*Bioinformatics*, 2017, 1–2
doi: 10.1093/bioinformatics/btx540
Advance Access Publication Date: 28 August 2017
Applications Note

OXFORD

Genome analysis

# NGS-pipe: a flexible, easily extendable and highly configurable framework for NGS analysis

**Jochen Singer[1,2,†], Hans-Joachim Ruscheweyh[1,2,3,†], Ariane L. Hofmann[1,2], Thomas Thurnherr[1,2], Franziska Singer[2,4], Nora C. Toussaint[2,4], Charlotte K. Y. Ng[5,6,7], Salvatore Piscuoglio[6], Christian Beisel[1], Gerhard Christofori[5], Reinhard Dummer[8], Michael N. Hall[9], Wilhelm Krek[10], Mitchell P. Levesque[8], Markus G. Manz[11], Holger Moch[12], Andreas Papassotiropoulos[13,14,15,16], Daniel J. Stekhoven[2,4], Peter Wild[12], Thomas Wüst[2,3], Bernd Rinn[2,3] and Niko Beerenwinkel[1,2,*]**

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland, [2]SIB Swiss Institute of Bioinformatics, Basel, Switzerland, [3]Scientific IT Services, ETH Zurich, Basel, Switzerland, [4]NEXUS Personalized Health Technologies, Zurich, Switzerland, [5]Department of Biomedicine, University of Basel, Basel, Switzerland, [6]Institute of Pathology, [7]Division of Gastroenterology and Hepatology, University Hospital Basel, Basel, Switzerland, [8]Department of Dermatology, University Hospital Zurich, Zurich, Switzerland, [9]Biozentrum, University of Basel, Basel, Switzerland, [10]Institute for Molecular Health Sciences, ETH Zurich, Zurich, Switzerland, [11]Division of Hematology, [12]Institute of Pathology and Molecular Pathology, University Hospital Zurich, Zurich, Switzerland, [13]Division of Molecular Neuroscience, Department of Psychology, [14]Transfaculty Research Platform Molecular and Cognitive Neurosciences, [15]Psychiatric University Clinics University of Basel, Basel, Switzerland and [16]Department Biozentrum, Life Sciences Training Facility, University of Basel, Basel, Switzerland

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Bonnie Berger

Received on March 30, 2017; revised on August 18, 2017; editorial decision on August 24, 2017; accepted on August 26, 2017

## Abstract

**Motivation:** Next-generation sequencing is now an established method in genomics, and massive amounts of sequencing data are being generated on a regular basis. Analysis of the sequencing data is typically performed by lab-specific in-house solutions, but the agreement of results from different facilities is often small. General standards for quality control, reproducibility and documentation are missing.
**Results:** We developed NGS-pipe, a flexible, transparent and easy-to-use framework for the design of pipelines to analyze whole-exome, whole-genome and transcriptome sequencing data. NGS-pipe facilitates the harmonization of genomic data analysis by supporting quality control, documentation, reproducibility, parallelization and easy adaptation to other NGS experiments.
**Availability and implementation:** https://github.com/cbg-ethz/NGS-pipe
**Contact:** niko.beerenwinkel@bsse.ethz.ch

## 1 Introduction

Advances in next-generation sequencing (NGS) have led to technologies capable of producing massive amounts of data at low costs. However, the analysis of these data is usually carried out using lab-specific in-house solutions. As a consequence, many different workflows are implemented for the same type of data, such that results are not easily

comparable and are often hard to reproduce. Several studies have shown that individual pipelines often have limited overlap in their results (Alioto *et al.*, 2015; Denroche *et al.*, 2015; Hofmann *et al.*, 2017), which impedes the potential of identifying true biological signals and of clinical applications. The developers of the Genome Analysis Toolkit attempt to stratify genome analysis by providing best practices (https://software.broadinstitute.org/gatk/best-practices/), but these recommendations are currently not fully implemented computationally.

Here, we introduce NGS-pipe, an automated and user friendly framework for the design of pipelines for the analysis of large-scale sequencing data, such as cancer genomics data. NGS-pipe allows to easily develop tailored workflows for the analysis of whole-exome (WES), whole-genome (WGS) and transcriptome (RNA-seq) sequencing data by providing building blocks to execute state-of-the-art tools, as well as appropriate error handling. An important goal of NGS-pipe is to overcome the common lack of automated procedures to ensure reproducibility. This is particularly important for clinical applications, where well documented and standardized protocols are a requirement (Aziz *et al.*, 2015).

## 2 Features of NGS-pipe

NGS-pipe incorporates tools for detecting single nucleotide variants (SNVs), insertions and deletions (indels) and copy number variants (CNVs), as well as for estimating gene expression levels. In addition to the primary read data analysis, NGS-pipe also generates runtime statistics and quality reports. It can be launched on a single computer or a cluster, where independent steps are executed in parallel. A practical introduction and examples can be found in the GitHub repository.

*Modularity*. NGS-pipe is implemented using the workflow management system Snakemake (Koster and Rahmann, 2012). In combination with a modular backbone, where the execution of each analysis step is controlled by a rule, NGS-pipe is a flexible, easily extendable and highly configurable framework for NGS analysis. By modifying a configuration file, users can easily adjust the parameters for each rule without changing its implementation and include or exclude complete analysis steps in order to adapt the pre-configured workflows to the specific needs of their own experiment.

*Workflows for WES, WGS and RNA-seq data*. To illustrate NGS-pipe, we have implemented and tested predefined workflows for the automated analysis for cancer WES, WGS and RNA-seq data (Fig. 1) to assist users inexperienced in data analysis or pipeline design. A description of these workflows, including the computational tools they integrate, can be found in the GitHub repository. Similar workflows can be implemented using NGS-pipe for other NGS applications.

*Quality control and statistics*. NGS-pipe supports quality control and provides statistics on each step of the analysis. Users can assess the quality of each sequencing file in the output of FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) or Qualimap2 (Okonechnikov *et al.*, 2016), and inspect basic statistics such as how many reads passed the individual analysis steps.

*Performance and scalability*. With NGS-pipe, samples can be analyzed independently of each other, providing full parallelization. For instance, we analyzed WES data from a tumor and matched normal sample comprising 60 million paired-end reads in 20 h and 10 such pairs in 22 h on a compute cluster [HPE ProLiant BL460c Gen9 – Two 12-core Intel Xeon E5-2680v3 processors (2.5–3.3 GHz)], where the two-hour overhead is due to waiting times of the local batch queuing system. Similarly, one RNA-seq dataset consisting of 80 million single-end reads and 10 such datasets were analyzed in 2.5 and 3 h, respectively.

*Reproducibility, documentation and error handling*. A high level of automation, a clear documentation of the pipeline and strict error handling facilitate reproducibility, a major goal of NGS-pipe. All parameters for all tools included in the analysis of an NGS experiment are documented in a configuration file. Using Snakemake functionality, there are several additional layers of documentation within NGS-pipe, e.g. logging the executed commands and generating graphical representations of the workflows. As NGS-pipe has been designed to analyze a large number of datasets in parallel, automatized error handling is a fundamental requirement. If one of the steps of the pipeline failed and produced incomplete or no results, the computation of all depending steps is halted and an error message is thrown, using Snakemake intrinsics. After the issue is resolved the pipeline independently resumes the analysis.

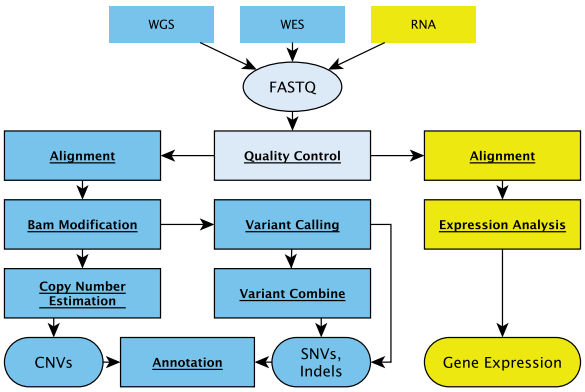


**Fig. 1.** Schematic overview of the different pre-configured pipelines available in NGS-pipe

## 3 Conclusion

NGS has become a standard genomics method in research labs and is currently implemented in clinical settings to aid patient diagnostics and treatment. NGS-pipe provides a Snakemake-based framework for analyzing such NGS data in a transparent and reproducible manner. The pre-configured workflows are easy to extend and adapt, extending the range of possible applications, including beyond cancer genomics.

## Funding

This work was supported by the European Research Council [ERC Synergy Grant No. 609883]; SystemsX.ch [RTD Grant 2013/150, IPhD Grant SXPHI0_142005 and SyBIT]; the Swiss Cancer League [KLS-2892-02-2012]; the Swiss National Science Foundation [Ambizione grant number PZ00P3_168165 to S.P.].

*Conflict of Interest*: none declared.

## References

Alioto,T.S. *et al.* (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.*, **6**, 10001.

Aziz,N. *et al.* (2015) College of american pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch. Pathol. Lab. Med.*, **139**, 481–493.

Denroche,R.E. *et al.* (2015) A cancer cell-line titration series for evaluating somatic classification. *BMC Res. Notes*, **8**, 823.

Hofmann,A.L. *et al.* (2017) Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. *BMC Bioinformatics*, **18**, 8.

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

Okonechnikov,K. *et al.* (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, **32**, 292–294.