BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 2 2012, pages 284–285
doi:10.1093/bioinformatics/btr613

Sequence analysis

Advance Access publication November 8, 2011

NARWHAL, a primary analysis pipeline for NGS data

R. W. W. Brouwer[1,2], M. C. G. N. van den Hout[1], F. G. Grosveld[1] and W. F. J. van IJcken[1,*]

[1]Center for Biomics, Department of Cell Biology, Erasmus Medical Center, Rotterdam and [2]Netherlands Bioinformatics Centre (NBIC), Nijmegen, The Netherlands

Associate Editor: Martin Bishop

ABSTRACT

Summary: The NARWHAL software pipeline has been developed to automate the primary analysis of Illumina sequencing data. This pipeline combines a new and flexible de-multiplexing tool with open-source aligners and automated quality assessment. The entire pipeline can be run using only one simple sample-sheet for diverse sequencing applications. NARWHAL creates a sample-oriented data structure and outperforms existing tools in speed.

Availability: https://trac.nbic.nl/narwhal/

Contact: w.vanijcken@erasmusmc.nl

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

Received on June 29, 2011; revised on October 17, 2011; accepted on November 2, 2011

1 INTRODUCTION

Massive parallel sequencing has reduced the costs of DNA sequencing to a fraction of that of traditional Sanger sequencing by performing many immobilized reactions in parallel. At the introduction of this technology in 2007, ~400 000 reactions could be performed in parallel per run (Johnson *et al.*, 2007). Now, with the introduction of the latest generation of sequencing equipment, over 100 million DNA fragments are sequenced per sample. Such a large number of sequences is not necessary for many applications and are therefore better distributed over multiple samples.

In multiplexing, individual DNA fragments are associated to unique DNA sequences (barcodes). These barcodes allow individual reads to be assigned to a single sample. By multiplexing many samples in a single sequencing run, the costs of functional sequencing experiments are lowered dramatically. As a result, the price of RNASeq is now similar to that of expression microarrays and the expectation is that the price of sequencing and thus RNA-Seq will lower even further. Due to these developments, the number of individual samples is expected to increase greatly.

As many researchers do not have access to an extensive IT-infrastructure, most sequencing centers perform a basic primary analysis of the produced data and align the sequences to a reference. With the increase in the number of samples and the amount of data, the standard software, Casava (http://www.illumina.com/), has difficulty coping both in terms of sample tracking and in terms of computational efficiency. The publically available aligners Bowtie and BWA (Li and Durbin, 2009; Langmead *et al.*, 2009) are able to align sequencing data more efficiently than the Eland aligner of Casava. These open-source aligners are implemented as stand-alone command-line tools with many settings influencing their results. To run these manually for large numbers of samples with differing applications and thus settings can become quite laborious and error prone.

In order to automatically process samples sequenced using the Illumina sequencers, we have developed the NARWHAL software pipeline. This pipeline combines a new and flexible de-multiplexing tool with open-source aligners and automated quality assessment.

2 NARWHAL DESCRIPTION

NARWHAL automates de-multiplexing, alignment and quality assessment; it provides various data export formats as well as extensive reports on the sequencing results. Download and installation instructions, a user-manual, sample configurations, and mailing lists for users and developers can be found in the supplementary information at https://trac.nbic.nl/narwhal/.

2.1 Requirements to run

NARWHAL has been designed to automate the primary analysis of sequencing data requiring minimal input from the operator. Only three input parameters are required to start a NARWHAL run, namely the location of the Illumina Qseq files, the output folder and a sample-sheet. This sample-sheet contains for each sample the sample name, lane number, barcode, barcode location in the read, reference genome and sequence application analysis type to be performed. It is also possible to skip de-multiplexing and only perform alignment and QC by specifying an input FastQ file in the sample-sheet. Based on these parameters, a NARWHAL run folder is generated containing parameter files that will govern the automated parallel processing.

2.2 De-multiplexing

NARWHAL collects the Qseq files from an Illumina BaseCalls folder and converts these files to the FastQ format using a new optimized C tool. This tool is very efficient and can be highly parallelized. The FastQ files are subsequently de-multiplexed using a custom C tool. The barcode can be present in any read at any location in the read. De-multiplexing is a two step process. First, index files are generated in which each read is assigned to a specific sample. Second, the FastQ files are separated per sample. Due to this approach, multiple FastQ files for multiplexed paired-end reads can be efficiently processed.

*To whom correspondence should be addressed.

2.3 Alignment

The reads in the sample-specific FastQ files are aligned to the reference sequence(s) by the alignment script. NARWHAL includes adjustable but standardized profiles for each sequencing application (RNASeq, ChIPSeq, ExomeSeq, etc.), which specify the aligner to be used, Bowtie or BWA, and its settings. For example, in ChIPSeq only the best alignment is generally of interest, and the speed of Bowtie is preferred. While in ExomeSeq, we use the slower but more accurate BWA. For flexibility, all settings from the profiles can be overridden in the sample-sheet. This approach reduces the chance of human error and required hands-on time. The final alignment results are in SAM format.

2.4 Quality assessment

Quality assessments are performed on sorted BAM files that are generated by SAMtools (Li *et al.*, 2009). In this assessment, we calculate and visualize the number of mapped reads per chromosome and the replication frequency. In addition, graphs are generated of the read-length and edit-distance distributions. From the edit-distance distribution, the edit-rate is calculated as the fraction of mismatches per aligned base. The QC information is assembled per sample in a single PDF file (Supplementary Material S1). The interpretation of these graphs is application dependent.

2.5 Technical implementation details

The NARWHAL control scripts are implemented in Python and BASH, while the data processing-intensive de-multiplexing tool is written in C. For the QC analysis, we use R and LaTeX together with a custom C++ read processor. Many of the tasks performed by NARWHAL can be parallelized to a large degree even on relatively modest hardware. NARWHAL uses the GNU parallel tool to perform parallelization on tasks that are I/O intensive such as data format conversions. For the alignments, NARWHAL utilizes the multi-threading support present in most alignment tools as this is more efficient than external parallelization strategies.

3 COMPARISON TO OTHER TOOLS

Numerous tools have been developed to process massively parallel sequencing data (Supplementary Table S2). These tools focus on performing quality control analyses (FastQC), format conversions (SAMtools) or provide complete processing pipelines for specific applications, e.g. NGS backbone (Blanca *et al.*, 2011) and GATK (McKenna *et al.*, 2010). Many of these tools have been included in the Galaxy framework (Giardine *et al.*, 2005). Even though this framework allows for a large flexibility using complex workflows, it is not able to handle external file locations and variable numbers of in- and output files. In contrast to Galaxy, NARWHAL is capable to analyze data from different applications with flexible settings in an automated fashion directly from the sequencer. Our pipeline increases the ease-of-use, reduces manual errors and hands-on time when processing large numbers of samples.

The function of NARWHAL overlaps mostly with the Illumina Casava software suite. Both tools perform de-multiplexing and alignment. We compared NARWHAL with Casava v1.7 for several characteristics (Table 1). In contrast to Casava, NARWHAL uses standard formats and open-source alignment tools making it easy to integrate with downstream analysis software. NARWHAL offers superior ease-of-use due to its simple configuration. NARWHAL processes the same dataset significantly faster than Casava on the same hardware.

Table 1. Comparison of NARWHAL and Casava

	NARWHAL	Casava
Computational time (CPU h)	200.85	597.4
Real runtime (h)	39.4	76.83
File formats	FastQ, SAM and BAM	Qseq, Export and Extended
Alignment methods	BWA, Bowtie	Eland, PhageAlign
Configuration	Sample-sheet	Sample-sheet and several configuration files

For both tools, the analysis times for a single paired-end 76 bp flow-cell (1 613 111 526 reads) with 16 indexed samples were determined using the GNU time tool. The comparison was performed on a server with 4 quad-core CPUs with 2 GB of memory per core. The tools were allowed a maximum of eight concurrent processes.

4 CONCLUSION

The NARWHAL sequence analysis pipeline allows for the automated processing and analysis of different and multiplexed datasets obtained from Illumina sequencers. NARWHAL allows easy alignment of individual samples to the reference genome of choice using predefined alignment profiles. The output files from NARWHAL are in the typical standard formats used in the field of sequencing (FastQ, SAM and BAM) and can be easily loaded into downstream analysis software. By taking these tools and pipelines into account with specific profiles, NARWHAL provides a coherent data analysis workflow with minimal hands-on time, a reduced chance of human error and faster analysis.

ACKNOWLEDGEMENT

The authors thank I. Palli and Dr H. Mei for their helpful support.

Funding: Netherlands BioInformatics Center as part of the BioAssist programme (to R.W.W.B.); NGI Booster grant from the Netherlands Genomics Initiative (to F.G.G.).

Conflict of Interest: none declared.

REFERENCES

Blanca,J.M. *et al.* (2011) ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using next generation sequence. *BMC Genomics*, **12**, 285.

Giardine,B. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

McKenna,A. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.