BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 22 2011, pages 3200–3201
doi:10.1093/bioinformatics/btr554

Genome analysis
Advance Access publication October 7, 2011

Knime4Bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME[†]

Pierre Lindenbaum[1], Solena Le Scouarnec[2], Vincent Portero[1] and Richard Redon[1,*]
[1]Institut du thorax, Inserm UMR 915, Centre Hospitalier Universitaire de Nantes, 44000 Nantes, France and
[2]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

Associate Editor: John Quackenbush

ABSTRACT

Summary: Analysing large amounts of data generated by next-generation sequencing (NGS) technologies is difficult for researchers or clinicians without computational skills. They are often compelled to delegate this task to computer biologists working with command line utilities. The availability of easy-to-use tools will become essential with the generalization of NGS in research and diagnosis. It will enable investigators to handle much more of the analysis. Here, we describe Knime4Bio, a set of custom nodes for the KNIME (The Konstanz Information Miner) interactive graphical workbench, for the interpretation of large biological datasets. We demonstrate that this tool can be utilized to quickly retrieve previously published scientific findings.

Availability: http://code.google.com/p/knime4bio/.

Contact: richard.redon@univ-nantes.fr

Received on August 11, 2011; revised on September 13, 2011; accepted on September 29, 2011

1 INTRODUCTION

Next-generation sequencing (NGS) technologies have led to an explosion of the amount of data to be analysed. As an example, a VCF (Danecek *et al.*, 2011) file (Variant Call Format—a standard specification for storing genomic variations in a text file) produced by the 1000 Genomes Project contains about 25 million Single Nucleotide Variants (SNV), [http://tinyurl.com/ALL2of4intersection (retrieved September 2011)], making it difficult to extract relevant information using spreadsheet programs. While computer biologists are used to invoke common command line tools—such as Perl and R—when analysing those data through Unix pipelines, scientific investigators generally lack the technical skills necessary to handle these tools and need to delegate data manipulation to a third party.

Scientific workflow and data integration platforms aim to make those tasks more accessible to those research scientists. These tools are modular environments enabling an easy visual assembly and an interactive execution of an analysis pipeline (typically a directed graph) where a node defines a task to be executed on input data and an edge between two nodes represents a data flow. These applications provide an intuitive framework that can be used by the scientists themselves for building complex analyses. They allow data reproducibility and workflows sharing.

Galaxy (Blankenberg *et al.*, 2011), Cyrille2 (Fiers *et al.*, 2008) and Mobyle (Nron *et al.*, 2009) are three web-based workflow engines that users have to install locally if computational needs on datasets are very large, or if absolute security is required. Alternatively, softwares such as the KNIME (Berthold *et al.*, 2007) workbench or Taverna (Hull *et al.*, 2006) run on the users' desktop and can interact with local resources. Taverna focuses on web services and may require a large number of nodes even for a simple task. In contrast, KNIME provides the ability to modify the nodes without having to re-run the whole analysis. We have chosen this latest tool to develop Knime4Bio, a set of new nodes mostly dedicated to the filtering and manipulation of VCF files. Although many standard nodes provided by KNIME can be used to perform such analysis, our nodes add new functionalities, some of which are described below.

2 IMPLEMENTATION

The java API for KNIME was used to write the new nodes, which were deployed and documented using some dedicated XML descriptors. A typical workflow for analysing exome sequencing data starts by loading VCF files into the working environment. The data contained in the INFO or the SAMPLE columns are extracted and the next task consists in annotating SNVs and/or indels. One node predicts the consequence of variations at the transcript/protein level. For each variant, genomic sequences of overlapping transcripts are retrieved from the UCSC knownGene database (Hsu *et al.*, 2006) to identify variants leading to premature stop codons, non-synonymous variants and variants likely to affect splicing. Some nodes have been designed to find the intersection between the variants in the VCF file and a various source of annotated genomic regions, which can be: a local BED file, a remote URL, a mysql table, a file indexed with tabix (Li, 2011), a BigBed or a BigWig file (Kent *et al.*, 2010). Other nodes are able to incorporate data from other databases: dbSNFRP (Liu *et al.*, 2011), dbSNP, Entrez Gene, PubMed, the EMBL STRING database, Uniprot, Reactome and GeneOntology (von Mering *et al.*, 2007), MediaWiki, or to export the data to SIFT (Ng and Henikoff, 2001), Polyphen2 (Adzhubei *et al.*, 2010), BED or MediaWiki formats. After being annotated, some SNVs (e.g. intronic) can be excluded from the dataset and the remaining data are rearranged by grouping the variants per sample or per gene as a pivot table. Some visualization tools have also been implemented: the Picard API (Li *et al.*, 2009) or the IGV

[*]To whom correspondence should be addressed.

[†]During the reviewing process of this article another solution based on KNIME but focusing on FASTQ data files was published by Jagla *et al* (Jagla *et al.*, 2011).

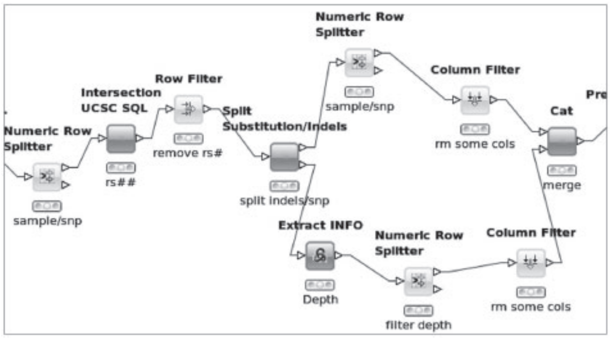

Fig. 1. Screenshot of a Knime4Bio workflow for the NOTCH2 analysis.

browser (Robinson *et al.*, 2011) can be used visualize the short reads overlapping a variation.

As a proof of concept, we tested our nodes to analyse the exomes of six patients from a previously published study (Isidor *et al.*, 2011) related to the Hajdu Cheney syndrome (Fig. 1). For this purpose, short reads were mapped to the human genome reference sequence using BWA (Li and Durbin, 2010) and variants were called using SAMtools mpileup (Li *et al.*, 2009). Homozygous variants, known SNPs (from dbSNP) and poor-quality variants were discarded, and only non-synonymous and variants introducing premature stop codons were considered. On a RedHat server (64 bits, 4 processors, 2 GB of RAM), our KNIME pipeline generated a list of six genes in 45 min: *CELSR1*, *COL4A2*, *MAGEF1*, *MYO15A*, *ZNF341* and more importantly *NOTCH2*, the expected candidate gene.[1]

3 DISCUSSION

In practical terms, a computer biologist was close to our users to help them with the construction of a workflow. After this short tutorial, they were able to quickly play with the interface, add some nodes and modify the parameters without any further assistance, but the suggestion or the configuration of some specific nodes (for example, those who require a snippet of java code). At the time of writing, Knime4Bio contains 55 new nodes. We believe Knime4Bio is an efficient interactive tool for NGS analysis.

ACKNOWLEDGEMENTS

We want to thank the Biostar community for its help, Jim Robinson and his team for the BigWig java API, and Dr Cedric Le Caignec for the *NOTCH2* data.

Funding: Inserm, the 'Centre Hospitalier Universitaire' of Nantes; the 'Fédération Française de Cardiologie' (FFC); 'Fondation pour la Recherche Médicale' (FRM). Solena Le Scouarnec is supported by the Wellcome Trust (Grant n WT077008).

Conflict of Interest: none declared.

REFERENCES

Adzhubei,I.A. *et al*. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Berthold,M.R. *et al*. (2007) Knime: the konstanz information miner. In Preisach,C. *et al*. (eds) *GfKl*, *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, pp. 319–326.

Blankenberg,D. *et al*. (2011) Integrating diverse databases into an unified analysis framework: a Galaxy approach. *Database*, **2011**, bar011.

Danecek,P. *et al*. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Fiers,M.W.E.J. *et al*. (2008) High-throughput bioinformatics with the Cyrille2 pipeline system. *BMC Bioinformatics*, **9**, 96.

Hsu,F. *et al*. (2006) The UCSC Known Genes. *Bioinformatics*, **22**, 1036–1046.

Hull,D. *et al*. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.

Isidor,B. *et al*. (2011) Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis. *Nat. Genet.*, **43**, 306–308.

Jagla,B. *et al*. (2011) Extending KNIME for next generation sequencing data analysis. *Bioinformatics*, **27**, 2907–2909.

Kent,W.J. *et al*. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.

Liu,X. *et al*. (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894–899.

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

Li,H. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.

Nron,B. *et al*. (2009) Mobyle: a new full web bioinformatics framework. *Bioinformatics*, **25**, 3005–3011.

Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.

Robinson,J.T. *et al*. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

von Mering,C. *et al*. (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.

[1] The workflow was posted on myexperiment.org at: www.myexperiment.org/workflows/2320.