CNAplot — Software for visual inspection of chromosomal copy number alteration in cancer using juxtaposed sequencing read depth ratios and variant allele frequencies

Hansen, Marcus H.; Hokland, Peter; Nyvold, Charlotte G.

Published in:
SoftwareX

DOI:
10.1016/j.softx.2020.100503

Publication date:
2020

Document version
Final published version

Document license
CC BY

Citation for published version (APA):

Terms of use
This work is brought to you by the University of Southern Denmark through the SDU Research Portal. Unless otherwise specified it has been shared according to the terms for self-archiving. If no other license is stated, these terms apply:
- You may download this work for personal use only.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk

Download date: 11. Jan. 2021
CNAplot — Software for visual inspection of chromosomal copy number alteration in cancer using juxtaposed sequencing read depth ratios and variant allele frequencies

Marcus H. Hansen a,*, Peter Hokland b, Charlotte G. Nyvold a

Hematology-Pathology Research Laboratory, Department of Hematology, Department of Clinical Research, Odense University Hospital and University of Southern Denmark, Odense, Denmark

Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

Article history:
Received 9 March 2020
Received in revised form 23 April 2020
Accepted 4 May 2020

Keywords:
Copy number alteration
Copy number variation
Whole exome sequencing
Next generation sequencing
Copy neutral loss of heterozygosity

Abstract
Cancer genomes are frequently struck by chromosomal lesions. Next generation sequencing provides potentially high-resolution assays in a single analysis to detect copy number alterations (CNA). The implementation and usage of sequencing may partly be hampered by the lack of transparency in copy number calling algorithms. Here, we present the software CNAplot for aligned visualization of variant allele frequencies and read depth ratios generated from paired whole exome sequencing samples. We implement transparent statistics to evaluate shifts in allele frequencies and sequencing read counts in order to support a copy number event, to complement other tools and to detect somatic copy-neutral loss of heterozygosity.

1. Motivation and significance
Cancer genomes are frequently marked by several chromosomal lesions, such as translocations where a part of one chromosome attaches to another, acquired copies of whole or parts of chromosomes or chromosomal losses. Specific aberrations are
often recurrent and hold diagnostic or prognostic value to clinicians or researchers. Until recently, analyses on these copy number alterations (CNA) have mainly been reserved to the field and laboratory techniques of cytogenetics. Next generation sequencing (NGS) has made it possible to explore these chromosomal aberrations in a single assay with a very high resolution as compared to several former techniques. In spite of the potentially valuable information and practical aspects, which can support the laboratory work-up and conclusions of the cytogeneticist or pathologist, the implementation may be hampered by the lack of transparency in copy number calling algorithms, lack of simplicity or applicability.

Whole exome sequencing (WES) has been utilized in the detection of CNA for a decade, since in-solution capture-based assays entered the scene [1]. Studies show that WES generally provides a lower resolution or an increase in noise compared to whole genome sequencing [2,3]. Also, the concordance between algorithms frequently implemented in published research studies, such as XHMM, CoNIFER, and CNVnator, ADTEx, CONTRA, Control-FREEC, EXCAVATOR, ExomeCNV, Varscan2 etc. have been found to be low [4–6]. The majority of tools are initiated through command-line interfaces with no graphical interaction, and often requires some basic knowledge of shell scripting or having more or less complicated workflows, such as CNVnator [7]. It is found that the few papers that offer a web-based tool are unavailable a few years after publication, such as exemplified by VCS (visualization of CNV or SNP). In either case, the possibility to assess changes in chromosomal copies by altered sequencing read depth and along with affected variant allele frequencies is not straightforward using the currently available command line tools. This further emphasizes the need for practical software with such capabilities.

Here, we present the software CNAPlot developed for the purpose of plotting variant allele frequencies juxtaposed read depth ratios between case sample and control sample from WES with a medium to high depth of coverage, and which is produced in the same sequencing run. With this tool, the chromosomal copy number alterations are visualized along with significantly altered allele frequencies for fast assessment of acquired copy alterations. In addition, nonparametric statistical tests of observed aberrant regions are provided to support copy change or even copy-neutral loss of heterozygosity by absence of altered read depths. Thus, the provided software may aid the researchers, or cytogeneticists, by directly complementing other analyses and by being easy to use directly on variant call formatted files (VCF) and depth of coverage of defined regions.

2. Software description

CNAPlot takes input generated from whole exome next generation sequencing reads, as illustrated by the example described below. The sample purity and sequencing must be of sufficient quality and coverage. In general, we do not recommend exome sequencing with less than 50 million reads per sample as it may lead to poor resolution. Two types of files are needed for each of the paired samples: (1) read counts retrieved from provided intervals (BED), such as gene regions as used in the example here, and (2) variants in the variant call format (VCF). Following import of data, the variants are visually aligned to the positions of respective calculated read depth ratios (Fig. 1). Significant shift between focus variant allele frequencies (VAF) and control sample VAF is calculated from Pearson’s χ² (Eqs. (1)–(2)) or, optionally, Fisher’s exact test for samples with lower coverage (not recommended). Based on initial testing of multiple sequencing samples from leukemia and lymphoma patients, we evaluated that McNei-
mar’s paired test for related sample pairs was not superior, when compared to χ². Hence, all sample pairs are analytically regarded as independent.

| Table 1 |
| 2 × 2 contingency table of case-control paired allele depths (D_{sample,alt}):

<table>
<thead>
<tr>
<th>Samples</th>
<th>Case</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>D_{a,a}</td>
<td>D_{a,b}</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The significance threshold is selected to match the number of variants, n, i.e. the number of multiple tests performed and the accepted threshold of false positives. The following summarizes the test for a 2 × 2 contingency table (Table 1, Eqs. (1)–(2)), given one degree of freedom and significance level of α = 10⁻⁴ applicable to testing hundreds of variants per chromosome. Here, the test evaluates the number of observed reference and alternate allele (O) given by its depth (D) to the expected (E):

\[ \chi^2_{\text{Pearson}} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} > 15.14 \]  

\[ \chi^2_{\text{Pearson}} = \frac{N(D_{C,A}D_{N,B} - D_{C,B}D_{N,A})^2}{(D_{C,A} + D_{C,B})(D_{N,A} + D_{N,B})(D_{C,A} + D_{N,A})(D_{C,B} + D_{N,B})} \]  

\[ p_{\text{compound}} = \prod_{i=1}^{n} p(i) \]  

Generally, we have observed that for high quality sequencing data a Bonferroni corrected threshold of α = \( \alpha_{uncorr} / n \) is compatible with the presented analytical workflow. The compound probability (Eq. (3)) is calculated from the longest stretch of significantly altered variants and provides supporting evidence of a large copy change. As example, the theoretical probability of coincidentally observing three successive, independently, altered variants, with the proposed α, is then 10⁻¹².

Statistical evaluation of difference between binned read counts extracted from specified intervals is performed with Wilcoxon rank-sum test (Mann–Whitney U [8]), thus treating both unmatched and matched samples, i.e. from the same individual, as independent samples. In comparison, the Wilcoxon signed- rank test [9,10] for related case-controls show in inferior performance for CNV detection. The implemented algorithm is thus:

1. Let list be a 1 × 2n matrix, where the first half, list(1 to n), is the case with n number of observations, and the second half is the normal control, list(1 + n to 2n), with each holding the same number of observations. Each element is the number of reads of a given interval normalized to the number of total reads for the respective sample. In the provided example, each row represents the read count for a given gene.
2. Add index column with \( i_1 \) to \( i_{2n} \) to the list
3. Sort this list matrix from the lowest to the highest value, removing tie rows.
4. Add a new index column with \( i_1 \) to \( i_{2n} \) to the list
5. Then the test value is the sum of the ranks, \( U = \text{Sum}(i') \) for \( i_1 \) to \( i_n \) or \( U = \text{Sum}(i') \) for \( i_{n+1} \) to \( i_{2n} \).

The selected number of observations in each bin is expected to be high, i.e n ≥ 20, thus U is approximately Gaussian [10] and the standardized rank statistics value is then evaluated as follows, for \( \alpha = 0.05 \) (Eqs. (4)–(6)):

\[ z = \frac{|U - \mu_U|}{\sigma_U} > 1.96 \]  

\[ \mu_U = \frac{n^2}{2} \]  

\[ \sigma_U = \frac{n^2}{2} \]
Fig. 1. CNAplot input requirements. Two file types are needed for each of the paired samples: (1) read counts in specified intervals (BED), such as gene regions, and (2) variants in the variant call format (VCF). Software executables are available for OS X, Linux and Windows.

\[
\sigma_u = \sqrt{\frac{n^2(2n + 1)}{12}}
\]  

2.1. Software architecture

CNAplot has been developed in the cross-platform programming environment Xojo (Xojo, Houston, TX, USA). Compiled executables are provided for OS X, Windows 10 and Linux (Ubuntu), with no additional software required to run the tool. The software implements parts of the proprietary chart and graph plotting library, ChartDirector (Advanced Software Engineering Ltd, Hong Kong) through MBS Xojo ChartDirector plugin (Christian Schmitz Software GmbH, Nickenich, DE). The imported data is stored in an in-memory SQLite database created at runtime. Thus, the RAM usage is proportional to number of variants and number of read counts. Memory usage was 125 MB under OS X for the provided example files.

2.2. Software functionalities

In summary, the major functionalities of the software are (1) visual alignment of variants and read counts for evaluation of VAF shifts and altered read counts for the detection of somatic chromosomal copy-affecting and copy-neutral events, (2) statistical inference and (3) export of plots to file (JPEG). The individual plots of each chromosome include an estimated position of the centromer relative to the flanking variants on the respective p-arm and q-arm. The positions have been retrieved from the UCSC Table Browser (http://genome.ucsc.edu, UCSC Genome Informatics Group, University of California, Santa Cruz, CA, USA) for the assemblies GRCh37/hg19 and GRCh38/hg38 (track: Chromosome Band, table: cytoBand). This data is preloaded in CNAplot (current version 1.2).

3. Illustrative example

In this section, we will provide an example of usage based on paired-end whole exome sequencing on the Illumina platform (San Diego, CA, USA) with a case of T-cell acute lymphoblastic leukemia (T-ALL) from a previously published report [11], and a control sample derived from a paired skin sample with approximately \(90 \times 10^6\) reads each. T-ALL is a malignancy caused by neoplastic proliferation of immature precursors of T-cells. It is often aggressive with a rapid progression.

The following provides a brief step-by-step description on how to generate the necessary input files for CNAplot under Linux. Note, these steps are not required to run the software with the provided examples files, nor is it the required workflow. Alternatively, user input files may be produced with bcftools [12] for variant calling and GATK 3.8 (Broad Institute, Cambridge, MA, USA) DepthOfCoverage command for obtaining read counts in specified regions (see supplied example input files).
Fig. 2. CNAplot requires variants and sequencing read counts in intervals of suitable sizes. The variants must be in single sample variant call format (VCF) as it parses the included genotype information to variant allele frequencies. Read counts input is provided in browser extensible data format (BED), and may be generated with bedtools multicov or other similar tools. See example files for correct input format. The data is loaded into memory upon import through SQLite (version 3).

3.1. Installation of necessary tools to generate CNAplot input files

We recommend that alignment of short paired-end reads from WES is performed with Burrows Wheeler aligner (BWA) [13]. BWA can be installed on Linux by opening a terminal window and typing `sudo apt-get install bwa` or as a conda package, `conda install -c bioconda bwa` (Anaconda Inc, Austin, TX, USA). Picard Tools (Broad Institute) and bedtools [14] are installed in the same manner. GATK4 is downloaded directly from the homepage (gatk.broadinstitute.org) or installed as a conda package.

3.2. Alignment of sequences to the human reference genome, sorting and indexing

Open a terminal window and type the following command to perform alignment to your provided reference genome and target regions:

```
bwa mem -M -R "@RG\tID:\tSM:\tPL:\tLB:\tPU: " \nreferencegenome.fasta $file1 $file2 | samtools view -L regions.bed -Sb -> aligned.bam
```

The resulting alignments can be inspected in Integrative Genomics Viewer (IGV) [15,16] after the sorting and indexing of the BAM file, e.g. with Picard or Samtools. Note that if your whole exome sequencing library preparation is amplicon-based instead of capture-based, hard filtering of PCR duplications with GATK4 MarkDuplicates is not recommended as it will remove a large fraction of the total reads.

```
PicardCommandLine SortSam INPUT=aligned.bam \ OUTPUT=aligned.sorted.bam SORT_ORDER=coordinate
PicardCommandLine BuildBamIndex INPUT=aligned.sorted.bam
```

3.3. Retrieval of sequencing counts

Obtaining read depths for provided intervals of your alignment can be performed using the `multicov` tool, which is a part of the bedtools utilities [14]. As example, interval regions (e.g. refGene) in the Browser Extensible Data format (BED) can be retrieved from the UCSC Table Browser at genome.ucsc.edu or the provided example BED-file can be downloaded from the data repository (http://dx.doi.org/10.7910/DVN/BWY6SK, Harvard Dataverse) and used as intervals (GRCh37 assembly). The resulting output contains read counts for a specific gene with two files created for each sequencing pair – e.g. case sample and control sample.

```
bedtools multicov -bams aligned.sorted.bam \ -bed regions.bed > aligned.sorted.bed
```

3.4. Variant calling as input for CNAplot allele frequency retrieval

We recommend GATK4 for variant calling [17–19]. Each alignment in the sample-control sequencing pair is processed through the HaplotypeCaller tool in the genome analysis toolkit. Although base quality score is optional with the BaseRecalibrator and ApplyBQSR tools, it is highly recommended for improved assessment of base qualities, e.g. for downstream filtering (See the current
Fig. 3. This figure shows all variants aligned to the read count ratios of specified intervals in supplied BED-files. Here, each ratio represents a gene. As the number of genes roughly approximates the number of short coding variants, it is highly practical for the comparison of copy number alteration affected variant allele frequencies and read counts. The example here reveals a trisomy 4 and a smaller significant allele frequency shift on chromosome 9. Significant copy of calls on 12, 14, 19 and 22 may be false-positives at the specified significance level, \( \alpha \approx 0.0001(\chi^2) \), and is expected at this threshold. The compound probability provided by the software can be used to support copy changes, where two or more successive allele frequency alterations occur.

GATK4 documentation for more details. \textit{HaplotypeCaller} can detect variants directly by providing the generated alignments as input:

\begin{verbatim}
gatk HaplotypeCaller \\
-R referencegenome.fasta \\
-L regions.bed \\
-I alignment.sorted.bam \\
-O variants.vcf
\end{verbatim}

The resulting files from the outlined steps, i.e. a paired set of BED-files and a paired set of VCF, are all what is needed as input in CNAplot (Fig. 2).

Note that genotype information has been removed from the supplied examples VCF files. Loading of the provided sample data shows two highly significant alterations: (1) a chromosome 4 trisomy and (2) a chromosome 9 copy-neutral deletion (Figs. 3 and 4). Significant changes in variant allele frequency is backed by performing a U-test on change in read depths of the focus sample compared to that of the control sample (Fig. 5). The combined assessment of variant allele frequencies and read depth ratios enables the researcher to locate copy-neutral loss of heterozygosity – a feature which may be highly relevant as a supplemental tool for proper characterization of diagnosis and follow-up in a subset of the cancer patients.

The latest addition to the CNAplot software (version 1.2) estimates the boundary between the p- and q-arm of the selected chromosomes, here exemplified by a case of relapsed mantle cell lymphoma with both copy loss and gain on chromosome 8 (Figs. 6 and 7, not included as example data on Harvard Dataverse).

3.5. Troubleshooting

We have observed very poor resolution for samples sequenced with less than 30 million reads, and therefore our recommendation is 50 million reads, or preferably more, from capture-based whole exome sequencing library prepared from high quality DNA. We have not tested sample amplicon-based library preparation kits. Also, note that the control sample must be sequenced in the same sequencing run. This is, empirically, important for both capture and amplicon-based sequencing libraries. It may not be necessary for the read depth control to be derived from the same individual. However, using paired control sample variants from the same individual with a similar amount of reads is necessary to generate optimal results.

Variant files from whole genome sequencing can attain very large file sizes, so it may be necessary to restrict variants to, for example, the coding regions. Also, both WGS and WES may contain a large number of false positive variants. Thus, filtering on the basis of quality is necessary in any downstream analysis on next generation sequencing data. We refer to the included example as a guideline for the number of observations. Currently, CNAplot implements analysis of autosomes only.

Note that the first launch under OS X it may be necessary to CTRL-click on the CNAplot icon and selecting “Open” to accept launching software from an unidentified developer. Alternatively, in some less restrictive user settings it is adequate to accept that the software is launched for the first time. In Windows 10 it is required to dismiss the warning from Windows Defender Smartscreen in order to run the software. CNAplot does not install
any files, but runs directly from the downloaded and unzipped executables.

Source code is provided at the data repository.
(http://dx.doi.org/10.7910/DVN/BWY6SK, Harvard Dataverse).

4. Impact

Investigation of acquired copy number alterations in cancer research and clinical laboratory diagnostics, prognostics and follow-up using next generation sequencing is highly relevant as it may support the cytogeneticist and clinical decision making. One prominent example is the deletion of 17q, which is recurrently observed together with TP53 mutations in leukemias [20,21], lymphomas and other types of cancer. While whole exome sequencing has now passed its first decade of existence [1], and has especially gained popularity and routine usage in diagnostic laboratories the last couple of years, the focus has mainly been on nucleotide variant detection. Detection of copy number alteration from WES data has also been an area of intense interest, but compared to variant calling, which to a large extent has been driven by the computational guidelines from the Broad Institute with its Genome Analysis ToolKit (GATK), a general consensus on the workflow of detecting chromosomal gains or losses has not been established. While GATK4 Best Practices Workflows on copy number variation are being developed, either the focus is germ line copy number calls, or it requires a very large Panel of Normals of suggested 40 control samples or more from a consistent library preparation and sequencing setup. Most often, this is not within reach for the majority of researchers.

While many different types of CNA calling software exist, not many tools combine allele frequencies and read depth ratios, nor do they support plotting of both together. Some tools only require a single sample, such as the sophisticated QDNAseq R tool [22] for whole genome sequencing which relies on depth of coverage analyses alone, and thus does not detect acquired copy-neutral loss of heterozygosity. As sequencing costs continue to drop, the use of paired case-control samples, allele frequency and read depth juxtaposition has several advantages: Foremost, it gives the ability to determine somatic aberrations. Secondly, it strengthens the evidence of a given chromosomal copy alteration and provides the opportunity to detect acquired uniparental disomy, which is not visible from read depth ratios. Finally, the inclusion of a normal control sample in the same sequencing run helps to diminish the batch variation and noise inherent to whole exome sequencing, and next generation sequencing in general, such as the change of coverage in GC-rich regions. Such variation is often reproduced in the paired sample, thereby attenuating large fluctuations in read depth ratios and dispersion of allele frequencies. As a matched control is most often used for the detection of somatic mutations, it is frequently included in a diagnostic setup or for a specific research purpose.

5. Conclusions

In conclusion, the presented lightweight plotting and analysis software provides fast means to evaluate copy number alterations, juxtaposing both variant allele frequencies and sequencing read depth ratios from control-paired cancer patient samples. This feature enables the evaluation of copy neutral chromosomal losses, which is still neglected to some extent in cancer research and clinical diagnostics. As very few tools with graphical user interface exist, and none are in the presented form, this software
Fig. 5. Observing a very low compound probability is a strong indicator of large chromosomal lesions. Thus, such values are unlikely in data of otherwise high quality data by chance if a reasonable significance threshold is selected, e.g. \( \alpha = 0.0001 \). Here, a copy-neutral loss of heterozygosity is evident from the short arm of chromosome 9: Variant allele frequencies have changed significantly, while the copy number is unchanged.

Fig. 6. This illustrative example shows chromosome 8 from a patient with mantle cell lymphoma (not provided as example data). The plot reveals that a large part of the p-arm has been deleted in either of the parental copies. Also, multiple copies of chromosome 8q are detected. Because a large part of the chromosomes contains one or more copy number alterations, the baseline has been shifted 20%. This is not needed in most cases as the baseline is set by the median value of the read depth ratios. A significant change in variant allele frequency is observed throughout chromosome 8.

may be valuable to many researchers. We hope that CNAplot may help to map the extent of copy-neutral events in various cancer forms and aid in the detection of copy number alterations, in general, by its graphical user interface and transparency. The software is available free of charge and can be downloaded through the provided link.
Fig. 7. A significant change in read depths of genes on chromosome 8 is observed for the mantle cell lymphoma sample. This conclusion is based on bins of 30 genes in each statistical test, using the Mann–Whitney U-test. As of CNAplot version 1.2 the location on the chromosome for each bin is shown.

Funding

Expenses for this project was covered privately by the first author (DK VAT/CVR: 28960638) with research support from the University of Southern Denmark and Odense University Hospital (OUH).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Dina Mohyeldeen, MD, for critical proofreading of the final draft. The first and senior author wish to thank the research colleagues at Hematology–Pathology Research Laboratory, OUH.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.softx.2020.100503.

References


