UPPSALA
UNIVERSITET

UPTEC X 18 034
Examensarbete 30 hp
Februari 2019

# Coverage Analysis in Clinical Next-Generation Sequencing

Anna Louise
Odelgard

UPPSALA
UNIVERSITET

**Teknisk- naturvetenskaplig fakultet**
**UTH-enheten**

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

# Abstract

## Coverage Analysis in Clinical Next-Generation Sequencing

*Anna Louise Odelgard*

With the new way of sequencing by NGS new tools had to be developed to be able to work with new data formats and to handle the larger data sizes compared to the previous techniques but also to check the accuracy of the data. Coverage analysis is one important quality control for NGS data, the coverage indicates how many times each base pair has been sequenced and thus how trustworthy each base call is. For clinical purposes every base of interest must be quality controlled as one wrong base call could affect the patient negatively. The softwares used for coverage analysis with enough accuracy and detail for clinical applications are sparse. Several softwares like Samtools, are able to calculate coverage values but does not further process this information in a useful way to produce a QC report of each base pair of interest. My master thesis has therefore been to create a new coverage analysis report tool, named CAR tool, that extract the coverage values from Samtools and further uses this data to produce a report consisting of tables, lists and figures. CAR tool is created to replace the currently used tool, ExCID, at the Clinical Genomics facility at SciLifeLab in Uppsala and was developed to meet the needs of the bioinformaticians and clinicians. CAR tool is written in python and launched from a terminal window. The main function of the tool is to display coverage breath values for each region of interest and to extract all sub regions below a chosen coverage depth threshold. The low coverage regions are then reported together with region name, start and stop positions, length and mean coverage value. To make the tool useful to as many as possible several settings are possible by entering different flags when calling the tool. Such settings can be to generate pie charts of each region's coverage values, filtering of the read and bases by quality or write your own entry that will be used for the coverage calculation by Samtools. The tool has been proved to find these low coverage regions very well. Most low regions found are also found by ExCID, the currently used tool, some differences did however occur and every such region was verified by IGV. The coverage values shown in IGV coincided with those found by CAR tool. CAR tool is written to find all low coverage regions even if they are only one base pair long, while ExCID instead seem to generate larger low regions not taking very short low regions into account. To read more about the functions and how to use CAR tool I refer to User instructions in the appendix and on GitHub at the repository anod6351

Handledare: Claes Ladenvall
Ämnesgranskare: Adam Ameur
Examinator: Jan Andersson
ISSN: 1401-2138, UPTEC BIO18034

# Populärvetenskaplig sammanfattning

Upptäckten av variation i DNA har revolutionerat både diagnostisering och behandling av sjukdomar som beror på förändringar i arvsmassan. Cancer är en sådan sjukdom som orsakas av genetiska förändringar så kallade mutationer vilket leder till okontrollerad celltillväxt. I takt med att bättre analysmetoder utvecklats kan man idag ofta analysera vilka mutationer som lett till sjukdomen. Det i sin tur möjliggör riktade diagnoser och patientanpassade behandlingar. (Schmidt *et al.* 2016) Exempelvis är vissa läkemedel enbart verksamma för specifika mutationer, i andra fall finns det läkemedel som är direkt farliga att ta om man har en viss genuppsättning. (Schmidt *et al.* 2016) Detta gör att patientanpassade behandlingar ofta är mer effektiva och det går att undvika onödiga behandlingar och bieffekter. Rätt behandling kan även spara tid och resurser som annars skulle ha lagts på mindre effektiva behandlingar.

Den genetiska analysen börjar med att blod- eller biopsiprov skickas till sjukhusets kliniker. Från vävnaden renar man fram DNA. Det processas sedan i maskiner som läser av individens DNA, så kallad sekvensering. Resultaten från sekvenseringen sparas i filer som innehåller flera korta sekvenskopior av originalprovet. (Nikiforova *et al.* 2013) I filerna kodas beståndsdelarna i DNA:t om från kemiska strukturer till bokstäverna A, T, G och C. Filerna behöver sedan processas och analyseras innan data kan tolkas. (Nikiforova *et al.* 2013) När sjukhusgenetiker sedan har tolkat vilken medicinsk betydelse de funna genetiska varianter har så skickas det vidare till behandlande läkare som diagnostiserar patienten. Se figur 1 för förtydligande av händelseförloppet.

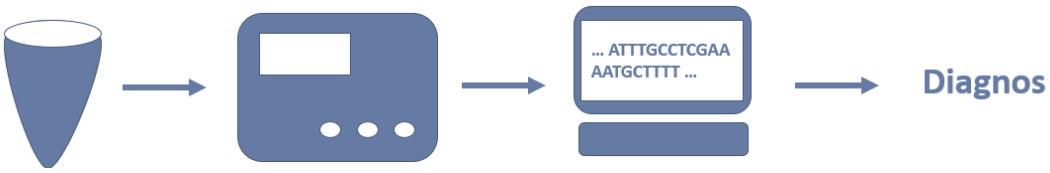



**Fig. 1** Övergripande flödesschema av diagnostiseringsförloppet av genetiska sjukdomar. DNA-provet sekvenseras med en sekvenseringsmaskin, baserna i DNA:t omvandlas till bokstäver som sparas i filer på en dator. Denna information används sedan som underlag för en diagnos.

Sekvenseringen blir inte alltid helt rätt för alla baspar därför är det viktigt att validera de data som genereras. För kliniska applikationer räcker det med endast ett felaktig avläst baspar för att ge fel diagnos och då även fel behandling. För att undvika misstag genomförs flera olika kvalitetskontroller, inklusive en täckningsanalys. (Merino *et al.* 2017) En av den kliniska täckningsanalysens viktigaste syften är att säkerställa att alla genetiska regioner som ska undersökas verkligen har blivit tillräckligt bra sekvenserade. Det vill säga om tillräckligt med data har erhållits för att kunna anta att avläsningen av DNA:t blivit korrekt. För tillfället är det ont om verktyg som genererar en täckningsanalytisk rapport för klinisk användning och att skapa detta verktyg har varit mitt examensarbete.

Det verktyg jag utvecklat tar in avläst data från sekvenseringen och en fil med en lista över de genetiska områden som ska kontrolleras. Vill man diagnostisera cancer så har man till exempel oftast en sjukdomsspecifik genpanel (lista över genetiska områden där cancerspecifika förändringar ofta förekommer). (Soukupova *et al.* 2018) Verktyget beräknar sedan täckningsdjup och täckningsbredd över dessa områden. Täckningsdjup är hur många sekvenserade kopior som täcker ett område av intresse i det mänskliga DNA:t och täckningsbredd är andelen basparpositioner som sekvenserats

över ett visst täckningsdjup. Resultaten presenteras som en rapport innehållandes tabeller, listor och bilder. Rapporten kommer främst att behandlas manuellt så det är viktigt att resultatet presenteras övergripande, men mjukvaran kommer även att användas som en del i automatiserade analyser. Verktyget har många valbara inställningar för att kunna tillämpas på genpaneler och data med olika behov. Flexibiliteten bidrar till möjligheten att skapa anpassade rapporter för användarens behov samtidigt som det sparar tid och lagringsutrymme då bara de nödvändiga funktionerna körs. Verktyget finns tillgängligt för alla som är intresserade på GitHub. Programmet ligger på kontot anod6351 i mappen CARtool.

# Table of Contents

# 1. Acronyms

| | |
|---|---|
| BAM | Binary Alignment/Mapping |
| BED | Browser Extensible Data |
| CAR | Coverage Analysis Report |
| DNA | DeoxyriboNucleic Acid |
| IGV | Integrative Genomics Viewer |
| IT | Information Technology |
| NGS | Next Generation Sequencing |
| PCR | Polymerase Chain Reaction |
| ROI | Region of Interest |
| SAM | Sequence Alignment Map |
| SS | Sanger Sequencing |
| QC | Quality Control |
| WES | Whole Exome Sequencing |
| WGS | Whole Genome Sequencing |

# 3. Introduction

A major breakthrough in healthcare is the application of knowledge on how variation in our genetic material may lead to disease and disability. The implementation is made possible by novel genetic technologies and IT solutions that promises better diagnosis and treatment of diseases that are caused by somatic or germline mutations. Using information on disease related mutations personalized treatments has become a possible option. Personalized treatments are better for the patient by treating the actual cause of the disease and excluding redundant or less effective therapies. (Schmidt *et al.* 2016)

The process of diagnosing using genetic material starts with a blood or a tissue sample from the patient. From this sample DNA is extracted and sequenced. To save time, cost and effort, sometimes only certain regions of interest (ROI) with respect to the disease are sequenced and examined for disease related mutations. This design option is commonly referred to as a gene panel. (Soukupova *et al.* 2018) In some cancers for example, certain genes are known to be involved in the disease development in many patients. By only sequencing those genes cost can be reduced while still being able to detect small cancer clones at good resolution. Just one mutation could make all the difference during diagnosis hence the accuracy of the data is vital. Quality control procedures in laboratories are in place to ascertain that the generated DNA sequence is equivalent to the individual's actual genetic material and that the data is of high quality. For Next Generation Sequencing methods an important quality control step is the coverage analysis. (Merino *et al.* 2017) A high coverage value over a position in the DNA sequence indicates a high probability of the position being sequenced correctly while a lower than expected value indicates missing information or an uncertainty in quality. What a high versus low value is depends on the methods used and what is studied.

The coverage analysis report is generated by a bioinformatic tool. Currently the number of available tools for clinical usage is limited and this is where my master thesis comes in. I have created CAR tool, a coverage analysis tool created for clinical usage. It produces a coverage analysis report from the NGS data. In comparison, the ExCID (ExCID, 2019) coverage tool, currently in use at the Clinical Genomics Uppsala facility, works, but is difficult to develop and improve due to long and very sparsely commented code. It also has several functions that are not used and is not maintained any longer by the developers. Creating the CAR tool enables the facility to tailor the tool for the needs of the clinicians and the bioinformaticians, making it easier to fit the analysis pipelines.

# 4. Background and theory

## 4.1 DNA sequencing

There are several sequencing technologies available today. It all started with the first-generation sequencing method called Sanger sequencing (SS) which was developed in the 1970s. (Sanger *et al.* 1977) Sanger sequencing is based on the generation of randomly terminated copies of the DNA sample. At the end of each copied sequence a color labeled nucleotide blocks further elongation. Sequences are then separated by size using gel electrophoresis and run through a detector reading the DNA sequence. Sanger sequencing was used for the human genome project, during this project it became evident that cheaper and faster methods would be needed if this technique was to be used. The request by the National Human Genome Research Institute resulted in Next Generation Sequencing (NGS) techniques. (van Dijk *et al.* 2014) The faster and cheaper means of sequencing by

NGS produces several more and much shorter reads (sequenced copies of the DNA sample) than SS. The greater number of reads and the shorter read length put pressure on the need for constructing bioinformatic tools and pipelines that would be able to manipulate and handle the larger sets of data. (van Dijk *et al.* 2014)

### 4.1.2 What to sequence

The large data sizes can, to some extent, be reduced by choosing what regions in the DNA to sequence. For example, it is not always necessary to sequence an individual's entire genetic material, referred to as whole genome sequencing (WGS). The genome (the entire set of an individual's DNA) largely contain regions that are not expressed. The small number of regions that are expressed encodes for proteins that are essential for our function. These regions can in turn be further divided into exons and introns, where only exons encode for the protein and the introns are removed from the final mRNA that is used for constructing the protein. Because exonic information is used to produce proteins there is much interest in only sequencing the exons, called whole exome sequencing (WES). Compared with WGS it is much faster and cheaper since less data is generated and analyzed. The exome is in fact only ~2% of the genome. One could however go even further and limit the sequencing to targeted panels, a collection of specific genomic regions of interest, ROI.(Koboldt *et al.* 2013) This approach is vastly used in the clinical setting for diagnosis of inherited disease and cancer. The design of these panels is flexible, most often they are chosen to consist of regions in the genome that is known to contain somatic or germline mutations causing a certain disease.

### 4.1.3 Short vs. long read sequencing

Major drawbacks with SS are the high cost and the small sequencing throughput. To overcome these problems NGS emerged, a faster and cheaper means of sequencing that resulted in shorter read lengths. Illumina is a popular short read sequencer with high throughput, low error rate and fast turn-around between 4 hours and 3.5 days. The short reads are just a couple of hundred base pairs long while another popular sequencer, PacBio, generates reads more than 10 times longer in just a few hours (Ardui *et al.* 2018). However, the shorter reads tend to be more accurate as well as cheaper compared to long-read sequencing.(Ardui *et al.* 2018) Bioinformatically, longer reads are easier to map correctly to a reference than shorter reads and are able to more accurately detect larger genetic variants. Also, the type of error during sequencing effects the coverage, random errors in reads generated by PacBio don't need as high coverage to remove as the biased errors from most short read techniques has. Another factor that effect the coverage needed is how the reads are generated and the libraries constructed, more on this in the section below.

### 4.1.4 Amplicon vs capture libraries

Several different manufacturers and chemistries are available for NGS sequencing. The first step after DNA isolation consist of preparing the DNA for sequencing, i.e creating a sequencing library. The library can be constructed using two different main techniques. An amplicon library consists of reads that have been generated by target PCR primers, or similar techniques, while a capture library instead has probes that capture randomly fragmented DNA. Regions of interest are then amplified for both techniques. According to a comparison study by Samorodnitsky and colleagues, one advantage of the capture libraries proved to be a more uniform distribution of the reads compared to the amplicon libraries. Also, some variants were not found with the amplicon library and the authors' explanation for this is among others low coverage. The amplicon libraries proved in turn to have better on target rates.(Samorodnitsky *et al.* 2015)

## 4.2 Bioinformatics

Bioinformatics is the field of biological data analysis. A vast amount of data from biological/medical samples is constantly generated that needs storage and most often also further processing before analysis. For instance, the raw sequence data produced during NGS needs to be further processed in several steps in a bioinformatic pipeline. The reads need to be QC'd, trimmed and aligned against a reference genome. Finally, genetic variants are called, annotated and filtered. In all steps the quality and probability of the data being correct needs to be assessed. In the clinical setting these bioinformatic steps are often followed by annotation and further filtering against population frequency databases, disease specific databases etc. Eventually the variants are interpreted by clinical staff and a clinical report is written to the referring physician with information on what analysis that has been performed, what variants that have been detected and what information is available about how the detected variants influence disease development.

### 4.2.3 File formats

There are several types of file formats to keep in mind when dealing with genetic data. The ones I have used during this project are BAM files and BED files. The BAM files are compressed binary files that contains information about the reads from the sequencing. (Li *et al.* 2009)To extract the data of the reads within specific regions in the DNA I have used gene panels stored in BED files to describe what regions are of interest. The regions in the BED file are specified by tab separated rows containing chromosome, start position and end position. Additional information about the region such as region name and score etc. could be added in the next columns but is not mandatory. (Zhang 2016) Important to know when dealing with BED files is that the standard for start position indices is 0-based while the end position is 1-based. This means that if the first base in a region would have the start value 0 then the end value is 1.

### 4.2.4 Coverage – NGS data quality measurement

Coverage is a quality measurement used for NGS data. There are two types of coverage measurements, coverage depth and breadth. Coverage depth is measured over a region or position at which the number of covering reads are counted. A coverage depth of 10X at one base position means that 10 sequenced reads cover that base. Coverage breadth is measured over a region and is the fraction of positions having XX or higher coverage depth (fig 2).

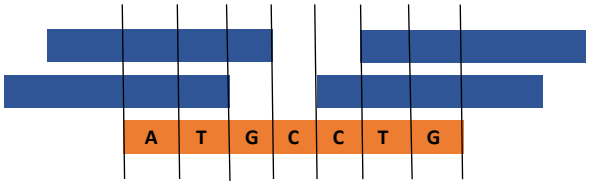



**Fig. 2** Coverage: A region of interest in the genome is colored orange and the sequenced reads in blue. Coverage depth over the base pair A is 2X. And the coverage breadth over the orange region A,T,G,C,C,T,G at 2X is 57% as 4 out of 7 base pairs are covered by 2 reads or more.

The limit on sufficient coverage depth and breadth values varies depending on what type of analysis is performed and what type of data that has been generated. For example, the coverage depth generally needs to be over a couple of hundreds for a cancer sample while it may be sufficient to go as low as 10X for inherited diseases. For cancer the fraction of cells in a sample that harbor decease causing mutations vary. This may be due to sample purity or to the presence of subclones and understand the needs for a high sequence coverage. Available tools for computing coverage are for example Bedtools and Samtools (Li *et al.* 2009, Quinlan 2014). These tools are fast, well used and provide useful coverage computations but lack a final processing step that generates a comprehensive coverage report, a key component in the clinical QC report.

### 4.2.5 Quality scores

During sequencing a detector registers the different bases of the reads. These signals must then be converted into bases by a base call software. To assess the quality of these base calls it is common to also generate phred scores. The phred score is a probability measurement of the base call being incorrect, it is defined as $-10\log_{10}(P)$ with P being the estimated error probability of the base call. (Ewing & Green 1998) . The quality score keeps several errors during detection in mind such as peak resolution, spacing of peaks as well as the ratio of peaks called and those uncalled. (Ewing & Green 1998)

After base call an alignment software is used to map the reads to a reference. The reads could be mapped to the wrong position by several reasons such as a high similarity to multiple regions, due to errors during sequencing or base call for example. MapQ is used to estimate the quality of the alignment, it is defined as $-10\log_{10}\Pr(\text{Read is wrongly mapped})$ (Li *et al.* 2008). Each mapQ score indicates how likely it is that the alignment is incorrect, a low mapQ score means a high probability of the alignment being incorrect and thus an alignment of low quality.

### 4.2.6 Tools for coverage calculations – samtools

Samtools is a software that can be used to extract information and manipulate alignment files in the formats SAM, BAM and CRAM. (Li *et al.* 2009) The tool has a very time efficient algorithm which makes it desirable to use by my tool. CAR tool uses Samtools depth function to calculate the coverage depth values for each base pair. This data is then further processed by CAR tool. Using Samtools it is also easy to filter the reads by quality and strand which are requirements for my tool. The coverage values generated by Samtools are saved to a file with one base coverage depth value per row. This file is further processed by CAR tool to generate an analytical report, for example by expressing coverage values per ROI:s and to create a short list of coverage depth values below a certain threshold value along with other useful information such as region name and transcript etc.

### 4.2.7 Containers

Moving a collection of scripts from one system to another while making sure all dependencies and the right versions of these are installed can be quite tricky. Containers can be used to solve this problem, Singularity is one such program that lets the user create a container as a single file that can be easily moved and used to replicate the run elsewhere. (Kurtzer *et al.* 2017) A singularity container was used to run the CAR tool software on the clinic's test server. The container, also called the image, was easily built by a definition file in which all dependencies and what to contain is specified. In my case Samtools 1.9, python 3.7.2 and the source code for CAR tool on Github were included. Once the image is built CAR tool can be run using singularity and the single singularity image file. The program is now easy to remove or move and run elsewhere in the same specified environment.

## 4.3 Clinical vs Research applications

The requirement for accuracy is the major difference in research and clinical trials. The stakes are much higher for clinical applications where one mistake in the most extreme case could be fatal. In research on the other hand a slight error rate is mostly accepted and generally do not really affect the outcome of a study. The acceptance of the result not being 100% accurate in research highlights the need to add functionality or additional steps when adopting research tools and techniques for clinical applications. Currently coverage analysis tools exist for research purposes but lack the extra accuracy needed for clinical usage. To be included in a clinical setting the clinical geneticist/physician needs to know that all nucleotides within a target region has been sequenced at sufficient depth to

exclude any pathogenicity from that region. Whenever an important genetic region is not covered at sufficient depth, an alternative sequencing technology may need to be used to complement the NGS analysis.

# 5. Method


## 5.1 List demands

The tool is supposed to be integrated in the current bioinformatic pipeline. To be of use it must generate suitable information for both the bioinformaticians and the clinicians. A list of requirements was created after screening for solutions online, examine the current tool at the facility and after holding interviews with both bioinformaticians and clinicians asking about their needs. A first draft was completed after discussion with my supervisor, Claes Ladenvall, one of the project clients. The draft was then used as a template for further discussions with the clinicians and bioinformaticians.

## 5.2 Design considerations

A list of requirements that the tool had to meet was compiled from the bioinformaticians and clinicians' input during the interviews and then rephrased to specific tasks for the implementation. Because there are tools available that can extract a lot of the information needed from the sequence data, a search for a suitable tool to use within this software was made. In the end I decided to use Samtools and its function depth as it allowed different filtering options as well as calculating the important coverage depth values. Keeping CAR tool dependent on as few tools/programs as possible will make it easier to use, as not as many updates and programs have to be installed for it to work. Other dependencies are the python library matplotlib for generating plots. When the prototype was ready clinicians and bioinformaticians were once again asked for feedback and input.

## 5.3 Control and Validation

To ensure that the tool work properly all functions were manually evaluated and checked off a control list. After all functions passed the control tests, the tool was evaluated by testing both larger and different data sets and compared with the currently used tool, ExCID, and the genomic browser IGV.

The most vital part of the tool is the short list of low coverage regions. This list will give a fast indication if any important region lacks sufficient coverage and needs to be re-sequenced. The list with low coverage regions needs to be evaluated by both making sure that the regions found actually are low regions (true positives) and that it does not miss regions found by ExCID (false positives). This evaluation was done by comparing the low regions start position, stop position and coverage value with the genomic browser IGV. By opening the read file in IGV one can graphically compare the read distribution along a reference DNA template. In IGV each base position has a calculated coverage value and these coverage values was summed up over the region defined by CAR tool and calculated

to a coverage mean. Only if all bases of the region have a value below the coverage threshold value and a sum equal to the one generated by CAR tool the region will be seen as a *true positive*. But proving that the list contains true positives is not enough, the list of low regions must also contain the regions already found by the currently used program ExCID. Meaning there can be no *false negatives* either. The validation set consisted of 12 patient sample files with an amplicon library and one with a capture-based library. The result was entered into tables and any dissimilarity of region intervals or mean values was evaluated with IGV.

## 5.4. Implementation

The tool was implemented and run on the development server at the clinic in a singularity container. The container holds all dependencies for the tool along with the actual program. The dependencies and path to the program file on GitHub is written into a description file that is used to build the container. When created, the singularity container is run along with the file program launcher that is the main script of CAR tool to generate the coverage analysis report.

# 6. Results

## 6.1 Coverage Analysis Report Tool

The CAR tool is a tool for assessment of per base quality of NGS data. The tool generates lists, tables and figures that can be used to evaluate coverage depth and coverage breadth over regions of interest. It is of great importance that each base pair in these regions is sequenced properly since only one mistake could lead to a different diagnosis. CAR tool both presents the data to decide on quality while also aid the clinicians by presenting the data in a compact and comprehensive way. Only the essential data is presented in the output to prevent the clinician to browse through large files with unnecessary data. CAR tool is especially created to deal with the lack of clinically appropriate coverage analytical tools. It is a comprehensive and easy to use tool for both clinicians and bioinformaticians. It is flexible, with multiple optional settings that adapt the output based on the user needs. Compared to the previously used tool it is easier to use being well documented and adapted to the clinic's needs.

The tool will be used by several different work groups both directly and indirectly. The bioinformaticians will carry out the coverage analysis as part of an automatic bioinformatic pipeline and send the resulting report together with additional material to the clinicians or hospital geneticists. They in turn will evaluate the list of variants detected, relate them to the type of diagnosis and forward a clinical report to the referring physician. The physician finally will decide on treatment based on the information from all relevant tests.

For further information on how to use the tool I refer to the usage instructions in the Appendix and on Github under the repository anod6351.

## 7.2 Functions and Settings

The tool is built to handle data sets with different demands on the tool. It will therefore be possible to turn on or off certain parts of the analysis and choose proper settings for each analysis. The tool

generates lists of tables and figures that are meant to aid the clinicians in rapidly gaining an overview of the data. The main task for this tool has been to extract regions in the sequenced DNA that is of insufficient sequencing depth. The user inputs a threshold coverage value and all parts of regions below this threshold among the ROI:s is written to a csv file. The tool also computes coverage breadth at three different coverage depth values and writes it into a csv file. These two files along with a log file is always generated, the rest of the functions are optional to keep the program flexible. For further usage instructions and to learn what setting that can be made see the usage instruction in the Appendix or on GitHub in the respiratory anod6351.

### 7.2.1 Sub regions and their mean coverage depth

Every base pair coverage depth value within the ROI positions is checked. All adjacent base pairs either above or below the entered coverage threshold will define a sub region in the list. The sub region is saved with start and stop positions, length of the sub region and the mean coverage depth value of the base pairs in it. Two mean coverage files are generated, one is referred to as the full list containing all sub regions while the second file only contains ROI:s with sub regions that has a *mean below* the chosen coverage threshold.

Each ROI is represented by the region name, the chromosome, start and stop positions. For each region a mean coverage value is calculated together with the length of the region. If the same ROI is divided into two or more sub regions, the new subregion is added to the right of the previous sub region on the same row. A mean coverage value of 0.0 indicates that the regions has not been sequenced and a value of for example 14.22 is the mean number of fragments covering each base pair. The list is displayed in table 1 below.

**Table 1.** Mean coverage depth list. Each row is a region of interest described with region name chromosome and start and stop. The mean coverage column represents mean number of fragments covering the bases of this region. Observe that the values are only created for further explanation and not real values.

| Region Name | Chr | Start | Stop | Mean Coverage | Length |
|---|---|---|---|---|---|
| ROI 1 | chr1 | 429486475 | 429486490 | 0.0 | 15 |
| ROI 2 | chr2 | 892647151 | 892647158 | 14.22 | 7 |
| ROI 3 | chr3 | 69242922 | 69243000 | 45.01 | 78 |

### 7.2.2 Statistical table with coverage breadth

For each region the coverage breadth is calculated at the three different coverage depth thresholds entered by the user. Each region then has three coverage breadth values. The statistics table contain 5 different columns as default. A data type column indicating if the data has been filtered by the program or not, followed by the region name and the three coverage breadth values at three given coverage depths. A coverage breadth value of 1.0 indicates 100% of the bases in this region are covered by at least the same number of fragments as the coverage depth threshold. For visual representation of the table see table 2. An extra column called validation is added if the validation option is activated, more on that below.

**Table 2.** Representation of the output file including the coverage breadth statistics. The first column indicates weather any changes has been done to the data when running the program. The second column gives the name of the ROI and the following columns contain the coverage breadth values at the three chosen coverage depth thresholds. Observe that the values are only created for further explanation and not real values.

| Data type | Region Name | 100X | 300X | 400X |
|---|---|---|---|---|
| Raw | ROI 1 | 1.0 | 0.5 | 0.4 |
| Raw | ROI 2 | 1.0 | 1.0 | 1.0 |
| Raw | ROI 3 | 1.0 | 1.0 | 0.8 |

### 7.2.3 Log file

A record of what settings was used, who ran the analysis and what files that was used is saved to the log file in order to replicate and keep track of the analysis. The first section in the file is provided by the tool or the user with information such as date, program version and what data files that has been used. The second part contains the mean calculations of coverage depth and breadth at the chosen coverage depth thresholds. And the final section lets the user know if any of the optional settings were activated such as combine ROI:s that comes from the same gene or if any figures were generated.

**Table 3.** Representation of the log file. The first part of the table contains information about the run, such as what data that has been used and who ran CAR. The second part contain mean calculations of coverage breadth and depth and the third part contains information about optional settings that has been triggered. Observe that the values are only created for further explanation and not real values.

| LOG | | | |
|---|---|---|---|
| Program Version 1.0 | | | |
| Date: | 2019-01-03 17:19:55.630104 | | |
| Coverage analysis run by: | Anna | | |
| Region file (BED file): | ../Regions.bed | | |
| Read file (BAM file): | ../Reads.bam | | |
| Output folder and file name: | CAR_patient1 | patient1 | |
| | | | |
| Coverage depth thresholds: | 100 | 700 | 800 |
| Mean Coverge Breadth: | 1.0 | 0.94 | 0.93 |
| Mean Coverage Depth: | 6338.27 X | | |
| | | | |
| Combine regions activated | | | |
| Figures generated | | | |

### 7.2.4 Per base coverage depth – optional

The detailed list contains all the coverage values for each and every position in the regions from the BED file. These values are computed with Samtools that prints one base pair coverage value per row. The result from Samtools is reshaped to contain all coverage values per region as well as additional information about the region such as region name, see table 4.

**Table 4.** The per base coverage file contains all ROI:s (represented by name, chromosome, start and stop position) with all the coverage depth values, one per base pair. Observe that the values are only created for further explanation and not real values.

| Region Name | Chr | Start | Stop | Coverage Depth | | |
|---|---|---|---|---|---|---|
| ROI 1 | chr1 | 12486197 | 12486286 | 3001 | 3001 | … |
| ROI 2 | chr2 | 7634965 | 76357586 | 1830 | 1830 | ... |

### 7.2.5 Validation list - optional

The validation list contains all regions with a coverage breadth value below 95% at the first coverage depth threshold, a representation of the validation list is seen in table 5. The regions with coverage breadth values below 95% are marked with stars in the statistics table, seen in table 6. This approach is currently used at the clinic by some of the bioinformaticians.

**Table 5.** Representation of the validation list. The region name is followed by the coverage breadth below 95% at the first coverage depth threshold, the chromosome, start position, stop position and length of the region. Observe that the values are only created for further explanation and not real values.

| Region Name | % Coverage at: 100 X | Chr | Start | Stop | Length |
|---|---|---|---|---|---|
| ROI 1 | 0 | chr1 | 971478 | 971499 | 21 |
| ROI 3 | 15 | chr3 | 8961247 | 8962237 | 990 |

**Table 6.** Representation of the Statistics table with the validation column added. The Regions with a coverage breadth value at the first coverage depth threshold is marked with *** and added into the validation list. Observe that the values are only created for further explanation and not real values.

| Data type | Region Name | 100X | 300X | 400X | Validation |
|---|---|---|---|---|---|
| Raw | ROI 1 | 0.8 | 0.5 | 0.4 | *** |
| Raw | ROI 2 | 1.0 | 1.0 | 1.0 | |
| Raw | ROI 3 | 0.7 | 0.4 | 0.4 | *** |

### 7.2.6 Figures - optional

Three different figures can be generated as PDF files. A pie chart, shown in figure 3, and a bar plot, shown in figure 4, show the distribution of low coverage versus high coverage bases. Yellow color depicts low coverage depth and blue stands for coverage depth values above the threshold. The pie chart gives an overview of the ROI coverage depth while the bar plot can be used to further investigate the coverage ratios per exon. To save memory and time in finding the low coverage ROI images, only ROI:s containing a low coverage regions are plotted and saved. These figures can be used to get a statistical understanding of the data in a graphical manner. The final graph seen in figure 5 display the region and marks the low region's areas by the same color scheme as the previous figures. This figure also has a table of low regions information next to each plot.

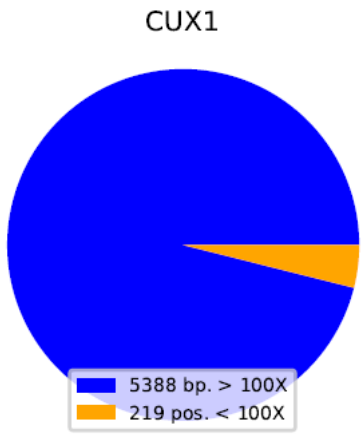



**Fig. 3.** A pie chart assessing the base pair coverage quality over the gene CUX1. The blue label represents all base pairs above or at 100X coverage while the yellow represents all other base pairs below the coverage threshold, set to 100X in this example.

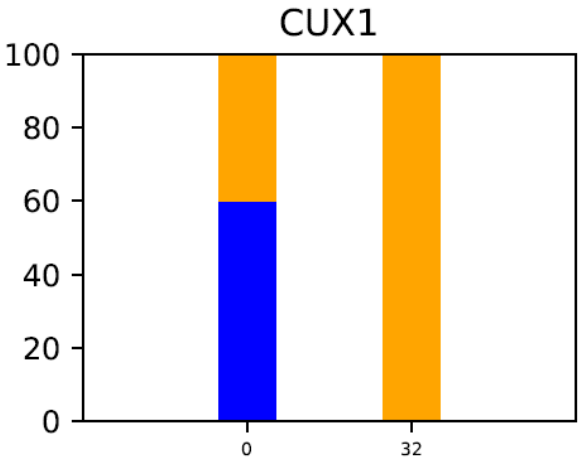
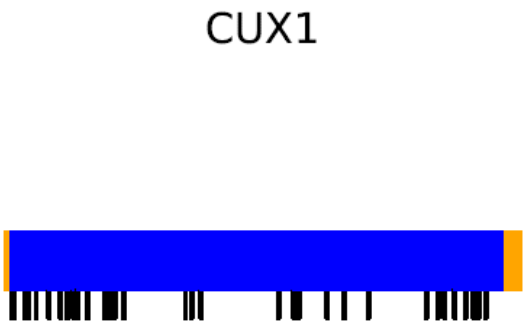



**Fig. 4**. A bar plot of the low coverage ROIs from the gene CUX1. The first ROI has around 40% of its bases below the coverage threshold while the second ROI has none of its bases above the threshold. Only the ROIs with coverage values below the threshold is displaced in the plot.




| | Chr | Start | Stop | Mean | Length |
|---|---|---|---|---|---|
| R1 | chr7 | 101459310 | 101459373 | 26.97 | 63 |
| R2 | chr7 | 101926226 | 101926382 | 0.48 | 156 |

**Fig. 5**. The region plot of the gene CUX1 to the left consists of several ROI:s, these are represented with dark lines. The blue region are bases above the coverage threshold while the two yellow parts represents bases below the coverage threshold. These yellow regions are seen in the table to the right. The first yellow region is defined as R1 and the second R2. With the default setting the low coverage region is further defined by chromosome, start, stop, mean coverage depth and length of the region.

### 7.2.7 Optional Settings

CAR is a flexible tool aimed to handle different needs. Apart from the optional settings described previously these settings do not change the visual representation of the output rather the data itself, such as a filtering option. The additional settings are listed below. For further usage instructions see the appendix.

- Pass on commands to Samtools
- Filtering of BAM file with mapQ and/or phred (Q) score, on all data or as a supplement
- Strand specific coverage breadth reporting
- Adding a new column in the coverage depth lists with information, yes or no, if a region belongs to a defined group of regions. Can be used to identify regions that are known to have low performance etc. This list of known regions is added as an input .bed file
- Adding transcript and exon information to the coverage depth lists

- Indicate position of hotspots using arrows within the region plot. This list of hotspots is added as input

## 7.3 Validation

The validation of the CAR tool was done by two comparison studies. The first study was done to validate the results produced by CAR and the second study was done to compare CAR with the currently used tool ExCID. For more details concerning the validation specification see table 7 below.

**Table 7.** Validation specification

| | |
|---|---|
| Method | Validation of CAR tool |
| Analysis/Hypothesis | How CAR tool performance compared to current tool |
| Reason for analysis | Design of new tool for Coverage Analysis |
| Validation against | ExCID and IGV |
| Sample material | BAM and BED files |
| Validation material | QC report generated from ExCID |
| Treatment of the material | On the test server or locally for non-sensitive data |

### 7.3.1 Validation using IGV

The validation study with IGV was made in order to check that the low coverage regions found by the CAR tool were real low coverage regions. All low coverage regions detected by CAR in the test data set were found to be true low regions, this by comparing the region coverage values with observed values in IGV. The validation was a success with no coverage value above the threshold for the base pairs and the coverage mean value of the region was coherent with observed values in IGV. See table 8 for the result of the validation study.

**Table 8.** Comparison study with IGV. The low coverage regions are defined by name, chromosome, start position and end position. The coverage depth column is calculated with CAR tool and the following column IGV are calculated from the coverage depth values in IGV. Data used: TruSightMyeloid_regions.bed, OncoSpan_180316.bam, Coverage depth threshold limit: 100X

| Region Name | Chr. | Start (0-based) | Stop (1-based) | Coverage depth | IGV | Concordance |
|---|---|---|---|---|---|---|
| DNMT3A | 2 | 25475061 | 25475066 | 0.0 | 0 | Yes |
| CUX1 | 7 | 101459310 | 101459373 | 26.97 | 27 | Yes |
| CUX1 | 7 | 101926226 | 101926382 | 0.48 | 0.48 | Yes |
| CDKN2A | 9 | 21994319 | 21994330 | 0.0 | 0 | Yes |
| BCOR | X | 39933741 | 39933770 | 33.17 | 33 | Yes |
| STAG2 | X | 123176417 | 123176495 | 23.24 | 23.24 | Yes |

### 7.3.2 Validation using ExCID

The second validation study with ExCID were made to make sure no low coverage regions were missed by the new tool CAR. For this study 12 amplicon-library data files and one capture-library data file were used as test data sets. For the amplicon-based library files the low coverage regions were very much the same between CAR tool and ExCID, see table 9. All low coverage regions found by ExCID were also found by CAR tool. However, some minor differences in the decimals of the mean coverage depth values could be seen in 14 out of 71 cases. 9 of the 14 differences in coverage values differed from 0.01 to 0.07 between the two tools while the other 5 values differed slightly more,

from 1.72 to 28.09. For this study the coverage threshold was set to 100X and with a high throughput, so the differences are in fact not that large in regard to the data size. All different values were evaluated with IGV and all of the coverage values seemed to agree with CAR tool. The 5 slightly larger differences also had other reported lengths in ExCID while sharing the same start and stop positions as the result in CAR which affect the computed mean value. I am not sure how this difference has occurred in ExCID. Another important thing to keep in mind is that ExCIDs start positions are 1 based while CAR tool uses 0-based start positions.

The low coverage regions detected from the capture-based library file showed greater differences between the two tools, see table 10. Differences such as region length and different regions found by the two tools. Either ExCID found low coverage regions that in fact were above the coverage threshold or had longer intervals compared to CAR tool. No filtering of the data was used when running the two tools. Out of the 37 low regions found by both tools 8 regions differed in start/stop position and mean value. The different results were evaluated with IGV which backed the regions provided by the CAR tool, ExCID reported some base pairs to have low coverage even if that was not the case. The longer regions could be due to ExCID using some sort of default smoothening on the data set and thus removing fragments. Overall the differences were no that large, differing with just a few numbers of fragments, also for the capture-based file a coverage threshold of 10X was used together with a lower throughput than for the amplicon files. More in detail the first region that differed had the same start position for the two tools but was reported as 7 base pairs longer in the results from ExCID. The last 7 base pairs were investigated in IGV and they were all in fact over the coverage depth threshold. The second differently found region were only found by ExCID, a single base pair with a coverage depth slightly below the coverage threshold while being slightly above in IGV. The third region were a large region found by ExCID, CAR tool instead found 3 smaller sub regions in the end of the larger region. This larger region found by ExCID had several coverage values above the threshold in it according to IGV that should have split the regions. And finally, the 8th region that differed had a slightly earlier start and later end position, the length differed 35 base pairs in total. The base pairs not found with CAR tool but with ExCID proved again to be over the threshold with IGV.

**Table 9**. Validation with ExCID using amplicon-based library files. Both tools found the same regions, however 9 minor differences in coverage value between 0.01 and 0.07 (not counted as a difference in the table due to the small difference) was seen and 4 larger differences (1.72, 7.79, 25.33 and 28.09).

| Number of alignment files (BAM) | Same low regions | Different low regions | Same coverage depth | Different read depth |
|---|---|---|---|---|
| 12 | 71/71 | - | 67/71 | 4/71 |

**Table 10.** Validation with ExCID using capture-based library files. 8 regions differed slightly in length and thus had different coverage depth values.

| Number of alignment files | Same low regions | Different low regions | Same coverage depth | Different read depth |
|---|---|---|---|---|
| 1 | 29/37 | 8/37 | 29/37 | 8/37 |

### 7.3.3 Confusion matrix and sensitivity

The performance of CAR tool is very good. All found low coverage regions proved to be real low regions, referred to as true positives, while also not missing to report any other low region. The high

coverage regions not reported are referred to as true negatives. The true positive values in the confusion matrix displayed below in table 11 come from 6 low coverage regions found by CAR tool on a test set evaluated by IGV, 71 low coverage regions found by CAR tool on the 12 amplicon library files evaluated with ExCID and finally 35 low coverage regions found in the capture-based file that were evaluated with IGV when the two tools differed. The true negative value is the sum of all regions specified in the BED files for each run minus the true positives for each run. 325 regions in the BED file used for the amplicon library files and 935 regions in total for the capture library file. Calculation of the true negative value: 325*12 – (71+6) + (935-35) = 5048. The sensitivity and specificity are both calculated to be equal to 1 by TP/(TP+FN) for sensitivity and TN/(TN+FP) for the specificity.

**Table 11.** A confusion matrix indicating 112 true positives and 5048 true negatives. The true positives represent the correctly found low coverage regions and the true negatives represents the correctly not flagged regions with sufficient coverage.

| *Actual data* | | **POSITIVE** | **NEGATIVE** |
|---|---|---|---|
| *CARtool* | **POSITIVE** | 112 | 0 |
| | **NEGATIVE** | 0 | 5048 |

# 8. Discussion

CAR tool emerged out of the need for a coverage analysis tool for clinical application. It is designed to find every subregion, within a list of given ROI:s, that fail to reach the coverage depth limit set by the user. The clinical personnel thus have necessary information to decide if analyses, like a re-sequencing, might be evident. It is important that each such low region is found. To evaluate this the result from CAR tool was compared to IGV:s coverage depth values over the specified low regions. All low regions were proven to correspond with the coverage values in IGV and being below the threshold.

Also, a comparison with ExCID results across multiple amplicon-based library files showed that no low region was missed by CAR tool that was found by ExCID. The sensitivity and specificity estimates when applying CAR tool on amplicon data in this evaluation are very promising. A comparison with a capture-based library file showed greater dissimilarities. The new CAR tool tended to find more short regions while ExCID provided longer low coverage regions or other regions only found by ExCID. To try to make sense out of these differences the region coverage values were compared to IGV, showing that the longer reads or the ones only found by ExCID actually contained coverage values above the coverage threshold. The two programs found low regions around almost the same areas but where CAR tool made several subregions, each ending at a high coverage value peak, ExCID provided longer regions even though some values where above the threshold. A possible explanation for this is that ExCID uses some form of smoothening on the reads. A smoothening strategy would primarily have an effect on the regions that have variation in coverage values close to the specified threshold limit. Such a variation is more likely to be present in capture-based libraries, compared with amplicon libraries, because of the randomly distributed reads. I have not been able to identify any smoothening algorithm in the source code for ExCID but still believe it to be a possible explanation. The source code is sparsely commented which makes it very difficult to understand. This

is also one of the major issues with the currently used program ExCID. I hope that my program will be easier to use with all its comments and additional documentation on how to use the tool properly. By creating a tool at the clinic, they can form the tool after the current needs and work with perfecting it after new demands.

The main usage for the tool is the short list of low coverage and calculation of coverage breadth values. The tool mainly brings out the information that ExCID currently provides for different clinical pipelines. A new feature are the images that can be used to get an idea of where the low regions can be found. For instruction on how to use the tool please see the user instructions in the appendix. It is very important how the BED file and region name is constructed in order for the program to extract the right information. During this project I realized that not all bed files have 0-based start position, so this is also very important to check when using BED files, CAR tool uses 0-based start positions. Another important note is the execution time, the execution times varies depending on what settings that are activated. For example, generation of images doubles the run time. The files I have run have an execution times around 10-15 minutes with the default settings, which are about the same for the currently used tool ExCID, but the execution time for this tool also differ when applied on different data sets and settings.

The design of the tool has changed during the project, in the beginning and during most of this project I have used bedtools and its wrapper PyBedtools to compute the coverage values. I found it to be very easy to work with in python and one was able to save Bedtools object rather than creating new files every time a coverage calculation was made. However, Bedtools did not provide the phred score and mapQ filtering I needed, so instead of having several dependencies I decided to use Samtools throughout. To provide users additional flexibility in the call to Samtool I added an extra input option that gives the user the possibility to write any Samtools command as input. Keeping the tool as dependency free and flexible as possible has been my goal all along combined with ease of usage. During this change I realized some differences between the two tools. Bedtools also counts indels into the coverage depth value while Samtools and IGV does not. Another observation during the validation was that the same low regions were to be found again and again on the amplicon-based library files. I was aware about this pattern from the beginning so that an extra column in the low coverage table could be incorporated indicating if a subregion is known to often be low or not, given that the users adds a list of low coverage sub-regions.

It is not until CAR tool is used for real that the program is put to the most difficult and thorough test. It will be run alongside with ExCID and evaluated for each run until it can successfully replace its precursor. It is my hope that the tool will continue to grow at the clinic and be both improved and adapted for new demands and to run efficiently as a part of the most important analysis pipeline. Possible improvements could be implementation of an optional smoothening function. The capture-based library alignment files generated multiple low subregions. It could be good option to have the opportunity to find longer regions instead of shorter ones. Also, a graphical interface would be nice to have, that makes it easier to work with the tool even without programming experience.

# 9. Acknowledgements

A big thank you to the clinical genomics team, I could not have done this without you! Thank you for offering me this super cool project to begin with while also making me feel like one in the team from day one. I would like to thank my wonderful supervisor, Claes Ladenvall, for all the encouragement and support during my master thesis and everyone in the bioinformatics group for both helping me with the tool but also for the great company. Also, a big thank you to all the clinicians that has helped me with the design of the tool.

I would like to thank everyone involved in my master thesis, my editor Adam Ameur, my program supervisor Jan Andersson and coordinator Lena Henriksson for your support on my project and help with the report. To Jan and Lena I also want to say thank you for the past 5 years, thank you for your encouragement and commitment to all the students and the program itself. And finally, I also need to thank my family and friends for the support and for believing in me. The day I started at Uppsala university was a new chapter in my life and now once more I am turning a chapter as a graduate in bioinformatics.

# 10. References

Ardui S, Ameur A, Vermeesch JR, Hestand MS. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic acids research 46: 2159–2168.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research 8: 186.

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013. The Next-Generation Sequencing Revolution and Its Impact on Genomics. Cell 155: 27–38.

Kurtzer GM, Sochat V, Bauer MW. 2017. Singularity: Scientific containers for mobility of compute. PloS one 12: e0177459.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research 18: 1851–1858.

Merino GA, Murua YA, Fresno C, Sendoya JM, Golubicki M, Iseas S, Coraglio M, Podhajcer OL, Llera AS, Fernández EA. 2017. TarSeqQC: Quality control on targeted sequencing experiments in R: MERINO et al. Human Mutation 38: 494–502.

Nikiforova MN, Wald AI, Roy S, Durso MB, Nikiforov YE. 2013. Targeted Next-Generation Sequencing Panel (ThyroSeq) for Detection of Mutations in Thyroid Cancer. The Journal of Clinical Endocrinology & Metabolism 98: E1852–E1860.

Peng Q, Vijaya Satya R, Lewis M, Randad P, Wang Y. 2015. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. BMC genomics 16: 589.

Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current Protocols in Bioinformatics 47: 11.12.1-11.12.34.

Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, Damodaran S, Bhatt D, Reeser JW, Datta J, Roychowdhury S. 2015. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. Human Mutation 36: 903–914.

Sanger F, Nicklen S, Coulson AR. 1977. DNA Sequencing with Chain-Terminating Inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74: 5463–5467.

Schmidt KT, Chau CH, Price DK, Figg WD. 2016. Precision Oncology Medicine: The Clinical Relevance of Patient-Specific Biomarkers Used to Optimize Cancer Treatment: Precision Oncology Medicine. The Journal of Clinical Pharmacology 56: 1484–1499.

Soukupova J, Zemankova P, Lhotova K, Janatova M, Borecka M, Stolarova L, Lhota F, Foretova L, Machackova E, Stranecky V, Tavandzis S, Kleiblova P, Vocka M, Hartmannova H, Hodanova K, Kmoch S, Kleibl Z. 2018. Validation of CZECANCA (CZEch CAncer paNel for Clinical Application) for targeted NGS-based analysis of hereditary cancer syndromes. PloS one 13: e0195761.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. Trends in Genetics 30: 418–426.

Zhang H. 2016. Overview of Sequence Data Formats. Methods in molecular biology (Clifton, NJ) 1418: 3.

# 11. Appendix

## 11.1 Usage instruction

CAR tool is a tool for assessment of per base quality of NGS data. The tool generates lists, tables and figures that can be used to evaluate coverage depth and breadth over regions of interest. It is of great importance that each base pair in these regions is sequenced properly since only one mistake could lead to a different diagnosis. CAR tool both presents the data to decide on quality while also aid the clinicians by presenting the data in a compact and comprehensive way. CAR tool is especially created to deal with the lack of clinically appropriate coverage analytical tools. It is a comprehensive and easy to use tool for both clinicians and bioinformatics. Flexible with multiple optional settings that adapt the output based on the users' needs with lists, tables and figures.

**Example run**: Coverage Analysis Report tool is launched from a terminal window.

$ python ProgramLancher.py -a myRegions.bed -b myReads.bam -c 30 50 70 -e NameOfUser -o OUTfolder OUTfilename

**Mandatory input:**

**-a** 1 BED file containing the regions of interest. The bed file of the regions should contain chromosome, start position, end position and region name in that order. The Start position is 0-based and the end position 1-based.

For additional options such as merging regions from the same gene in the report the region name must start with the name of the gene followed by a dot. If adding exon and transcript information to the report the name should be the following; RegionName.exon.#.transcript.#.chr#..... Other words than exon and transcript is okay to use but the order of the exon number and transcript must be as the order above.

**-b** 1 BAM file with the reads
**-c** 3 coverage depth threshold values
**-e** Enter name of the person running the analysis
**-o** Enter the name of the output folder and name of the output files

**Optional input:**
**-p** Phred score and mapQ filtering of the bam file. Scores below the chosen values will not be used for the coverage calculation. With option (-p all value1 value2) the filtered BAM file is used throughout the analysis while (-p value1 value2) is used only for an additional column in the statistics table.

**-k** Combine rows in the BED file, the rows in the bed file that has the same for example gene name will be combined and the tables and figures will be calculated on per gene information. For this option the region name must be placed first followed by a dot.

**-f** Create figures, to save time and memory the images are only generated if called.

**-s** Strand specificity, this generates two additional statistics tables one with only positive reads and one with only negative reads. All other computations are done regardless of the strand type.

**-v** Validation, creates an additional column in the statistics table. A star indicates that the gene or ROI had a coverage breadth below 95% at the first threshold value for the statistics table. This will also generate a validation list with all those genes or ROIs below 95%.

**-t** Enter a hot spot list of base pairs of interest. These positions will be pointed out in the region figure by arrows. The hot spot list should be a BED file with the columns chromosome, start, stop and region name.

**-m** Adds exon and transcript information to the mean lists. Make sure that the region name is written as follows; RegionName.exon.#.transcript.#.chr#..... Other words than exon and transcript is okay to use but the order of the exon number and transcript must be as the order above.

**-l** Enter a list of regions that are known to be low. A column in the mean short list and mean full list is added that indicate if the regions are usually low. And the table in the Region plot will have the first column in the row colored red. The input list should be a BED file with the columns chromosome, start, stop and region name.

**-d** Returns the detailed per base pair coverage list

**-i** Tailor the command sent to Samtools to calculate coverage depth. The command will end with adding the bam and bed files and saving the resulting file. Observe that Samtools uses a cut off value for the reads that effects the coverage if more reads are in the bam file. CAR tools set the cut of value to 30000.

For example the default is set to be “samtools depth -a -d 30000”. This string can be changed by the user in the input followed by the flag -i. The new string is then merged with “-b theRegionsfile.bed Reads.bam and the output is saved to the correct file and folder.

**Standard output:**

**Full mean list** of sections below and over the first entered coverage threshold value. All base pair coverage values are evaluated for every ROI or gene and are used to create new sub regions. The sub region consists of all adjacent base pair position either above or under the threshold. For each sub region the coverage mean is calculated.

**Short mean list** is extracted from the full mean list. Only sub regions below the coverage threshold are saved in this shorter list. If the ROI or gene has no sub regions below the coverage threshold they will not be in the list.

**Statistics table** with coverage breadth values for each ROI or gene. The coverage breadth is calculated from the 3 threshold values in the input. The original statistics table is not strand specific, but if this choice is activated two additional statistical tables one for each option are generated.

**Log file,** a record of what files that was used, who run the analysis, used settings together with calculations of the total coverage mean value and total mean of coverage breadth values.

**Optional output:**

**Validation list**, list of ROIs that has less than 95% coverage in the first column of the statistics table. These are marked with stars in the statistics table if the validation option is turned on.

**Strand specific tables**, two additional statistics table is generated one for only positive reads and one for only negative reads. The last column in each table indicates the difference in coverage breadth.

**Phred score and mapQ filtering of the bam file**. The filtered coverage depths are either used for the whole analysis or only in the statistics table, added as a row under each ROI. The option, -p all value1 value2, is used if the whole analysis is to be run with the filtered bam file while -p value1 value2 only uses the filtered bam file as additional rows in the statistics table.

**Figures**
A pie chart of positions above or under the coverage threshold per ROI.
A region plot that visualize where the coverage is lower than the threshold together with a table of the low regions name, chromosome, start position, stop position, mean and length.
A bar plot with low coverage exons displayed as bars with amount of positions above or under the coverage threshold. These figured are only created if combine rows is activated.

**Dependencies:**
Samtools – for per base coverage depth calculations.

**Source code:**
The source code can be found at Github in the repository anod6351/CARtool