Computational and Structural Biotechnology Journal 15 (2017) 471–477

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows

Pranav Kulkarni, Peter Frommolt *

Bioinformatics Core Facility, CECAD Research Center, University of Cologne, Germany

ARTICLE INFO

Article history:
Received 13 July 2017
Received in revised form 29 September 2017
Accepted 6 October 2017
Available online 25 October 2017

ABSTRACT

While Next-Generation Sequencing (NGS) can now be considered an established analysis technology for research applications across the life sciences, the analysis workflows still require substantial bioinformatics expertise. Typical challenges include the appropriate selection of analytical software tools, the speedup of the overall procedure using HPC parallelization and acceleration technology, the development of automation strategies, data storage solutions and finally the development of methods for full exploitation of the analysis results across multiple experimental conditions. Recently, NGS has begun to expand into clinical environments, where it facilitates diagnostics enabling personalized therapeutic approaches, but is also accompanied by new technological, legal and ethical challenges. There are probably as many overall concepts for the analysis of the data as there are academic research institutions. Among these concepts are, for instance, complex IT architectures developed in-house, ready-to-use technologies installed on-site as well as comprehensive Everything as a Service (XaaS) solutions. In this mini-review, we summarize the key points to consider in the setup of the analysis architectures, mostly for scientific rather than diagnostic purposes, and provide an overview of the current state of the art and challenges of the field.

© 2017 Kulkarni, Frommolt. Published by Elsevier B.V. on behalf of the Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Next-Generation Sequencing (NGS) has emerged as a standard technology for multiple high-throughput molecular profiling assays. Among these are transcriptome sequencing (RNA-Seq), whole-genome and whole-exome sequencing (WGS/WXS) for instance for genome-wide association studies (GWAS), chromatin immunoprecipitation or methylated DNA immunoprecipitation followed by sequencing (ChIP-Seq or MeDIP-Seq), as well as a multitude of more specialized protocols (CLIP-Seq, ATAC-Seq, FAIRE-Seq, etc.). NGS is actually a subordinate concept for a number of comparatively new technologies. This review is focused on the analysis of data generated by the most widely used Illumina sequencing machines. Other technologies include the Sequencing by Oligonucleotide Ligation and Detection (SOLiD) method (Applied Biosystems), the 454 sequencing (Roche) and IonTorrent (ThermoFisher) machines as well as sequencers of the third generation manufactured by Oxford Nanopore and Pacific Biosciences. All these technologies are capable of generating tremendous amounts of information at base-level resolution, within relatively short time, and at low cost. These recent developments have turned the methods used in research projects into systems-wide analysis tools on organisms and diseases, which has revolutionized the paradigms followed in the life sciences in general. The appropriate selection of the right approaches to the analysis of the data is therefore a key discipline of this new era. In particular, there is a big need for clever ways to organize and process all the data within reasonable time [26] and in a sustainable and reproducible way (Fig. 1). Across most research projects as well as in many clinical environments, NGS analysis workflows share a number of steps which are the same for many use cases. Scientists around the globe have therefore established highly standardized analysis pipelines for basic NGS data processing and downstream analysis. The analysis workflows must be highly standardized, but at the same time flexible enough to also do tailored analyses and quickly adopt novel analysis methods that are developed by the scientific community.

For academic institutions, there are a number of good reasons to make investments into their own data analysis infrastructure instead of relying on commercial out-of-the-box solutions. First, academic institutions are strongly interested in having full control over the algorithms that are used for the analysis of their data. Second, the commercial solutions which do exist are either less flexible regarding their extension with additional analysis features or lack functionality and scalability to organize multiple analyses from many laboratories and very diverse research projects at a time. Third, the issue of ownership and access permissions to the data can be very important for confidential research data and patient data. This is a strong argument to not bring the data

* Corresponding author.
E-mail address: peter.frommolt@uni-koeln.de (P. Frommolt).

https://doi.org/10.1016/j.csbj.2017.10.001

2001-0370/© 2017 Kulkarni, Frommolt. Published by Elsevier B.V. on behalf of the Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

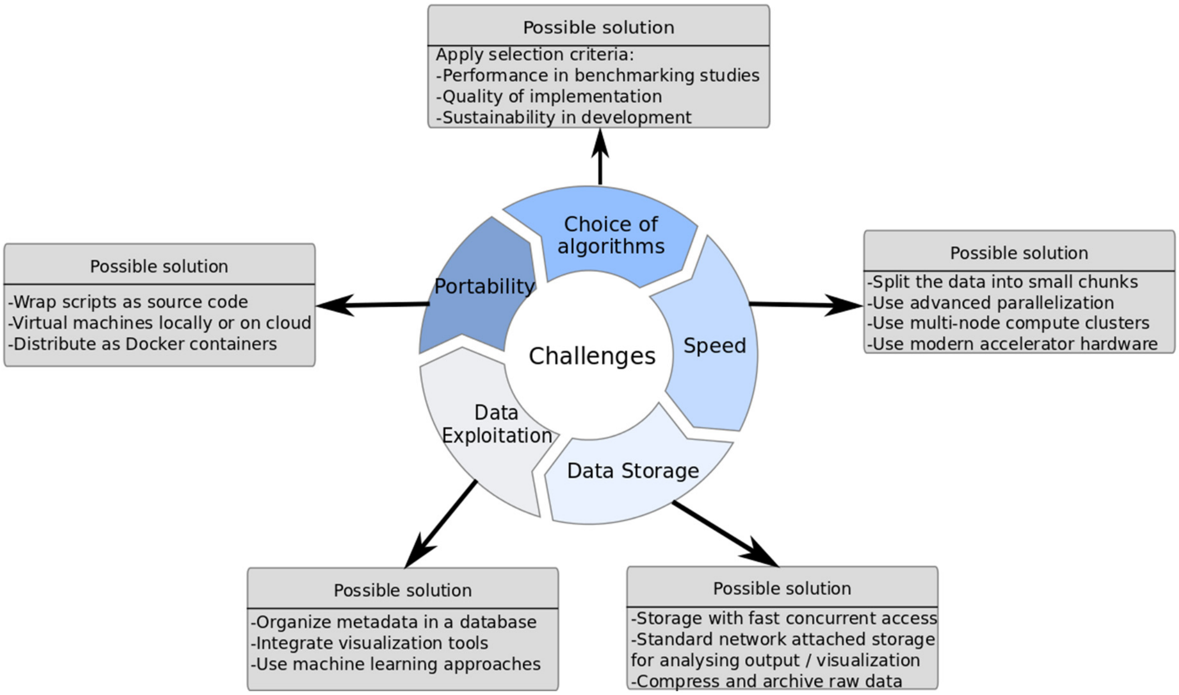
Possible solution
Apply selection criteria:
-Performance in benchmarking studies
-Quality of implementation
-Sustainability in development
Choice of algorithms
Portability
Speed
Challenges
Data Exploitation
Data Storage
Possible solution
-Wrap scripts as source code
-Virtual machines locally or on cloud
-Distribute as Docker containers
Possible solution
-Split the data into small chunks
-Use advanced parallelization
-Use multi-node compute clusters
-Use modern accelerator hardware
Possible solution
-Organize metadata in a database
-Integrate visualization tools
-Use machine learning approaches
Possible solution
-Storage with fast concurrent access
-Standard network attached storage for analysing output / visualization
-Compress and archive raw data

Fig. 1. Overview of the most important challenges in the design and implementation of NGS analysis workflows and suggestions how these challenges can be addressed.

beyond the institution's firewall, e.g. by giving them to external suppliers following an Everything as a Service (XaaS) model. The requirements regarding reproducibility, validation, data security and archiving are particularly high where NGS technology is being used for diagnostic purposes. Finally, a data analysis infrastructure developed in-house provides improved flexibility in the design of the overall architecture and allows, for instance, the quick integration of novel scientific methodology into bioinformatics pipelines. On the downside, an in-house data analysis infrastructure requires significant investments regarding personnel, time and IT resources.

The optimal way to setup NGS analysis workflows highly depends on the number of samples and the applications for which data processing needs to be pipelined. The diversity of challenges in the setup of such a system are reflected by the fact that over the last years, a whole research field has emerged around new approaches to all aspects of NGS data analysis. For a small research laboratory (<20 scientists) with a very narrow research focus, the setup may require analysis pipelines for highly specialized scientific questions at a maximum of flexibility. In contrast, an academic core facility typically needs to process data from dozens of laboratories covering multiple research fields at a sample throughput in the thousands per year. Key features of a successful analysis workflow system in such an environment are therefore resource efficiency in data handling and processing, reproducibility, and sustainability.

2. Data Processing

The unique challenge, but also the big chance in the NGS analysis field lies in the tremendous size of the data for every single sample analyzed. The raw data typically range in dozens of gigabytes per sample, depending on the application. For whole-genome sequencing, the size of the raw data can be even up to 250 GB (Fig. 2). Given sufficient computational resources, the overall workflows can be streamlined and highly accelerated by establishing centralized standard pipelines through which all samples analyzed at an academic institution are processed. State-of-the-art version control on the underlying pipeline scripts greatly improves the reproducibility of NGS-based research results in such an environment. A commonly used system for both software development and version control is *git*: the software version used for a particular analysis can, for instance, be controlled by tracking the ID of the latest git commit before the analysis has started (https://github.com).

A very basic decision is whether to build the data processing pipelines up from scratch or whether to leverage one of the existing frameworks for large-scale NGS analysis pipelining. Among the most prominent of these are the frameworks *GenePattern* [32] and *Galaxy* [12]. Both are open source platforms for complex NGS data analyses operated on cloud-based compute clusters linked to a front-end web server which enables facilitated user access. They can be used to run computational biology software in an interactive graphical user interface (GUI)-guided way and make these tools accessible also to scientists without extensive skills in programming and on a UNIX-like command line. They offer many off-the-shelf pipeline solutions for commonly used analysis tasks, which can however be modified in a flexible and interactive way. Other publicly available analysis workflow systems include QuickNGS [7,40], Chipster [15], ExScalibur [5], and many others (Table 1). Regarding the setup of the overall architecture, the daily operation and the choice of the particular tools, especially in customized pipelines, a user of any of these systems still heavily relies on the help of experts with IT and computational biology background. Thus, the decision whether a data analysis infrastructure is build up from scratch or based on one of the aforementioned frameworks is mostly a trade-off between flexibility and the necessary investments at all levels.

2.1. Typical Steps in an NGS Analysis Pipeline

Raw data are usually provided in FastQ, an ASCII-based format containing sequence reads at a typical length of up to 100 base pairs. Alongside with the sequences, the FastQ format provides an ASCII-coded quality score for every single base. After initial data quality checks and filtering procedures, the first step in most analysis workflows is a sequence alignment to the reference genome or transcriptome of the organism of origin. For de novo sequencing of previously uncharacterized genomes, a reference-free de novo genome assembly is required. These are the most data-intensive and time-consuming

Download English Version:

https://daneshyari.com/en/article/8408575

Download Persian Version:

https://daneshyari.com/article/8408575

Daneshyari.com