OXFORD

Databases and ontologies

AlmostSignificant: simplifying quality control of high-throughput sequencing data

Joseph Ward[1], Christian Cole[1,*], Melanie Febrer[2,†] and Geoffrey J. Barton[1,*]

[1]Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, DD1 5EH, UK and
[2]Molecular Medicine, School of Life Sciences, University of Dundee, Dundee, DD1 5RH, UK

[†]Present address: Illumina Cambridge Ltd, Granta Park, Great Abington, CB21 6GP, Cambridge, UK
*To whom correspondence should be addressed.

Abstract

Motivation: The current generation of DNA sequencing technologies produce a large amount of data quickly. All of these data need to pass some form of quality control (QC) processing and checking before they can be used for any analysis. The large number of samples that are run through Illumina sequencing machines makes the process of QC an onerous and time-consuming task that requires multiple pieces of information from several sources.

Results: AlmostSignificant is an open-source platform for aggregating multiple sources of quality metrics as well as run and sample meta-data associated with DNA sequencing runs from Illumina sequencing machines. AlmostSignificant is a graphical platform to streamline the QC of DNA sequencing data, to store these data for future reference together with extra meta-data associated with the sequencing runs not typically retained. This simplifies the challenge of monitoring the volume of data produced by Illumina sequencers. AlmostSignificant has been used to track the quality of over 80 sequencing runs covering over 2500 samples produced over the last three years.

Availability and Implementation: The code and documentation for AlmostSignificant is freely available at https://github.com/bartongroup/AlmostSignificant.

Contacts: c.cole@dundee.ac.uk or g.j.barton@dundee.ac.uk

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

1 Introduction

The current generation of high-throughput, short-read DNA sequencing technologies quickly generate large datasets, but all sequencing technologies produce errors, whether systematic, technical, or simple human mistakes. Tracking the quality of the sequence data is critical to a successful experiment, but can become time consuming for all but the smallest projects. Monitoring the success of a sequencing run requires dedicated software located on the sequencing machine or standalone software provided by its manufacturer. These are often used in conjunction with tools such as FastQC (www.bioinformatics. babraham.ac.uk/projects/fastqc) for assessing the quality of the output fastq files and FastQScreen (www.bioinformatics.babraham.ac. uk/projects/fastq_screen) to check for contaminating DNA. Although

these tools are simple to use and interpret for a handful of samples or a single project, this is difficult for sequencing facilities that handle many sequencing runs simultaneously.

AlmostSignificant is open-source software designed to simplify the aggregation of quality statistics for sequencing runs from Illumina MiSeq, NextSeq and HiSeq machines. The single, consistent interface for viewing these data helps to quickly and efficiently identify potential issues with sequencing runs, lanes within the run, or with the original samples. Aggregating information into a single system has the added bonus of allowing extraneous meta-data to be gathered automatically and the overall quality trends in sequencing runs to be viewed. Thus, the data production and run-protocol for each machine can be tracked over time and so guide lab-side decision making and
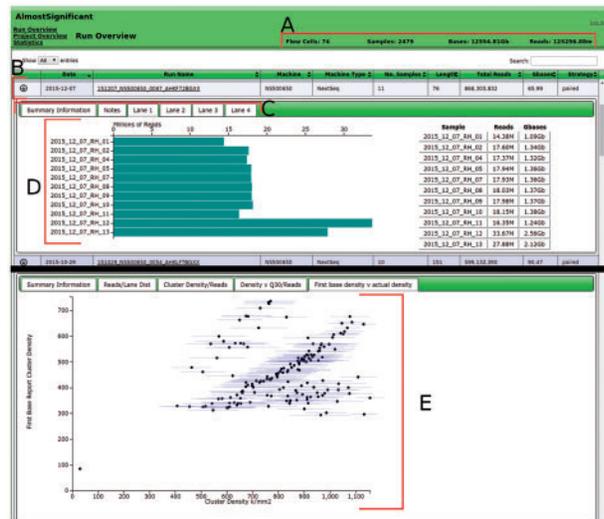
Fig. 1 A composite view of the AlmostSignificant interface. Shows *Run-Detail* view, with an expanded row showing the distribution of data produced for all of the samples in a lane. (**A**) A selection of summary statistics relevant to the current view, in this case showing total number of runs and amount of data. (**B**) A row showing information about a single sequencing run. (**C**) Tabs for the expanded drop-down for the selected run, accessed by clicking the arrow on the left of each row. (**D**) Plot showing the data distribution for the samples in the selected lane. (**E**) The relationship between cluster density for a run and the quality of data produced from the *Statistics-View* (Original version of this figure can be found in the Supplementary data.)

problem solving. In contrast to services such as Illumina's BaseSpace, AlmostSignificant uses local resources, is open-source, and all data remains local. While there are tools with some similarity to AlmostSignificant (e.g. StatsDB; Ramirez-Gonzalez *et al.*, 2014, MultiQC; Ewels *et al.*, 2016), there are distinguishing differences. AlmostSignificant includes data produced as part of the bcl2fastq Illumina pipeline (e.g. cluster density, Fig. 1E, and lane de-multiplexing) and sample meta-data rather than being limited to read data only. AlmostSignificant has been in production for over three years and currently tracks >80 runs covering >2500 samples.

2 Software description

2.1 Interface
The AlmostSignificant interface is designed to be consistent regardless of the data being viewed but still able to show the information pertinent for the task at hand. The interface is separated into five different views; the *Run Overview*, the *Run-details View*, the *Project Overview*, the *Project-details View* and the *Statistics View*.

The *Run Overview* (Fig. 1) shows a table of all the runs, with important metrics for the run shown at the top of the page. Additional information for each run can be accessed but clicking the arrow on the left of each row. This additional information allows at-a-glance checking of the top indexes that could not be assigned to samples (indicative of being given the wrong indexes) and bar-plots to identify any pooling inconsistencies and to ensure each sample produced the required amount of data. The *Run-Detail view* shows a list of all reads for all samples of the run. The interface allows output from FastQC or FastQScreen to be viewed in a single tab. Thus the user can obtain an overview of the quality of the current sample at a single glance, rather than having to open multiple files or scroll down a large HTML page.

The *Statistics View* shows slightly different information to the other views. The rows give a break-down for each length of run for

each type of machine used. For example, 100 bp HiSeq runs. It shows the number of flow cells that have been run for the run-protocol as well as the mean reads and bases produced for a whole flow-cell or a single lane. Expanding the row shows more details about the run-protocol, including the minimum and maximum output seen in reads and bases, as well as plots comparing the cluster density for the lane to the amount of data produced. Over-clustering flowcells has a detrimental effect on the read quality and output. The statistics view can reveal this and allows for optimization of laboratory protocols.

The *Project View* gathers together information on samples from individual projects, regardless of how many runs they were spread over. This allows a simple at-a-glance view of the amount of data produced for each project, similar to the way the *run overview* does for runs. The project name is a link to the *Project-detail View* summary table for all samples associated with the project regardless of which run they were sequenced on. This gives an easy way to view the samples for a project without having to visit views for multiple runs. An example server and dataset is available at http://www.compbio.dundee.ac.uk/almostSignificant.

2.2 Data loading
The script dataLoading.py requires access to three folders as input. The first is the processed-run folder generated by the Illumina bcl2fastq program; this provides information about de-multiplexing and sequences with undetermined indexes as well other meta-data. The second input is the run folder that is output from the sequencing machine (including sample sheet); this provides information on the run such as the cluster densities and other sequencing-run related meta-data. The third input is a single folder containing all of the quality information output by FastQC and/or FastQScreen. This folder can contain sub-folders but the names for the files should match the fastq files found in the processed run folder. Optional arguments include selecting the instrument type (defaults to NextSeq) and a flag to tell the script to identify the undetermined indexes found in the Undetermined Index files produced by the sequencing machine. Data loading is performed independently of the interface allowing it to be adapted to suit new formats without detriment to AlmostSignificant's user interface.

2.3 Requirements
AlmostSignificant requires python 2.7+ and Django 1.8+. For optimal usage, the illuminate python package (available via pip) should be installed, and pdflatex should be available from the command line. AlmostSignificant has been tested on Ubuntu 16.04 and Centos 6.6.

References
Ewels,P. *et al.* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 2016 Jun 16, btw354.
Ramirez-Gonzalez,R.H. *et al.* (2014) StatsDB: platform-agnostic storage and understanding of next generation sequencing run metrics. *F1000Research*, 2, 248.