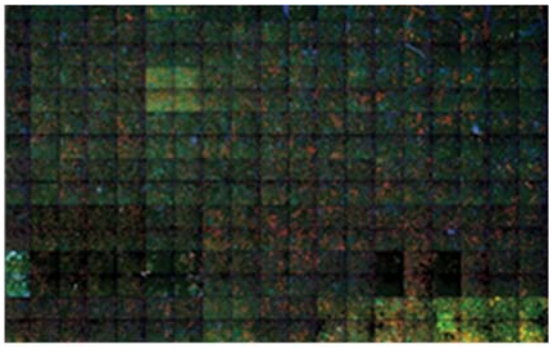
Preliminary Overview of Cyber-infrastructure Needs for Light Microscopy

Claire Brown, McGill University, Advanced BioImaging Facility (ABIF)
Pina Colarusso, University of Calgary, Live Cell Imaging Facility (LCI)
Stephen Ogg, University of Alberta, Cell Imaging Centre (CIC)

Summary: Light microscopy is applicable to fields ranging from engineering, basic sciences, life sciences and health sciences. Every researcher has a need to "see" the system they are studying. Light microscopy technology continues to develop at a rapid pace, enabling us to acquire images with higher temporal and spatial sampling rates, or using automation so that larger and more specimens can be imaged. Coupled with increases in the number of distinct probes that can be imaged, these improvements have generated an explosion of large, high- resolution multi-dimensional (3D, multi-colour, time-lapse) image data sets. These multi-dimensional data sets contain tremendous amounts of quantitative information and need to be analyzed, collated, summarized and interpreted. The sheer size of these datasets creates a challenge in image management. The ability to visualize, store, move, backup, archive and interrogate images, is a perquisite to extracting meaningful quantitative information from these datasets.

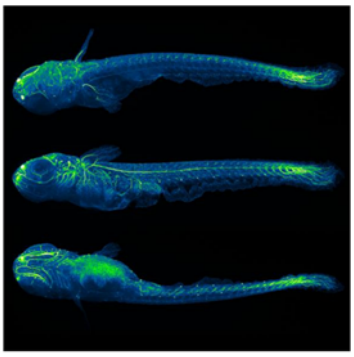
Techniques from the life sciences field that will be particularly demanding on cyber-infrastructure, each of which is described in detail below, include 1) high content screening microscopy – automatically performs multi-colour, multi-position fluorescence imaging in multi-well plates, 2) light-sheet, also known as selective plane illumination microscopy - the ability to image millimetre sized specimens at micrometre resolution, 3) super resolution using localization microscopy – the ability to acquire and post-process more than 10,000 images into a rendered image with 10-20 nm resolution, and 4) whole slide scanning of large tissue samples with visible histochemical stains or multicolour fluorescence stains. In addition to the massive image management challenge inherent in collecting this type of data, to be useful, these data sets also demand a tremendous amount of CPU and GPU time for proper analysis. Current cyber-infrastructure available at universities across the country does allow the effective use of these cutting edge research technologies, putting Canadian scientists at a severe competitive disadvantage.



http://www.photonics.com/Article.aspx?AID=38553

High Content Screening: Five colour imaging of 96 well plates. Routine experiment could be with one plate, four images per well and one timepoint.
Case Study: Five colour imaging of cancer cells screening a small chemical library. Examining translocation from the nucleus to the cytosol.
Storage: Each image is 2048 x 2048 pixels, 16-bit (2-bytes) = **8.4 Mb per image**. Four images per well and five colours so total storage per plate=96 wells x 4 images/well x 8.4 MB/image x 5 colours = **16 Gb** per plate. Ten plates per day so ~2500 plates per year or **40 Tb/year**.
Computation: Segmentation and translocation assay for **4.8 million images per year**. Note some analysis will be much more intensive than translocation.

Graphics: Calculations can be done using GPU to speed up processing.



https://www.janelia.org/lab/keller-lab

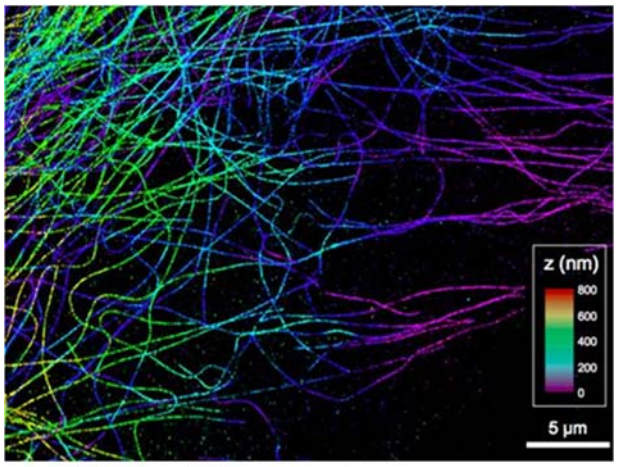
Light Sheet Microscopy: Rapid, multi-colour, samples up to millimeters in size for fixed and cleared tissue with micrometer resolution. This technology generates the largest file sizes at this time and is the most computationally demanding application described here.

Case Study: Morphology of entire mouse brain with 3 µm resolution. Sample is 10 mm x 10 mm x 10 mm.

Storage: Each image is 2048 x 2048 pixels, 16-bit image (2 bytes) = 2048x2048x2 = **8.4 Mb** per image. With a pixel size corresponding to 1.0 µm on the sample, a single image is ~2.0 mm x 2.0 mm of tissue. So need 5x5 image tiles or 25 images for that encompasses one plane of the entire brain. With a z-stack (depth) spacing of 2 µm, 5000 planes would be required per colour to include the whole 3D brain tissue. Thus, two colour imaging would need: 25 images x 5000 z-stack images x 2 colours = 250,000 images per colour which is **2 Tb per experiment**. Ten experiments per week at 48 weeks per year for **960 Tb/year**.

Computation: Surface rendering of 2048x2048x5000x2 = **4.2×10^{10} pixels** per data set.

Graphics: Calculations can be done using GPU to speed up processing.



http://www.microlat.com.ar/

Single Molecule Localization Microscopy:

Single molecule localization microscopy relies on taking an image of hundreds of single fluorophore molecules in a single image frame, localizing each of these and recording its position at high precision to within ~ 10 nm and repeating this for thousands of image frames. This dataset is subsequently used to generate one super resolution image frame by summation, containing many millions of single molecule localization points.

Case Study: Super resolution imaging of a data set with a control sample and samples with three different treatments. Ten images for each sample and the experiment will be done in triplicate. A high resolution image will be generated from 10,000 single molecule images. Two colours per sample.

Storage: Each image is 2048 x 2048 pixels, 16-bit image (2 bytes) = 2048x2048x2 = 8.4 Mb per image. A 63x/1.4 NA oil immersion lens will sample each pixel at ~0.1 µm in x and y, generating images with a field of view of 200 µm x 200 µm. Therefore, the entire dataset will be **6.7 Tb**(10,000 images, 2 colours, 8.4 Mb per image, 4 samples, 10 images per sample. Five data sets per week for **33.6 Tb/week** with 48 weeks per year would be **1610 Tb/year**.

Computation: Calculations of millions of single molecule localizations per image per colour.

Graphics: Calculations can be done using GPU to speed up processing. Read out and display of millions of localizations to give a single super resolution image.

https://www.mskcc.org/research-advantage/core-facilities/

Whole Slide Fluorescence Imaging: Five colour imaging of large tissue samples with a slide scanner at high resolution using tile scanning. Samples can be as large as 50 mm x 20 mm and resolution needs to be at a level of 0.1 µm. Also used for tissue microarrays stained with multiple fluorescent probes and of a similar size and at a similar resolution.

Case Study: Five colour imaging of breast tissue including normal, tumor and tumor proximal tissues. Images of the whole slide at 50 mm x 20 mm at a conservative resolution of 1.0 µm. This is a reasonable estimate because some applications may only have 3 colours but others will require a resolution higher than 1.0 µm per pixel.

Storage: Each image is ~2048 x 2048 pixels, 16-bit image (2-bytes) = 2048x2048x2 = **8.4 Mb per image**. With pixel samples set at 1.0 µm in x and y, one image is 2 mm x 2 mm. Therefore, tissue is 1000 mm^2 so 250 images per slide per colour are necessary to image the whole tissue area. Thus, the total storage space for imaging of one 5 colour slide will be = 8.4 Mb x 250 images x 5 colours = **10.5 Gb** per slide. With 100 slides a week or 5000 slides per year that would correspond to **52.5 Tb per year**.

Computation: Segmentation and image analysis of **6.25 million images per year**.

Graphics: Calculations can be done using GPU to speed up processing.

Summary:

For an optical microscopy facility with these four instruments, the imaging data sets would correspond to 40 Tb (HCS), 960 Tb (Light Sheet Microscopy), 1610 Tb (Single Molecule Localization Microscopy) and 52.5 Tb (slide scanner). **This yields a total of 2660 Tb per year.** The computation needs are less well defined but will scale with the amount of data as computations are quite complex for all four microscopy platforms.

Note that ABIF at McGill will have these four instruments by the end of 2016, and the LCI at the University of Calgary has two of these instruments and funding is being sought for a third, the CIC at the University of Alberta currently has two of these technologies and funding is being sought for a third.

Possible Solutions:

It would be ideal to tap into the expertise and infrastructure of Compute Canada. This organization's storage, CPU and GPU resources to allow Canadian researchers to benefit from the existing national light microscopy infrastructure, positioning us to efficiently reach important conclusions by data analysis rather than image management. There would still be a need for local storage at each instrument but the bulk of long term data storage and computations could be conducted on Compute Canada infrastructure. A schematic of possible linkages to the Compute Canada infrastructure is presented below. Each microscope has an acquisition computer that

would be connected to a data analysis and local storage workstation via fiber optic. Then each local workstation would be connected to Compute Canada infrastructure via fiber optic.

Constraints:

1. Funding for institutional infrastructure needs to be secured to ensure that there are fiber optic connections between microscope workstations and compute Canada resources.
2. NIH guidelines state the images need to be maintained for 3-years. Most researchers assume 3-years after publication so retention is more likely to be for 5-7-years total. https://ori.hhs.gov/education/products/clinicaltools/data.pdf

References:

1) Helpful discussions with Douglas Richardson, Harvard University and Mark Sanders, University of Minnesota. More detailed information is available upon request.
2) Almada P, Culley S, Henriques R. "PALM and STORM: Into large fields and high-throughput microscopy with sCMOS detectors." Methods. 88:109-121 (2015).