Check for updates

SOFTWARE TOOL ARTICLE

Chanjo: Clinical grade sequence coverage analysis [version 1; peer review: 1 approved, 1 approved with reservations]

Robin Andeer¹, Måns Magnusson ²⁻⁴, Anna Wedell^{2,3}, Henrik Stranneheim^{2,3}

¹Science for Life Laboratory, Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, Stockholm, Sweden

²Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden

³Centre for Inherited Metabolic Diseases, Karolinska University Hospital, Stockholm, Sweden

⁴Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Stockholm, Sweden

V1 First published: 16 Jun 2020, 9:615
<https://doi.org/10.12688/f1000research.23605.1>

Latest published: 16 Jun 2020, 9:615
<https://doi.org/10.12688/f1000research.23605.1>

Abstract

Coverage analysis is essential when analysing massive parallel sequencing (MPS) data. The analysis indicates existence of false negatives or positives in a region of interest or poorly covered genomic regions. There are several tools that have excellent performance when doing coverage analysis on a few samples with predefined regions. However, there is no current tool for collecting samples over a longer period of time for aggregated coverage analysis of multiple samples or sequencing methods. Furthermore, current coverage analysis tools do not generate customized coverage reports or enable exploratory coverage analysis without extensive bioinformatic skill and access to the original alignment files. We present Chanjo, a user friendly coverage analysis tool for persistent storage of coverage data, that, accompanied with Chanjo Report, produces coverage reports that summarize coverage data for predefined regions in an elegant manner. Chanjo Report can produce both structured coverage reports and dynamic reports tailored to a subset of genomic regions, coverage cut-offs or samples. Chanjo stores data in an SQL database where thousands of samples can be added over time, which allows for aggregate queries to discover problematic regions. Chanjo is well tested, supports whole exome and genome sequencing, and follows common UNIX standards, allowing for easy integration into existing pipelines.

Chanjo is easy to install and operate, and provides a solution for persistent coverage analysis and clinical-grade reporting. It makes it easy to set up a local database and automate the addition of multiple samples and report generation. To our knowledge there is no other tool with matching capabilities. Chanjo handles the common file formats in genetics, such as BED and BAM, and makes it easy to produce PDF coverage reports that are highly valuable for individuals with limited bioinformatic expertise. We believe Chanjo to be a vital tool for clinicians and researchers performing MPS analysis.

Open Peer Review

Approval Status ? ✓

	1	2
version 1	?	✓
16 Jun 2020	view	view

1. **Ryan M. Layer**, University of Colorado
 Boulder, Boulder, USA
Michael Bradshaw , University of
 Colorado Boulder, Boulder, USA

2. **Ksenia Lavrichenko**, Oslo University
 Hospital, Oslo, Norway

Any reports and responses or comments on the article can be found at the end of the article.

Keywords

Genomics, QC, MPS, Coverage analysis, Clinical analysis

Corresponding author: Måns Magnusson (mans.magnusson@scilifelab.se)

Author roles: **Andeer R:** Methodology, Software, Validation, Visualization, Writing – Review & Editing; **Magnusson M:** Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Wedell A:** Conceptualization, Funding Acquisition, Investigation, Project Administration, Writing – Review & Editing; **Stranneheim H:** Conceptualization, Investigation, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The work was supported by the Swedish Research Council (2019-01154), the Karolinska Institutet, the Stockholm County Council (20170022), and the Knut & Alice Wallenberg Foundation (KAW 2014.0293).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Andeer R *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Andeer R, Magnusson M, Wedell A and Stranneheim H. **Chanjo: Clinical grade sequence coverage analysis [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2020, 9:615 <https://doi.org/10.12688/f1000research.23605.1>

First published: 16 Jun 2020, 9:615 <https://doi.org/10.12688/f1000research.23605.1>

Introduction

Compared to extensive serial Sanger sequencing, exome sequencing can be done at a small fraction of the cost per sample (the same order of magnitude as one average-sized gene) and whole exome sequencing (WES) has been more or less established in the clinic for a few years¹. Sequencing technologies are continuing to improve at a rapid pace. The Illumina HiSeq X system, and the more recent Illumina Novaseq 6000, reduce the cost of whole genome sequencing (WGS) to slightly more than \$1,000 for sequencing a human genome to 30x coverage². For the first time it is now possible to analyze complete human genomes within reasonable time and cost. This will further increase the pace of implementation of massively parallel sequencing (MPS) in new areas, such as diagnostics of inherited genetic disease. When analyzing the enormous data volume from WES and WGS, it is important to identify underrepresented genomic regions by calculating and tracking coverage quality control metrics. This is particularly true if the sequence data is used in diagnostics, since low or not exposed regions can lead to false positive or false negative results^{3,4}. There are a number of tools that provide basic coverage and overlap annotation functionalities: PicardTools⁵, BEDTools⁶, Sambamba⁷ and GATK⁸. These tools are excellent for comparing the overlap of two feature sets using single operations. However, they do not offer a solution for coverage analysis with different thresholds and different genomic and biological features. Moreover there are many laboratories around the world that work in a production setting where new samples are sequenced and analyzed every week. There is no present solution for persistent storage of coverage data that makes comparisons of hundreds or thousands of samples possible. This is essential to locate genomic regions that are hard to sequence and where the local sequencing pipeline gives insufficient information.

Furthermore, to our knowledge there are no tools that support dynamic report generation. To address these needs, we have developed **Chanjo**, a fast and flexible toolkit for seamless coverage analysis of genomic and biological features across multiple samples. Chanjo has been incorporated into the clinical analysis pipeline at Clinical Genomics Science for Life Laboratory and analysed more than 4500 rare disease WES and WGS samples to date. We believe Chanjo to be a vital tool for clinicians and researchers performing MPS analysis.

Methods

Implementation

Chanjo is written in Python (3.2+). It follows UNIX conventions and is built around text streams that can be incorporated into pipelines (Figure 1). Chanjo is distributed via GitHub, installation is simple and robust thanks to extensive tests. Chanjo loads output files from Sambamba depth⁷ and stores coverage related statistics, e.g. average coverage and completeness in a SQL database. Completeness is defined as the percentage of bases meeting a user-defined coverage threshold for each genomic interval (Figure 2). Chanjo does not aim to analyze the whole genome, but will limit the analyses to predefined genomic regions of interest defined in BED format.

Figure 1. Chanjo + Sambamba workflow.

Figure 2. Completeness description.

Supported features

Chanjo supports any genomic intervals as long as they adhere to the BED format with two optional columns for linking exons to transcripts and genes. Hence, it is easy to set-up independent databases using different gene and transcripts definitions, e.g. ccds, refseq or ensembl. When the genomic intervals are defined and added to the database, it is simple to add additional samples.

API

Chanjo uses a predefined database schema where exons are organized into transcripts and genes (Figure 3). Extracting basic coverage metrics such as “average coverage”, “overall completeness”, etc. for different transcripts and genes is easily done through the Python application programming interface (API). Setting up a database with the “init” subcommand only needs to be done once. After the basic structure of the database is in place, the user can add an arbitrary number of samples to the database and include or exclude samples, as preferred from the downstream analysis. The SQL schema has been designed to be a powerful tool on its own for studying coverage. It allows for quick aggregate metrics across multiple samples and can be used as a general coverage API for accompanying tools. One example of such a tool is Chanjo-Report, a clinical-grade coverage report generator for Chanjo output developed to be used in a clinical setting. The report can be tailored to include any number of samples, genes or transcripts (Figure 4). The report can be exported as a PDF and is assembled via a web interface using Chanjo-Report together with a Chanjo database. This works as a powerful bridge between bioinformaticians and professionals that work with analysis who often lack programming skills.

Operation

Chanjo requires a installation of Python version 3.2 or above. In a production setting, a MySQL database is preferred, however it is possible to use sqlite, which comes with the python installation. Chanjo has been installed and tested on Linux and Mac OS environments while there should not be any problems to install on Windows. Performance wise, a standard computer will be sufficient to run Chanjo.

Use cases

Workflow

It is straight forward to setup a working demo of Chanjo to test how to use the tool:

```
chanjo init --demo ./ chanjo -demo
for file in *. coverage .bed
do
    echo "${file}"
    chanjo load --group group1 "${file}"
done
chanjo calculate mean --pretty
```

Figure 3. Chanjo database schema.

Figure 4. Chanjo Report example.

The first command initializes a SQL database with tables, it will also use a reduced bed file to link exons, transcripts, and genes according to the definitions in "hgnc.min.bed". Chanjo uses output generated by "sambamba depth" which includes average coverage and completeness data for each exon defined in "hgnc.min.bed". The for loop loads data for 3 samples into the database for persistent storage. Finally, the CLI can be used to execute simple queries and output the results in JSON format.

```
chanjo calculate mean sample_1
```

```
{
  'metrics': {
    'completeness_10': 90.38,
    'completeness_20': 90.92,
    'mean_coverage': 193.85
  },
  'sample_id': 'sample_1'
}
```

Chanjo is intended to be used with a central database where samples are continuously added over time. This facilitates aggregate statistics, e.g. trending coverage metrics, poorly covered regions of the genome and comparing gene panel coverage across samples.

Chanjo report

In many settings the workflow for a sample starts in the lab where DNA is prepared and sequenced. After that bioinformaticians prepare the output for analysis and hand it over to a researcher or clinician. Chanjo-Report is

developed as a tool to present data generated by the bioinformaticians to the end-user. Here one can specify a subset of regions, in many cases a gene panel that is specific for a disease group, and get a well structured report of the fraction of transcripts/genes that are fully covered and which are not. This has proven essential when performing clinical tests with MPS data.

Summary

We have developed a novel tool, Chanjo, for continuous and accurate coverage analysis for multiple samples, ideal for WES as well as WGS. We believe Chanjo will be useful for sequencing facilities in general and clinical facilities in particular, where stringent quality control is required. The user only needs to initialize a database once by using a definition of regions and links, e.g. transcripts and genes, in the BED format. Samples are then added by loading data from sambamba depth into the database. Chanjo has been implemented to be easily included in an existing workflow of sequencing analysis. Furthermore, we introduce Chanjo Report to present coverage reports that are easy to generate and to interpret for non-bioinformaticians. This enables, e.g. estimation of accuracy of negative analyses, indicating regions that may require resequencing or investigation using alternative technology.

To our knowledge there are no software freely available today capable of continuous coverage analysis across multiple samples with dynamic report generation like Chanjo.

Software availability

1. Source code available from: <https://github.com/Clinical-Genomics/chanjo>
2. Archived source code as at time of publication: <http://doi.org/10.5281/zenodo.32664>⁹
3. Software license: MIT

Acknowledgements

Thanks to Valterri Wirta at Clinical Genomics for allowing this project to develop.

References

1. Yang Y, Muzny DM, Reid JG, *et al.*: **Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders.** *N Engl J Med.* 2013; **369**(16):1502–11.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Erika Check Hayden: **Technology: The \$1,000 genome.** *Nature News.* 2014; **507**(7492): 294.
[Publisher Full Text](#)
3. Brownstein CA, Beggs AH, Homer N, *et al.*: **An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge.** *Genome Biol.* 2014; **15**(3): R53.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Vrijenhoek T, Kraaijeveld K, Elferink M, *et al.*: **Next-generation sequencing-based genome diagnostics across clinical genetics centers: Implementation choices and their effects.** *Eur J Hum Genet.* 2015; **23**(9): 1142–1150.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Broad Institute: **Picard tools.**
[Reference Source](#)
6. Quinlan AR, Hall IM: **BEDTools: A flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Tarasov A, Vilella AJ, Cuppen E, *et al.*: **Sambamba: Fast processing of NGS alignment formats.** *Bioinformatics.* 2015; **31**(12): 2032–2034. ISSN 1367-4803.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. DePristo MA, Banks E, Poplin R, *et al.*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet.* 2011; **43**(5): 491–498.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Andeer R, Magnusson M, Kern J, *et al.*: **chanjo: Chanjo 3.0.0.** (Version v3.0.0) *Zenodo.* 2015.
<http://www.doi.org/10.5281/zenodo.32664>

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 24 March 2022

<https://doi.org/10.5256/f1000research.26048.r127509>

© 2022 Lavrichenko K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ksenia Lavrichenko

Department of Medical Genetics, Oslo University Hospital, Oslo, Norway

Andeer and colleagues present a novel software tool Chanjo that aims at DNA variant quality control via coverage analyse in the context of high-throughput sequencing in diagnostics. Streamlining of the software promises easy incorporation into genomic diagnostic pipelines and accessibility for non-bioinformatic personnel.

The paper is well-articulated and easy to follow overall. The software is very well documented and distributed via package repositories which makes it easy to adapt. I have a few minor comments, aimed at making the advantages of using the tool more transparent:

- The main one is that I am missing an actual use case and demonstration on why the tool is "highly valuable", even "vital", in the clinic. The current 'Use case' section belongs to the technical tutorial while in the paper I expect to see an actual example - perhaps one or a couple of genes for which among the 4500 cases (or respective reference samples) analysed the tool helped indicate a false positive or a false negative, or preferably both scenarios.
- In the aforementioned use case, it could be relevant (and highly recommended) to compare Chanjo performance with other mentioned "state of art" tools in the chosen loci.
- All the figures should be accessible outside of the context of the main text of the paper, therefore they need a caption with a respective description (and not just a figure title).
- Please, do add the annotation of what the colors mean (as graphical legend or in the caption of the figure) or drop the colors if they do not convey a specific meaning, to reduce information noise.
- If you do not mention the elements of a figure in the paper text, please remove them from the figure (see Figure 2, exon-intron)
- Sentence 6 in the Introduction "This is particularly true if the sequence data is used in diagnostics, since low or not exposed regions can lead to false positive or false negative

results": you mean "low-covered" regions? and "not exposed" as in "heterochromatin"?

- Could you explicitly mention some assumptions made or existing potential biases, and future improvements?
- Can this tool also be useful for newer technologies, e.g. long-read sequencing?

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, genomics, structural variants, short and long read sequencing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 20 September 2021

<https://doi.org/10.5256/f1000research.26048.r92794>

© 2021 Layer R et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ryan M. Layer

BioFrontiers Institute, University of Colorado Boulder, Boulder, CO, USA

Michael Bradshaw

BioFrontiers Institute, University of Colorado Boulder, Boulder, CO, USA

Chanjo is a tool that allows for the tracking and review of sequencing coverage statistics over time and across many samples and groups. While numerous tools exist to measure such statistics in a single batch, there are not any that enable tracking these over time, which can be very important for a diagnostics laboratory.

Commentary on the manuscript:

The manuscript is by and large well written and clear.

All figures are missing descriptions, these need to be added.

Figure 1:

- What do the shapes and colors mean?
- What does load do?
- What are you calculating?

Figure 2:

- Is this something that appears on one of the reports or just a figure for describing what completeness is?
- What do the different parts of the figures represent?
- What is the actual % completeness of this example?

Figure 3:

- This is the one figure that might be fine without a description, but it might be worth explaining for those not familiar with relational databases that there are 3 tables used in Chanjo (table1, table2, and table3) and that name1 references information found in table2 and table3.

Figure 4:

- How was this generated? Is it a stock report or a customized one? What is this screenshot of a report supposed to be showing me?

In the manuscript it is described how Chanjo has been extensively tested, on the GitHubs for Chanjo and Chanjo Report, they have some automated checks, but the most recent commit in both of these reports is causing the automated test to fail. It would be worth fixing whatever is causing these tools to fail their own test cases.

When you say “dynamic report generation” what exactly does this mean?

Consider rewording part of this sentence: “This is particularly true if the sequence data is used in diagnostics, since low or not exposed regions can lead to false-positive or false-negative results.”

Commentary on the documentation and tool:

- The tool is pip-installable, thank you, that is phenomenally user-friendly.

- The documentation of how to use the tool is in need of serious improvement. What is on the Github README vs what is on the <https://clinical-genomics.github.io/chanjo/> do not match.
- The documenting of this tool does not appear to be complete, or mostly complete. Documentation needs to be consistent and complete prior to sharing Chanjo publicly.
- There is a nice demo of how to use Chanjo but it doesn't fully explain what is being done or why.
- What does the link command do? Why do we need that?
- The file "chanjo.yaml" is required to use the tool, but is not a required parameter. It is mentioned that Chanjo will just look for that file in the current working directory if not otherwise specified, but Chanjo is installed as a globally accessible tool, to be used anywhere so this feature doesn't make great sense.
- "chanjo load", in the demo it is shown how to use the "--group" option, but in the introduction of the documentation, you show that this is not a required option in order to load a sample into Chanjo. But in order to use chanjo report a group is required. There seems to be a disconnect here in what is required for the functionality of the tools.
- Reading through the Github issue for Chanjo I learned there is a "chanjo remove" command. Which for me as well as the issue creator could have been very useful during the setup stages after accidentally adding numerous samples without groups. The remove command no longer appears to be a command in chanjo, but could be very useful. Samples cannot be re-loaded twice and cannot be removed so it seems the only option here is to delete the database, which seems like an extreme example.
- In the manuscript, it is described how Chanjo can be used with an SQL database and server or with SQLite. The demo shows how to use the SQLite version but I cannot find a description of how to use a full SQL database.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, computational genomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research