coriandR (ChrOmosomal abeRration Identifier AND Reporter in R)

Tool documentation

Dr. med. Vera Koch

coriandR (ChrOmosomal abeRration Identifier AND Reporter in R)

Institution

`coriandR` was designed and implemented as a part of a doctoral thesis in the Department of Hematology, Oncology and Immunology at the university hospital at Philipps-University Marburg.

Contributors:

Dr. med. Vera Koch

Dr. rer. nat. Clemens Thölken

Supervisors:

PD Dr. Elisabeth K. M. Mack - Department of Hematology, Oncology and Immunology, Philipps-University Marburg, Baldingerstraße, 35043 Marburg, Germany

Prof. Dr. Ho-Ryun Chung - Institute of Medical Bioinformatics and Biostatistics, Philipps-University Marburg, Hans-Meerwein-Straße 6, 35032 Marburg, Germany

`coriandR` Publications

Koch, V. Optimierung Und Vergleich Bioinformatischer Methoden Zur Kalkulierten Karyotypisierung Der Akuten Myeloischen Leukämie Mittels Next Generation Sequencing. Philipps-Universität Marburg, 2024. https://doi.org/10.17192/z2024.0288.

Tarawneh, T.S.; Rodepeter, F.R.; Teply-Szymanski, J.; Ross, P.; Koch, V.; Thölken, C.; Schäfer, J.A.; Gremke, N.; Mack, H.I.D.; Gold, J.; et al. Combined Focused Next-Generation Sequencing Assays to Guide Precision Oncology in Solid Tumors: A Retrospective Analysis from an Institutional Molecular Tumor Board. Cancers 2022, 14, 4430. https://doi.org/10.3390/cancers14184430.

Kremer J.; Koch V.; Thölken C.; Chung H.-R.; Thiede C.; Neubauer A.; Mack E.K.M. (2021): Risk stratification of acute myeloid leukemia based on calculated karyotyping by next generation sequencing. Gemeinsame Jahrestagung der Deutschen, Österreichischen und Schweizerischen Gesellschaften für Hämatologie und Medizinische Onkologie (Hybrid-Kongress), 01.-04. Oktober 2021: Abstracts. In: Oncol Res Treat 44 (Suppl. 2), Artikel V592, S. 1–335. DOI:10.1159/000518417.

`coriandR` Licence

Table of contents

1. Introduction to `coriandR` and background

Cancer is a prevalent disease that can be treated with a standard first-line medical treatment if the cancer type is common and well-understood. An increasing number of tumours is characterised based on mutation profiles (Vogelstein et al. 2013). For some diseases like acute myeloid leukemia, genetical alterations play an important role in classification and are substantially involved in patients prognosis (Döhner 2022).

At the molecular tumour board at the university hospital at Philipps-University Marburg, medical treatment options and clinical trial participation opportunities for patients with advanced or rare tumours are adressed based on molecular profiling results. Here, the frequently used methods are molecular testing by next-generation sequencing (NGS) including gene panels for the detection of short-sequence variants and copy-number alterations as well as gene fusion panels. Immunohistochemistry for microsatellite instability and PD-L1 expression complement NGS (Tarawneh 2022). Further analysis contains an ultra-low-coverage ($<$ 0.2x) whole-genome sequencing for detection of additional copy-number alterations outside the panel's target regions with `coriandR` (ChrOmosomal abeRration Identifier AND Reporter in R, Koch 2024).

`coriandR` can be used for estimation of calculated karyotype and copy number variations in hamatological malignances and solid tumours in research. `coriandR` requires unpreprocessed paired-end samples in FASTQ file format. For statistical testings, it is nessesary to generate a Panel of Normals (PON) from sequencing data. The PON samples come from the same tissue type (blood or histological tumor-free tissue samples) and were processed under the same conditions as the tumour samples and have a normal karyotype. Estimation of the calculated karyotype for the tumour samples is based on a two-tailed normal distribution test. The mapping statistics and the results of calculated karyotyping are displayed in a PDF report with genome overview plots, table with calculated karyotype and estimation of CNVs in deviating regions as well as in chromosome overview plots.

2. Methods

2.1 Tools

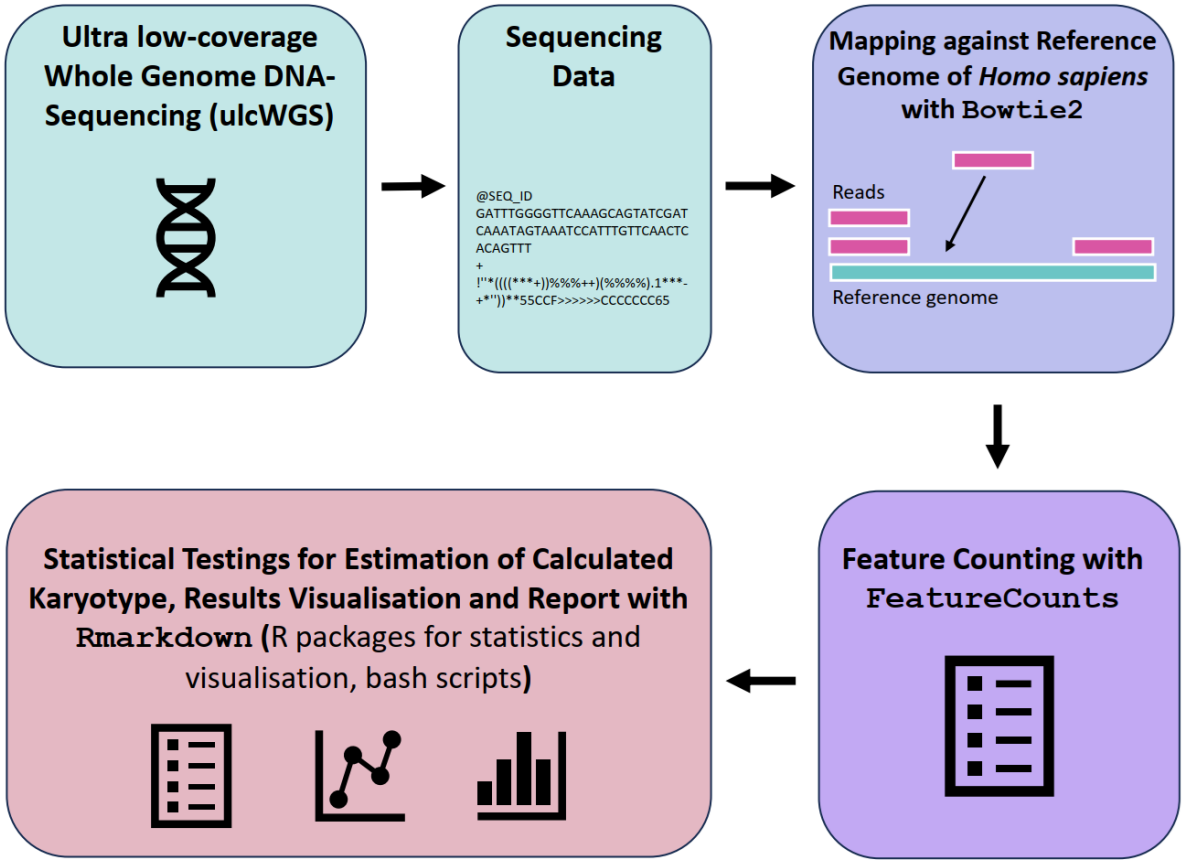


Figure 1: The steps of sequencing data processing with `coriandR` for estimation of calculated karyotype and copy number variations

2.1.1 Bowtie2 `Bowtie2` (version 2.5.2, Langmead und Salzberg 2012) was used to align the sample genome to human reference genome in paired-end mode. Here, we used the version GRCh38.p13 of human reference genome (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.39/, accessed November, 19 2023). For the first analysis, a reference genome index must be build with `bowtie2-build` (see `Bowtie2` manual under http://bowtie-bio.sourceforge.net/manual.shtml, accessed November, 19 2023).

2.1.2 SAMtools `SAMtools` (Li et al. 2012) in version 1.20, which implements various utilities for post-processing alignments in the SAM and BAM formats, was used to convert SAM files to BAM and sort them by genomic coordinates. We used a script `sam2bam` (https://github.com/thoelken/bioinfo-toolbox, accessed November, 19 2023) by Dr. Clemens Thölken for sorting and converting from SAM to BAM.

2.1.3 FeatureCounts `FeatureCounts` (Liao et al. 2014) in version 2.0.6 was used to count reads in non-overlapping predefined bins of 1.000.000 bp length according to the annotation file `bins.gtf`.

2.1.4 R `RStudio` in version 2024.09.0 and programming language `R` in version 4.3.3 were used to create `R Markdown` scripts for statistical analysis of the `FeatureCounts` output table and plotting. The packages `base` (version 4.3.3), `datasets` (version 4.3.3), `methods` (version 4.3.3), `stats` (version 4.3.3), `utils` (version 4.3.3)

were used for data structures and statistical testings, `graphics` (version 4.3.3), `grDevises` (version 4.3.3) for visualisation, `knitr` (Xie 2014, version 1.48), `tinytex` (Xie 2019, version 0.53) and `rmarkdown` (Allaire et al. 2014, version 2.28) to create a report in PDF file format from a `RMarkdown` script.

2.2 `coriandR` Methods

`coriandR` can be started in two modes - one for generation of a Panel of Normals with the `bash` script `pon_creator.sh` and one for calculated karyotyping with the `bash` script `coriandr.sh`.

2.2.1 Panel of Normals A panel of normals (PON) contains sequencing data from individuals with a normal karyotype who are representative for the analysed population. Multiple samples are used to compensate for or identify technical artefacts and normal biological variability. It is also important that the samples are obtained and processed under the same conditions (same sample preparation methods like DNA extraction and sequencing technology) as the test samples. Using the read depth method, distribution of the reads in samples can be calculated for PON and test samples, which can be used to estimate aberrations like CNVs.

The script `pon_creator.sh` calls up the `RMarkdown` script `report.create.a.pon.and.stats.Rmd` to collect all paired-end PON samples in FASTQ file format in the directory `sample.pon/` and to create a PON with the name `sample.pon`. `pon.creator.sh` can be used with the following command:

```
bash pon_creator.sh sample.pon sample.pon/ sample.pon/pon.meta.csv
```

To generate a Panel of Normals, the reads of all PON samples are summed up after counting per bin with `FeatureCounts`. The reads on autosomal chromosomes are used for the analysis of both genders, while the reads on the X and Y chromosomes are only used as a statistical reference for the same gender. In the next step, the PON is normalised by median per library size and per megabase for the whole genome, including X and Y chromosomes.

A possible error source in NGS is the distortion of the gc content, which is an irregularity between the proportion of guanine (g) and cytosine c bases in a genomic region and means the number of fragments in a region (Dohm et al. 2008). The gc content is particularly high in gene-rich regions and therefore not evenly distributed across the entire genome (Dohm et al. 2008). In the at-rich regions, the coverage (sequencing depth) increases with increasing gc content, while coverage in gc-rich regions decreases with increasing gc content due to the greater stability of the gc pairs to higher temperatures (Borisova et al. 1993) and, consequently, the faster denaturation of the at-rich regions. The maximum genomic coverage is observed in the regions with 0.4 to 0.55 of the gc content (Benjamini and Speed 2012). For `coriandR`, the range from 0.35 to 0.6 was defined as the optimal comparable range for the gc content which was calculated using `Bedtools` (Quinlan 2014). Bins with gc content less than the 0.275 percentile genome-wide assumed normal distribution were considered to be extreme and masked from the PON. Additionally, bins with the highest 1% variance genome-wide and those that meet both requirements were masked from the PON.

A report for creating a panel of normals is then generated with the plots for sequencing depth of individual PON subjects, normalizing PON with a scaling factor equal to 2 (ploidy) for autosomes and equal to 1 for X and Y in male individuals and PON chromosomal overview with masking bins.

2.2.2 Calculated Karyotyping In the first step of calculated karyotyping, the sequencing data of a tumour sample are normalised by the median sequencing depth per bin. Thereafter, we used standardisation with calculation of the pseudo z-values of the distribution of the bins. Later they will be compared with the theoretical normal distribution.

$Z_i = \frac{x_i - 1}{\sigma_(X)}$

where x_i represents the reads in the bin, $\sigma_(X)$ is the standard deviation of the bin estimated from the PON, Z_i is the pseudo z-score of the bin.

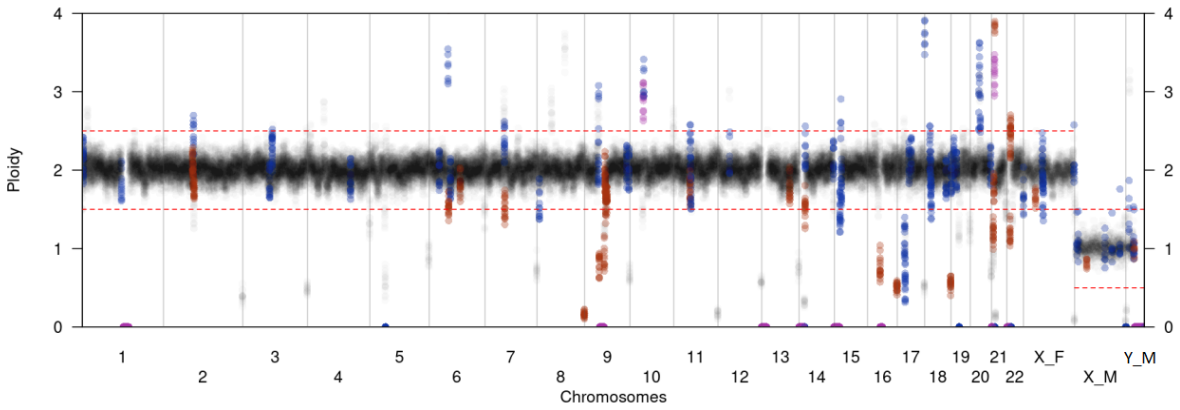


Figure 2: Visualisation of masked bins in a normalised PON with 19 samples. Bins with extremely low gc-content are marked in red, bins with a high variance in blue, bins that meet both requirements in purple. The bins without deviations in variance or gc content are displayed in black.

In addition, we tested pseudo z-scores against a normal distribution with parameters of the PON in a two-tailed test. The obtained p-values were adjusted using the Benjamini-Hochberg method (Benjamini and Hochberg 1995) in control of the false discovery rate.

$Z_i \sim N(\overline{Z_i^{PON}}, \overline{\sigma_i^{PON}})$,

where Z_i is the z-score of the bin in the sample derived from a normal distribution, N, with mean z-score $\overline{Z_i^{PON}}$ and variance of this bin in the PON $\overline{\sigma_i^{PON}}$.

In consideration of the adjusted p-values, the deviating bins are calculated. A deletion in a bin is detected if the normalised value for that bin is below the median for all normalised PON samples with a significance level of $\alpha = 0.05$. An amplification leads to a value above the median.

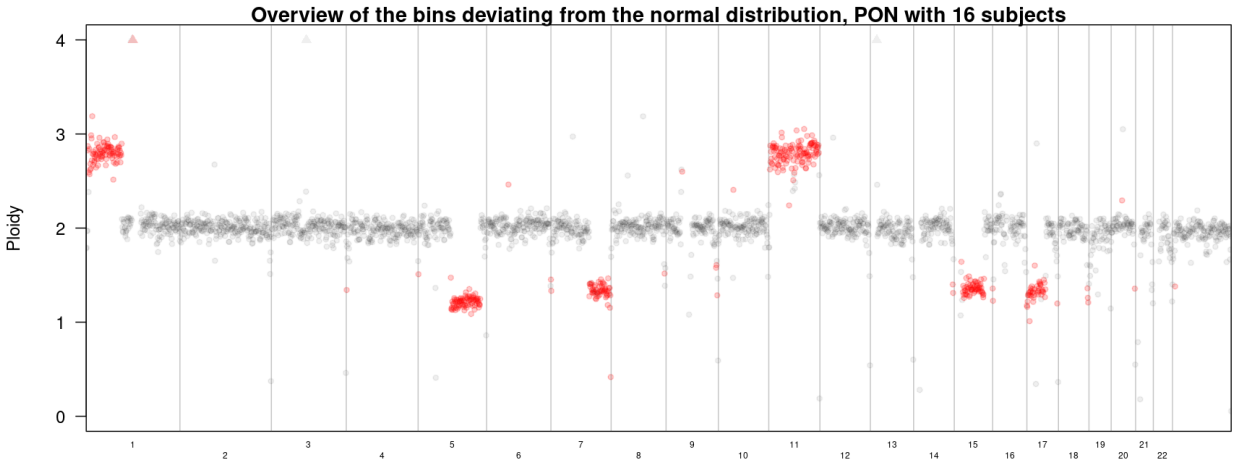
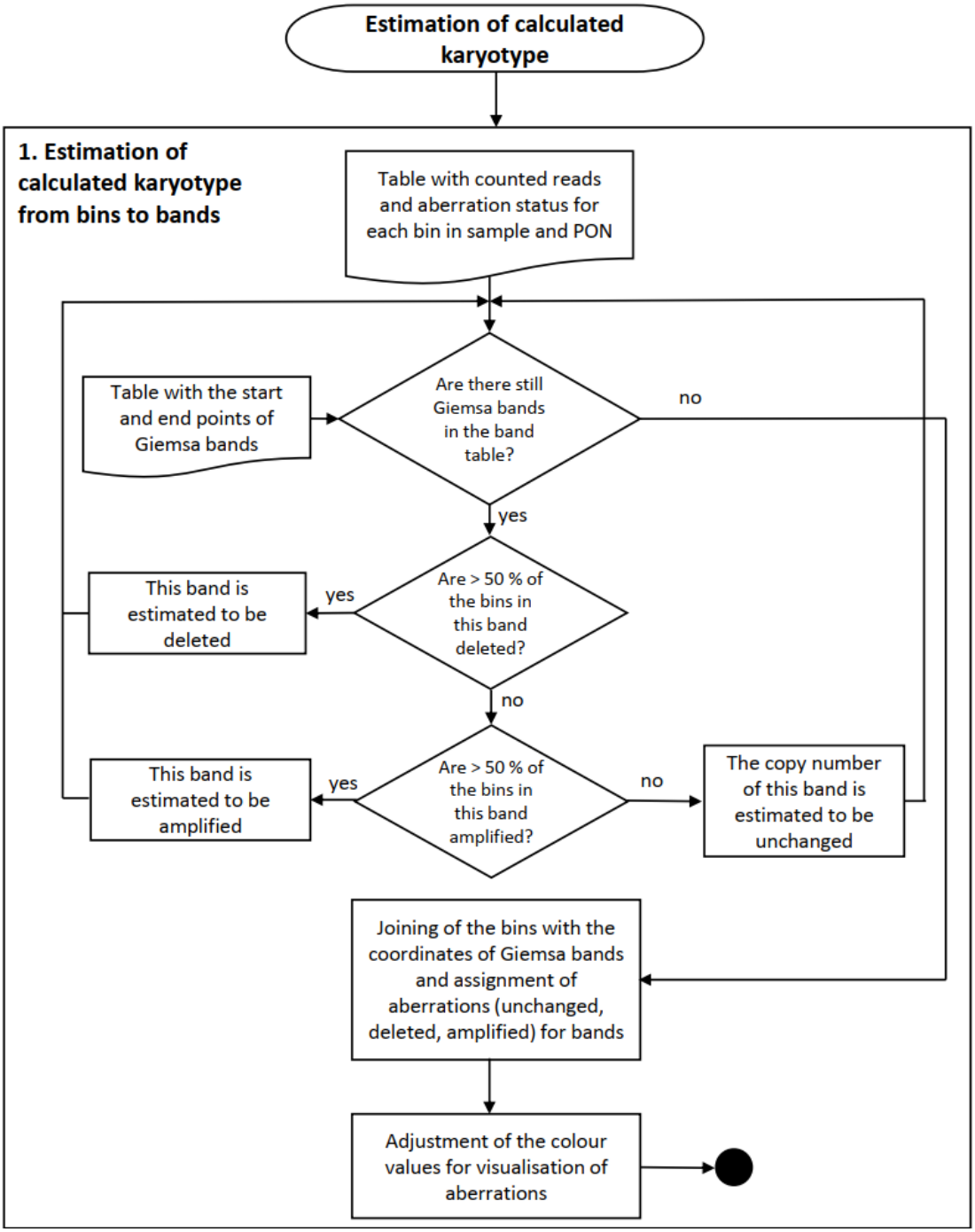


Figure 3: Visualisation of a calculated karyotype (on the level of bins) of a sample with a complex karyotype. Bins with deviation from PON are marked in red.

For the estimation of the calculated karyotype, we used the genomic coordinates of the G-bands from the cytogenetic landmarks (Cheung et al. 2001). Since the G-bands are longer than one megabase we make sure that the start and end points of the bins overlap with the start and end positions of the bands as much as possible. We determined the start point of a bin that was as close as possible to the start position of a band.

A bin with the closest possible to end point must correspond the end point of a band. The masked bins were not included here. If half of all bins in a band have the status “deleted”, that band is also considered to be deleted. The chosen cut-off of half of the bins also applies to the band amplification. All other bins are classified as non-aberrant.

To simplify the interpretation of the report for `coriandR` users, we decided to group the bands together according to the ISCN nomenclature (ISCN 2020). To merge the bands, we make sure whether they are on the same chromosome and contiguous and whether they have the same aberration status. Examining the coordinates for the G-bands we then define the loss or amplification of whole chromosome arms or even whole chromosomes. If more than 80% of the G-bands of a chromosome arm or a whole chromosome have aberrations of the same type, we call it deleted or amplified.


Estimation of calculated karyotype
1. Estimation of calculated karyotype from bins to bands
Table with counted reads and aberration status for each bin in sample and PON
Table with the start and end points of Giemsa bands
Are there still Giemsa bands in the band table?
no
yes
Are > 50 % of the bins in this band deleted?
yes
This band is estimated to be deleted
no
Are > 50 % of the bins in this band amplified?
yes
This band is estimated to be amplified
no
The copy number of this band is estimated to be unchanged
Joining of the bins with the coordinates of Giemsa bands and assignment of aberrations (unchanged, deleted, amplified) for bands
Adjustment of the colour values for visualisation of aberrations

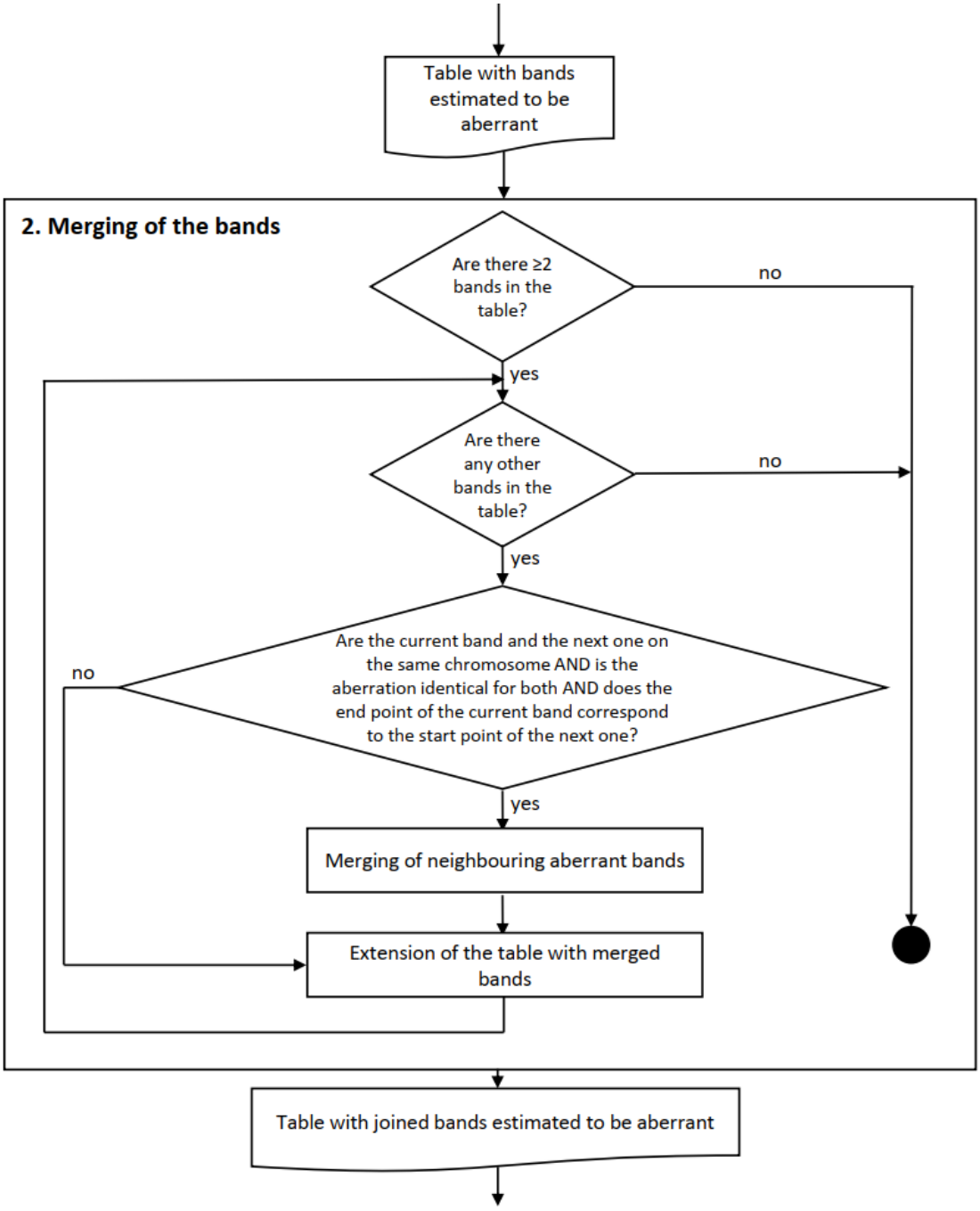
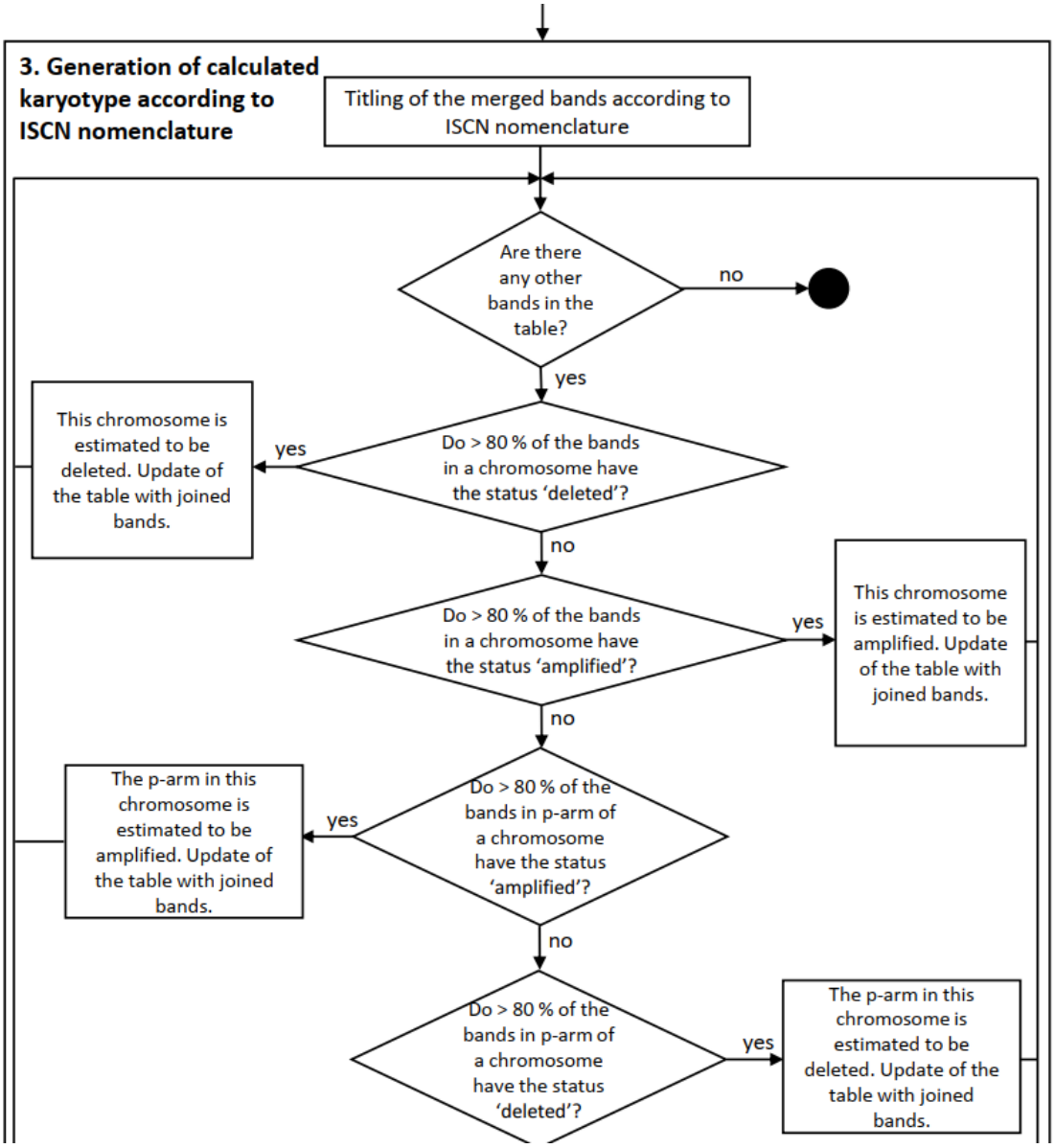


Table with bands estimated to be aberrant
2. Merging of the bands
Are there ≥2 bands in the table?
no
yes
Are there any other bands in the table?
no
yes
Are the current band and the next one on the same chromosome AND is the aberration identical for both AND does the end point of the current band correspond to the start point of the next one?
no
yes
Merging of neighbouring aberrant bands
Extension of the table with merged bands
Table with joined bands estimated to be aberrant


3. Generation of calculated karyotype according to ISCN nomenclature
Titling of the merged bands according to ISCN nomenclature
Are there any other bands in the table?
no
yes
Do > 80 % of the bands in a chromosome have the status 'deleted'?
yes
This chromosome is estimated to be deleted. Update of the table with joined bands.
no
Do > 80 % of the bands in a chromosome have the status 'amplified'?
yes
This chromosome is estimated to be amplified. Update of the table with joined bands.
no
Do > 80 % of the bands in p-arm of a chromosome have the status 'amplified'?
yes
The p-arm in this chromosome is estimated to be amplified. Update of the table with joined bands.
no
Do > 80 % of the bands in p-arm of a chromosome have the status 'deleted'?
yes
The p-arm in this chromosome is estimated to be deleted. Update of the table with joined bands.

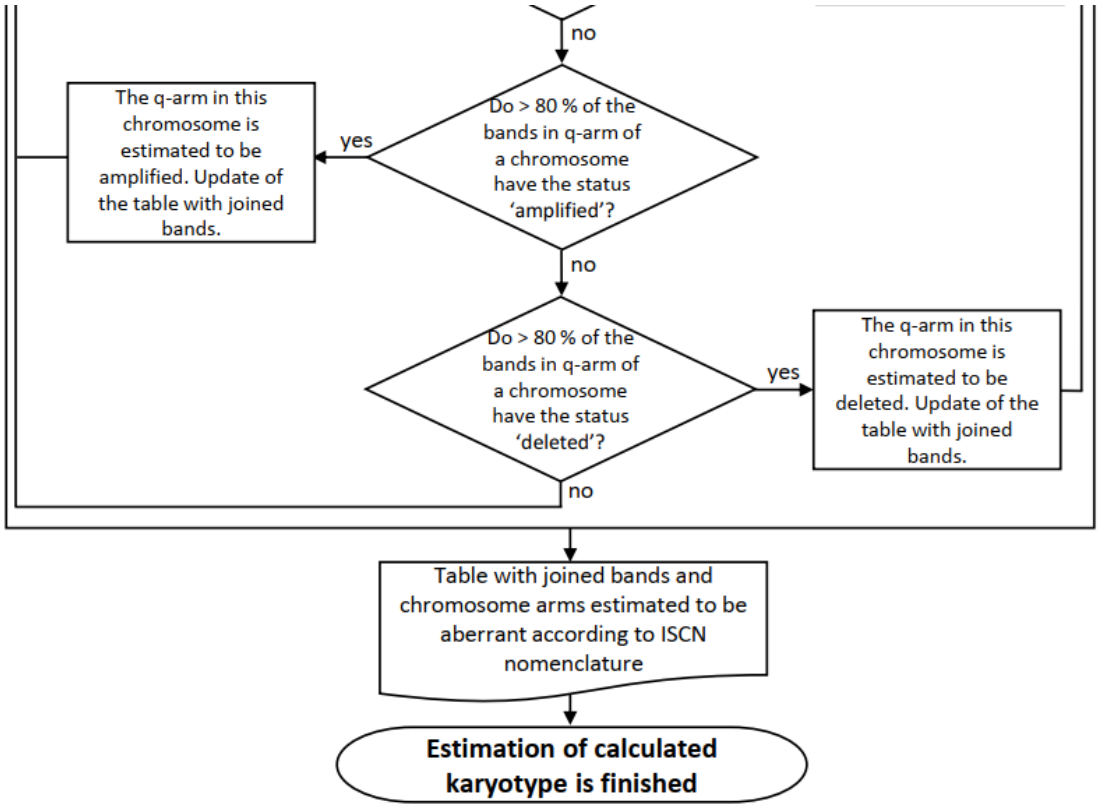


Figure 4: Visualisation of the calculated karyotyping process according to the ISCN nomenclature in form of a flowchart. The creation of calculated karyotype is divided into three steps ('Estimation of calculated karyotype from bins to bands', 'Merging of bands', 'Generation of calculated karyotype according to ISCN nomenclature'), characterised by the rectangles. The ellipses represent the start and end points of the calculated karyotyping process. The arrows visualise data flow. The rectangles with wavy lower edges are the tables that are used for estimations. The rhombs symbolise a condition that can be met (arrow 'yes') or not (arrow 'no'). The black circles represent the end points of each step.

After estimation of deviating bins (deletions or amplifications) with bin size of 1.000.000 bp, genes of interest located in deviating chromosomal regions can be estimated. `coriandR` contains a genes list with cancer driver genes (Bailey et al. 2018) and genes that play an important role in disease development of Acute Myeloid Leukemia (Papaemmanuil et al. 2016), since `coriandR` was originally developed for estimation of calculated karyotype in acute myeloid leukemia samples.

The DNPM (German Network for Personalized Medicine, ger. *Deutsches Netzwerk für Personalisierte Medizin*, https://dnpm.de/, accessed September 23, 2024) created a list of genes of interest based on the research in germline tumour-detected variants in 49,264 cancer patients (Kuzbari et al. 2023) and genetic dysfunction across all human cancers (Sondka et al. 2018) as well as the database for FDA-recognised human genetic variants `OnkoKB` (Sarah et al. 2024, https://www.oncokb.org/, accessed September 23, 2024) which can be used in the search for potential targets for the treatment of patients with solid tumours. The choice of genes list depends on the `coriandR` using mode: `standard` or `solid` by call up the programme in the command line.

The table with genes of interest contains the information about gene ID, the chromosome on which the gene is located, the chromosomal coordinates as well as the type of CNVs: deletion or amplification.

Finally, Giemsa bands and bins of each chromosome in sample are visualised with a chromosome plot. Here, black data points represent the unchanged bins on a chromosome, their position is normalised to the ploidy of the chromosome. The thin grey lines visualise the error bars. The G-bands are located in the lower part of the chromosome plots.

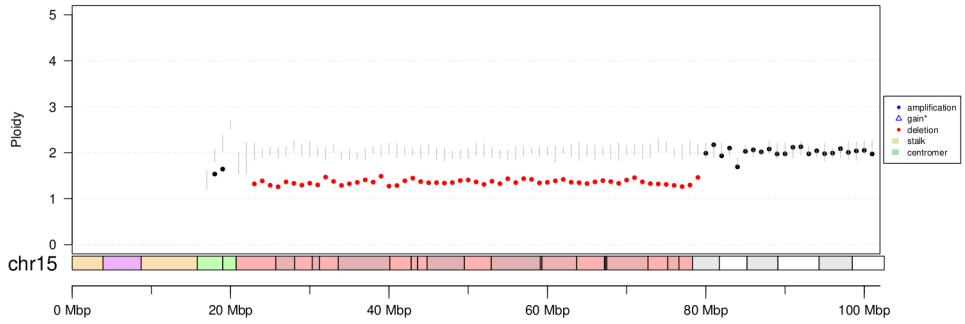


Figure 5: Plot of the chromosome 5 in a sample with a complex karyotype, especially with a 5q-deletion. Each data point in black represents a bin in chromosome 5, normalised to the ploidy of 2. The differently grey-coloured G-bands are located in the lower part of the figure. The chromosomal coordinates are displayed in megabase size. The legend refers to the colours of the G-bands in the lower part of the figure and bands which are estimate to be deleted (red colouring of the bands) or amplified (blue colouring of the bands). The centromeres are shown in green. 'stalk' areas are marked in yellow and represent a connection between two chromosomes or chromosome arms that can occur during mitosis or meiosis. 'gain' indicates an amplification with an estimated copy number of more than 5.

3. coriandR Installation

3.1 Native Installation for UNIX/Linux OS

To run calculated karyotyping with coriandR, you need following tools to be installed on your computer with UNIX/Linux OS:

- Bowtie2 (version 2.3.5.1, Langmead und Salzberg 2012)
- SAMtools (Li et al. 2012) in version 1.20
- FeatureCounts (Liao et al. 2014) in version 2.0.6
- RStudio in version 2024.09.0 and programming language R in version 4.3.3 with packages base (version 4.3.3), datasets (version 4.3.3), methods (version 4.3.3), stats (version 4.3.3), utils (version 4.3.3), graphics (version 4.3.3), grDevises (version 4.3.3), knitr (Xie 2014, version 1.48), tinytex (Xie 2019, version 0.53) and rmarkdown (Allaire et al. 2014, version 2.28).

1. Install all dependencies.
2. Clone the coriandR git repository.
3. Open terminal and change into coriandR directory with `cd [path to coriandR folder]`.
4. Make sam2bam.sh script executable with `chmod +x sam2bam.sh`.
5. You can start calculated karyotyping with coriandR with `bash coriander.sh [Sample ID] [path to patient.meta.tsv] [input/[read_1].fastq.gz] [input/[read_2].fastq.gz] [Usage mode: 'standard' or 'solid']` (see 4.3 How to Generate a Calculated Karyotyping Report of a Tumour Sample).

3.2 Installation as Docker Container

Calculated karyotyping

1. Clone the coriandR git repository.
2. Open terminal and change into coriandR directory with `cd [path to coriandR folder]`.
3. Build a Docker image of coriandR with `sudo docker build .. Docker` will use the Dockerfile in coriandR root directory to install all dependencies and tools. This process will take some time, but you only need to build the image once.
4. Check the ID of your image with `sudo docker image ls`.
5. Copy IMAGE ID of coriandR image.
6. Create a folder input in coriandR directory and copy your FASTQ files (read 1 and 2) into it.
7. Finally, you can use coriandR container for calculated karyotyping with the command `sudo docker run -v .:/coriandr [IMAGE ID] [Sample ID] [path to patient.meta.tsv] [input/[read_1].fastq.gz] [input/[read_2].fastq.gz] [Usage mode: 'standard' or 'solid']`

Generation of a new Panel of Normals

1. Clone the coriandR git repository.
2. Save the Dockerfile for calculated karyotyping in the root directory somewhere else. Put Dockerfile from pon_creator_docker folder into the coriandR root directory.
3. Open terminal and change into coriandR directory with `cd [path to coriandR folder]`.
4. Build a Docker image of coriandR with `sudo docker build .. Docker` will use the Dockerfile to install all dependencies and tools. This process will take some time, but you only need to build an image once.
5. Check the ID of your image with `sudo docker image ls`.
6. Copy IMAGE ID of coriandR image.
7. Create a folder input in coriandR directory and copy your FASTQ files (read 1 and 2) for PON into it + tabl with meta information.
8. Finally, you can use coriandR for generation of a new Panel of Normals with the command `sudo docker run -v .:/coriandr [IMAGE ID] [PON ID] input/ [path to pon.meta.csv]`

4. coriandR Usage

4.1 What to adapt for your analysis

You need to change the paths in the file `config.txt` to absolute paths in your file system:

- `index` which means the index of the reference genome in `Bowtie2`. You need to create an index of your version of reference genome first.
- `gtf` where you need to set the path to the bin annotation file in your cloned repository like `/home/user/coriandR/tables/genome/bins.gtf`.
- `pon` where you need to set the path to the Panel of normals (PON) table like `/home/user/coriandR/tables/pon.tsv`. In 4.2 How to Create a Panel of Normals, you can learn how to create a new PON.
- `gccontent` where you need to set the path to the gc content file of the human reference genome in you cloned repository like `/home/user/coriandR/tables/GRCh38.p13.genome.1M.nucl`.

To use `coriandR` in `Docker`, you need to copy the `Bowtie2` index into the `coriandR` directory, like `index="tables/genome/bowtie_index/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index"`. Likewise for `Bowtie2` index, `coriandR` in `Docker` will only accept relative paths to needed tables and sources: `gtf="tables/genome/bins.gtf"`, `pon="tables/pon.tsv"`, `gccontent="tables/GRCh38.p13.genome.1M.nucl"`.

The file `patient.meta.tsv` contains information about the `sampleID` and patientsgender‘. You need to change them according to your data.

4.2 How to Create a Panel of Normals

1. Prepare a table with meta data: this table contains columns “sample” and “gender”. The samples with IDs like `sample1.fastq.gz` and the genders of the samples (`M`/`F`) separated by `,` will be used for further calculations. You can use the premade table `pon.meta.csv`.
2. Create a folder with only the paired-end FASTQ PON samples and table with meta information.
3. Open the `coriandR` folder in terminal.
4. To start the tool `pon.creator.sh` enter the following mandatory parameters in terminal:

- name of the new panel of normals;
- path to folder with paired-end FASTQ files;
- path to meta table (gender table).

Example use:

```
bash pon_creator.sh pon data/sample.pon/ data/sample.pon/pon.meta.csv
```

4.3 How to Generate a Calculated Karyotyping Report of a Tumour Sample

You can use `coriandR` native or in a container. The script `coriandr.sh` requires 5 mandatory arguments.

1. Sample ID
2. Path to sample meta file
3. FASTQ1: path to FASTQ file with read 1
4. FASTQ2: a path to FASTQ file with read 2
5. Usage mode: ‘standard’ for displaying of all aberrations or ‘solid’ for estimation of only high level amplifications (> 5 copies) and deletions (< 0.5 copies)

Example use:

```
bash coriander.sh 101010 /data/101010.meta.tsv /data/Fastq/101010_R1.fastq /data/Fastq/101010_R2.fastq
standard
```

5. Limitations of coriandR

5.1 Read alignment bias due to ultra-low-coverage:

We developed a tool for estimation of calculated karyotype in ultra-low-coverage (< 0.2x) whole-genome sequencing (ulcWGS) data and used the read depth method for statistical testings since the low coverage would not provide us with enough information to allow the usage of other methods like breakpoints analysis. Read alignment can cause a bias, where repetitive regions in the reference genome could lead to ambiguous alignment of a significant number of reads (Pirooznia et al. 2015). To avoid it, please pay attention to `Bowtie2` logs after mapping (file `logs.bowtie.txt` in your output folder) to identify samples with a low percentage of uniquely mapped reads.

5.2 Absolute ploidy not reported:

It is not possible to make statements about the absolute ploidy of the samples (complete duplications or triplications of whole genomes), although genome duplications can occur early in the oncogenesis in more than 30 % of human tumours (Prasad et al. 2022). By assuming the ploidy of special genomic regions for normalisation (ploidy of 1 for male gonosomes, ploidy of 2 - for autosomes and the X chromosome of women), a complete genome multiplication would result in ploidy for somatic regions as ploidy of 2 and on the male gonosomes - of 1, in report of `coriandR`.

5.3 Only copy number variations (CNVs) reported:

The structural rearrangements can not be reported with `coriandR` due to inadequate information about genomic breakpoints because of chosen very low coverage.

5.4 No statements about the clonal heterogeneity in a sequenced sample:

Several tumour clones can coexist simultaneously in tumour tissue or bone marrow and are not recognised with `coriandR` or can lead to errors in statistical testings. If your chromosome plots do not have biologically explainable distribution of reads like in the example below, it may be helpful check the existence of tumour clones with other methods.

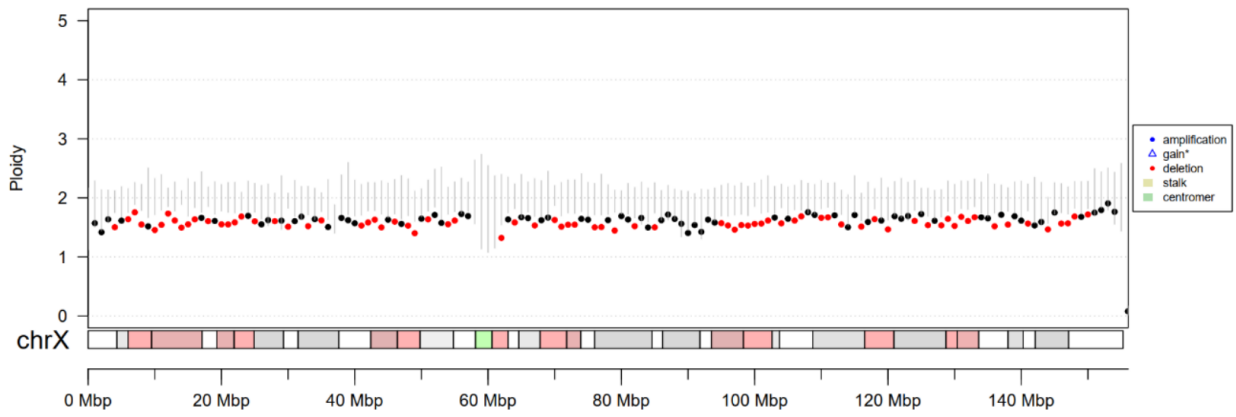


Figure 6: X chromosome plot in a sample with multiple tumour clones with karyotype 45, X, -X; 46, XX.

5.5 For research only:

`coriandR` was developed for evaluation of karyotype as a part of a combined approach which includes gene panels for the detection of short-sequence variants and copy-number alterations as well as gene fusion panels. Estimation of calculated karyotype with read depth approach in ulcWGS do not provide medical specialists with enough information for diagnosis or treatment options.

6. References

Allaire J, Xie Y, Dervieux C, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2024). *rmarkdown: Dynamic Documents for R*. R package version 2.28, https://github.com/rstudio/rmarkdown.

Bailey, Matthew H.; Tokheim, Collin; Porta-Pardo, Eduard; Sengupta, Sohini; Bertrand, Denis; Weerasinghe, Amila et al. (2018): Comprehensive Characterization of Cancer Driver Genes and Mutations. In: Cell 173 (2), 371-385.e18. DOI: 10.1016/j.cell.2018.02.060

Benjamini, Yoav; Hochberg, Yosef (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. In: Journal of the Royal Statistical Society: Series B (Methodological) 57 (1), S. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Benjamini, Yuval; Speed, Terence P. (2012): Summarizing and correcting the GC content bias in high-throughput sequencing. In: Nucleic Acids Research 40 (10), e72. DOI: 10.1093/nar/gks001.

Borisova, O. F.; Shchyolkina, A. K.; Chernov, B. K.; Tchurikov, N. A. (1993): Relative stability of AT and GC pairs in parallel DNA duplex formed by a natural sequence. In: FEBS letters 322 (3), S. 304–306. DOI: 10.1016/0014-5793(93)81591-m.

Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, Furey TS, Kim UJ, Kuo WL, Olivier M et al. Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature. 2001 Feb 15;409(6822):953-8. PMID: 11237021

Dohm, Juliane C.; Lottaz, Claudio; Borodina, Tatiana; Himmelbauer, Heinz (2008): Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. In: Nucleic Acids Research 36 (16), e105. DOI: 10.1093/nar/gkn425.

Döhner, Hartmut; Wei, Andrew H.; Appelbaum, Frederick R.; Craddock, Charles; Dinardo, Courtney D.; Dombret, Hervé et al. (2022): Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. In: Blood 140 (12), S. 1345–1377. DOI: 10.1182/blood.2022016867.

International Standing Committee on Human Cytogenomic Nomenclature and McGowan-Jordan, J. and Hastings, R.J. and Moore, S. (2020): ISCN 2020: An International System for Human Cytogenomic Nomenclature. In: An International system for human cytogenetic nomenclature. Karger Publishers. ISBN: 9783318067064.

Koch, V. Optimierung Und Vergleich Bioinformatischer Methoden Zur Kalkulierten Karyotypisierung Der Akuten Myeloischen Leukämie Mittels Next Generation Sequencing. Philipps-Universität Marburg, 2024. https://doi.org/10.17192/z2024.0288.

Kuzbari Z, Bandlamudi C, Loveday C, Garrett A, Mehine M, George A, Hanson H, Snape K, Kulkarni A, Allen S, Jezdic S, Ferrandino R, Westphalen CB, Castro E, Rodon J, Mateo J, Burghel GJ, Berger MF, Mandelker D, Turnbull C. Germline-focused analysis of tumour-detected variants in 49,264 cancer patients: ESMO Precision Medicine Working Group recommendations. Annals of Oncology. 2023 Mar 1;34(3):215-227. doi: 10.1016/j.annonc.2022.12.003.

Langmead B, Salzberg SL (2012): Fast gapped-read alignment with Bowtie 2. In: Nat Methods 9 (4), S. 357–359. DOI: 10.1038/nmeth.1923.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PMID: 19505943; PMCID: PMC2723002.

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923-30. doi: 10.1093/bioinformatics/btt656. Epub 2013 Nov 13. PMID: 24227677.

Papaemmanuil, Elli; Gerstung, Moritz; Bullinger, Lars; Gaidzik, Verena I.; Paschka, Peter; Roberts, Nicola D. et al. (2016): Genomic Classification and Prognosis in Acute Myeloid Leukemia. In: The New England journal of medicine 374 (23), S. 2209–2221. DOI: 10.1056/NEJMoa1516192.

Pirooznia, Mehdi; Goes, Fernando S.; Zandi, Peter P. (2015): Whole-genome CNV analysis: advances in computational approaches. In: Frontiers in genetics 6, S. 138. DOI: 10.3389/fgene.2015.00138.

Prasad, Kavya; Bloomfield, Mathew; Levi, Hagai; Keuper, Kristina; Bernhard, Sara V.; Baudoin, Nicolaas C. et al. (2022): Whole-Genome Duplication Shapes the Aneuploidy Landscape of Human Cancers. In: Cancer research 82 (9), S. 1736–1752. DOI: 10.1158/0008-5472.CAN-21-2065.

Quinlan, Aaron R. (2014): BEDTools: The Swiss-Army Tool for Genome Feature Analysis. In: Current protocols in bioinformatics 47, 11.12.1-34. DOI: 10.1002/0471250953.bi1112s47.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sarah P. Suehnholz, Moriah H. Nissan, Hongxin Zhang, Ritika Kundra, Subhiksha Nandakumar, Calvin Lu, Stephanie Carrero, Amanda Dhaneshwar, Nicole Fernandez, Benjamin W. Xu, Maria E. Arcila, Ahmet Zehir, Aijazuddin Syed, A. Rose Brannon, Julia E. Rudolph, Eder Paraiso, Paul J. Sabbatini, Ross L. Levine, Ahmet Dogan, Jianjiong Gao, Marc Ladanyi, Alexander Drilon, Michael F. Berger, David B. Solit, Nikolaus Schultz, Debyani Chakravarty; Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. Cancer Discov 1 January 2024; 14 (1): 49–65. https://doi.org/10.1158/2159-8290.CD-23-0467

Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018 Nov;18(11):696-705. doi: 10.1038/s41568-018-0060-1. PMID: 30293088; PMCID: PMC6450507.

Tarawneh TS, Rodepeter FR, Teply-Szymanski J, Ross P, Koch V, Thölken C, Schäfer JA, Gremke N, Mack HID, Gold J, et al. Combined Focused Next-Generation Sequencing Assays to Guide Precision Oncology in Solid Tumors: A Retrospective Analysis from an Institutional Molecular Tumor Board. Cancers. 2022; 14(18):4430. https://doi.org/10.3390/cancers14184430

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013 Mar 29;339(6127):1546-58. doi: 10.1126/science.1235122. PMID: 23539594; PMCID: PMC3749880.

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Xie Y (2019). "TinyTeX: A lightweight, cross-platform, and easy-to-maintain LaTeX distribution based on TeX Live." *TUGboat*, *40*(1), 30-32. https://tug.org/TUGboat/Contents/contents40-1.html.