Journal of Biotechnology 261 (2017) 53–62

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

OTP: An automatized system for managing and processing NGS data

Eva Reisinger[a,b,*], Lena Genthner[a], Jules Kerssemakers[a], Philip Kensche[a], Stefan Borufka[a], Alke Jugold[a], Andreas Kling[a], Manuel Prinz[a], Ingrid Scholz[a], Gideon Zipprich[a], Roland Eils[a,b,c,d], Christian Lawerenz[a], Jürgen Eils[a]

[a] *Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Germany*
[b] *Heidelberg Center for Personalized Oncology, DKFZ-HIPO, DKFZ, Heidelberg, Germany*
[c] *Institute of Pharmacy and Molecular Biotechnology and Bioquant Center, University of Heidelberg, Heidelberg, Germany*
[d] *Translational Lung Research Center Heidelberg (TLRC), German Center for Lung Research (DZL), University of Heidelberg, Heidelberg, Germany*

ARTICLE INFO

Keywords:
Next-generation sequencing
Data management
Data processing
Automation
User interface
Standardization

ABSTRACT

The One Touch Pipeline (OTP) is an automation platform managing Next-Generation Sequencing (NGS) data and calling bioinformatic pipelines for processing these data. OTP handles the complete digital process from import of raw sequence data via alignment of sequencing reads to identify genomic events in an automated and scalable way. Three major goals are pursued: firstly, reduction of human resources required for data management by introducing automated processes. Secondly, reduction of time until the sequences can be analyzed by bioinformatic experts, by executing all operations more reliably and quickly. Thirdly, storing all information in one system with secure web access and search capabilities. From software architecture perspective, OTP is both information center and workflow management system. As a workflow management system, OTP call several NGS pipelines that can easily be adapted and extended according to new requirements. As an information center, it comprises a database for metadata information as well as a structured file system. Based on complete and consistent information, data management and bioinformatic pipelines within OTP are executed automatically with all steps book-kept in a database.

1. Introduction

Human genetics and genomics research has massively expanded in the past decade and constant cost reduction made it feasible to introduce the Next-Generation Sequencing (NGS) technology into daily hospital routine and large cohort analyses. Both the size and diversity of datasets produced in health-care and medical-research have increased, which challenges the capabilities of life science data centers to correctly and efficiently process and store data. To cope with the flood of data, large storage pools and compute facilities, comprehensive processing platforms and automatic data handling are urgently needed to reduce human work. Through automation the large quantities of genomic data can be processed with small staff expenditure and fewer faults then by manual interaction. In a research context, automatically managed workflow initialization and maintenance allows the bioinformatics experts to focus on developing new analysis methods and interpretation rather than processing the data. In clinical context, automation ensures fast provision of results relevant for treatment. How NGS data processing is managed at a large data provider such as the German Cancer Research Center (DKFZ) and the affiliated National Center for Tumor Diseases (NCT) (NCT Heidelberg, 2017) is presented in this publication.

The DKFZ is one of the largest biomedical research centers and genome sequencing facilities in Europe. By now, most of the enormous data flow produced within the DKFZ is processed automatically via the One Touch Pipeline (OTP). OTP is a platform for structured data storage, data access management as well as processing of NGS data. It is designed to meet the data management needs in a diverse scientific environment and clinical research. Data processing capacity is provided to DKFZ's in-house projects like the Heidelberg Center for Personalized Oncology (HIPO) (HIPO, 2017) as well as to partners in the German Network for Translational Cancer Research (DKTK) (DKTK, 2017) and the German contributions to the International Cancer Genome Consortium (ICGC) (ICGC, 2017). In most of these projects, several hundreds of genomes have been processed and managed via OTP, such that a total of about 20.000 samples covering more than 11.000 donors have been accrued since the beginnings of OTP in 2012 (Fig. 1). In 2013, the NCT established sequencing for many of their cancer patients. This shift towards clinical application led to new requirements in order to provide a solution as a basis for treatment decision in clinical cancer care. In particular, clinical data processing demands faster and more automated

[*] Corresponding author at: Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Germany.
E-mail address: e.reisinger@dkfz-heidelberg.de (E. Reisinger).

http://dx.doi.org/10.1016/j.jbiotec.2017.08.006
Received 22 February 2017; Received in revised form 3 August 2017; Accepted 7 August 2017
Available online 10 August 2017

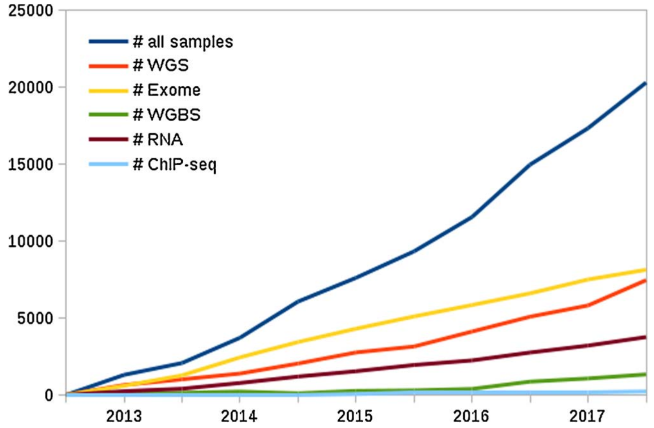

Fig. 1. An overview of all samples imported into OTP within the last five years is shown. The blue line represents all samples loaded in OTP. Each of the other lines represents another sequencing type. The graphic shows that the import of Exome and RNA data was quite constant while the import of whole genome sequencing (WGS) data increased. This can be explained by the fact that since late 2015 an increasing number of samples are produced by the Illumina HiSeq X Ten sequencers at the DKFZ. The reason for the rise of imported whole genome bisulfite (WGBS) samples is that a methylation calling pipeline is offered since 2016. The processing of ChIP-seq data was recently implemented. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

processing, improved data protection and security as well as streamlined quality assurance.

The automated management of recurrent, standardized analyses of diverse NGS data types is the key feature of OTP. Starting at automated meta- and raw data import, OTP handles the complete digital process from quality control and sequence alignment to the identification of single-nucleotide and structural genomic events, without manual interaction. Throughout this complete process OTP provides data provenance, quality monitoring and alerting, automated notification of processing status and automated error handling. To fulfill the requirements in the field of both, clinical and basic cancer research, a so called fast-track procedure was implemented which enables OTP to handle the data relevant for treatment decisions with a higher priority than the research data. Furthermore, since data security is a major concern in the human genomics field, OTP allows sustainable user management. Raw, processed and result files as well as access to according web sites is exclusively provided to authorized users.

Several other platforms also seek to provide such integration and automation and therefore prominent examples will be discussed and compared to OTP in chapter 9. It will be shown that OTP is strong in project data organization, performance and automation.

2. Dataflow in OTP

The general dataflow in OTP consists of several steps. First, the data is imported into OTP whereby raw data is stored on the file system and corresponding metadata in the database. Next, the alignment and the calculation of quality control values for the FASTQ files are triggered. After the alignment is completed, further analyses like variant calling are performed. To execute most of these analyses, OTP integrates Roddy, an external job execution framework (Heinold, 2016), which contains pipelines developed by on-site bioinformaticians. At the end of each step the user is automatically notified about the current status. OTP is able to process whole genome sequencing (WGS), whole exome sequencing (WES), whole genome bisulfite sequencing (WGBS), ChIP-seq and RNA sequencing data. A scheme of the dataflow is shown in Fig. 2.

2.1. Data import

The sequencing data and the corresponding metadata need to be registered in the OTP database and stored on the file system in a defined and consistent way. To minimize manual work during import and to allow import at night or during weekends, the process is fully automated. This was achieved by integrating the service management software OTRS (OTRS, 2017), which was adapted to call a URL on the OTP server as soon as a notification about new sequencing data arrives. For manual import from sequencing data providers without defined interface to OTP, a GUI is implemented that allows data import with a few clicks. In both manual and automatic import, OTP validates if the metadata file contains all information required for processing and storage on the file system. In case of successful validation, the actual import process is initiated, meaning that the metadata is stored in the OTP database and the sequencing files are installed on the file system. If validation fails due to ambiguities or missing information, manual interaction is required to clarify the metadata with the provider.

2.2. Processing of sequencing files

After the import of the sequencing data, quality control values are calculated for all FASTQ files using FastQC (Andrew, 2010). Simultaneously, several processing steps, covering alignment and corresponding quality control, single nucleotide variation (SNV) calling (Jones et al., 2012; Jones et al., 2013), insertion and deletion (InDel) calling (Rimmer et al., 2014), structural variation (SV) calling (Toprak et al., 2017), and copy number variation (CNV) calling (Kleinheinz et al., 2017), are applied. In many of these pipelines additional QC values are produced using in-house developed calculations. An overview of the different pipelines and integrated tools is given in Table 1.

Whether and how pipelines are executed depends on the sequencing type and on configurations defined per project. For execution three different conditions must be fulfilled: (1) the configuration for the specific pipeline and the project must be specified, (2) input parameters, like adapter sequence or library preparation kit, must be stored in the OTP database and (3) processing thresholds, like a minimal genome coverage have to be reached. The results of the pipelines are stored on the file system and QC values and further result information are stored in the database and provided via the GUI. For reproducibility purposes, the exact commands and parameters for the processing are stored in the database. After each processing step OTP ensures that only authorized users can access the result data on the file system by setting the access permissions based on the Unix group defined for the project. Data submitter and registered project members are automatically notified about successful termination of the process via email. The notification contains information, like the location of processed data on the file system, the link to the result data in the GUI, and subsequent analyses which will be performed.

2.3. OTP components

For the processing of data in a computing environment, OTP uses a number of separate components (Fig. 3). All compute intense processing tasks on genomic data are submitted to a high-throughput cluster running a batch processing system. OTP offers direct submission or indirect submission through the execution of the external job execution system Roddy. For simple tasks, including file management tasks and executing simple bioinformatics analysis tools, like FastQC (Andrew, 2010) or bwa aln (Li and Durbin, 2010), OTP submits the jobs directly to the cluster. More complex workflows, like the variant calling analyses, are submitted indirectly via Roddy. OTP collects all required input and configuration information from the database and executes Roddy providing this information. In turn, Roddy triggers the required analysis, submits all compute intense jobs to the cluster and returns the corresponding job identifiers to OTP. To determine whether an analysis