Data Pipelines at the Gene Center

A. Hauser, S.Wickles, K. Akman, A. Graf, S. Krebs, C. Ungewickell, O. Berninghausen, T. Becker, D. Schulz, J. Massier, G. Witte, P. Wendler, U. Gaul, H. Blum, R. Beckmann

Deep Sequencing

Illumina GA IIx

Up to 150 GByte are produced per day (without images).
A whole run takes three to ten days and can produce nearly 1 TByte of raw data.

Dedicated Fileserver

35TB on RAID 6 of 24 * 2 TB HDDs with 5 GBit/s I/O speed.
20 TB are used.

Illumina Cluster

The sequencing cluster provides a rich software environment for processing sequence data, from raw conversion using Illumina software to open source analysis packages in R.

Visualization and data management are partially covered by a gbrowse installation, a web-based genome browser similar to Flybase.
The fileserver contains 34 TB of data.

Convert raw cluster intensities to text

```
$ sge-bcl2qseq Run028/Data qseq/
```

-> Qseq text containing sequence reads

Convert to more commonly accepted format

```
$ qseq2fastq lane1_1_*.qseq > lane1_1.fastq
```

-> FastQ text format, e.g. 8GB per lane

Basic quality control

```
$ fastqc -t 8 lane_1_1.fastq -o fastqc/
```

-> Simple to understand HTML output

Map reads to reference genome

```
$ sge-bowtie-single hg18-chromosomes lane1.fastq
```

-> SAM format storing the reads mapped positions

Compress and index for faster access

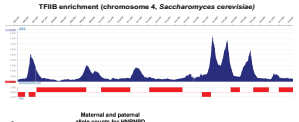
```
$ sge-samtools lane1.sam
```

-> indexed BAM format: less storage, faster access

Optionally extract positions of pair centers

```
$ bam2paired_end_centers.pl yeast.fasta lane1.bam
```

-> produces readily usable tab-separated files



- RNA-Seq
- MNase-Seq
- SNP-Detection
- Allele Counting
- ...

Data Flow

Acquisition

Expensive scientific instruments like Deep Sequencers and Cryo-Electron-Microscopes produce a huge amounts of raw data.

Data are directly written to dedicated fileservers or copied by hand. Scripts automatically copy the data to the clusters for processing and archive them to the LRZ.

Data Processing

Compute clusters

In-house Compute clusters provide the power and flexibility to process the data.
Apart from the computing speed, the I/O performance plays an ever increasing role.
Automated daily backups are not possible without a huge impact on the I/O performance.

Automated raw data steps

Raw data usually need to be converted to formats suitable for the tools used for further processing.
Data cleaning and quality control is partially included.
Scripts are in place to handle these steps automatically using the entire cluster.

Semi-automated further steps

Further processing of the data is semi-automated but could be fully automated if minimal config files in a strict format are provided with the raw data.
Simple interactions may be necessary for filtering low quality data.
These common steps can often be achieved with different programs (imagic vs. ctffind; bowtie vs. bwa).

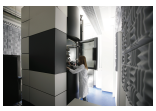
Analysis

The diverse ways to make use of the data, the multitude of parameters and the iterative process of analysis usally only allow for providing templates, documentation and consulting.

Cryo-EM

Titan Krios

In production, around 200 GB of images are produced daily.
Up to 1 TByte of data is accumulated in sessions from one to five days.

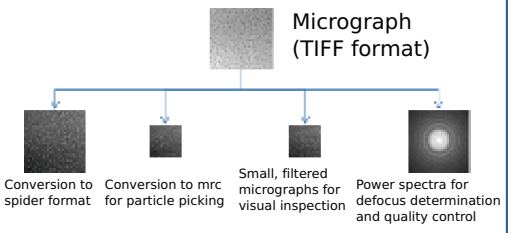


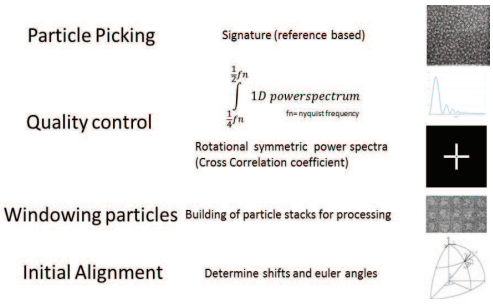
Dedicated Fileserver

16TB on RAID 6 of 20 * 1 TB HDDs with 4 GBit/s I/O speed.
12TB are used.

Cryozero Cluster

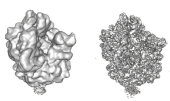
The biggest cryo cluster consists of 30 compute nodes from two generations with over 300 CPU cores.
Spider, Sparx and other Cryo-EM tools are installed in multiple versions and setup to make use of the entire clusters, e.g. via MPI.
The fileserver contains 32TB of data.





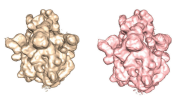
Refinement

Improving the resolution of the density maps needs a lot of user interaction. The best alignment parameters have to be found to converge ito a high resolution structure.



Sorting

Heterogenity in the data set can prevent high resolution reconstructions. The user has to sort for different functional states and use the most homogeneous projections for reconstruction.



Backup and Archiving

LRZ Tivoli Archive

More than 10 Petabytes stored on dozens of servers with 1000s of Tapes.
The gene center uses over 100TB.

With the amount of data and the dynamics of scientific careers organized backup and archiving is essential to fullfill the requirements of good scientific conduct.

Housedata

Two servers directly mirror each other providing highly available storage for data exchange and backup.