Bioinformatics, 33(10), 2017, 1565–1567
doi: 10.1093/bioinformatics/btx003
Advance Access Publication Date: 9 January 2017
Applications Note

Genome analysis

KNIME4NGS: a comprehensive toolbox for next generation sequencing analysis

Maximilian Hastreiter, Tim Jeske, Jonathan Hoser, Michael Kluge, Kaarin Ahomaa, Marie-Sophie Friedl, Sebastian J. Kopetzky, Jan-Dominik Quell, H.-Werner Mewes and Robert Küffner*

Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on September 7, 2016; revised on December 14, 2016; editorial decision on January 2, 2017; accepted on January 4, 2017

Abstract

Summary: Analysis of Next Generation Sequencing (NGS) data requires the processing of large datasets by chaining various tools with complex input and output formats. In order to automate data analysis, we propose to standardize NGS tasks into modular workflows. This simplifies reliable handling and processing of NGS data, and corresponding solutions become substantially more reproducible and easier to maintain. Here, we present a documented, linux-based, toolbox of 42 processing modules that are combined to construct workflows facilitating a variety of tasks such as DNAseq and RNAseq analysis. We also describe important technical extensions. The high throughput executor (HTE) helps to increase the reliability and to reduce manual interventions when processing complex datasets. We also provide a dedicated binary manager that assists users in obtaining the modules' executables and keeping them up to date. As basis for this actively developed toolbox we use the workflow management software KNIME.

Availability and Implementation: See http://ibisngs.github.io/knime4ngs for nodes and user manual (GPLv3 license)

Contact: robert.kueffner@helmholtz-muenchen.de

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

1 Introduction

Success of large-scale data analysis depends on sophisticated bioinformatic support to process, integrate, analyze and interpret Big Data volumes. In order to cope with the increased throughput of massive data generating experiments we create structured and reusable processing workflows for various analyses that are easy to apply and can be shared among the community. Although other comparable workflow management tools such as Galaxy (Goecks *et al.*, 2010) exist, we selected the open source platform KNIME (Konstanz information miner, Berthold *et al.* (2008)), as it offers an intuitive graphical user interface, is user friendly and easily extendable. With a very strong and active community, a remarkable number of new functions have been incorporated into KNIME. For instance, Knime4Bio (Lindenbaum *et al.*, 2011) has been developed for the interpretation of biological NGS datasets starting from variant calls.

Here, we describe an extension of KNIME by adding the functionality for essential NGS data processing. We developed a comprehensive linux-based KNIME toolkit including well-documented modules (nodes) for important steps like read preprocessing, read mapping, variant calling, detection of differential expression and annotation. Complementary to previously existing nodes, our toolbox now facilitates the assembly of basic building blocks into a wide range of customized NGS analysis workflows.

2 Implementation

The nodes provided by our extension can be connected smoothly by using their generic interfaces. They are well documented and assist users in the robust configuration of the underlying

Table 1. Collection of tools that are included in the KNIME4NGS extension

Whole exome/genome sequencing	*Description*
FastQC[a] and RawReadManipulator	Analysis and filtering of raw NGS reads
BWA, Bowtie2 and Segemehl	Mapping against reference genome
Samtools and Picard Utilities	File conversion, PCR duplicate removal as well as several auxiliary functionalities
GATK Utilities	GATK Best Practices and walkers for VCF manipulation
KGGSeq, SNPsift/eff, VEP	Variant annotation and filtration
RNA-Seq	***Description***
STAR	Splice-aware read aligner
FeatureCount	Assigning sequence reads to features
DESeq, EdgeR and Limma	Detection of differential expressed genes

For complete list and RawReadManipulator see user manual.

[a]Modified Version.

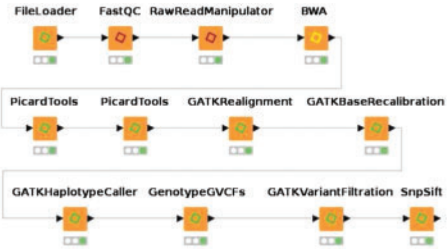

Fig. 1. Examplary workflow for analysing whole-exome data. This includes all steps from raw read quality control up to variant filtration and annotation

executables. Beyond the individual nodes we also added new layers of functionality that support the management and update of the underlying software packages as well as their high-throughput execution.

2.1 KNIME nodes

Our KNIME node extension for NGS currently contains 42 nodes (excerpt shown in Table 1. See comprehensive user manual for complete list and full description). Besides a number of auxiliary nodes, this extension provides building blocks and wrappers for well-established tools like BWA (Li and Durbin, 2010), DESeq (Anders and Huber, 2010) and GATK (McKenna *et al.*, 2010). The nodes can be easily combined to create standardized workflows starting from initial preprocessing of raw data, up to the variant calling and biological annotation (Fig. 1). This enables expert and non-expert users to set up reliable processing workflows quickly without the need of dealing with command-line tools. We also extended KNIME to use NGS specific file types. This improves workflow robustness enabling automated checks of tool configuration parameters and versions. To organize the required software binaries, we developed a lightweight binary management system accessible via the internal preference page. Thus, distributed nodes do not need to include the underlying executables, which simplifies keeping track with frequently updated software packages. The binary manager (See manual section 6) obtains necessary executables and integrates them into the workflow, minimizing repetitive node configuration.

NGS datasets are typically very large. Parallelizing the workflows and controlling data and results is crucial. Therefore we designed and tested our toolkit to work with corresponding KNIME extensions like openBIS (Bauch *et al.*, 2011), for storing and managing datasets, or the KNIME parallel chunk environment and cluster execution, for high-performance computing.

2.2 High-throughput executor

In the analyses of large datasets, even established tools are prone to spurious premature termination. This aborts the entire workflow and requires extensive human intervention to ensure that all of the parallel executing nodes/samples finish successfully. To reduce this effort as much as possible, we developed the High-Throughput Executor (HTE) as extension of the standard KNIME NodeModel. The HTE model collects process termination data of our nodes (e.g. error logs, execution time) and stores them in a local database provided by the system. Depending on the configuration and the process termination status, it automatically retries the execution of the failed node. The database allows the user to keep track and reduce the effects of unexpected software behaviour caused by for instance insufficient memory or randomly occurring errors. Additionally, the HTE model ensures by dedicated lock-files that not the entire workflow, but only those nodes are re-executed that depend on failed previous steps.

3 Discussion

Analysis of Big Data is often difficult to maintain and reuse as it depends on the correct application of multiple processing steps. As a solution, compiling the necessary tools into standardized graphical workflows makes the analysis pipeline more explicit, transparent and adaptable and can therefore stimulate collaboration between computational and wet lab biologists. The publication of scientific results together with the generating workflows furthermore has the potential to improve representation, reproducibility and dissemination of findings substantially.

We have presented a modular toolkit for the construction of NGS data processing workflows based on the KNIME environment that, in several ways, extends generic solutions e.g. the SeqAn KNIME library (Döring *et al.*, 2008). First of all, our toolkit comprises a set of nodes for individual NGS processing steps that can be combined into various workflows to serve as an extensible basis for many advanced NGS projects. The provided nodes and derived workflows are suitable for experts as well as less experienced users due to their integrated automatic parameter validation and comprehensive node descriptions accessible through the KNIME user interface. The modularity of our nodes allows fast and easy workflow adjustments by adding nodes or creating alternate paths. NGS projects are now characterized by their huge data volume and their need for massive parallel data processing. To improve the smooth handling of such Big Data workflows, we also provide a set of crucial technical KNIME extensions. These improve the robust setup, maintenance and configuration of workflows as well as their reliable execution and debugging with minimal human intervention. Taken together, KNIME4NGS can substantially lower the effort for scientists entering into the areas of NGS and Big Data.

Conflict of Interest: none declared.

References

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Bauch,A., *et al*. (2011) openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*, **12**, 1.

Berthold,M.R. *et al*. (2008) KNIME: The Konstanz Information Miner. In: Preisach,C. *et al*. (eds), *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*. Springer Berlin Heidelberg, pp. 319–326.

Döring,A. *et al*. (2008) SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

Goecks,J. *et al*. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*., **11**, 1.

Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.

Lindenbaum,P., *et al*. (2011) Knime4Bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME. *Bioinformatics*, **27**, 3200–3201.

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