

Pipeline IT requirements

The pipeline does not have a strict set of IT requirements. We have found that many of our early access customers have different opinions on these matters and different existing IT setups, which in turn has led them to different storage solutions and different data management. Therefore we are reluctant to dictate a "correct" solution. Nevertheless, this page attempts to give some guidelines.

Disk space and network infrastructure

The way the Solexa machine works is that images are acquired and stored on the instrument and must then be transferred to an external computer to be analysed by the analysis software, which handles image processing, base-calling and sequence alignment. The main issue to be aware of is that each instrument run generates ~1TB of data during a full 2-3 day run. However, ~70% of this is TIFF image data that can potentially be sent to tape after a run is finished and you are satisfied and a reanalysis is not required. These large data volumes mean that you will need

1. A high-throughput ethernet connection (1 Gb or more recommended) or other data transfer mechanism.
2. A suitably large holding area for the images and analysis output (1 TB per run). As there will almost certainly be some overlap between copying/analysis/possible reanalysis, 2-3 TB should be an absolute minimum. We generally recommend 10 TB, which should be enough to hold a few weeks' worth of runs online. As the analysis pipeline processes large amounts of data, fast disk access is important too.
3. You need to consider which parts of the data you want to back-up and what infrastructure you want to provide for that. If you want to keep image data, then half a terabyte per run is required. We're working on export utilities and more efficient data formats for the analysis output. Furthermore, the pipeline provides an option to perform a (lossless) compression of the data.

Analysis computer

The analysis software was designed to run on a range of computer architectures and to fit into existing IT infrastructures as seamlessly as possible. The following guidelines may be helpful:

Operating system

The software should run on all common Unix/Linux variants. It also runs on Windows using the cygwin emulation, and on MacOS X. However, we do not intend to support any platform other than Linux at the moment. For your reference, internally we are running CentOS 4.1, even though there is no reason to believe that any other Linux distro would be less suited.

Hardware

A high-end box (e.g. a dual-processor, dual-core Intel Xeon or AMD opteron) should be adequate as an analysis computer; on this type of hardware you can expect to turn around the image analysis and base-calling of a full run in ~ 1 day. Sequence alignment would take additional time; depending on which alignment program you're running probably somewhere between a few hours (using our fast short-read whole-genome alignment program Eland) and days (using more traditional alignment programs).

Pipeline parallelisation is built around the multi-processor facilities of the make utility and scales very well to well beyond 8 nodes; substantial speed-up is expected even for parallelisation across several hundred CPUs. The pipeline can make use of multi-processor systems (SMP) and clusters (it is compatible with Sun Grid Engine and LSF make). Internally we are using Sun Grid Engine, and a full 1 GB run-folder is typically analysed in around 3-4 hours on 7 dual-processor, dual-core Opterons.

Comments

Might be worth quoting that on 7 dual CPU, dual core opteronms a full 1G runfolder takes about 3 hrs.