Bioinformatics, 36(6), 2020, 1983–1985
doi: 10.1093/bioinformatics/btz850
Advance Access Publication Date: 14 November 2019
Applications Note

Databases and ontologies

DigestiFlow: from BCL to FASTQ with ease

Manuel Holtgrewe [1,2,*], Clemens Messerschmidt[1,2], Mikko Nieminen[2,3] and Dieter Beule[3]

[1]Core Unit Bioinformatics, Berlin Institute of Health, Berlin 10178, Germany, [2]Charité — University Medicine Berlin, Berlin 10117, Germany and [3]Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin 13125, Germany

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on July 29, 2019; revised on October 29, 2019; editorial decision on November 8, 2019; accepted on November 13, 2019

Abstract

Summary: Management of raw-sequencing data and its pre-processing (conversion into sequences and demultiplexing) remains a challenging topic for groups running sequencing devices. They face many challenges in such efforts and solutions ranging from manual management of spreadsheets to very complex and customized laboratory information management systems handling much more than just sequencing raw data. In this article, we describe the software package *DigestiFlow* that focuses on the management of Illumina flow cell sample sheets and raw data. It allows for automated extraction of information from flow cell data and management of sample sheets. Furthermore, it allows for the automated and reproducible conversion of Illumina base calls to sequences and the demultiplexing thereof using bcl2fastq and Picard Tools, followed by quality control report generation.

Availability and implementation: The software is available under the MIT license at https://github.com/bihealth/digestiflow-server. The client software components are available via Bioconda.

Contact: manuel.holtgrewe@bihealth.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Laboratories operating modern-sequencing facilities face a multitude of challenges. These include sample tracking, *raw data preprocessing* (conversion of raw sequencer output into sequences and demultiplexing of pooled experiments which is usually done in the same step), quality control of sequencing results and delivery to the requesting party. Although there is no clear consensus of what comprises a Laboratory Information Management System (LIMS), the term LIMS is often used to describe systems supporting these step. Simple 'pure peopleware' implementations consist of spreadsheets on network shares while comprehensive commercial packages such as Illumina BaseSpace Clarity LIMS offer highly adjustable but very expensive solutions. A number of academic and open solutions fall in between, offering a variable number of features and degrees of customizability.

The general lack of agreement of what a LIMS should cover or not cover stems from the fact that sequencing laboratories alone differ greatly. Areas of difference include the types of samples accepted (tissues/blood, DNA/RNA, final libraries/pools or a subset thereof), and the type of data generated (raw base calls (BCLs), sequences, aligned reads or bioinformatics analytical reports). In addition, the surrounding information technology (IT) infrastructure varies greatly as does the degree of integration with such additional IT systems.

In this article we present our approach *DigestiFlow* (DF) that addresses the different needs of organizations by focusing on a small, well-defined subset of tasks: management of Illumina flow cell and sample sheet information and orchestrating the step converting BCLs to sequences and demultiplexing pooled sequencing runs. To the best knowledge of the authors, in this domain DF offers unparalleled functionality. Flow cells can be filled with an arbitrary combination of libraries using any combination of index and molecular barcode reads. DF also supports the barcode being part of the template sequence. DF provides extensive features for sanity checking and comparison of expected indexing reads with those actually seen in the raw BCL data.

This is particularly important in an era where technologies such as single cell and low input sequencing require an ever-growing complexity of barcoding and indexing schemes and the amount of sequencer throughput is growing dramatically. We have encountered flow cells with more than 600 libraries and expect this to grow with increasing sequencer throughput.

A fundamental link to central IT is the integration with existing authentication infrastructure via directory servers, e.g. Microsoft ActiveDirectory (AD). DF supports linking accounts to central AD instances as well as using user accounts that only exist within the system. Beyond this, the system provides its functionality through a REST API (representational state transfer application programmable interface application programing interface) such that other services can be easily integrated. Instead of covering all possible functionality and sample tracking schemes, DF avoids the complexity of a monolithic system and can be integrated as a part of a modular system.

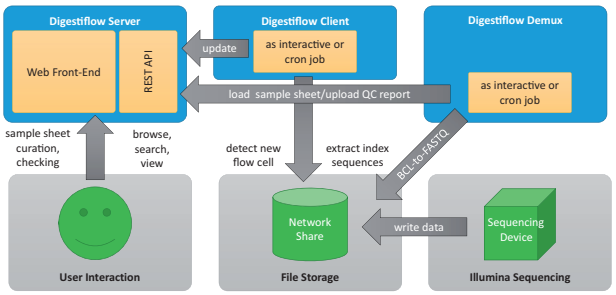

Fig. 1. Architectural overview. Sequencing instruments write data to a specified file system storage. A periodically running DF Client detects new flow cells and registers them with the DF Server. Once sequencing is complete and sample sheet information has been approved by the operator, DF Demux performs the conversion to FASTQ files and creates all QC reports. Users can not only browse and view, but also manage and curate flow cells and their sample sheets through DF Server

However, it can also just be standalone without integration with any other system.

2 Materials and methods

DF consists of three major components. The architecture of the system is shown in Figure 1. The figure also shows interaction with a minimal set of external systems.

2.1 DF server

DF Server is a web app implemented with the SODAR-Core (https://github.com/bihealth/sodar_core) and Django frameworks in the Python programing language. It uses a PostgreSQL database system. It allows for the curation of flow cells and libraries together with arbitrarily complex index and barcoding schemes. Barcodes can be organized in barcode sets such that their sequence can be entered once and subsequently be referred to by name. Furthermore, sequencing machines can be registered with their main properties (e.g. whether the second barcode read needs to be reverse-complemented, depending on the paired indexing workflow used). DF Server allows the visualization of barcodes detected by DF Client in the BCL files (see Fig. 2) and compares them to the libraries and barcodes entered by the users into the flow cell sample sheet. DF Server provides a number of sanity checks for both barcodes from raw-sequencing data, including a barcode frequency distribution and recognizing expected spike-ins such as PhiX sequence. It can also cross check between sample sheets and barcodes in the raw sequences, detecting barcodes present in one but missing in the other. Furthermore, users can add comments to flow cells and attach arbitrary files, which is useful for exchanging spreadsheets or concentration measurement reports from the wet lab.

2.2 DF client

DF Client is meant to be called periodically via a cron job to monitor the storage volume where sequencer(s) write output data. It reads the metadata files and registers any new flow cell with DF Server (or alternately, flow cells can be pre-registered in DF Server and their properties are then updated by DF Client). Once the barcodes have been sequenced completely, the DF Client extracts and evaluates their sequences and posts this information to DF Server. The client also detects when sequencing has succeeded (and various failure conditions) and updates the information in the server. DF Client is written in the Rust programing language.

2.3 DF Demux

DF Demux is also meant to watch the storage volume where the sequencers write their output data. Once sequencing of a flow cell is complete and marked as ready in DF Server, it starts the pre-processing by first obtaining the flow cell information from DF Server. Flow cells can be marked for delivery as BCL files, (possibly) demultiplexed sequences, or both. If raw BCL files are to be delivered, DF Server simply creates a TAR (tape archive) file for each lane that contains all the information required for demultiplexing this one lane.

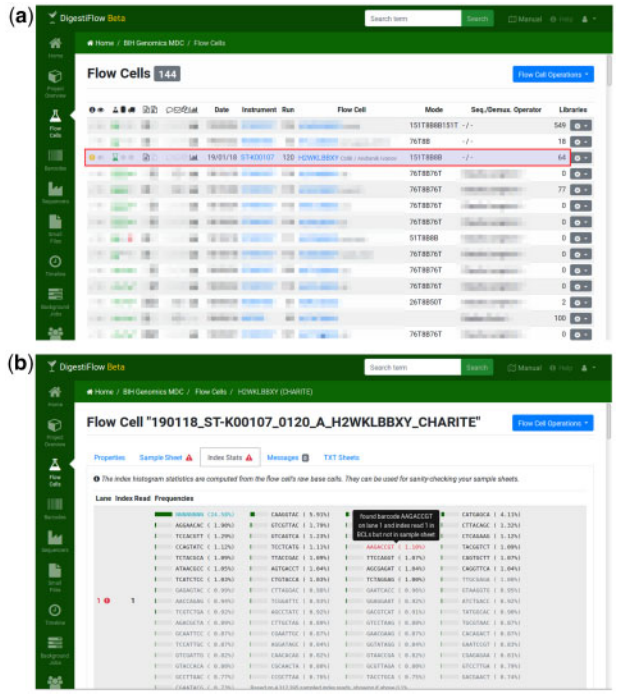

Fig. 2. When adding the sample sheet (not shown), the operator made a small mistake. The adapter P37 is given twice for the same lane in the sample sheet while the adapter sequence 'AAGACCGT' occurs in the raw BCLs but not in the sample sheet. This information can then be used for debugging sample sheet information. This is highlighted in the sample sheet (**a**) and the display of the adapters read from the raw BCL data (**b**)

For pre-processing, it first checks whether the flow cell can be processed by simply calling the Illumina vendor software *bcl2fastq* (version 1 or 2, depending on the needs of the raw data) and calls the program accordingly. Otherwise, it generates a series of calls to bcl2fastq and Picard Tools (http://broadinstitute.github.io/picard/) to perform the required pre-processing. An example for this is the Agilent XT protocol where molecular barcode sequences are stored in the second barcode read which bcl2fastq does not support. We refer to homogenous flow cell loads that can be processed with the bcl2fastq as *basic pre-processing* while *flexible pre-processing* allows arbitrary combination of library indexing and barcoding schemes.

Once pre-processing is complete, FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is run on the results and the quality control results are be collected with MultiQC (Ewels *et al.*, 2016). Finally, the MultiQC report is posted as a message to the flow cell in DF Server using the REST API together with the log files as attachment. This then allows the sequencing staff to review the results and react accordingly. DF Demux is implemented in the Python programing language with a Snakemake (Köster and Rahmann, 2012) workflow using Bioconda (Grüning *et al.*, 2018) for software package management.

3 Results

3.1 The states of a flow cell

DF tracks three components of sequencing: (i) the sequencing process itself, (ii) pre-processing and (iii) data delivery. The possible state values differ for each of these three steps. See the manual in the Supplementary Material for full details, but they can be summarized as follows.

Table 1. Comparison important properties and features in commercial and free software for the management of Illumina flow cells information popular in the sequencing community based

Metric	DigestiFlow	BaseSpace Clarity LIMS	OpenBIS LIMS-ELN	MendeLIMS	MISO	Parkour LIMS
License	MIT	commmercial	free for non-commercial	free for non-commercial	GPL	GPL
Self-hosted	✓	—	✓	✓	✓	✓
LDAP auth	✓	✓	✓	✓	✓	—
(REST) API	✓	✓	✓	—	✓	✓
Sample tracking	minimal	advanced	basic	basic	basic	advanced
Basic pre-proc.	✓	✓	✓	✓	—	✓
Flexible pre-proc.	✓	—	—	—	—	—
Sheet checks	✓	—	—	—	—	—
BCL checks	✓	—	—	—	—	—

pre-proc., preprocessing

Each step starts in the initial state. For pre-processing, an operator has to set the state of the pre-processing step to ready which signals DF Demux to start. Once the client detects that the sequencing of a flow cell has started the sequencing state changes to running. Similarly, once DF Demux has started, the pre-processing state changes to running. If sequencing or demultiplexing fails or succeeds, the corresponding state is updated accordingly (failed/success). For both, a human operator can set a special confirmed failure/success state manually. For example, a confirmed failure state will be set manually after they determine pre-processing failed due to overall low sequencing quality and it is not possible to 'rescue' pre-procesisng by fixing a sample sheet. Or a confirmed success state may be set after a human operator determines that QC passes visual inspection. A special success with warning state allows users to flag situations such as sequencing which succeeded for all but one lane due to technical issues.

For delivery, a human user has to set the state explicitly. This built-in system for keeping track of the delivery state is particularly useful if more than one user is handling data delivery, especially when used in conjunction with the message feature for leaving notes on flow cells.

3.2 Comparison with existing methods

Existing methods include the following: openBIS ELN-LIMS (Barillari *et al.*, 2016), which builds on top of openBIS (Bauch *et al.*, 2011) and has a high number of features for sample submission and -tracking yet also has a large number of dependencies, MendeLIMS (Grimes and Ji, 2014) which has basic sample tracking functionality yet is bound to a rigid data processing workflow, Managing Information for Sequencing Operations (MISO)-LIMS (https://github.com/miso-lims/miso-lims) which offers basic sample tracking functionality yet does not include features for pre-processing, and Parkour LIMS (Anatskiy *et al.*, 2019) which provides extensive sample tracking and advanced lab notebook features yet also does not integrate automated pre-processing. Although being out of scope of this article, we note that DF could be integrated with other software packages as long as they provide an API with additional code. The integration with Parkour LIMS appears particularly appealing as it is based on the same technology as DF (Python/Django) and has few other dependencies itself. Table 1 contains a comparison of the listed tools given some important features.

3.3 Features for improving sequencing results

3.3.1 Sample sheet validation

Based on practical experience, we greatly appreciate the automated comparison of observed adapter sequence content and sample sheet. Unexpected sequence in either set is an indication for possible errors. DF Server provides fine-grained control to acknowledge and suppress inconsistency warnings (after either fixing errors or accepting errors and then excluding corresponding data). Furthermore, common artifacts such as PhiX sequence are automatically recognized and show up as information rather than warnings or errors. Figure 2 shows an example.

3.3.2 Reproducibility, automation and quality control

The DF Client and Demux components are available from Bioconda as Conda packages and Docker images, thus allowing for future proof installations and creating reproducible workflows. By offering REST APIs and two useful client applications, DF greatly supports sequencing and demultiplexing operators in automating their work. Further automation can be added later as the APIs are open. Automated quality control using FastQC and aggregation using MultiQC also allows users to spot problems earlier (together with the sample sheet adapter checks described earlier). In our experience this allows for the early detection of many common issues. For example, from time to time, it occurs that the same adapter was used for two different libraries in the same lane. This error might be hard to spot on paper or in spreadsheets but applications such as DF Server can easily detect and report such problems similar to the example shown in Figure 2.

Acknowledgements

We thank all members of the sequencing facilities of BIH, Charité and MDC, especially Tatiana Borodina and Marten Jäger for their valuable input and feedback on DigestiFlow functionality and design. We thank Nina Thiessen for language editing.

Funding

All authors were funded as staff of the Berlin Institute of Health (BIH).

Conflict of Interest: none declared.

References

Anatskiy,E. *et al.* (2019) Parkour LIMS: high-quality sample preparation in next generation sequencing. *Bioinformatics*, **35**, 1422.

Barillari,C. *et al.* (2016) openBIS ELN-LIMS: an open-source database for academic laboratories. *Bioinformatics*, **32**, 638–640.

Bauch,A. *et al.* (2011) openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*, **12**, 468.

Ewels,P. *et al.* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.

Grimes,S.M. and Ji,H.P. (2014) MendeLIMS: a web-based laboratory information management system for clinical genome sequencing. *BMC Bioinformatics*, **15**.

Grüning,B. *et al.* (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, **15**, 475–476.

Köster,J. and Rahmann,S. (2012) Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.