SOFTWARE TOOL ARTICLE

CoverView: a sequence quality evaluation tool for next generation sequencing data [version 1; referees: 2 approved]

Márton Münz¹, Shazia Mahamdallie¹,², Shawn Yost¹,², Andrew Rimmer¹, Emma Poyastro-Pearson¹,², Ann Strydom¹,², Sheila Seal¹,², Elise Ruark¹,², Nazneen Rahman¹-³

¹Division of Genetics & Epidemiology, The Institute of Cancer Research, London, SM2 5NG, UK
²TGLclinical, The Institute of Cancer Research, London, SM2 5NG, UK
³Cancer Genetics Unit, Royal Marsden NHS Foundation Trust, London, SM2 5PT, UK

Abstract

Quality assurance and quality control are essential for robust next generation sequencing (NGS). Here we present CoverView, a fast, flexible, user-friendly quality evaluation tool for NGS data. CoverView processes mapped sequencing reads and user-specified regions to report depth of coverage, base and mapping quality metrics with increasing levels of detail from a chromosome-level summary to per-base profiles. CoverView can flag regions that do not fulfil user-specified quality requirements, allowing suboptimal data to be systematically and automatically presented for review. It also provides an interactive graphical user interface (GUI) that can be opened in a web browser and allows intuitive exploration of results. We have integrated CoverView into our accredited clinical cancer predisposition gene testing laboratory that uses the TruSight Cancer Panel (TSCP). CoverView has been invaluable for optimisation and quality control of our testing pipeline, providing transparent, consistent quality metric information and automatic flagging of regions that fall below quality thresholds. We demonstrate this utility with TSCP data from the Genome in a Bottle reference sample, which CoverView analysed in 13 seconds. CoverView uses data routinely generated by NGS pipelines, reads standard input formats, and rapidly creates easy-to-parse output text (.txt) files that are customised by a simple configuration file. CoverView can therefore be easily integrated into any NGS pipeline. CoverView and detailed documentation for its use are freely available at github.com/RahmanTeamDevelopment/CoverView/releases and www.icr.ac.uk/CoverView

Keywords

NGS, Quality Sequencing Minimum, QSM, quality assurance, quality control, depth of coverage, base quality, mapping quality, genetic testing
Corresponding author: Nazneen Rahman (rahmanlab@icr.ac.uk)

Author roles: Münz M: Conceptualization, Software, Writing – Original Draft Preparation; Mahamdallie S: Conceptualization, Methodology, Writing – Review & Editing; Yost S: Conceptualization, Methodology, Writing – Review & Editing; Rimmer A: Conceptualization, Software; Poyastro-Pearson E: Methodology; Strydom A: Supervision; Seal S: Methodology, Supervision; Ruark E: Conceptualization, Validation; Rahman N: Conceptualization, Formal Analysis, Funding Acquisition, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: The work was supported by the Wellcome Trust [200990].

Copyright: © 2018 Münz M et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Münz M, Mahamdallie S, Yost S et al. CoverView: a sequence quality evaluation tool for next generation sequencing data [version 1; referees: 2 approved] Wellcome Open Research 2018, 3:36 (doi: 10.12688/wellcomeopenres.14306.1)

First published: 04 Apr 2018, 3:36 (doi: 10.12688/wellcomeopenres.14306.1)
Introduction

Next generation sequencing (NGS) has transformed genetic medicine by dramatically reducing the cost and time of genetic testing, which has led to a rapid global expansion in clinical genetic testing\(^1\). It is obviously crucial that these tests are accurate, with low false positive and false negative error rates. Robust quality assurance and quality control are required to minimise these errors and thus ensure reliable test results\(^2\).

Three primary metrics are used to evaluate sequence quality in NGS data: depth of coverage (how many sequence reads are present at a given position), base quality (have the correct bases been called in sequence reads) and mapping quality (have the reads been mapped to the correct position in the genome). False negative errors are often caused by insufficient depth of coverage\(^3\), and it is vital that regions with low coverage are flagged and reviewed, not least because they may require additional interrogation\(^4\). Coverage evaluation is also useful for comparing different NGS library generation strategies, to identify regions with suboptimal performance\(^5\) and for probe design optimisation\(^6\).

Poor base and poor mapping quality can cause false negative and false positive errors. Systematic base calling errors often occur in regions with high or low GC content or in homopolymer stretches\(^7\), while mapping errors often occur in regions of high homology (e.g. pseudogenes) that result in ambiguously mapped reads\(^8\). As a starting point for evaluating base and mapping quality, base callers and read mappers provide Phred quality scores that quantify base and mapping quality. These scores give the probability that a particular base has been identified incorrectly (base quality score, BQ\(^9\)) or a read has aligned to the wrong genomic position (mapping quality score, MQ\(^10\)).

Several tools and packages for analysing depth of coverage, base and mapping quality exist, for example GATK DepthOfCoverage and DiagnoseTargets\(^11\), QualiMap\(^12\), FastQC\(^13\), samtools depth\(^14\) and pysamstats\(^15\). These tools have different strengths, weaknesses and trade-offs, particularly in relation to flexibility and usability. Most available tools were developed for the research environment rather than the medical setting and have not focused on the needs and requirements of genetic testing in patients.

Here we present CoverView, a quality evaluation tool for NGS data that was designed to be user-friendly, fast, flexible and easy to integrate into NGS analysis pipelines. CoverView was developed to provide the quality assurance and quality control information required by clinical NGS testing laboratories, though we believe it is equally useful for research use. We recently proposed the Quality Sequencing Minimum (QSM) to deliver comprehensive, consistent, transparent NGS quality assurance information about depth of coverage, base and mapping quality, and we use CoverView to evaluate fulfilment of a QSM in our laboratory\(^16\). We also use CoverView as the quality control tool for all our research and clinical NGS analyses and it is integrated into our exome analysis tool, OpEx (Optimised Exome)\(^17\).

Methods

Implementation

CoverView is implemented as an easy-to-use tool that can process the read count, BQ and MQ of mapped sequencing reads. It reports a series of informative quality control (QC) metrics with increasing levels of detail from a chromosome-level summary to per-base profiles. It also flags regions that do not pass user-defined quality requirements. The tool is implemented in Python v.2.7.13 and Cython v.0.25.2, with its graphical user interface (GUI) developed in Flask v. 0.12.1, HTML5 v.5.1 and jQuery v.3.1.1.

CoverView requires a BAM file (containing the mapped reads) as its input with the corresponding .BAI file\(^18\). A BED file is also required with each record in the BED file defining the user-specified genomic region of interest (e.g. an exon of a gene) for which depth of coverage and sequence quality metrics will be reported. The BAM file may optionally contain reads marked as duplicates as CoverView can generate metrics with duplicate reads either included or excluded.

CoverView generates four output (.txt) files that provide different information about the quality of the input BAM dataset (for CoverView input files see Data and software availability section)\(^19\). At the most detailed level, per-base profiles of position specific metrics are reported for each region (Table 1). This base-level resolution is important because regions may only partially fail quality metrics, for example one part of an exon may have high quality depth of coverage whilst another part is poorly covered.

### Table 1. Position specific metrics reported as per-base profiles in the specified genomic regions.

<table>
<thead>
<tr>
<th>Position specific metrics</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage (COV)</td>
<td>The number of mapped reads covering the position</td>
</tr>
<tr>
<td>Median Base Quality (MEDBQ)</td>
<td>Median base quality of all read bases mapping to the position</td>
</tr>
<tr>
<td>Fraction of Low Base Quality (FLBQ)</td>
<td>Fraction of read bases mapping to the position with a base quality lower than a user-specified threshold</td>
</tr>
<tr>
<td>Median Mapping Quality (MEDMQ)</td>
<td>Median mapping quality of all reads covering the position</td>
</tr>
<tr>
<td>Fraction of Low Mapping Quality (FLMQ)</td>
<td>Fraction of reads covering the position with a mapping quality lower than a user-specified threshold</td>
</tr>
<tr>
<td>Quality Coverage (QCOV)</td>
<td>Number of mapped reads covering the position with read mapping quality and base quality higher than user-specified threshold</td>
</tr>
</tbody>
</table>
The coverage profile (COV) provides information on how the depth of coverage changes across the region, whilst the FLBQ and FLMQ metrics describe the fraction of coverage of low base or mapping quality scores at each position, respectively (Table 1). This is important because at positions with high FLBQ or FLMQ values it may not be possible to call variants with confidence even if the COV threshold is met. In addition, mapping and base quality scores are summarised by their median in the per-base MEDDBQ and MEDDMQ profiles. Finally, the QCOV profile integrates coverage, base and mapping quality information for each base by counting the number of covering reads that satisfy the user-defined quality requirements.

As systematic base calling errors can be strand-specific, QC metrics for forward and reverse reads separately can be of interest. CoverView can, optionally, output the described profiles calculated for forward (+) and reverse (-) reads only, facilitating detection of strand-specific biases in coverage or base and mapping quality.

Summary statistics derived from the per-base profiles are reported for each user-specified region to describe the overall quality of the region (Table 2). MEDCOV and MEDQCOV summarise the coverage Profiles by their median across positions and MINCOV and MINQCOV provide information about the least covered position. Finally, the MAXFLMQ and MAXFBLQ metrics describe the lowest sequence quality positions in the region. CoverView can also output region-level metrics calculated for forward and reverse reads separately.

Users can define the minimal requirements to ‘pass’ the quality test and if this is not met the region of interest is ‘flagged’. Defining these minimum requirements for depth of coverage, base and mapping quality are the basis of the QSM that is described in detail in the accompanying paper. In CoverView a minimum or maximum value can be specified for any of the metrics in Table 2. For example, users may set a lower threshold to MINCOV and an upper threshold to MAXFLBQ; i.e. a region will be tagged with “FLAG” if coverage is below the defined value or FLBQ exceeds the maximum value at any position within the region, otherwise the region will be tagged as “PASS”. A separate column in the output file indicates the pass/flag status of each region.

For correct variant annotation, interpretation and follow-up it is essential to know which gene transcripts a flagged region overlap with. A simple script (ensembl.db) is included for creating a transcript database using any Ensembl release. In a separate output file, CoverView optionally reports the transcript coordinates (CSN coordinates) of all genomic intervals that are covered by less than 15 high quality reads (QCOV<15), which are referred to as “poor quality intervals”. Users may wish to review positive or negative calls within such regions, to ensure confidence in the calls.

CoverView also provides a chromosome-level summary. This is important because the input BAM file may contain unmapped reads or reads that are mapped outside the targeted regions. In order to quantify the fraction of sequencing data that are not useful for variant calling within regions in the BED file, CoverView creates a chromosome-level summary that reports the total read counts (RC) and the read counts of on-target (RCIN) and off-target regions (RCOUT) for each chromosome. In addition, the outputted table includes the mapped, unmapped and total read counts calculated for the whole genome.

Finally, CoverView provides a GUI with multiple views (‘Analysis View’, ‘Genes View’, ‘Regions View’, ‘Profiles View’), that allow users to intuitively explore the results. The Analysis View shows metadata of the analysis such as the names of input files and key configuration options. The Genes View offers a clickable bar plot displaying per-chromosome read counts (both on-target and off-target) and a filterable table providing information on the number of flagged regions for each gene on the selected chromosome (Figure 1). The Regions View displays a scrollable, searchable and filterable table of region-specific metrics values (Table 2) with flagged regions and the metric(s) underlying the flag highlighted (Figure 2). The Profiles View provides an interactive table and graph of per-base quality profiles for a selected region aligned with the corresponding reference genome sequence. Users can change the metrics displayed, zoom in, or navigate along the sequence with

<table>
<thead>
<tr>
<th>Table 2. Summary metrics for the specified genomic regions.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary metrics of region</strong></td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td>Read count (RC)</td>
</tr>
<tr>
<td>Median coverage (MEDCOV)</td>
</tr>
<tr>
<td>Minimum coverage (MINCOV)</td>
</tr>
<tr>
<td>Median quality coverage (MEDQCOV)</td>
</tr>
<tr>
<td>Minimum quality coverage (MINQCOV)</td>
</tr>
<tr>
<td>Maximum fraction of low mapping quality (MAXFLMQ)</td>
</tr>
<tr>
<td>Maximum fraction of low base quality (MAXFBLQ)</td>
</tr>
</tbody>
</table>
Figure 1. CoverView GUI Genes View for TSCP analysis in GIAB sample. CoverView GUI screenshot of the Genes View for TruSight Cancer Panel (TSCP) data generated for the Genome in a Bottle (GIAB) sample shows a bar plot of per-chromosome read counts and the list of flagged regions. Nine regions in eight genes were flagged for falling below MINQCOV $\geq 50$ in the GIAB sample.

Figure 2. CoverView GUI Regions View for PMS2 in GIAB sample. CoverView GUI screenshot of Regions View for PMS2 data in the GIAB sample showing the summary metrics per region. Metrics that fall below user-defined thresholds are shown in red and flagged. The MINQCOV for PMS2 exons 12 and 15 are below MINQCOV $\geq 50$. 
Quality minimum thresholds overlaid as dashed horizontal lines (Figure 3, Figure 4). The GUI is a Flask application that runs in the web browser on port 5000.

**Operation**

CoverView can be installed by running a simple Bash script. Installation requires Python v.2.7.9 or later (Python2 series), GCC v.4.2.1, GNU make v.3.81 and virtualenv v.15.1.0.5. Note that Python v.2.7.9 and later include pip by default. Additional dependencies (e.g. the Python module pysam) are automatically downloaded and set up in an isolated Python virtual environment by the installation script.

Once installed, the tool can be customised with a simple INI configuration file and run from Linux/Unix command line (see Documentation for details). CoverView can be easily integrated into NGS analysis pipelines, as shown for the OpEx (Optimised Exome) pipeline. The CoverView documentation is hosted on GitHub Pages at https://rahmanteamdevelopment.github.io/CoverView/

CoverView v.1.4.3 documentation is also included in Supplementary File 1.

**Use case**

We use the TruSight Cancer Panel (TSCP) for testing cancer predisposition genes in both research and clinical settings. We generated TSCP data on the National Institute of Standards and Technology (NIST) Reference Material (RM) 8398, for which there are experimentally validated genotyped sites provided by the NIST-led Genome in a Bottle (GIAB) Consortium. We mapped paired-end Illumina sequencing reads to the human reference genome (GRCh37) using Stampy v.1.0.20 with BWA v.0.7.5a for pre-mapping. Duplicate reads were marked with Picard v.1.90. The resulting BAM file was analysed with CoverView v.1.4.3 with a BED file specifying the TSCP targeted regions. Duplicate reads were included in the analysis. In this use case we set the minimal requirement for a region to ‘PASS’ as MINQCOV ≥ 50: i.e. all base positions in a targeted region were required to be covered by at least 50 reads of good quality (MQ ≥ 20 and BQ ≥10). The CoverView analysis was completed in 13 seconds (for CoverView Output files see Data and software availability section).

Nine of the 1471 targeted regions, affecting eight genes, were flagged as not fulfilling the MINQCOV ≥50 requirement (Figure 1). Here we focus on PMS2 exon 12 as an example of

![Figure 3. CoverView GUI Profiles View showing PMS2 exon 12 COV and QCOV data in GIAB sample.](image)
Figure 4. CoverView GUI Profiles View showing PMS2 exon 12 QCOV and FLMQ data in GIAB sample. CoverView GUI screenshot of Profiles View for PMS2 exon 12 data in the GIAB sample showing the quality coverage QCOV (blue), fraction of read bases with low mapping quality FLMQ (red) and the minimum QCOV threshold as dashed horizontal lines, across the region. This shows that the majority of reads mapping to this region have low mapping quality.

The Regions View provides region-level metrics values for PMS2 exon 12 (Figure 2). The table shows that 3404 reads were mapped to this region. Although the least covered base has 173 reads (MINCOV=173), one part of the region was covered with only 30 reads of good quality (MINQCOV=30). The MINQCOV column is therefore highlighted in red, indicating that its value is below the pre-set quality requirement. The table also suggests that the large difference between MINCOV and MINQCOV is due to poor mapping quality because the fraction of low mapping quality reads in the region is very high (MAXFLMQ=0.9; at a given position 90% of reads did not fulfill MQ ≥ 20). This is further supported by the Profiles View which shows the COV and QCOV metrics together (Figure 3). These two depth of coverage profiles along the entire exon are substantially different, and the FLMQ profile shows that the mapping quality is consistently poor (FLMQ>0.8) along the whole exon (Figure 4). This explains the striking difference between the COV and QCOV profiles since low quality reads are not counted as part of quality depth of coverage. PMS2 has a nearby pseudogene with strong homology to exons 9, 11–15 that causes ambiguous mapping and it is not possible to robustly analyse exon 12 by TSCP data alone. However, the CoverView outputs show that every base in 1462/1471 (99%) TSCP regions in the GIAB sample pass the MINQCOV ≥50 quality threshold.

Conclusion

Next generation sequencing data are error-prone, subject to random errors affecting individual samples and systematic errors, due to sequence contexts and biases of sequencing platforms, affecting many samples. Stringent, comprehensive quality management is therefore essential when using NGS for clinical applications. CoverView is a freely available NGS quality evaluation tool that provides quality metrics at the highest possible resolution by outputting per-base profiles, together with informative summary metrics that highlight which areas require further review. The CoverView outputs can be integrated into NGS pipelines so that regions that pass user-defined thresholds can automatically proceed and regions flagged as falling below user-defined thresholds can be further evaluated. The CoverView GUI provides a simple, visual interface with which to explore CoverView outputs and to investigate flagged regions.

We developed CoverView to be easy to install and use and we believe it can be quickly integrated into any NGS pipelines. CoverView is now the quality evaluation tool we use for all our clinical and research NGS analyses.

Data and software availability

CoverView input and output files for TSCP analysis in GIAB sample are available at: http://doi.org/10.17605/OSF.IO/87K6S
Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

CoverView is available at: github.com/RahmanTeamDevelopment/CoverView/releases and www.icr.ac.uk/CoverView

CoverView documentation is available at: https://rahamanteamdevelopment.github.io/CoverView/

Latest source code: https://github.com/RahmanTeamDevelopment/CoverView

Archived source code as at time of publication: http://doi.org/10.5281/zenodo.1206100

Software license: MIT

Supplementary material

Supplementary File 1. CoverView v.1.4.3 documentation.

Click here to access the data.

References


22. virtualenv: Virtual Python Environment builder. Reference Source
Reference Source

24. Illumina: TruSight Cancer: targeting genes previously linked to a predisposition towards cancer.
Reference Source

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

30. Picard.
Reference Source

Data Source
Open Peer Review

Current Referee Status: ✔ ✔

Somak Roy
Department of Pathology, Division of Molecular & Genomic Pathology, University of Pittsburgh, Pittsburgh, PA, USA

The authors have described their software solution, CoverView, to analytically and visually profile Next Generation Sequencing (NGS) data for use in a clinical environment. Specifically, the manuscript describes quality control parameters derived from sequencing depth, base quality, and mapping quality score.

The described software solution appears to have great utility as a pluggable tool in the clinical workflow of QC review of NGS data. The graphical user interface (GUI) appears to be intuitive and user friendly for use by laboratory personnel who may not have experience with command line and other programming tools. This tool can potentially be a useful resource for validation of NGS assays where understanding the coverage and sequencing quality profile across sequenced region is important.

I summarize my comments below that should help further improve the content of the manuscript.

1. The authors have described sequencing depth, base quality, and mapping quality scores as the primary measurements for CoverView. However, the drafted manuscript reads such that these three parameters are the only indices to be evaluated for quality. Since there are additional parameters from sequenced reads that reflect quality of NGS data, the authors should list these under discussion section of the manuscript for the readers to have a holistic view of NGS QC.

2. Scalability: the manuscript demonstrates the application performance (computational time, visual rendering, and user interaction) based on using data from targeted panel. Parsing and analyzing data from BAM files can be computationally expensive, especially as the sequenced region increases (targeted panels with several hundred genes, exomes, etc). The authors reference to integrating CoverView in their clinical exome sequencing pipeline (OpEx). May I request the authors to provide benchmark for computational time, visual rendering, and user interaction comparing OpEx and TSCP to demonstrate the scalability of the software solution. It will be interesting to know how the rendered user interface adapts to large array of data points while preserving user experience.

3. Input BAM file and multiple aligners: Can the authors discuss their experience with regards to the input BAM file when generated by multiple sequence alignment algorithms? In the use case, the authors describe using Stampy and BWA. I am assuming that each of the BAM files were separately analyzed by CoverView. Did the authors face any challenges when using BAM files from different alignment algorithms?
4. Quality Sequencing Minimum (QSM): The authors reference QSM as being used with CoverView in their laboratory. It will be helpful if some description of QSM is included in the manuscript.

References

*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?*
Yes

**Competing Interests:** No competing interests were disclosed.

**Referee Expertise:** Molecular diagnostics, cancer genomics, bioinformatics, clinical informatics.

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

doi:10.21956/wellcomeopenres.15567.r32740

Richard Bagnall
Agnes Ginges Centre for Molecular Cardiology, Centenary Institute, Sydney Medical School, The University of Sydney, Sydney, NSW, Australia

Coverview is presented as a computational tool to assess the amount, and quality, of aligned next generation sequencing data reads at user-defined regions. Additionally, a data visualisation tool that runs in a web-browser window displays the results and allows the user to navigate the data and investigate regions that are flagged as failing a quality metric. Coverview is designed with clinical laboratories in
mind, where the assurance that target regions have been adequately sequenced is vital. The tool and underlying computer code is made freely available, and a detailed user manual is provided.

A number of tools and approaches are already available to measure and visualise the amount and quality of NGS read data, as these are important factors in determining the reliability, sensitivity and specificity of variant calling. However, Coverview brings the analysis and visualisation of results together in a single package, requires very basic knowledge of the command line, and should be a useful addition.

The manuscript is clear and presents the features of Coverview, with presentation of a use case using a validated gene panel (TruSight Cancer Panel). This use case shows how high depth sequencing of an established gene panel yields only a few regions of low coverage for interrogation. Although, in principal, Coverview could be used to investigate an exome, in practice the number of flagged regions may become unwieldy.

General comments:

A description of the QSM would be helpful, as most people will not know what this is.

The number of bases below QCOV for each region on the ‘Regions View’ is not displayed, but could help the user prioritise regions for further interrogation.

Thank you for introducing CoverView.

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?
Yes

Competing Interests: No competing interests were disclosed.

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