Data processing on a large scale

Chris Lawerenz[1], Sylwester Radomski[2] and Jürgen Eils[1]

[1] German Cancer Research Center, Germany
{c.lawerenz, j.eils}@dkfz.de
[2] University of Heidelberg, Heidelberg, Germany

Abstract. Within the last decade the high costs and complexity of Next Generation Sequencing (NGS) data organization put pressure on NGS data centres to organize convenient IT service infrastructures for automatic data management, processing and analyses. Our market analysis showed that existing applications processing NGS data were insufficiently documented, not extensible or strongly dependent on the underlying technical system. Thus, we were motivated to develop an automated job control system, called One Touch Pipeline (OTP), to ensure highest quality and cost reduction of data processing in terms of man power and time.

The functionality of OTP encompasses all the relevant steps of the whole pipeline from sequence acquisition to data analysis using high-performance computing. The OTP provides a flexible solution for automatic processing of NGS sequence data generated by sequencing centres. User friendly web pages and the platform independence of the OTP application guarantees a sustainable solution in the long term.

The crucial strengths of OTP are:
•Automatic processing including alignment of NGS data.
•Management and coordination of sequencing runs and associated metadata.
•Generation of quality control (e.g. fastqc results, coverage rate) and scores.
•Web based support for principal investigators, sequencing centres etc.
•Monitoring of job activities and quality control.
•Interfaces for automated export of raw data and results to ICGC, EGA and ENA repositories.
•Automated distribution of jobs across the cluster system with 1600 nodes
•Storage capacity: approximately 10 Petabytes.

Projects using OTP application: the German ICGC projects PedBrain, MMML and Early Onset Prostate Cancer, the "Deutsches Epigenom Programm" (DEEP) as part of the International Human Epigenome Consortium (IHEC), the National Genome Research Network (NGFN), the Heidelberg Initiative for Personalized Medicine (HIPO).