on movie-mode data collection that we applied. The spectrum calculations were based on the raw movie frames *prior* to the movie-alignment procedures (Figures 2&4). The CTF-determination is based on MSA class averages of preprocessed amplitude patches of 1280x1280 pixels (Louys et al. 1989; van Heel et al. 2000; Sander et al. 2003; van Heel et al. 2012). One important aspect of applying the CTF correction on the full input data set is that the delocalization of the information stemming for an individual particle can easily extend to an area ten time larger than that covered by the particle (see Figure 2 or: (Louys et al. 1989; van Heel et al. 2000; Sander et al. 2003; van Heel et al. 2012)) implying that huge areas around the individual particles must be included during all processing. Moreover, if the particle images are not first CTF corrected, they cannot be directly compared/correlated to each other, complicating the processing.

How well our amplitude spectrum procedures work is illustrated by the unexpected anisotropic magnification problems diagnosed during the CTF assessment (Figure 2). A discussion emerged in the 3DEM mailer in April 2014 on possible anisotropic magnification encountered on KRIOS microscope data under different experimental conditions (van Heel 2014). We had found up to 7% magnification anisotropy in a dataset collected on our Cs-corrected FEI Titan Krios microscope under "normal" magnification conditions as evidenced by the elliptical shape of ice/water rings. Even in the presence of astigmatism, the water rings at ~3.6 Å and at ~2.2 Å spatial frequency are supposed to be rotationally symmetrical because of the diffraction power being concentrated at these specific spatial frequencies. The 2.2 Å water ring is clearly discerned in eigenvector analysis in spite of it being at the Nyquist frequency (Figure 2) where the DQE values have dropped significantly. We can even locate the "super resolution" (beyond Nyquist) ~1.8 Å water ring folding back into the image ("aliasing"). This can only emerge due to the "full data set" aspect of the eigenvector-based CTF analysis which integrates all available information into the main eigenvectors of the dataset, bringing even the weakest of effects to statistical significance. Note that this Falcon 2 camera is not aimed at "super resolution" single-electron detection as is, for example, the Gatan K2 Summit or the Falcon 3 camera (Kuijper et al. 2015), but our analysis shows that significant "super resolution" information has nevertheless been recorded.

Since that discussion in 3DEM, two papers appeared suggesting methods for correcting this resolution limiting problem (Grant and Grigorieff 2015; Zhao et al. 2015) on a test sample and then correction of this anisotropy on any data set collected with the EM instrument under those conditions. In our case, the measurement stems directly from the dataset itself. The 7% anisotropy we had found was corrected by a re-alignment of the instrument by the manufacturer. We thus thought that the issue had been resolved, but when we failed to refine the structure to a resolution better than ~5 Å we re-examined the CTF spectra in detail. A residual 2.6% anisotropic magnification in an approximately diagonal direction was thus found. Moreover, we observed different anisotropic magnification parameters in different datasets collected on the same Cs-corrected FEI Titan Krios instrument, implying that one cannot consider them instrumental constants to be used for all data sets. We associate this anisotropic-magnification effect primarily with a misalignment of the Cs corrector, in conjunction with the correction of astigmatism using the Cs-corrector optics. The microscope manufacturer (FEI Company) has informed us of a forthcoming revision of their alignment protocol, designed to eliminate this magnification anisotropy.

The problem of reference bias in single-particle analysis was identified decades ago (Boekema et al. 1986) and various "reference-free", "unbiased" solutions have been proposed, such as (unsupervised) invariant classification (Schatz and van Heel 1990) and alignment by classification ABC (Dube et al. 1993) . The concept of "reference-free" has however since suffered serious devaluation by the introduction of confusing (and conflicting) nomenclatures. For example, in (Penczek et al. 1992) their 2D alignment algorithm is misleadingly called "reference-free" because the single reference image used to start their procedures is chosen by a random generator rather than by the human operator. Also "2D classification" schemes used in connection with maximum likelihood approaches (Sigworth 1998; Scheres et al. 2005) are not "reference free" but are rather associated with a random selection of starting references, and these procedures are claimed to introduce less bias (by using "bias-free" seeds). When others refer to these ML approaches, however, those procedures are often claimed to be "reference free" (van Heel 2013).

We will here prove by inference, that the angular reconstitution (AR) approach is equal or better than the projection matching (PM) approach. There is much misunderstanding in the cryo-EM field on the difference between the AR - and the PM approach for Euler-angle assignment. AR is based on finding "common projection lines" (CPL) between an input image and (a few) other images which have already received an Euler orientation assignment (van Heel 1987; van Heel et al. 2012). These other images can be noise-free class averages, or an "anchor set" derived from a 3D reconstruction. Projection matching (van Heel 1984), in contrast, requires a multi-reference alignment (van Heel and Stöffler-Meilicke 1985) of the input image with respect to all possible projection images of the current 3D reconstruction

(van Heel et al. 2000; Grigorieff 2007; Bai et al. 2015; Sigworth 2015). This represents a massive computational effort.

Let us now take a step back and re-iterate the concept of isotropic ("instrumental") resolution achievable from a given number of noise-free projection images of a 3D object (De Rosier and Klug 1968; Crowther 1971; van Heel and Harauz 1986; Sigworth 2015): $N \times N_{sym} \geq 2D/g$. (With: N the number of projection images needed to achieve a uniform spatial-frequency resolution of g for a complex with diameter D, and N_{sym} the number of asymmetric units for the given pointgroup symmetry). Thus, to achieve a uniform 3 Å resolution for our 300 Å-diameter D6 worm hemoglobin structure we only need: N ≥17 noise-free projection images (and only half that number if we are only refining at the ~6 Å resolution level). Note that these low numbers correspond to the typical number of projection images one uses as an "anchor set" for angular-reconstitution. Thus if we have those "17" projection images (at 3 Å resolution) – now inversing the argumentation – we can always generate the full 3D structure, and from that generate the thousands of projection images needed for a PM Euler-angle assignment. In AR we use those "17" projection images *directly* to find the Euler angles, without the detour over the 2D projection-matching references.

This argumentation shows that the AR Euler-angle assignment uses at least the *same* information as is used for PM. Although there can be differences in implementation details (Fourier space filtering, weightings applied, etc.) the AR approach is thus – at least – as good as the PM approach in this refinement context. AR is, in fact, superior to PM because the measurable data in single particle cryo-EM are the 2D projection images (or class averages) which are directly used for AR Euler angle assignments. The 3D reconstruction, required for generating the many PM references, in contrast, may not yet be filled uniformly with information at the early stages of processing. This missing information (large missing wedges in Fourier space) hampers the projection matching procedures leading to unnecessary wrong local-minimum convergences (van Heel and Harauz 1986; Sanz-García et al. 2010). It is thus no wonder that "starting models" used for PM-based refinements procedures in packages like "Frealign" (Grigorieff 2007) and "Relion" (Scheres 2012) are normally generated by AR or its various derivatives (Sigworth 2015; Nogales 2016).

Interestingly, PM was first introduced in conjunction with random Euler angle assignments for starting up 3D reconstructions (Harauz and Ottensmeyer 1984; van Heel 1984; van Heel and Harauz 1986; van Heel et al. 2000) but the emphasis of our developments in subsequent years moved to the AR approach which is robustly based on measured data alone. Indeed, our routine of starting up the Euler-angle assignments at random, followed by an AR refinement of those angles, demonstrates how much more effective AR is compared to PM.

In terms of final high-resolution refinements, AR achieves significant speed increases compared to PM-based refinements. Note that PM has historically always been used to align

the dataset in the form of multi-reference alignments (MRA) (van Heel and Stöffler-Meilicke 1985), where AR was used for the actual Euler angle assignment (Schatz et al. 1995; Klaholz et al. 2004). In our present ABC approach, we avoid these MRA procedures to avoid reference bias and to speed up the computations.

A further anisotropic effect we encountered (not to be confused with the anisotropic magnification discussed above) is the anisotropic distribution of the particle orientations in the data set. We found many more "top views" than "side views" which implies that the top views will overall accumulate more statistical significance in the 3D analysis. The primary consequence of these preferred orientations will be that in the best sampled directions, the statistics are better than in other directions. The reproducible resolution (as measured by the FSC) should thus be higher in the top-view orientations than in the poorer sampled side-view directions. For structures with a dihedral (Dn) or cyclic (Cn) pointgroup symmetry this results in a higher number of top views than side views (or vice versa). We have introduced simple FSC variants, the $FSC_{\beta<60°}$ and the $FSC_{\beta>60°}$, to assess this resolution anisotropy.

Anisotropic resolution is normal in electron tomography due to the limitations in tilting the sample holder (missing wedge/cone) and special versions of the FSC have been proposed for tomography (see, for example (Diebolder et al. 2015)). Here we use top- and side-view FSCs to evaluate the better sampling of the top views versus the side views. In an earlier publication it was found that the presence of "over abundant" top-view particles in a worm-hemoglobin dataset led to an extension of the 3D reconstructions in the top-view direction (Boisset et al. 1998). Our reconstructions performed with the linear weighted back-projection algorithm (Harauz and van Heel 1986), shows that those earlier results were due to the use of the non-linear, iterative "SIRT" reconstruction algorithm. We suggest that the top- and side-view FSCs are useful routine metrics in cryo-EM for all oligomeric structures with dihedral or cyclic pointgroup symmetries.

We have here reconstructed the *L. terrestris* giant hemoglobin to a global resolution FSC_{tot} of ~3.8 Å, comparable to the resolution levels achieved by X-ray crystallography for this giant hemoglobin (Royer et al. 2006); (Ruggiero Bachega et al. 2015). The cryo-EM resolution in the best parts of the map are sufficient to distinguish between different types of large side chains. The Asp-36 and Glu-32 residues of linker chain O (Figure 13c) have apparently lost most of their sidechain density, as reported in other studies (Allegretti et al. 2014; Bartesaghi et al. 2014). The resolution achieved is also sufficient to elucidate the environment of the heme groups with densities for the proximal and distal histidines (Figure 13b,c).

As a rule of thumb, the resolution achievable in X-ray crystallography is a function of the size of the crystallographic asymmetric unit. The smaller the unit cell, the higher the resolution because the molecules are more rigidly constrained within a limited space. Thus, a 2.6 Å resolution structure could be achieved from a crystal with only part of the 1/12th subunit in the

asymmetric unit (only a trimer of *abcd* tetramers is present in the asymmetric unit, no linker chains) (Strand et al. 2004). The *G. paulistusa* structure at 3.2 Å (Ruggiero Bachega et al. 2015) was based on crystals containing three protomers in the crystallographic asymmetric unit. The X-rays crystal of *L. terrestris* hemoglobin contains two full D6 molecules in the asymmetric unit (24 protomers) and was resolved to 5.5 Å (Royer et al. 2000); The structure was later refined to 3.5 Å by merging the information from different types of crystals (Royer et al. 2006). From this we conclude that the cryo-EM resolution we achieved is in line with the asymmetric-unit size-related resolution in X-ray crystallography.

The global resolution we achieved by cryo-EM is apparently thus limited by the overall size and flexibilities of the complex rather than by the instrumentation or data processing. In X-ray crystallography, this limitation is fundamental in that all variations occurring within the crystallographic unit cell are averaged out during data collection. In single-particle cryo-EM, in contrast, each molecule remains accessible – as an individual independent measurement – for further refining. This equally true for any fraction of the complex that can deviate from its average position within the complex; one can thus further refine separate parts of the complex in cryo-EM. This type of refining would compare to having only a smaller part of the complex in the asymmetric unit in X-ray crystallography. In cryo-EM as well as in X-ray crystallography this could thus yield a higher resolution local structure than that achieved for the full complex. An excellent example is given by a recent study on the functioning of the mitochondrial ATP synthase complex (Zhou et al. 2015).

In our current methodological paper, we restrict ourselves to the overall hemoglobin structure without entering into the issue of local optimizations. The most stable part of our cryo-EM map is clearly the central linker-chain core of the complex; the heme-containing globins on the outer periphery have a lower resolution due to local flexibilities/movements. In X-ray crystals containing the entire complex in the asymmetric unit, the situation appears reversed: the inner part of the L1/L2/L3 triple coiled-coil helix are not well resolved due to local flexibilities (Royer et al. 2000; Royer et al. 2006) compared the outer globin domains that more directly involved in the crystal contacts. Based on our rigid-core cryo-EM findings, our current working hypothesis is that oxy-deoxy conformational changes are localized in this outer layer of the worm hemoglobin (work in progress). It remains a challenge to understand the structural details of the very high oxygen-binding cooperativity would for the giant worm hemoglobin (Fushitani et al. 1986).

CONCLUSIONS

The quality of data collection in single-particle cryo-EM has dramatically improved by the introduction direct electron detectors. This, in combination with decades of developments in instrumentation and data-processing, have led to a true "resolution revolution" in the

field. The processing of the cryo-EM data however, still largely follows reference-based "projection-matching" approaches which are prone to reference bias. We argue that the angular reconstitution approach for Euler-orientation is necessarily better and faster than the projection matching approach, introduced some decades ago. The alignment-by-classification (ABC) approach used to achieve the best overall alignment of the full data set avoids the use of specific references. Our extension of the ABC approach to four dimensions (ABC-4D) yields a robust, reference-free pipeline for studying the structure of biological complexes in mixed conformational states by single-particle cryo-EM. We demonstrated the power of our approach by solving the structure of the giant *L. terrestris* hemoglobin to near-atomic resolution. The anisotropy of the orientations of the particles leads to a corresponding anisotropy in the reproducible resolution (FSC) in different orientations. The $FSC_{\beta<60°}$ and the $FSC_{\beta>60°}$ are straightforward metrics for assessing this effect. We have, for the first time, elucidated a heme group's quasi-atomic environment within the heme-binding pocket of a hemoglobin without crystallographic restraints.

MATERIALS & METHODS

Sample preparation

Blood of earthworm (*L. terrestris*) was extracted from the seventh segment of the body of some 20 individual worms and the (pooled) blood was diluted in 0.1 M Tris-HCl buffer, pH 7.0, 1 mM EDTA. Sample in the volume of 3 µl was applied on a Quantifoil grid (R2/2, Quantifoil Micro Tools GmbH) and plunge-frozen in liquid ethane using Vitrobot Mark IV (FEI) at 22 ºC, 100% humidity with a 2.5 s blot time.

EM data collection

The 4096x4096-pixel micrographs were collected using EPU software under low-dose conditions on the Titan Krios microscope (FEI, NeCEN), equipped with XFEG, Cs-corrector and Falcon II camera at 300 keV, magnification of 59000x (resulting in the final pixel size of 1.12 Å) at a defocus range -1.2 to -1 µm. The image acquisition was performed under control of the EPU software (FEI) in movie-mode, collecting series of 7 frames per movie (5235 movies in total) with a total electron dose of ~40 e-/$Å^2$. The average pixel size, originally calibrated at 1.12 Å/pixel, was later corrected to an estimated 1.11 Å/pixel to compensate for the re-interpolation applied to correct for (2.6%) anisotropic magnification found during the analysis.

Cryo-EM image processing

All EM image processing was performed in the context of the IMAGIC-4D software package (van Heel et al. 1996; van Heel et al. 2012). The more number-crunching tasks like MSA eigenvector data compression and 4D refinements were performed on MPI clusters running LINUX, typically using 32-128 cores. Interactive test runs, general preprocessing, classifications, etc. were typically performed on desktop/notebook computers under Windows 10 or

Ubuntu 14.04 operating systems.

Model building and refinements

Fitting of the atomic co-ordinates (PDB ID: 2GTL; (Royer et al. 2006)) into the cryo-EM densities was performed (and refined) using REFMAC5 (Vagin et al. 2004) and Coot software (Emsley et al. 2010). Graphic representations were performed using the PyMOL and UCSF Chimera packages (Pettersen et al. 2004; Schrodinger 2015).

Acknowledgements

We thank: Garib Murshudov for discussions on the use of the REFMAC software; Chris Diebolder and Peter Peters for discussions; Gert Oostergetel for providing samples; Élen Tomazela for administrational support; and Ralf Schmidt of Image Science GmbH, Berlin, for programming support. We acknowledge crucial discussions with Max Haider of CEOS GmbH and Gijs van Duinen of FEI Company on the issue of anisotropic magnification. We thank Richard Garrett and Eduardo Horjales of USP São Carlos for help with the ongoing interpretation of the map and for providing structural data prior to publication. Our research was financed in part by grants from: from the Dutch ministry of economic affairs Cyttron II FES-0908; HTS&M Initiative: FES-0901; by NanoNextNL of the Government of the Netherlands and 130 partners; from the BBSRC (Grant: BB/G015236/1)); from the Netherlands Organization for Scientific Research (NWO grant: 016.072.321); the Brazilian science foundations: CNPq (Grants CNPq-152746/2012-9 and CNPq-400796/2012-0), and the Instituto Nacional de C,T&I em Materiais Complexos Funcionais (INOMAT). We acknowledge the use of NeCEN electron microscopes (Leiden University) funded by NWO and the European Regional Development Fund of the European Commission.

REFERENCES

Adrian, M., Dubochet, J., Lepault, J. and McDowall, A.W. (1984) Cryo-electron microscopy of viruses. Nature 308(5954): 32-36.

Afanasyev, P., Ravelli, R.B., Matadeen, R., De Carlo, S., van Duinen, G., Alewijnse, B., Peters, P.J., Abrahams, J.P., Portugal, R.V., Schatz, M. and van Heel, M. (2015) A posteriori correction of camera characteristics from large image data sets. Sci Rep 5: 10317.

Allegretti, M., Mills, D.J., McMullan, G., Kühlbrandt, W. and Vonck, J. (2014) Atomic model of the F420-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector. Elife 3: e01963.

Amunts, A., Brown, A., Bai, X.C., Llacer, J.L., Hussain, T., Emsley, P., Long, F., Murshudov, G., Scheres, S.H. and Ramakrishnan, V. (2014) Structure of the yeast mitochondrial large ribosomal subunit. Science 343(6178): 1485-1489.

Bai, X.C., McMullan, G. and Scheres, S.H. (2015) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40(1): 49-57.

Bartesaghi, A., Matthies, D., Banerjee, S., Merk, A. and Subramaniam, S. (2014) Structure of beta-galactosidase at 3.2-A resolution obtained by cryo-electron microscopy. Proc Natl Acad Sci U S A 111(32): 11709-11714.

Boekema, E.J., Berden, J.A. and van Heel, M.G. (1986) Structure of mitochondrial F1-ATPase studied by electron microscopy and image processing. Biochimica et Biophysica Acta (BBA) - Bioenergetics 851(3): 353-360.

Boekema, E.J. and van Heel, M. (1988) Molecular shape of Lumbricus terrestris erythrocruorin studied by electron microscopy and image analysis. Biochim Biophys Acta 957(3): 370-379.

Boisset, N., Penczek, P.A., Taveau, J.-C., You, V., de Haas, F. and Lamy, J. (1998) Overabundant single-particle electron microscope views induce a three-dimensional reconstruction artifact.

Ultramicroscopy 74(4): 201-207.

Borland, L. and van Heel, M. (1990) Classification of image data in conjugate representation spaces. Journal of the Optical Society of America and Optics Image Science and Vision 7(4): 601-610.

Brilot, A.F., Chen, J.Z., Cheng, A., Pan, J., Harrison, S.C., Potter, C.S., Carragher, B., Henderson, R. and Grigorieff, N. (2012) Beam-induced motion of vitrified specimen on holey carbon film. J Struct Biol 177(3): 630-637.

Campbell, M.G., Cheng, A., Brilot, A.F., Moeller, A., Lyumkis, D., Veesler, D., Pan, J., Harrison, S.C., Potter, C.S., Carragher, B. and Grigorieff, N. (2012) Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. Structure 20(11): 1823-1828.

Cheng, Y. (2015) Single-particle cryo-EM at crystallographic resolution. Cell 161(3): 450-457.

Cheng, Y., Grigorieff, N., Penczek, P.A. and Walz, T. (2015) A primer to single-particle cryo-electron microscopy. Cell 161(3): 438-449.

Corrected Electron Optical Systems GmbH (2015) Cs-corrector.

Crewe, A.V. (1983) High-resolution scanning transmission electron microscopy. Science 221(4608): 325-330.

Crowther, R.A. (1971) Procedures for three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs. Philos Trans R Soc Lond B Biol Sci 261(837): 221-230.

de Haas, F., Zal, F., You, V., Lallier, F., Toulmond, A. and Lamy, J.N. (1996) Three-dimensional reconstruction by cryoelectron microscopy of the giant hemoglobin of the polychaete worm Alvinella pompejana. J Mol Biol 264(1): 111-120.

De Rosier, D.J. and Klug, A. (1968) Reconstruction of three dimensional structures from electron micrographs. Nature 217(5124): 130-134.

Diebolder, C.A., Faas, F.G., Koster, A.J. and Koning, R.I. (2015) Conical Fourier shell correlation applied to electron tomograms. J Struct Biol 190(2): 215-223.

Dube, P., Tavares, P., Lurz, R. and van Heel, M. (1993) The portal protein of bacteriophage SPP1: A DNA pump with 13-fold symmetry. EMBO J 12(4): 1303-1309.

Elmer, J., Palmer, A.F. and Cabrales, P. (2012) Oxygen delivery during extreme anemia with ultra-pure earthworm hemoglobin. Life Sci 91(17-18): 852-859.

Emsley, P., Lohkamp, B., Scott, W.G. and Cowtan, K. (2010) Features and development of Coot. Acta Crystallographica Section D 66(4): 486-501.

Fabrigar, L.R. and Wegener, D.T. (2011) Exploratory factor analysis (Oxford University Press).

Faruqi, A.R., Henderson, R., Pryddetch, M., Allport, P. and Evans, A. (2005) Direct single electron detection with a CMOS detector for electron microscopy. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 546(1–2): 170-175.

Fushitani, K., Imai, K. and Riggs, A.F. (1986) Oxygenation properties of hemoglobin from the earthworm, Lumbricus terrestris. Effects of pH, salts, and temperature. J Biol Chem 261(18): 8414-8423.

Grant, T. and Grigorieff, N. (2015) Automatic estimation and correction of anisotropic magnification distortion in electron microscopes. J Struct Biol.

Grigorieff, N. (2007) FREALIGN: high-resolution refinement of single particle structures. J Struct Biol 157(1): 117-125.

Harauz, G. and Ottensmeyer, F. (1984) Direct three-dimensional reconstruction for macromolecular complexes from electron micrographs. Ultramicroscopy 12(4): 309-319.

Harauz, G. and van Heel, M. (1986) Exact filters for general geometry three dimensional reconstruction. Optik 73: 146-156.

Henderson, R. (1995) The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. Quarterly reviews of biophysics 28(02): 171-193.

Henderson, R. (2013) Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. Proc Natl Acad Sci U S A 110(45): 18037-18041.

Hirsch, R.E., Jelicks, L.A., Wittenberg, B.A., Kaul, D.K., Shear, H.L. and Harrington, J.P. (1997) A first evaluation of the natural high molecular weight polymeric Lumbricus terrestris hemoglobin as an oxygen carrier. Artif Cells Blood Substit Immobil Biotechnol 25(5): 429-444.

Image Science: Michael Schatz (2015) Image Science - em2em 3DEM conversion program. 2015.

Kastner, B., Fischer, N., Golas, M.M., Sander, B., Dube, P., Boehringer, D., Hartmuth, K., Deckert, J., Hauer, F., Wolf, E., Uchtenhagen, H., Urlaub, H., Herzog, F., Peters, J.M., Poerschke, D., Luhrmann, R. and Stark, H. (2008) GraFix: sample preparation for single-particle electron cryomicroscopy. Nat Methods 5(1): 53-55.

Klaholz, B.P., Myasnikov, A.G. and Van Heel, M. (2004) Visualization of release factor 3 on the ribosome during termination of protein synthesis. Nature 427(6977): 862-865.

Kühlbrandt, W. (2014) Biochemistry. The resolution revolution. Science 343(6178): 1443-1444.

Kuijper, M., van Hoften, G., Janssen, B., Geurink, R., De Carlo, S., Vos, M., van Duinen, G., van Haeringen, B. and Storms, M. (2015) FEI's direct electron detector developments: Embarking on a revolution in cryo-TEM. J Struct Biol.

Kunath, W., Weiss, K., Sackkongehl, H., Kessel, M. and Zeitler, E. (1984) Time-resolved low-dose microscopy of glutamine-synthetase molecules. Ultramicroscopy 13(3): 241-252.

Li, X., Mooney, P., Zheng, S., Booth, C.R., Braunfeld, M.B., Gubbens, S., Agard, D.A. and Cheng, Y. (2013) Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat Methods 10(6): 584-590.

Louys, M., Schatz, M. and van Heel, M. (1989) Classification of rotational amplitude spectra of individual molecular images: a study of focus conditions of ice-embedded preparations. Eur J Cell Biol 49 (Suppl. 27): 58.

McMullan, G., Chen, S., Henderson, R. and Faruqi, A.R. (2009) Detective quantum efficiency of electron area detectors in electron microscopy. Ultramicroscopy 109(9): 1126-1143.

McMullan, G., Faruqi, A.R., Clare, D. and Henderson, R. (2014) Comparison of optimal performance at 300keV of three direct electron detectors for use in low dose electron microscopy. Ultramicroscopy 147: 156-163.

Milazzo, A.-C., Leblanc, P., Duttweiler, F., Jin, L., Bouwer, J.C., Peltier, S., Ellisman, M., Bieser, F., Matis, H.S., Wieman, H., Denes, P., Kleinfelder, S. and Xuong, N.-H. (2005) Active pixel sensor array as a detector for electron microscopy. Ultramicroscopy 104(2): 152-159.

Mindell, J.A. and Grigorieff, N. (2003) Accurate determination of local defocus and specimen tilt in electron microscopy. J Struct Biol 142(3): 334-347.

Mouche, F., Boisset, N. and Penczek, P.A. (2001) Lumbricus terrestris hemoglobin—The architecture of linker chains and structural variation of the central toroid. J Struct Biol 133(2–3): 176-192.

Nogales, E. (2016) The development of cryo-EM into a mainstream structural biology technique. Nat Meth 13(1): 24-27.

Penczek, P., Radermacher, M. and Frank, J. (1992) Three-dimensional reconstruction of single particles embedded in ice. Ultramicroscopy 40(1): 33-53.

Perutz, M. (1978) Hemoglobin structure and respiratory transport. Scientific American 239: 92-125.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. Journal of Computational Chemistry 25(13): 1605-1612.

RCSB Protein Data Bank (2015) Hemoglobin. 2015.

Roche, C.J., Talwar, A., Palmer, A.F., Cabrales, P., Gerfen, G. and Friedman, J.M. (2015) Evaluating the capacity to generate and preserve nitric oxide bioactivity in highly purified earthworm erythrocruorin: A giant polymeric hemoglobin with potential blood substitute properties. Journal of Biological Chemistry 290(1): 99-117.

Rosenthal, P.B. and Henderson, R. (2003) Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J Mol Biol 333(4): 721-745.

Royer, W.E., Jr., Sharma, H., Strand, K., Knapp, J.E. and Bhyravbhatla, B. (2006) Lumbricus erythrocruorin at 3.5 A resolution: architecture of a megadalton respiratory complex. Structure 14(7): 1167-1177.

Royer, W.E., Jr., Strand, K., van Heel, M. and Hendrickson, W.A. (2000) Structural hierarchy in erythrocruorin, the giant respiratory assemblage of annelids. Proc Natl Acad Sci U S A 97(13): 7107-7111.

Ruggiero Bachega, J.F., Vasconcelos Maluf, F., Andi, B., D'Muniz Pereira, H., Falsarella Carazzollea, M., Orville, A.M., Tabak, M., Brandao-Neto, J., Garratt, R.C. and Horjales Reboredo, E. (2015) The structure of the giant haemoglobin from Glossoscolex paulistus. Acta Crystallogr D Biol Crystallogr 71(Pt 6): 1257-1271.

Sander, B., Golas, M.M. and Stark, H. (2003) Automatic CTF correction for single particles based upon multivariate statistical analysis of individual power spectra. J Struct Biol 142(3): 392-401.

Sanz-García, E., Stewart, A.B. and Belnap, D.M. (2010) The random-model method enables ab initio 3D reconstruction of asymmetric particles and determination of particle symmetry. J Struct Biol 171(2): 216-222.

Schatz, M., Orlova, E.V., Dube, P., Jäger, J. and van Heel, M. (1995) Structure of Lumbricus terrestris hemoglobin at 30 Å resolution determined using angular reconstitution. J Struct Biol 114(1): 28-40.

Schatz, M. and Van Heel, M. (1990) Invariant classification of molecular views in electron micrographs. Ultramicroscopy 32(3): 255-264.

Scheres, S.H., Valle, M., Nunez, R., Sorzano, C.O., Marabini, R., Herman, G.T. and Carazo, J.M. (2005) Maximum-likelihood multi-reference refinement for electron microscopy images. J Mol Biol 348(1): 139-149.

Scheres, S.H.W. (2012) RELION: Implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol 180(3): 519-530.

Scherzer, O. (1949) The theoretical resolution limit of the electron microscope. Journal of Applied Physics 20(1): 20-29.

Schrodinger, L. (2015) The PyMOL molecular graphics system, version 1.8.

Sigworth, F. (1998) A maximum-likelihood approach to single-particle image refinement. J Struct Biol 122(3): 328-339.

Sigworth, F.J. (2015) Principles of cryo-EM single-particle image processing. Microscopy (Oxf).

Stewart, A. and Grigorieff, N. (2004) Noise bias in the refinement of structures derived from single particles. Ultramicroscopy 102(1): 67-84.

Strand, K., Knapp, J.E., Bhyravbhatla, B. and Royer, W.E., Jr. (2004) Crystal structure of the hemoglobin dodecamer from Lumbricus erythrocruorin: allosteric core of giant annelid respiratory complexes. J Mol Biol 344(1): 119-134.

Subramaniam, S. (2013) Structure of trimeric HIV-1 envelope glycoproteins. Proc Natl Acad Sci U S A 110(45): E4172-4174.

Suloway, C., Pulokas, J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., Stagg, S., Potter, C.S. and Carragher, B. (2005) Automated molecular microscopy: the new Leginon system. J Struct Biol

151(1): 41-60.

Vagin, A.A., Steiner, R.A., Lebedev, A.A., Potterton, L., McNicholas, S., Long, F. and Murshudov, G.N. (2004) REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta Crystallographica Section D 60(12 Part 1): 2184-2195.

van Bruggen, E.F. and Weber, R.E. (1974) Erythrocruorin with anomalous quaternary structure from the polychaete Oenone fulgida. Biochim Biophys Acta 359(1): 210-214.

van Heel, M. (1982) Detection of objects in quantum-noise-limited images. Ultramicroscopy 7(4): 331-341.

van Heel, M. (1984) Three-dimensional reconstructions from projections with unknown angular relationship. 8th European Congress Electron Microscopy Budapest. 2: 1347-1348.

van Heel, M. (1987) Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy 21(2): 111-123.

van Heel, M. (1989) Classification of very large electron microscopical image data sets. (Reutlingen, Allemagne Elsevier), 13.

van Heel, M. (2013) Finding trimeric HIV-1 envelope glycoproteins in random noise. Proc Natl Acad Sci U S A 110(45): E4175-4177.

van Heel, M. (2014) [3dem] Magnification anisotropy at low mag settings on Titan Krios.

van Heel, M. and Frank, J. (1981) Use of multivariates statistics in analysing the images of biological macromolecules. Ultramicroscopy 6(1): 187-194.

van Heel, M., Gowen, B., Matadeen, R., Orlova, E.V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M. and Patwardhan, A. (2000) Single-particle electron cryo-microscopy: towards atomic resolution. Quarterly reviews of biophysics 33(4): 307-369.

van Heel, M. and Harauz, G. (1986) Resolution criteria for three dimensional reconstruction. Optik 73(3): 119-122.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R. and Schatz, M. (1996) A new generation of the IMAGIC image processing system. J Struct Biol 116(1): 17-24.

van Heel, M., Portugal, R., Rohou, A., Linnemayr, C., Bebeacua, C., Schmidt, R., Grant, T. and Schatz, M. (2012) Four-dimensional cryo electron microscopy at quasi atomic resolution: IMAGIC 4D. International Tables for Crystallography F: 624-628.

van Heel, M., Portugal, R. and Schatz, M. (2009) Multivariate statistical analysis in single particle (cryo) electron microscopy. Handbook on DVD 3D-EM in Life Sciences.

van Heel, M. and Schatz, M. (2005) Fourier shell correlation threshold criteria. J Struct Biol 151(3): 250-262.

van Heel, M., Schatz, M. and Orlova, E. (1992) Correlation functions revisited. Ultramicroscopy 46(1): 307-316.

van Heel, M. and Stöffler-Meilicke, M. (1985) Characteristic views of E. coli and B. stearothermophilus 30S ribosomal subunits in the electron microscope. EMBO 4(9): 2389.

van Heel, M.G. (1978) Imaging of relatively strong objects in partially coherent illumination in optics and electron optics Optik 49(4): 389-408.

Zal, F., Toulmond, A. and Lallier, F. (2001) Utilisation comme substitut sanguin d'une hemoglobine extracellulaire de poids moleculaire eleve. (Google Patents).

Zhao, J., Brubaker, M.A., Benlekbir, S. and Rubinstein, J.L. (2015) Description and comparison of algorithms for correcting anisotropic magnification in cryo-EM images. J Struct Biol 192(2): 209-215.

Zhou, A., Rohou, A., Schep, D.G., Bason, J.V., Montgomery, M.G., Walker, J.E., Grigorieff, N. and Rubinstein, J.L. (2015) Structure and conformational states of the bovine mitochondrial ATP

synthase by cryo-EM. Elife 4.

5

Single-particle cryo-EM based on alignment by classification (ABC): *Lumbricus terrestris* hemoglobin at near-atomic resolution

CHAPTER 6

CHALLENGES IN SINGLE-PARTICLE CRYO-EM: HETEROGENEOUS AND SMALL PROTEIN ESPB, SUBSTRATE OF TYPE VII SECRETION SYSTEM

Pavel Afanasyev[1,2,3], Musa Sani[3], Nicole van der Wel[3], Raimond B.G. Ravelli[1], Massimiliano Maletta[3], Florence Pojer[4], Stewart T. Cole[4], Marin van Heel[2,5,6], Peter J. Peters[1,3]*

In preparation

[1]The Maastricht Multimodal Molecular Imaging Institute, Maastricht University, 6229 ER Maastricht, The Netherlands

[2]Institute of Biology Leiden, Leiden University, 2333 CC Leiden, The Netherlands.

[3]The Netherlands Cancer Institute, 1066 CX, Amsterdam, The Netherlands

[4]Global Health Institute, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

[5]Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, UK.

[6]Brazilian Nanotechnology National Laboratory – LNNano, CNPEM, C.P. 6192, 13083-970 Campinas SP, Brasil.

*Corresponding author

ABSTRACT

The virulence of *Mycobacterium tuberculosis* is largely determined by the ESX-1 (type VII) secretion systems. The ESX-1 is known to secrete several proteins with different roles in virulence. Here we are aiming to reveal the structure of the EspB protein, which is secreted into the host cell in an unknown manner. Secreted EspB can form oligomers detected in culture filtrates and cell lysates of *M. tuberculosis*. We obtained low-resolution 3D-reconstructions of the cleaved form of EspB in hexameric and heptameric oligomeric states. Due to the small size of EspB and heterogeneity of the sample, obtaining atomic resolution of the EspB structure by single-particle cryo-EM is challenging. Our data provides insights towards revealing the type VII secretion systems structure.

INTRODUCTION

Mycobacteria species are dangerous human pathogens, causing severe diseases like tuberculosis and leprosy. Tuberculosis is the second only to HIV/AIDS as the greatest killer worldwide due to a single infectious agent, *M. tuberculosis* (WHO 2014). The efficacy of the commonly administrated vaccine for tuberculosis (BCG, prepared from a strain of *M. bovis*) is highly variable (Andersen and Woodworth 2014). About one-third of the world's population still has latent tuberculosis (WHO 2014). A troubling development is the identification of multiple-drug-resistant strains, which represents a severe danger to the human population (Fogel 2015). Understanding virulence on a molecular and structural level is essential for the development of new drugs and vaccines for tuberculosis.

The virulence of *M. tuberculosis* is largely determined by systems of protein transport. Besides the more conserved Sec- and Tat-mediated protein secretion pathways (Natale et al. 2008), mycobacteria developed their own secretion systems, the ESX (named ESX-1 to ESX-5; also known as "type VII") secretion systems (Abdallah et al. 2007). One of these secretion systems, ESX-1, is of paramount importance for pathogenesis. It is responsible for many functions, including phagosome permeabilisation (de Jonge et al. 2007; Watson et al. 2012) and phagosomal escape into the cytosol during bacterial infection (van der Wel et al. 2007). The group of ESX secretion systems is unique: proteins, encoded in the ESX loci, show low homology to any of the known proteins, composing type I - VI secretion systems of Gram-negative bacteria (van Pittius et al. 2001). This difference is related to an outer membrane bilayer (so-called "mycomembrane"), which makes *mycobacteria* also different to most of Gram positive bacteria (Zuber et al. 2008). Besides, the mycobacterial capsular layer with its unique composition forms a barrier for protein translocation (Sani et al. 2010; Forrellad et al. 2013). All these characteristics suggest that the whole mechanism of the protein secretion of *mycobacteria* should be different to both Gram-positive and Gram-negative bacteria.

Though details of the secretion mechanism and the whole structure of the type VII secretion

(T7S) apparatus are still mostly unknown, some components and substrates of ESX systems have been well studied. Importantly, the core of ESX-1 system is composed of the conserved T7S transmembrane proteins EccC, EccB, EccD, EccE and a membrane-bound protease MycP (Houben et al. 2014). Studies of the two highly immunogenic secreted proteins, EsxA (ESAT-6) and EsxB (CFP-10) established their essential role in the virulence (Stanley et al. 2003) and determined their structures (Renshaw et al. 2005). These proteins form heterodimers, belonging to the family of small helical proteins, known as WxG100 proteins (Pallen 2002). The WxG100 proteins are similar to another families of secreted proteins: PE and PPE proteins, which also form heterodimers, structurally similar to the EsxAB complex.

The recently reported crystal structures of another virulence factor, the EspB protein, are similar to the EsxAB and PE-PPE complexes (Korotkova et al. 2015; Solomonson et al. 2015). Moreover, these studies showed that EspB can form oligomeric structures. Earlier studies showed that EspB is necessary for virulence and growth of bacteria in macrophages (Xu et al. 2007). Interestingly, EspB is required for the secretion of EsxA and EsxB and inversely, EsxA and EsxB are essential for the secretion of EspB (Xu et al. 2007). During the secretion process, the C-terminal domain (~10 kDa) of the full-length EspB protein (~50 kDa) is being cleaved by $MycP_1$ protease (Ohol et al. 2010). The mechanism of secretion of all substrates and the exact role of EspB in the infection process remain unknown.

Most of the structural studies of the ESX-1 components were performed using NMR spectroscopy or X-ray crystallography. The first method has a limitation on the size of the protein (NMR typically requires samples of <40 kDa), whereas the second requires proteins to be crystalized, which is a challenge in many cases. Recent progress in the developments of cryogenic electron microscopy (cryo-EM) instrumentation (Kühlbrandt 2014; Cheng 2015; Kuijper et al. 2015) allows one to study protein complexes of various size without having to obtain crystals. In our study we use single-particle cryo-EM to study the truncated (residues 1-338) oligomeric EspB complexes. We were able to identify the presence of $EspB_{1-338}$ in two isoforms: hexameric and heptameric and obtain low-resolution 3D-reconstructions of these proteins.

RESULTS

The initial biochemical analysis of the full-length recombinant EspB complex ($EspB_{1-460}$) by SDS PAGE revealed presence of the processed form of the protein ($EspB_{1-338}$) in the sample (data not shown). Therefore, to reduce the complexity in the dataset and avoid this heterogeneity, we focused in our study on the analysis only of the truncated form of EspB ($EspB_{1-338}$).

To perform a structural analysis of the $EspB_{1-338}$ oligomers we used the FEI Titan Krios microscope (NeCEN), equipped with the FEI Falcon II camera and operated in the movie-mode. We acquired a single-particle cryo-EM dataset of 207 movies. The particles in the micrographs have a size

of about 8.5 nm in diameter. In this dataset, the EspB particles showed a preferred "top-view" orientation (Figure 1a), where most of the particles have a donut-like shape. However, some rectangular-shaped and elongated side views can be identified in the raw single particles as well (Figure 1b). Those views were essential to create a reliable 3D-reconstruction. We tried several approaches to increase the number of side views (charging grids, poly-L-lysine coating etc.), however they appeared to be unsuccessful.

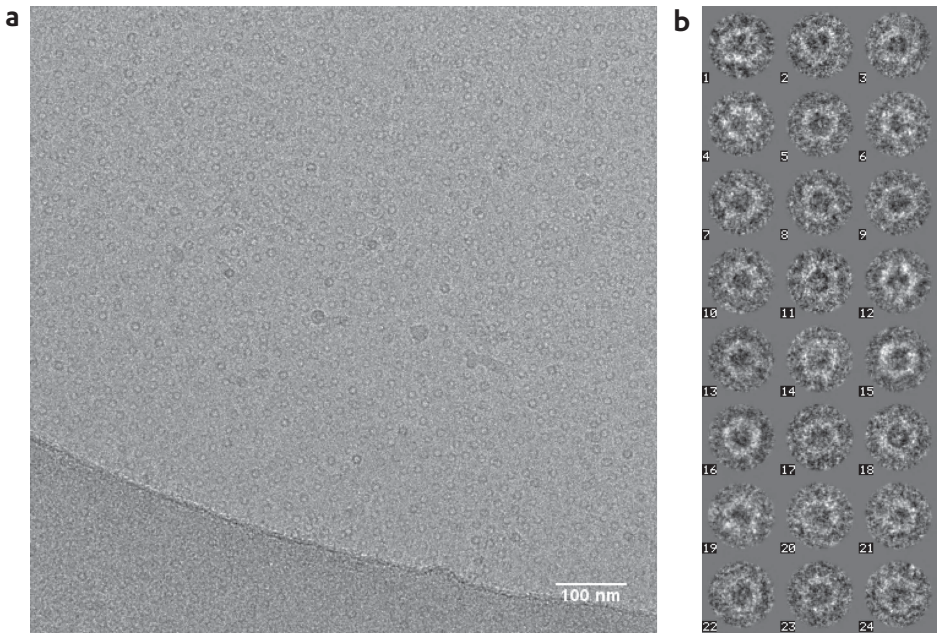

Figure 1. (a) A typical raw micrograph of $EspB_{1\text{-}338}$ oligomers (dark particles on light background). Most of the particles have a donut shape. **(b)** A gallery of the masked picked particles of ~8.5 nm from the prepared aligned movie-sums (contrast of the particles is inverted compared to (a)).

Using our methodology for reference-free particle picking (Chapters 4,5), we picked and initially selected 39304 good particles from the far-from-focus micrographs and performed multivariate statistical (eigenimage) analysis (van Heel and Frank 1981). Multivariate statistical data compression allows representing data as a linear combination of eigenimages (eigenvectors). The eigenimages describe the variance within the dataset and thus facilitate interpretation of the data. This approach is commonly used to identify the symmetry of cyclic oligomers (Dube et al. 1993; van Heel et al. 1996; White et al. 2004).

The eigenimages of the unaligned $EspB_{1\text{-}338}$ particles are presented in the Figure 2a. The first eigenimage closely resembles a total average of all the particles in the dataset. Due to the preferred orientation, the first eigenimages characterize prevailing top views. The eigenimages #2 and #3 clearly indicate a cyclic 7-fold (C7) symmetry in the dataset. Similarly, the eigenimages #5 and #6 indicate presence of a 6-fold (C6) symmetry; eigenimage #4

represents the difference in size between the different oligomeric states. Thus, the $EspB_{1-338}$ complex was shown be found in two oligomeric populations: heptamers and hexamers, where the first prevail. This was in agreement with the results of gel filtration, showing presence of more than one oligomeric state of the complex. The two subsequent eigenimages (#7 and #8) contain 8-fold component, corresponding to the tilted views of the heptamers. Eigenimages #9 and #10 are likely related to the side views.

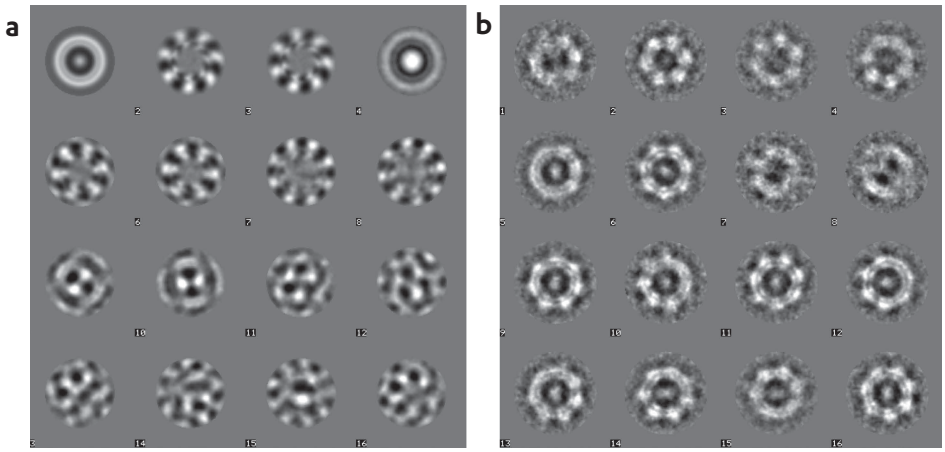

Figure 2. (a) Eigenimage analysis of the reference-free picked unaligned particles of the $EspB_{1-338}$ oligomers. The eigenimages represent presence of the two different oligomeric $EspB_{1-338}$ conformations in the dataset – heptameric (eigenimages #2,3; C7-symmetry) and hexameric (eigenimages #5,6; C6-symmetry). Eigenimage #1 represents the total average of all the particles within the dataset, whereas #4 reflects the difference in size between two oligomeric states. The 8-fold component of the eigenimages #7 and #8 corresponds to the tilted views of the heptamers. Eigenimages #9 and #10 may represent the side-views of the oligomers. **(b)** Examples of the class-averages of the $EspB_{1-338}$ particles resulting from hierarchical ascendant classification based on the first 10 eigenimages. Particles, corresponding to different classes can roughly be split into three subsets: "C7 top views" (classes 4-6, 9-16); "C6 top views" (classes 2,3); "side views" (classes 1,7,8).

We used hierarchical ascendant classification (van Heel 1989) based on the first 10 eigenimages to split the heterogeneity within the dataset. The resulting classes (Figure 2b) could easily be separated into three subsets: of "hexamers" (corresponding to C6-top and C6-tilted characteristic views), "heptamers" (corresponding to C7-top and C7-tilted characteristic views) and "side-views" (corresponding to both C6-side and C7-side characteristic views). We analyzed rotationally averaged class averages of hexamers and heptamers and found a difference in their diameter of ~10%. Due to the lack of side views, we could not split the hexamers side views from the heptamers side views.

Each subset was processed in parallel independently. Multivariate statistical analysis of each of the two subsets yielded three groups of two-dimensional class averages. We used the smaller size "side-views" class averages together with selected top views of "hexamers" to create initial 3D-reconstructions of hexamers of $EspB_{1-338}$. Similarly, the larger "side-views"

class averages together with selected top views of "heptamers" were used to create initial 3D-reconstructions of heptamers of $EspB_{1-338}$. Angular reconstitution was performed by random startup procedure (van Heel 1987) with imposed C6-symmetry for the hexamer 3D-reconstruction and C7-symmetry for the heptamers 3D-reconstruction. We performed a 4D-refinement of both reconstructions using all class averages in a competitive way, so that each class would find a 3D-reconstruction it fits best to (van Heel et al. 2012). This refinement yielded in two 3D-reconstructions of the heptamer and the hexamer of 16 Å and 13 Å resolution respectively, estimated by a half-bit criterion (van Heel and Schatz 2005). Figure 3a represents hexameric oligomeric state of $EspB_{1-338}$ and Figure 3b represents the heptameric oligomeric state of $EspB_{1-338}$. Unfortunately, the particle images in this dataset did not contain enough low-frequency information (representing general features of the molecules) to perform accurate classification and alignments, therefore the further refinements did not help to obtain a higher resolution.

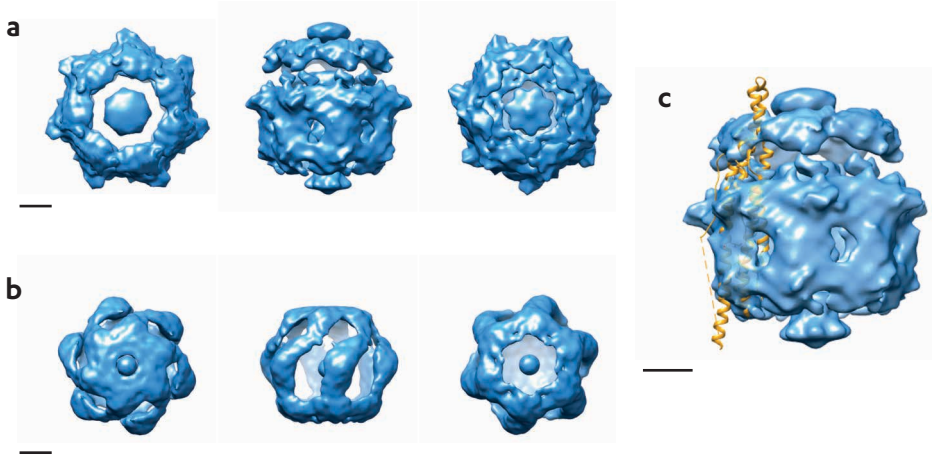

Figure 3. (a) Low-resolution map of the heptameric form of $EspB_{1-338}$ (left to right: top view, side-view and bottom view). **(b)** Low-resolution map of the hexameric form of $EspB_{1-338}$ (left to right: top view, side-view and bottom view). **(c)** Rigid-body fitting of the crystal structure of a monomer of $EspB_{7-278}$ (Korotkova et al. 2015) into a subunit of the heptameric form of $EspB_{1-338}$. The overall structure is in agreement with the 30 Å map of the full-length EspB of (Solomonson et al. 2015). The missing density at this threshold level might be either caused by flexible regions of the protein or large conformational changes during oligomerization. Scale bars correspond to 20 Å.

DISCUSSION

Cryo-EM is a rapidly developing structural biology technique for solving structures of protein complexes in their close-to-native conditions. The latest technological developments of direct electron detectors improved their detective quantum efficiency and facilitated solving structures of protein complexes up to a near-atomic resolution (Kühlbrandt 2014; Cheng 2015; Cheng et al. 2015). A powerful advantage of cryo-EM among other structural biology

techniques is in its ability for identifying different oligomeric and conformational states of the protein complexes in the same solution.

Using single-particle cryo-EM we demonstrated that $EspB_{1-338}$ can be present in two different oligomeric states – hexa- and heptameric, in comparison to the only heptameric one, recently identified in two independent studies (Korotkova et al. 2015; Solomonson et al. 2015). In those studies, the electron microscopy was performed by the negative staining technique (Brenner and Horne 1959; van Bruggen et al. 1960), which yields high-contrast images but limits the achievable resolution and may introduce artefacts. In our case we used the single-particle cryo-EM approach, in which the sample is imaged at its close-to-native conditions and the structure of the protein complex can be potentially solved to a near-atomic resolution. Our EspB structure of the heptameric state of EspB is of higher resolution and is in agreement with the results of (Solomonson et al. 2015).

In practice, achieving high-resolution 3D-reconstructions for small and heterogeneous complexes might be tricky due to a number of limitations and problems (for details see Chapter 1).

(i) First, the results of the cryo-EM data analysis greatly depend on the size of the protein complex. The majority of the solved structures with a near-atomic resolution by cryo-EM corresponds to the large-size proteins (ribosomes, proteasomes, viruses etc.). Protein complexes of smaller than ~500 kDa are still a great challenge due to the lack of low-resolution information in cryo-EM images, representing particle projections. This hampers alignments and correct angular determination.

(ii) Dealing with impure, heterogeneous or flexible samples requires advanced methodology in single-particle processing. Particles, corresponding to different conformations, can affect the quality or resolution of 3D-reconstructions if they are not classified properly. Samples, containing flexible regions of the biomolecules, are particularly tricky. Often those regions have a lower resolution in the final maps or even a missing density.

(iii) The sample preparation is of paramount importance for archiving a near-atomic resolution. In case of a strong preferred orientation of particles, the missing information from the rare characteristic views might not allow obtaining high-resolution 3D-reconstruction.

In our work we encountered all three problems at once for obtaining high-resolution 3D-reconstruction. The $EspB_{1-338}$ complex is small protein (240 kDa hexameric and 280 kDa heptameric form) of a barrel-like shape. Moreover, our sample was found in two oligomeric states with a preferred top-view orientation. Nevertheless, we were able to obtain two low-resolution 3D-reconstructions, representing different oligomeric states.

Such an oligomeric organization of the monomers of EspB might be important for facilitating the secretion of the other proteins. *In vivo*, the secreted EspB was shown to form oligomers, which were detected in culture filtrates and cell lysates of *M. tuberculosis* (Korotkova et al.

2015). Interestingly, EsxA and EsxB proteins (which require EspB for the secretion) have a similar size to the EspB cleft. This might allow to speculate that EspB might serve as a chaperone protein for ExsA and/or EsxB, although such a role of EspB is to be defined based on the high-resolution data. This concept could easily fit the fact, that during the secretion, the full-length EspB is being cleaved by the $MycP_1$ protease (Ohol et al. 2010), supposedly for the release of the transported protein. However, to prove this hypothesis and reveal the exact role of EspB in the secretion process, as well as the whole secretion mechanism, further structural studies are required.

The 3D-reconstructions we obtained, as well as the heptameric map from (Solomonson et al. 2015), contain artefacts: densities in the middle of each oligomer. This may result from an inaccurate angular determination and the wrong assignment of the 3D-membership. To overcome this problem and current limits in the refinement we are working on the optimization of the protein purification and the sample preparation. Studying the full-length EspB requires an especially pure sample, which particularly has no any form of present truncated EspB. Besides, we are also planning to facilitate our image analysis by collecting data with the Gatan K2 Summit direct electron detector. This camera is known to perform better for small proteins due to its higher digital quantum efficiency at low special frequencies (McMullan et al. 2014). Besides, to increase the contrast, we consider using cryo-EM phase plates (Glaeser 2013; Danev et al. 2014).

Solving the structures of the whole and cleaved oligomers with atomic resolution would reveal the mechanism of the EspB secretion and might identify the exact role of EspB in the bacterial virulence. This also would increase our overall knowledge on the type VII secretion, which is crucial for the pathogenicity. At last, information about the structure and mechanism of the functioning of each component of the T7S systems will contribute to the rational development of new drugs/vaccines against diseases, caused by *Mycobacteria* species.

MATERIALS AND METHODS

The oligimeric fractions of N-terminal fragment EspB (residues 1-338) and full-length EspB protein the were purified as described in (Wagner et al. 2013; Korotkova et al. 2015) and kindly provided by the group of S. Cole (EPFL, Lausanne, Switzerland). The size of the protein samples were verified by SDS-PAGE prior to the plunge-freezing. The samples (3 µl in 50mM Tris pH 7.5 and 500mM NaCl) were applied on the Quantifoil grids (Quantifoil Micro Tools GmbH), 200 mesh, coated with QC R2/2 thin double films. The grids were plunge-frozen in liquid ethane using Vitrobot Mark IV (FEI) at 22 ºC, 100% humidity with a 2.5 s blot time. The micrographs were collected using EPU software under low-dose conditions on the FEI Titan Krios microscope (NeCEN), equipped with FEG and Falcon II camera at 300 keV, magnification of 59000x (resulting in the final pixel size of 1.34 Å) at a defocus range -1.5 to -2 µm. The image

acquisition was performed during 1.5 seconds in a movie-mode in series of 7 frames with a total electron dose of ~30 e-/Å^2.

The processing of the data was performed in IMAGIC-4D software (van Heel et al. 2012) according to our standard methodology (described in Chapters 2, 3, 4 and 5) with differences in the advanced classification procedure as described in the main text.

The rigid-body fitting of the atomic structure (PDB entry 4XXX) was performed in the UCSF Chimera package (Pettersen et al. 2004).

ACKNOWLEDGEMENTS

We thank Willem Tichelaar for the sample preparation for the cryo-negative staining data; S. De Carlo and Rishi Matadeen for the sample preparation and data collection at NeCEN facility; Michael Schatz and Ralf Schmidt of Image Science GmbH, Berlin, for programming support and B. Alewijnse for the computational support.

REFERENCES

Abdallah, A.M., Gey van Pittius, N.C., Champion, P.A., Cox, J., Luirink, J., Vandenbroucke-Grauls, C.M., Appelmelk, B.J. and Bitter, W. (2007) Type VII secretion-mycobacteria show the way. Nat Rev Microbiol 5(11): 883-891.

Andersen, P. and Woodworth, J.S. (2014) Tuberculosis vaccines – rethinking the current paradigm. Trends in Immunology 35(8): 387-395.

Brenner, S. and Horne, R.W. (1959) A negative staining method for high resolution electron microscopy of viruses. Biochim Biophys Acta 34: 103-110.

Cheng, Y. (2015) Single-particle cryo-EM at crystallographic resolution. Cell 161(3): 450-457.

Cheng, Y., Grigorieff, N., Penczek, P.A. and Walz, T. (2015) A primer to single-particle cryo-electron microscopy. Cell 161(3): 438-449.

Danev, R., Buijsse, B., Khoshouei, M., Plitzko, J.M. and Baumeister, W. (2014) Volta potential phase plate for in-focus phase contrast transmission electron microscopy. Proceedings of the National Academy of Sciences 111(44): 15635-15640.

de Jonge, M.I., Pehau-Arnaudet, G., Fretz, M.M., Romain, F., Bottai, D., Brodin, P., Honoré, N., Marchal, G., Jiskoot, W. and England, P. (2007) ESAT-6 from Mycobacterium tuberculosis dissociates from its putative chaperone CFP-10 under acidic conditions and exhibits membrane-lysing activity. J Bacteriol 189(16): 6028-6034.

Dube, P., Tavares, P., Lurz, R. and van Heel, M. (1993) The portal protein of bacteriophage SPP1: A DNA pump with 13-fold symmetry. EMBO J 12(4): 1303-1309.

Fogel, N. (2015) Tuberculosis: A disease without boundaries. Tuberculosis (Edinb).

Forrellad, M.A., Klepp, L.I., Gioffré, A., Sabio y Garcia, J., Morbidoni, H.R., Santangelo, M.d.l.P., Cataldi, A.A. and Bigi, F. (2013) Virulence factors of the Mycobacterium tuberculosis complex. Virulence 4(1): 3-66.

Glaeser, R.M. (2013) Invited Review Article: Methods for imaging weak-phase objects in electron microscopy. The Review of Scientific Instruments 84(11): 111101.

Houben, E.N., Korotkov, K.V. and Bitter, W. (2014) Take five - Type VII secretion systems of Mycobacteria. Biochim Biophys Acta 1843(8): 1707-1716.

Korotkova, N., Piton, J., Wagner, J.M., Boy-Rottger, S., Japaridze, A., Evans, T.J., Cole, S.T., Pojer, F.

and Korotkov, K.V. (2015) Structure of EspB, a secreted substrate of the ESX-1 secretion system of Mycobacterium tuberculosis. J Struct Biol 191(2): 236-244.

Kühlbrandt, W. (2014) Cryo-EM enters a new era. Elife 3: e03678.

Kuijper, M., van Hoften, G., Janssen, B., Geurink, R., De Carlo, S., Vos, M., van Duinen, G., van Haeringen, B. and Storms, M. (2015) FEI's direct electron detector developments: Embarking on a revolution in cryo-TEM. J Struct Biol.

McMullan, G., Faruqi, A.R., Clare, D. and Henderson, R. (2014) Comparison of optimal performance at 300keV of three direct electron detectors for use in low dose electron microscopy. Ultramicroscopy 147: 156-163.

Natale, P., Brüser, T. and Driessen, A.J.M. (2008) Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—Distinct translocases and mechanisms. Biochimica et Biophysica Acta (BBA) - Biomembranes 1778(9): 1735-1756.

Ohol, Y.M., Goetz, D.H., Chan, K., Shiloh, M.U., Craik, C.S. and Cox, J.S. (2010) Mycobacterium tuberculosis MycP1 protease plays a dual role in regulation of ESX-1 secretion and virulence. Cell Host & Microbe 7(3): 210-220.

Pallen, M.J. (2002) The ESAT-6/WXG100 superfamily–and a new Gram-positive secretion system? Trends in microbiology 10(5): 209-212.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. Journal of Computational Chemistry 25(13): 1605-1612.

Renshaw, P.S., Lightbody, K.L., Veverka, V., Muskett, F.W., Kelly, G., Frenkiel, T.A., Gordon, S.V., Hewinson, R.G., Burke, B., Norman, J., Williamson, R.A. and Carr, M.D. (2005) Structure and function of the complex formed by the tuberculosis virulence factors CFP-10 and ESAT-6. EMBO J 24(14): 2491-2498.

Sani, M., Houben, E.N.G., Geurtsen, J., Pierson, J., de Punder, K., van Zon, M., Wever, B., Piersma, S.R., Jiménez, C.R., Daffé, M., Appelmelk, B.J., Bitter, W., van der Wel, N. and Peters, P.J. (2010) Direct visualization by Cryo-EM of the Mycobacterial capsular layer: A labile structure containing ESX-1-secreted proteins. PLoS Pathog 6(3): e1000794.

Solomonson, M., Setiaputra, D., Makepeace, Karl A.T., Lameignere, E., Petrotchenko, Evgeniy V., Conrady, Deborah G., Bergeron, Julien R., Vuckovic, M., DiMaio, F., Borchers, Christoph H., Yip, Calvin K. and Strynadka, Natalie C.J. (2015) Structure of EspB from the ESX-1 type VII secretion system and insights into its export mechanism. Structure 23(3): 571-583.

Stanley, S.A., Raghavan, S., Hwang, W.W. and Cox, J.S. (2003) Acute infection and macrophage subversion by Mycobacterium tuberculosis require a specialized secretion system. Proceedings of the National Academy of Sciences 100(22): 13001-13006.

van Bruggen, E., Wiebenga, E. and Gruber, M. (1960) Negative-staining electron microscopy of proteins at pH values below their isoelectric points. Its application to hemocyanin. Biochim Biophys Acta 42: 171-172.

van der Wel, N., Hava, D., Houben, D., Fluitsma, D., van Zon, M., Pierson, J., Brenner, M. and Peters, P.J. (2007) M. tuberculosis and M. leprae translocate from the phagolysosome to the cytosol in myeloid cells. Cell 129(7): 1287-1298.

van Heel, M. (1987) Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy 21(2): 111-123.

van Heel, M. (1989) Classification of very large electron microscopical image data sets. (Reutlingen, Allemagne Elsevier), 13.

van Heel, M. and Frank, J. (1981) Use of multivariates statistics in analysing the images of biological macromolecules. Ultramicroscopy 6(1): 187-194.

van Heel, M., Orlova, E.V., Dube, P. and Tavares, P. (1996) Intrinsic versus imposed curvature in

cyclical oligomers: the portal protein of bacteriophage SPP1. EMBO J 15(18): 4785-4788.

van Heel, M., Portugal, R., Rohou, A., Linnemayr, C., Bebeacua, C., Schmidt, R., Grant, T. and Schatz, M. (2012) Four-dimensional cryo electron microscopy at quasi atomic resolution: IMAGIC 4D. International Tables for Crystallography F: 624-628.

van Heel, M. and Schatz, M. (2005) Fourier shell correlation threshold criteria. J Struct Biol 151(3): 250-262.

van Pittius, N.G., Gamieldien, J., Hide, W., Brown, G.D., Siezen, R.J. and Beyers, A.D. (2001) The ESAT-6 gene cluster of Mycobacterium tuberculosis and other high G+ C Gram-positive bacteria. Genome Biol 2(10): 44.41-44.18.

Wagner, J.M., Evans, T.J., Chen, J., Zhu, H., Houben, E.N.G., Bitter, W. and Korotkov, K.V. (2013) Understanding specificity of the mycosin proteases in ESX/type VII secretion by structural and functional analysis. J Struct Biol 184(2): 115-128.

Watson, R.O., Manzanillo, P.S. and Cox, J.S. (2012) Extracellular M. tuberculosis DNA targets bacteria for autophagy by activating the host DNA-sensing pathway. Cell 150(4): 803-815.

White, H.E., Saibil, H.R., Ignatiou, A. and Orlova, E.V. (2004) Recognition and separation of single particles with size variation by statistical analysis of their images. J Mol Biol 336(2): 453-460.

WHO (2014) Global Tuberculosis Report 2014 (World Health Organization).

Xu, J., Laine, O., Masciocchi, M., Manoranjan, J., Smith, J., Du, S.J., Edwards, N., Zhu, X., Fenselau, C. and Gao, L.-Y. (2007) A unique Mycobacterium ESX-1 protein co-secretes with CFP-10/ESAT-6 and is necessary for inhibiting phagosome maturation. Mol Microbiol 66(3): 787-800.

Zuber, B., Chami, M., Houssin, C., Dubochet, J., Griffiths, G. and Daffé, M. (2008) Direct visualization of the outer membrane of mycobacteria and corynebacteria in their native state. J Bacteriol 190(16): 5672-5680.

CHAPTER 7

SUMMARIZING DISCUSSION

Introduction of the new generation of direct electron detectors, automatic data collection and development of the new cryo-EM methodology in the last decades, allowed solving structures of various protein complexes with near-atomic resolution. The new detectors, operating in movie-mode (collecting several movie frames during the overall exposure), require new accurate methodologies for automatic processing of the large datasets . We developed, tested and demonstrated a broad spectrum of new approaches for modern cryo-EM data analysis. Our developments, based on the philosophy of reference-free image processing, comprise a robust, fast and effective strategy for the analysis of large single-particle cryo-EM datasets.

In **Chapter 1** we introduced the concepts of single-particle cryo-EM, and emphasized its role in the field of structural biology. In our historical overview of technological and methodological developments, we outlined important milestones starting from the design of electron microscope by Ernst Ruska and Max Knoll, and ending with remarkable examples of the recently solved protein structures at near-atomic resolution. Among the methodological developments, we primarily covered pioneering and fundamental works in the field, which today are routinely used in data collection and image processing. We listed the main factors limiting the attainable resolution in cryo-EM and emphasized the methodological issues, investigated in this thesis. Thanks to the instrumentational developments and the computational power now available, we could elaborate on the original ideas (some already 20-40 years old) to improve single-particle cryo-EM analysis.

Chapter 2 illustrates the fundamental advantages of working with large datasets. We showed, that a cryo-EM dataset itself already contains statistically important information, which can be used retrospectively to improve the quality of the data. One can thus remove the fixed pattern from each digitally recorded image. In movie-mode cryo-EM data this fixed-pattern background hampers the movie alignments. *A posteriori* corrected images align better, and contain more high-resolution information. Our idea of the correction of the detectors by the data itself is simple, general and elegant. We showed, that it can be applied to images from all fields of digital image processing, including: astronomy, medical imaging and light microscopy.

In **Chapter 3** we present advanced ideas of movie alignments and implemented an iterative, over-relaxation algorithm for the alignment of full-frame (or sub-frame) cryo-EM movie-data. We also introduced a novel "P-spectrum" approach, which we use for assessing the quality of the movie alignments, and which can serve as input for CTF-determination. Chapters 2 and 3 together demonstrate our improved approaches for initial "pre-processing" of large, single-particle cryo-EM datasets.

In **Chapter 4** we focus on the next fundamental step in single-particle cryo-EM: particle picking. Particle picking can introduce serious reference bias if misused. The danger of using external references for the particle picking has been discussed extensively in the recent literature (Henderson 2013, Mao, Castillo-Menendez et al. 2013, Subramaniam 2013, van Heel 2013),where the controversial results by the group of Sodroski (Mao, Wang et al. 2012, Mao, Wang et al. 2013) were challenged. The issue has, however, not yet been resolved formally. We here continue this discussion by analysing the full newly deposited (2015) original dataset. We found further evidences to support our earlier hypothesis that the results of Sodroski and co-workers (Mao, Wang et al. 2012, Mao, Wang et al. 2013) are invalid. This controversy demonstrates the importance of using a clean methodology in single-particle cryo-EM data analysis. We suggest a reliable reference-free approach for particle picking as a solution to avoid reference bias.

Chapter 5 describes our whole reference-free methodology for single-particle cryo-EM data processing. We provided a detailed step-by-step protocol for solving structures of macromolecular complexes with near-atomic resolution. It includes *a posteriori* camera correction, movie-alignments, CTF-determination of (patches of) micrographs, reference-free particle picking, multivariate statistical eigenvector data compression, alignment-by-classification (ABC), angular reconstitution, and 4D processing. Our methodology is completely reference-free, as opposed to reference-based projection matching approaches, which are prone to reference bias. We demonstrated that our angular reconstitution approach for Euler angles determination is necessarily better and faster than the projection matching. Our extension of the ABC approach to four dimensions (ABC-4D) yields a robust, reference-free pipeline for studying the structure of biological complexes in mixed conformational states.
The power of our approach was illustrated by solving the structure of the giant hemoglobin of *Lumbricus terrestris* to a near-atomic resolution. We elucidated individual heme group's quasi-atomic environment within the heme-binding pocket without crystallographic restraints.

In **Chapter 6** we performed a single-particle cryo-EM analysis to reveal the structure of the EspB protein, a substrate of the type VII secretion system of *Mycobacterium tuberculosis*. The dataset we were working with, illustrated challenges in the single-particle cryo-EM. We here confronted a small and heterogeneous dataset which exhibited strong preferred orientations of the sample. These factors precluded us to obtain a near-atomic resolution structure, although we were able to identify two oligomeric states, of the EspB complex: a heptameric and a hexameric one. This study data provides structural insights of the secretion systems of *Mycobacterium tuberculosis*.

We expect that single-particle cryo-EM will become an easy and automatic technique for solving protein structures. Already today, due to an ease of the sample preparation and data analysis, many research groups, working with macromolecular complexes using X-ray crystallography techniques, have completely switched to single-particle cryo-EM (Kühlbrandt 2014, Bai, McMullan et al. 2015, Cheng 2015). In the near future we expect a breakthrough in solving a number of unknown structures of small biomolecules and complicated heterogeneous biomolecular complexes. Moreover, we anticipate developments of a number of drugs and vaccines based on the rational structure-based drug design using cryo-EM. We believe our work will contribute significantly to the success of cryo-EM.

REFERENCES

Bai, X.C., McMullan, G. and Scheres, S.H. (2015) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40(1): 49-57.

Cheng, Y. (2015) Single-particle cryo-EM at crystallographic resolution. Cell 161(3): 450-457.

Kühlbrandt, W. (2014) Cryo-EM enters a new era. Elife 3: e03678.

Mao, Y., Castillo-Menendez, L.R. and Sodroski, J.G. (2013a) Reply to Subramaniam, van Heel, and Henderson: Validity of the cryo-electron microscopy structures of the HIV-1 envelope glycoprotein complex. Proceedings of the National Academy of Sciences 110(45): E4178-E4182.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Desormeaux, A., Finzi, A., Xiang, S.H. and Sodroski, J.G. (2013b) Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. Proc Natl Acad Sci U S A 110(30): 12438-12443.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Xiang, S.H., Haim, H., Yang, X. and Sodroski, J. (2012) Subunit organization of the membrane-bound HIV-1 envelope glycoprotein trimer. Nat Struct Mol Biol 19(9): 893-899.

Subramaniam, S. (2013) Structure of trimeric HIV-1 envelope glycoproteins. Proc Natl Acad Sci U S A 110(45): E4172-4174.

van Heel, M. (2013) Finding trimeric HIV-1 envelope glycoproteins in random noise. Proc Natl Acad Sci U S A 110(45): E4175-4177.

APPENDIX &

SOFTWAREINSTRUMENTATIE VOOR DE STRUKTUURBEPALING VAN BIOLOGISCHE COMPLEXEN OP QUASI-ATOMAIR NIVEAU MIDDELS CRYOGENE ELECTRONENMIKROSCOPIE

Samenvattende discussie

De introductie van de nieuwe generatie "directe" elektronen detectoren, de invoering van automatische datacollectie, en de ontwikkeling van cryo-EM methodologie over de laatste decennia, maken het nu mogelijk om de structuur van onder andere proteïnecomplexen te bepalen met een quasi atomair oplossend vermogen. De nieuwe detectoren, die ook in "film modus" gebruikt worden (het opdelen van de totale belichting over meerdere "frames"), vereisen nieuwe methodologieën om ook deze grotere en gedetailleerde datasets optimaal te verwerken. We ontwikkelden en testten een breed spectrum aan technieken voor moderne cryo-EM data-analyse. Onze ontwikkelingen, gebaseerd op een filosofie van onbevooroordeelde beeldverweking, vormen een robuuste en snelle strategie om grote cryo-EM datasets te verwerken.

In **Hoofdstuk 1** introduceren we de basisprincipes van "single-particle cryo-EM", en diens belang in de structuurbiologie. In het historische overzicht van technische en methodologische ontwikkelingen noemen we belangrijke mijlpalen: van de ontwikkeling van het electromicroscoop door Ernst Ruska en Max Knoll, tot aan de structuren van biologische macromoleculen die met quasi-atomair oplossend vermogen zijn gereconstrueerd. Bij de methodologische ontwikkelingen noemen we de belangrijkste publicaties, die geleid hebben tot de huidige routinematige datacollectie en beeldverwerking. We noemen de belangrijkste factoren die het bereiken van hoge resolutie in de weg staan, en we benadrukken de belangrijkste nieuwe methodes die in deze scriptie worden behandeld. Door de ontwikkelingen in de instrumentatie en de steeds snellere computers, kunnen we ook de originele ideeën in single particle cryo-EM, waarvan sommigen 20-40 jaar oud zijn, verfijnen en beter uitbuiten.

Hoofdstuk 2 illustreert enkele fundamentele voordelen van het werken met grote datasets. We demonstreren dat een grote cryo-EM datasset statistisch significante informatie bevat over de sensor die gebruikt is, deze informatie kan wederom gebruikt worden om de dataset zelf te verbeteren. Onder andere worden vaste patronen weggehaald, die storend kunnen zijn bij de "movie alignment" (het oplijnen van de lichte verschuivingen tussen imdividuele movie frames). Door een succesvolle *a posteriori* correctie, kunnen betere eindresultaten (met hogere resolutie) worden bereikt. Het concept van het corrigeren van een grote dataset op basis van alleen de data zelf is simpel, algemeen, en elegant. We demonstreerden dat deze techniek toegepast kan worden in velerlei gebieden, onder andere astronomie, medische

beeldvorming, en lichtmicroscopie.

In **Hoofdstuk 3** presenteren we nieuwe ideeën over movie alignment en implementeren we een iteratief over-relaxatiealgoritme voor het oplijnen van gehele beelden (of deelbeelden) in datasets die in film modus zijn opgenomen. We introduceerden een nieuw P-spectrum voor het evalueren van de kwaliteit van zulke oplijningen; P-spectra kunnen ook dienen als invoer voor CTF-bepalingen. Hoofdstukken 2 en 3 samen, demonstreren een verbeterde voorbewerking van grote single-particle cryo-EM datasets.

Hoofdstuk 4 is gericht op een fundamentele stap in het single-particle cryo-EM verwerkingproces: het automatisch vinden van de moleculen in de beelden ("particle picking"). Verkeerd gebruikt kan particle-picking "reference bias" veroorzaken, wat leidt tot het vinden van niet-bestaande informatie. Deze reference-bias problemen hebben recentelijk tot heftige discussies in de literatuur geleid (Mao, Castillo-Menendez et al. 2013, Subramaniam 2013, van Heel 2013). De controversie ontstond over de resultaten van de onderzoeksgroep van Sodroski (Mao, Wang et al. 2012, Mao, Wang et al. 2013). Deze kwestie is echter nog niet formeel de wereld uit. De discussie wordt hier voortgezet in vorm van een analyse van een recentelijk gedeponeerde dataset. De nieuwe analyse ondersteunt onze eerdere hypothese/ conclusie dat de resultaten van Sodroski en collega's (Mao, Wang et al. 2012, Mao, Wang et al. 2013) niet kloppen. De controversie demonstreert het belang van een degelijk onderbouwde methodologie voor de beeldverwerking in single-particle cryo-EM. We stellen een degelijke methodologie voor voor particle picking die geen reference bias vertoont.

Hoofdstuk 5 omschrijft een complete methodologie die reference bias geheel vermijdt. We omschreven stap-voor-stap een protocol voor het verfijnen van structuren van macromoleculaire complexen naar een quasi-atomair oplossend vermogen. Dit protocol omvat de eerdergenoemde a posteriori camera correctie, de oplijning van film-modus data, de CTF-bepaling van de EM-beelden of delen daarvan, een particle picking zonder referenties, multivariate statistische eigenvector datacompressie, "alignment by classification" (ABC), "angular reconstitution", en 4-dimensionale dataverwerking. Onze metholdologie vermijdt bijvoorbeeld de bias die de "projection matching" techniek kan introduceren doordat deze van nature met referenties werkt. Onze uitbereiding van de ABC-techniek naar 4D verwerking (ABC-4D) resulteert in een robuuste, referentieloze pijplijn voor de studie van biologische structuren in heterogene conformaties.
De kracht van onze aanpak wordt geïllustreerd door het oplossen van de structuur van de hemoglobine van *Lumbricus terrestris* tot quasi-atomaire resolutie. De quasi-atomaire omgeving van een enkele heemgroep dient als voorbeeld van de resolutie die behaald kan

worden zonder kristallografische beperkingen.

In **Hoofdstuk 6** wordt een single-particle cryo-EM analyse toegepast op data van de EspB proteïne, een substraat van type VII afscheidingssysteem van *Mycobacterium tuberculosis*. Deze dataset illustreert enkele uitdagingen in single-particle cryo-EM. De dataset is relatief klein, en heterogeen, met een sterke voorkeur in de orientaties van de deeltjes. Deze factoren verhinderden het behalen van een quasi-atomaire resolutie. Het was wel mogelijk om twee oligomerische varianten van de EspB multimeer te onderscheiden: een heptamerische en een hexamerische variant. De bestudeerde data geven structurele inzichten op de secretiesystemen van *Mycobacterium tuberculosis*.

We verwachten dat single-particle cryo-EM een gemakkelijke en automatische techniek kan worden voor het oplossen van proteïnestructuren. Er zijn al meerdere onderzoeksgroepen die historisch gezien macromoleculaire complexen oplossen met behulp van de Röntgenkristallografie, die door de eenvoudigere prepareertechnieken en data-analyse methodes overstappen naar single-particle cryo-EM (Kühlbrandt 2014, Bai, McMullan et al. 2015, Cheng 2015). In de nabije toekomst verwachten we een doorbraak in het oplossen van kleinere biologische complexen en van gecompliceerde heterogene complexen. Bovendien anticiperen we ontwikkelingen van nieuwe medicijnen en vaccins gebaseerd op rationeel medicijnontwerp vanuit de cryo-EM. We denken dat ons werk een belangrijke bijdrage zal leveren tot het succes van single-particle cryo-EM.

REFERENCES

Bai, X.C., McMullan, G. and Scheres, S.H. (2015) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40(1): 49-57.

Cheng, Y. (2015) Single-particle cryo-EM at crystallographic resolution. Cell 161(3): 450-457.

Kühlbrandt, W. (2014) Cryo-EM enters a new era. Elife 3: e03678.

Mao, Y., Castillo-Menendez, L.R. and Sodroski, J.G. (2013a) Reply to Subramaniam, van Heel, and Henderson: Validity of the cryo-electron microscopy structures of the HIV-1 envelope glycoprotein complex. Proceedings of the National Academy of Sciences 110(45): E4178-E4182.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Desormeaux, A., Finzi, A., Xiang, S.H. and Sodroski, J.G. (2013b) Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. Proc Natl Acad Sci U S A 110(30): 12438-12443.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Xiang, S.H., Haim, H., Yang, X. and Sodroski, J. (2012) Subunit organization of the membrane-bound HIV-1 envelope glycoprotein trimer. Nat Struct Mol Biol 19(9): 893-899.

Subramaniam, S. (2013) Structure of trimeric HIV-1 envelope glycoproteins. Proc Natl Acad Sci U S A 110(45): E4172-4174.

van Heel, M. (2013) Finding trimeric HIV-1 envelope glycoproteins in random noise. Proc Natl Acad Sci U S A 110(45): E4175-4177.

VALORIZATION

Relevance

Cryogenic electron microscopy is a rapidly growing and powerful technique for elucidating the structure of biological macromolecules. It is becoming a popular technique in the field of structural biology, and is now used in structure-based drug design and vaccine development. The aim of our research was to develop, test and optimize novel methodologies for a reliable single-particle analysis of cryogenic electron microscopy data. Our reference-free approach improves the effectiveness and speed of the high-resolution single-particle analysis. Thus, our results hold great promise in the battle against some of the world's major healthcare problems thus potentially saving numerous human lives.

In Chapter 4, we challenged the results of (Mao et al. 2012; Mao et al. 2013b), which we find to be the result of manipulative data processing. Our studies triggered serious discussions in the fields of immunology, structural biology, medical biology, and about the functioning of the academic press (Cohen 2013; Henderson 2013; Mao et al. 2013a; Subramaniam 2013; van Heel 2013). This discussion is still ongoing, and has stimulated different parties in cryo-EM to pay attention to the issue of validation of the results. In particular, use of reference-free image processing. Thus, Electron Microscopy Database (EMDB; EBI-EMBL) now stimulates and demands a more detailed description of the processing details. A number of new online validation tools have been made available and a new database was initiated to make large raw datasets publically accessible. Moreover, this controversy opened up fundamental issues, related to the journal and referee responsibilities in the academic press. We propose further open discussions of the controversial results; transparency and open access to the original published data.

Target groups

The results of the work, presented in this thesis might be interesting to:

- research institutions, applying or starting to apply the cryo-EM methodologies in solving the structures of biological complexes
- pharmaceutical companies using structure-based drug design and vaccine development
- companies, developing and producing cryo-EM equipment, particularly electron microscopes, electron detectors and operating software
- manufacturers of various digital detectors for photography, medical imaging or astronomy
- academic press; funds, sponsoring HIV research and journalists, covering controversies in science and all the people interested in the development of the HIV vaccines

Activities/Products

Our methodological developments had or have been integrated into the popular scientific software package IMAGIC-4D (Image Science Software GmbH). This software pioneered many aspects of single-particle cryo-EM image processing and has been continuously developed for more than 35 years. The software is distributed semi-commercially on the basis of its maintenance costs.

Our *a posteriori* camera normalization, can be applied in various other fields of digital image processing, including photography, astronomy and medical imaging. The ideas are, however, simple and are thus easy to implement in any extensive image processing system.

The first cryo-EM structure of the worm hemoglobin, solved to near-atomic resolution, opens up new perspectives for studies of the process of the worm-hemoglobin oxygen binding, which in a long-term perspective could contribute to the development of the artificial blood.

Our cryo-EM study of the EspB protein open up new insight into the organization of the type VII secretion systems of *Mycobacteria* and contributes to our understanding of its virulence. This knowledge might facilitate the development of the new vaccine against *Mycobacterium tuberculosis*.

Innovation

In our work we suggested innovative ideas and tools in single-particle cryo-EM image processing (like a posteriori camera correction, P-spectrum, Fourier-space classification etc.). Apart from the theoretical and ideological considerations covered in the main text, our suggested and implemented methodologies are practically different from the analogues. Our approach is absolutely reference-free and thus has a less chance for obtaining invalid and reference-biased results; it is cheap and fast in term of computational power (all the details in chapter 5).

Schedule & Implementation

The suggested methodologies have been implemented in our IMAGIC-4D software and used for processing of the real challenging cryo-EM data for obtaining a near-atomic resolution. The ideas have been or will be published in the open literature without restrictions.

References

Cohen, J. (2013) Is high-tech view of HIV too good to be true? Science 341(6145): 443-444.

Henderson, R. (2013) Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein

from noise. Proc Natl Acad Sci U S A 110(45): 18037-18041.

Mao, Y., Castillo-Menendez, L.R. and Sodroski, J.G. (2013a) Reply to Subramaniam, van Heel, and Henderson: Validity of the cryo-electron microscopy structures of the HIV-1 envelope glycoprotein complex. Proceedings of the National Academy of Sciences 110(45): E4178-E4182.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Desormeaux, A., Finzi, A., Xiang, S.H. and Sodroski, J.G. (2013b) Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. Proc Natl Acad Sci U S A 110(30): 12438-12443.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Xiang, S.H., Haim, H., Yang, X. and Sodroski, J. (2012) Subunit organization of the membrane-bound HIV-1 envelope glycoprotein trimer. Nat Struct Mol Biol 19(9): 893-899.

Subramaniam, S. (2013) Structure of trimeric HIV-1 envelope glycoproteins. Proc Natl Acad Sci U S A 110(45): E4172-4174.

van Heel, M. (2013) Finding trimeric HIV-1 envelope glycoproteins in random noise. Proc Natl Acad Sci U S A 110(45): E4175-4177.

CURRICULUM VITAE

Pavel Valerievich Afanasyev was born in Leningrad, USSR (Saint-Petersburg, Russia) in 1988. He started his higher education in 2005 at the faculty of Physics and Mechanics of Saint-Petersburg State Polytechnic University (SPbSPU) at the Department of Biophysics. During his bachelor study he worked at the Department of Physiology of the Research Institute for Experimental Medicine of the Russian Academy of Medical Science, where he carried out a research project on the calpain overactivity over during the development of experimental autoimmune encephalomyelitis of rats. His master studies were performed at the Research Centre Nanobiotechnologies (SPbSPU) on developments of single-molecule techniques, particularly magnetic and optical tweezers. During an internship at the University of Colorado in Boulder, Richard McIntosh inspired Pavel to develop a career in the field of cryo-EM. In 2011 Pavel joined the group of Peter Peters at the Netherlands Cancer Institute as a PhD student. From 2013 the PhD project focused on single-particle cryo-EM methodology in collaboration with Marin van Heel. The latter work was largely performed at Leiden University, at the Netherlands Centre for Electron Nanoscopy (NeCEN). During his PhD, Pavel participated in a number of international cryo-EM workshops, courses and schools as an instructor and was invited to various research institutes to present seminars on single-particle cryo-EM. In 2016 Pavel accepted a position as postdoc in the group of Paula da Fonseca at MRC-LMB, Cambridge, UK.

LIST OF PUBLICATIONS

Afanasyev P., Ravelli R.B.G., Matadeen R., Carlo S.D., Duinen G.V., Alewijnse B., Peters P.J., Abrahams J.P., Portugal R.V., Schatz M., van Heel M. (2015). *A posteriori* correction of camera characteristics from large image data sets. *Scientific Reports*, 5, 10317.

Afanasyev P., Linnemayr-Seer C., Ravelli R.B.G, Matadeen R., Carlo S.D., Alewijnse, B., Portugal R.V., Pannu N.S., Schatz M., van Heel M. Single-particle cryo-EM based on alignment by classification (ABC) *Lumbricus terrestris* hemoglobin at near-atomic resolution. *In preparation*

Afanasyev P., van Heel M., Can 670,000 HIV-1 envelope trimer particles be extracted from the EMPIAR 10003 full data set. *In preparation*

van Heel M., **Afanasyev P.**, Schatz M. Assessing movie-data in single-particle cryo-EM. *In preparation*

Vázquez-Fernández E., Vos M.R., **Afanasyev P.**, Cebey L., Sevillano A.M., Vidal E., Rosa I., Renault L., Ramos A., Peters P.J., van Heel M., Young H.S., Requena J.R., Will H. Structure of an Infectious Mammalian Prion. *Submitted*

Tame M.A., Raaijmakers J.A., **Afanasyev P.**, Medema, R. H. (2016). Chromosome misalignments induce spindle-positioning defects. *EMBO reports*, e201541143.

PHD PORTFOLIO

Courses and workshops

- Instructor for the cryo-EM data processing workshop at the Paul Scherrer Institute, Switzerland, March 2016.
- Instructor for the cryo-EM data processing at the FEI cryo-TEM workshop at NeCEN, November, 2015
- Instructor for the "Single-particle cryo-EM data processing workshop", Maastricht University, July 2015
- Short talk at the "INSTRUCT course on Computational Tools Combining Atomic and Volume Data"; Harwell, UK, 2015
- Instructor for the cryo-EM data processing at the FEI cryo-TEM workshop at NeCEN, March, 2015
- Instructor for single-particle cryo-EM data processing at "The sixth School for Single-particle Cryo-EM", Sao-Paulo, Brazil, August 2014
- Poster at the "Cryo-EM 3D Image Analysis Symposium 2014", Lake Tahoe, CA, USA, 2014
- Poster at the EMBO Practical Course on Image Processing for cryo-EM, London, UK, 2013
- Short talk at the "The fifth School for Single-particle Cryo-EM", Sao-Paulo, Brazil, 2012
- Poster at the workshop "Computational Challenges in Structural Biology", ESBS, Strasbourg, France 2012
- Short talk at the course "Cytoskeleton in Cell Migration and Invasion", Institute Curie, Paris, France, 2012

Conferences and seminars

- Seminar (as an invited speaker by Dr. B. Klaholz) at the Institut Génétique Biologie Moléculaire Cellulaire), Strasbourg, France, January, 2016.
- Seminar (as an invited speaker by Prof. J-M. Carazo) at the National Center for Biotechnology, Madrid, Spain, October 2015.
- Seminar (as an invited speaker by Dr. A. Amunts) at the MRC-LMB Cambridge, UK, April 2015
- Poster at the GRC"Three Dimensional Electron Microscopy", Girona, Spain, 2014
- Poster at the GRC "Three Dimensional Electron Microscopy", New London, NH, USA 2013
- Participant at the International conference "From molecules to cells - new frontiers in 3D electron microscopy", Bonn, Germany, 2013
- Oral presentation "Developments of the Nanochamber for light and electron microscopy of mammalian cells", at the "Nano-Imaging under Industrial Conditions" meeting in Hilversum, The Netherlands, 2012
- Poster presenter on "Migration of mammalian cells in a nanochamber; its potential use for electron cryo tomography" at "The Sixth International Congress on Electron Tomography", EMBL, Heidelberg, 2011

Other graduate-student events

- OOA "English Writing and Presenting" course, NKI, 2013
- OOA Graduate Student Retreat, Zeeland, 2013
- 3rd Intercity Young Scientist Meeting "Cellular Dynamics and Signaling regulation in health

and disease" Utrech-Bunnik, 2012

- OOA course 'In the footsteps of Antoni van Leeuwenhoek", NKI, VU, AMC, Amsterdam, 2012
- OOA workshop "How to Write High Impact Papers" VU, Amsterdam, 2012
- OOA course "Basic Medical Statistics", NKI, Amsterdam, 2012
- OOA Graduate Student Retreat, Texel, 2011
- OOA Graduate Student Retreat, Ermelo, 2012

Student supervision (jointly with Marin van Heel)

- Katie Riciluca, PhD candidate at São Paulo University, 6-month internship at Leiden University, 2015. Project: Single-particle cryo-EM analysis of the hemocyanin from Acanthoscurria rondoniae
- Vitor Hugo Balasco Serrão, PhD candidate at São Paulo University, 6-month internship at Leiden University, 2015. Project: Single-particle cryo-EM analysis of the SelA-tRNASec-SelD ternary complex

ACKNOWLEDGEMENTS

My PhD has been a long and winding marathon. It is a result of efforts, support, contributions, discussions and motivating examples of many people. I am truly grateful to everyone, who contributed into its completion! Particularly, I would like to thank:

Mihoko Tame; Marin van Heel; Peter Peters; Bart Alewijnse; Michael Schatz; Ralf Schmidt; Sacha De Carlo; Rishi Matadeen; Max Maletta; Raj Pannu; Charlotte Linnemayr-Seer; Raimond Ravelly; Willem Tichelaar; Chris Diebolder; Paul van Schayck; Garib Murshudov; Pavol Skubák; Raffaella Tassoni; Kimberley Zwiers; Alicia Lammerts van Bueren; Musa Sani; Nicole van der Wel; Hans Janssen; Karin de Punder; Maaike van Zon; Noor Bakker; Sue Godsave; Pekka Kujala; Mary Morphew; Nico Ong; Axel Siroy; Giancarlo Tria; Delei Chen; Hirotoshi Furusho; Jérémie Piton; Florence Pojer; Stewart Cole; Sjaak Neefjes; Arnoud Sonnenberg; Kees Jalink; Frans Ramaekers; Ron Heeren; Élen Tomazela; Maria Miu; Susanne Roodhuijzen; Marjoleine van Egeraat; Anne Laan; Anna Kruip; Livia Smits; Karolina Skraskova; Jan Pieter Abrahams; Patrick Voskamp; Igor Nederlof; Ludovic Renault; Ariane Briegel; Roman Koning; Bram Koster; Thom Sharp; people from the NKI, particularly the Department of Cell Biology; members of M4I division (Maastricht University); Pleun Dona; Luigi Mele; Hervé-William Rémigy; Ester Vazquez Fernandez; Holger Wille; Patrick Bron; Rene Medema; Alexey Amunts; Paula da Fonseca; Bruno Klaholz; Elena Orlova; Jose Maria Carazo; Richard McIntosh; Alexander Myasnikov; Igor Orlov; Rosilene de Souza van Heel; Katie Riciluca; Vitor Serrão; Rodrigo Portugal; participants of single-particle cryo-EM Brazil schools in 2012 and 2014, FEI workshops in Leiden 2015; Maastricht cryo-EM workshop 2015; PSI workshop 2016; the Faculty of Physics and Mechanics of Saint-Petersburg State Polytechnic University, Department of Biophysics; Ekaterina Grishchuk; Nikita Gudimchuk; Vladimir Volkov; Anton Sabantsev; Yuri Rykov; Alexey Nazarov; Andrey Ilin; Evgeny Kuznetsov; Tim Craig; Arina Afanasyeva; George Pobegalov; Roman Dray; Amy Diallo; Emiliya Pogosyan; Amy Dohmen; Melinda Aprelia; Katharina Witting; Stephan Scherpe; Núria Sola Tapias; Anna Miquel-Cases; Alba Llopis Gómez; Michael Uckelmann; Rui Lopes; Rita Maia; Alessia Amore; Marion Libouban; Sander Timmer; Ella Nirmala; Li Zhang; Ksenia Shcherbakova; Bart van den Bogaard; Laurissa Gillespie; Laura Marija Uljane; Joni Prokkola; Adilia Dagkesamanskaya; Misha Grigoriev; Mikhail Khodorkovskii; Marina Karpenko; Viktor Klimenko; Oleg Bocharov; Nina Yurova; the Dutch culture, people and the country; McDonalds; Sublimefm.

Я благодарен маме и папе, Максиму, Ане и всем родственникам, в частности Даше Афанасьевой и коту Тиме; друзьям; конторе, береги заряд; Екатерине Тарасовой; радио Эхо Москвы и телеканалу Дождь, а также Шурке и Лаврентию Августовичу Пысину, не дававшими соскучиться.

And thanks to everyone, who I unintentionally missed in this list!

With love, Pavel.