MetaCNV - a consensus approach to infer accurate copy numbers from low coverage data

Stefanie Friedrich (Stefanie.Friedrich@scilifelab.se)
Stockholms Universitet https://orcid.org/0000-0002-3889-5589
Remus Barbulescu
Stockholms Universitet
Thomas Helleday
Karolinska Institutet
Erik LL Sonnhammer
Stockholms Universitet

Software

Keywords: Human genome analysis, copy number calling, low coverage data

Posted Date: February 4th, 2020

DOI: https://doi.org/10.21203/rs.2.15757/v2

Version of Record: A version of this preprint was published at BMC Medical Genomics on June 1st, 2020.
See the published version at https://doi.org/10.1186/s12920-020-00731-y.

MetaCNV - a consensus approach to infer accurate copy numbers from low coverage data

Stefanie Friedrich[1], Remus Barbulescu[1], Thomas Helleday[2], Erik LL Sonnhammer[1]

[1] Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Box 1031, 17121 Solna, Sweden

[2] Science for Life Laboratory, Department of Oncology-Pathology, Karolinska Institutet, Solna, Sweden

Abstract

Background: The majority of copy number callers requires high read coverage data that is often achieved with elevated material input, which increases the heterogeneity of tissue samples. However, to gain insights into smaller areas within a tissue sample, e.g a cancerous area in a heterogeneous tissue sample, less material is used for sequencing, which results in lower read coverage. Therefore, more focus needs to be put on copy number calling that is sensitive enough for low coverage data.

Results: We present MetaCNV, a copy number caller that infers reliable copy numbers for human genomes with a consensus approach. MetaCNV specializes in low coverage data, but also performs well on normal and high coverage data. MetaCNV integrates the results of multiple copy number callers and infers absolute and unbiased copy numbers for the entire genome. MetaCNV is based on a meta-model that bypasses the weaknesses of current calling models while combining the strengths of existing approaches. Here we apply MetaCNV based on ReadDepth, SVDetect, and CNVnator to real and simulated datasets in order to demonstrate how the approach improves copy number calling.

Conclusions: MetaCNV, available at https://bitbucket.org/sonnhammergroup/metacnv, provides accurate copy number prediction on low coverage data and performs well on high coverage data.

Keywords: Human genome analysis, copy number calling, low coverage data

Background

An important aspect of genome analysis is the study of genetic alterations between individuals in a cohort, or between samples from one individual, for instance to understand cancer progression. One type of genetic alteration is copy number variation (CNV), which describes the fact that a segment of a genome, for example spanning one or more genes, is amplified or deleted.

Up to 10% of the human genome has been estimated to contribute to CNVs, and abnormal copy numbers have been linked to mutation-prone diseases like cancer [1]. Knowledge about CNVs is not only crucial to understanding such diseases, especially with regards to their evolution, but also their effects on various phenotypes.

Next-generation sequencing has revolutionized the possibilities to study inter- and intra-individual genome alterations. For accurate CNV analysis, typically tissue samples with 200-500 ng of DNA are required for sequencing with high coverage [2]. However, new generations of sequencing techniques arise and new approaches to sequence even single cell genomes are being developed that only require 50 ng of DNA [2–4]. Copy number callers applicable to sequenced samples with very small amounts of DNA would create the possibility to investigate CNVs of smaller areas within a larger tissue sample, such as regions microdissected by laser, and thus offer spatial information of CNVs within formerly bulk-sequenced samples.

CNVs can be detected experimentally with various assays, such as comparative genomic hybridization (CGH, e.g. bacterial artificial chromosomes (BAC) array), fluorescent in situ hybridization (FISH), and genotyping array. Another option is calling CNVs from genome sequences,

although calling correct copy numbers this way is a challenging task. To develop a solution, four different approaches have emerged, each built on certain assumptions and each with advantages and disadvantages. [5, 6]

Callers purely based on a read coverage approach (i) predict copy number changes using the read coverage of a segment of the genome relative to the coverage of the whole genome. To achieve good results for both deletions and amplifications, this approach requires high sequencing depth (Figure S1). Further, short CNVs are often missed. The majority of existing copy number callers apply this concept, for example ReadDepth [7], CNV-seq [8], cn.MOPS [9], Control-FREEC [10], and CNVnator [11].

Callers based on a paired-end mapping, also termed read-pair, approach (ii) predict copy numbers based on changes in the insert size of paired-end reads. This approach requires paired-end sequenced data and only considers those pairs of reads where both ends of a pair , i.e. concordantly, have been mapped. This decreases the number of reads and thus the amount of data that can be used for copy number calling. Further, the detectable deletions and amplifications depend on the given insert size of the concordantly mapped read-pair [12]. With a paired-end mapping approach, amplifications larger than the insert size cannot be discovered. However, the paired-end mapping approach is relatively independent from the sequencing depth. Further, this approach is appropriate to identify structural variants in general (e.g. inversions, inter and intra-chromsomal translocations) [5, 13]. A popular package is BreakDancer [14].

The split-reads approach (iii) takes only discordant mapped reads; that is, only one mate was aligned concordantly. For the unmapped read, alignment to a reference genome is reattempted by splitting the read and aligning both parts separately. A popular package doing this is PINDEL [15].

Callers that apply combinations of the above described approaches, i.e. hybrids, represent the fourth type (iv). Combining a paired-end read and a read depth approach seems to be the most beneficial [5], for instance used by the callers SVDetect [16] and CNVer [17].

The rapidly growing field of single cell DNA (scDNA) sequencing challenges calling of variants, e.g. copy numbers, for individual cells having coverages ~1x. Examples of such callers specified for scDNA are Ginkgo [18] and SCNV [19]. They are both based on a read depth approach and correct for GC content bias. A limitation with these methods is that in order to deal with technical artefacts introduced by single cell sequencing leading to high noise, they require pools of at least 3 cells for calibration or normalisation. SCNV is further an example of adapting an established bulk method, SeqCBS, to scDNA data [19]. Another method is Lumpy [20], a structural variant caller applicable to low coverage data. However, it only outputs cnv types, i.e. deletion or amplification.

However, in general callers require relatively high sequence coverage to achieve good results, or if specialised in scDNA, callers require multiple cells to be applied on. Although higher sequencing depth can be achieved with technologies like polymerase chain reaction (PCR), this leads to unevenly distributed copies of unique molecules, and fewer unique molecules with increasing sequencing depth influencing mutation calling. We here introduce MetaCNV, a method that combines different approaches in order to create a caller that is sensitive enough to detect CNVs in low coverage data (below 10x), even one single cell, but also works on normal and higher coverage data (above 30x and 100x, respectively). It is a generally applicable method that in contrast to previous low coverage methods can be applied to single samples.

Algorithm and Implementation

Implementation

MetaCNV v1.4 is intended for use on unix operating systems. The graphical user interface was created with the GTK+ toolkit, a free library available for the majority of current unix distributions.

MetaCNV algorithm

MetaCNV combines the prediction of copy number callers based on different approaches and builds a consensus to achieve higher prediction accurateness. Some copy number callers (e.g. SVDetect) require a matched sample, for example to be able to distinguish somatic (only detected in the primary sample) from germline mutations (also found in the matched sample, e.g. blood). One of the current input callers for MetaCNV, SVDetect, was run with three different matched sample versions: a matched blood sample, a simulated normal sample with 20x read coverage, and a simulated null (simNull) alignment with constant zero read coverage, to test if this choice affects the prediction accurateness and increases sensitivity especially on low coverage data.

Running MetaCNV comprises five steps (Figure 1).

Choice of current input callers

The current input callers for MetaCNV were chosen due to their relatively good prediction accurateness on low coverage data (Figures 4 & S13). Further, they belong to different calling approaches which perform differently depending on the type of a CNV. In low coverage data, we observed that ReadDepth v0.9.8, which applies a read depth approach, predicts mid-sized (10^6-10^8 bp) and larger deleted segments (>10^8 bp) more accurately than other callers, whereas amplifications were more accurately predicted by SVDetect v1.3, a representative of the hybrid approach.

Finally, a caller's prediction coverage (Figure S7A, Table S3) is considered. Only ReadDepth and CopyCat [21] predict gapfree CNVs for an entire genome. Gapfree means that for each base pair of an investigated genome, a copy number is calculated which resulted in a high prediction coverage. SVDetect predicted for more than 90% of the genomes. The prediction coverage of Control-FREEC and CNVnator varied in cancer cell data sets from 33 to 65% and 35 to 80%, respectively. ReadDepth is built on the coverage-based approach, predicting copy numbers gapfree for the whole genome. It applies a negative binomial distribution to approximate an overdispersed Poisson distribution [7]. Further, it outputs absolute copy numbers for equally-sized, non-overlapping bins (Table S15). The optimal bin size is calculated by ReadDepth, but can be indirectly adapted by changing the false discovery rate.

SVDetect is a tool used to detect general structural variants and is based on the hybrid approach for copy number calling, meaning both the coverage and any change in the insert size of read-pairs are considered when inferring copy numbers [16]. SVDetect copy numbers are not predicted gapfree, and in contrast to ReadDepth and CNVnator it requires for copy number calling a matched normal sample with which to relate the studied sample (Table S16).

CNVnator v0.3.2, like ReadDepth, is based on the coverage approach, but with an additional mean-shift approach that produces a probability distribution function from the coverage data, and links each data point, i.e. bin, to its maxima. Further, CNVnator is calibrated using the extensive validation done by the 1000 Genomes Project [11, 22]. Similar to SVDetect, it does not calculate copy numbers for the entire genome. As with SVDetect, the bin size can be modified by the user. Despite achieving good prediction accurateness on high coverage data, CNVnator predicts extremely high copy numbers on low coverage data (Figures S8, Table S17). However, it still produced reliable classification of segments into deletions and amplifications for low coverage data as accurately as SVDetect and

ReadDepth, and complements their predictions (Figure S13, Table S17). Based on this, ReadDepth, SVDetect, in combination with CNVnator as a referee for conflicts between ReadDepth and SVDetect were chosen as input callers for MetaCNV.

Matched sample in SVDetect

When SVDetect was run as part of MetaCNV, a simulated null sample with zero read coverage (simNull) was used. The simNull alignment as matched sample is a novel idea to increase sensitivity and to remove an additional source of noise for low coverage data. If comparing a low coverage sample with a matched sample having high coverage it will disturb a correct copy number calling. Ideally, a matched sample should have the same constant coverage as the sample to investigate. With current tools this constant coverage on each base pair of the genome is not achievable (Figure S21). Therefore, we developed the novel idea of a simulated null alignment containing zero coverage, and thus no noise.

When SVDetect was run outside of MetaCNV for comparison, two different types of such matched samples were tested: a matched normal sample (matchedNormal) and a simulated normal sample with constant read coverage (simNormal).

The matched normal was taken from the cancer cell lines HCC1187 and HCC2218 (Table 1) for which such samples are available (blood samples HCC1187BL and HCC2218BL). The simulated normal was inferred using Pirs [23].

Segmentation of the genome

ReadDepth divides a given genome into bins of a calculated minimum or a multiple of the minimum bin size. The bin size can only be manipulated indirectly by increasing the false discovery rate. For both, the high and low coverage study of sequenced single cells, it was set to 0.01 (default). SVDetect accepts a bin size given by the user. Several bin sizes were tested for SVDetect but the best results

were achieved with a bin size of 400 bp and no bin overlap. Based on the calculated bin sizes from both callers hidden in the start and end position in the output files, new bins were calculated with different sizes while also considering all breakpoints given by the input callers (Figure 2). At the end of the MetaCNV prediction process, bins with a similar copy number to one decimal place were merged into one segment.

MetaCNV model

The MetaCNV model contains rules to be applied to each bin, depending on the predicted copy numbers of the input callers representing different approaches of copy number calling; the strength of rules is in the decreased risk of overfitting. In general, MetaCNV accepts deletions and normal copy numbers if they are predicted by a coverage approach (current ReadDepth), and amplifications if they are predicted by a read-pair approach or hybrid approach of coverage and read-pair (current SVDetect). More detailed rule descriptions are listed in Table S18. To distinguish between deletions and normal copy numbers from amplifications, two thresholds *T1* and *T2* are introduced. *T2* is the local minimum of the frequency of copy numbers produced by ReadDepth between 2 and 2.3; the other threshold *T1* is calculated as 2 * *P* - *T2* where *P* is the ploidy value (Figure 3).

Due to gaps in the predictions from at least one current input caller (no copy number predicted) or conflicting predicted copy numbers from both callers, the MetaCNV model comprises additional rules and includes CNVnator as the referee in such conflicting cases.

Normalisation of input data

When developing a consensus approach, the results of the considered input callers need to be normalised in order to compare their results correctly. The detected bias in the predicted copy numbers of each input caller was corrected with a copy-number-dependent normalisation. The bias (systematic error) is the difference between frequency peak and ploidy (Figure 3, equations 1), where CN is the predicted copy number by a caller, CN_{norm} is the normalised CN, and *P* is the ploidy value which is for the autosomes 2 and for the allosomes either 2 if female or 1 if male. The maximum absolute bias correction is 0.5.

$$CN_{norm} = CN + (factor \times bias) \quad (1)$$

$$bias = P - CN\ at\ max\,(frequency(CN)) \quad (1a)$$

$$factor = min\left(\frac{1}{2} \times CN\ ;\ 1\right) \quad (1b)$$

Converting log2 values into absolute copy numbers

In general, callers predict copy numbers either as absolute copy numbers or as log ratio values which can be converted into absolute copy numbers with

$$CN_{corrected} = P \times 2^{CN_{log2,corrected}} \quad (2)$$

Absolute copy numbers reflect the number of repeats of the sequence [7]. The $\log_2$ values represent $\log_2$ transformed ratios to a matched sample or the ploidy *P* [6].

When MetaCNV is run with input from SVDetect using a simulated normal sample, the formula above (equation 2) is applied to convert $\log_2$ values into absolute copy numbers for each bin. Before this, the $\log_2$ values CN_{log2} are corrected with a genome wide median:

$$CN_{log2,corrected} = CN_{log2} - \text{median}\,(CN_{log2}) \quad (2a)$$

When MetaCNV is run with input from SVDetect using a simulated null alignment, the predicted values are accepted as copy numbers. This produces CN_{log2} values that are always non-negative, because in the case of the simulated null alignment, they are relative to 0. Further, a distortion was observed which increases exponentially with the true amplification. This distortion is corrected with an equalizer factor *q* (Figure 20, equation 3). The net effect corresponds to a back log transformation that is calibrated to result in absolute copy numbers.

$$CN_{meta} = CN_{log2} \times q \quad (3)$$

$$q = \left(1 + \frac{1}{100} \times CN_{log2}\right)^{0.75\ CN_{log2}} \quad (3a)$$

Calculating error scores

Each MetaCNV bin is annotated with an error score e mirroring the prediction similarity of the input callers. The score depends on the consensus of the input callers per bin and is averaged for the optimised segments considering the bin length per segment. The error score per bin is calculated as the squared absolute copy number difference between ReadDepth and SVDetect:

$$e_{bin} = (CN_{RD} - CN_{SV\,Detect})^2 \quad (4)$$

For bins where ReadDepth did not predict a copy number, we assume CN = 0. ReadDepth predicts copy numbers gapless for a genome. Although, rare large gaps happen to occur but they are caused by non-sequenced regions which not contain genetic regions according to the reference genome GRCh38 [30]. If SVDetect did not predict a copy number, CN = *P* is used. SVDetect does not predict copy numbers gapless for a genome although the coverage is sufficient in case of the high coverage data and no other obvious reason could be identified. (Pre-processing and application of calling methods, Suppl.)

Accurateness evaluation

To evaluate the accurateness of a caller, its result needs to be compared to true copy number variations. Different ways were taken to set such a gold standard: simulated data, for example for the validation of CNV-seq; clinical data with experimentally confirmed CNV, e.g. for the validation of CNVnator; or cancer cell lines, e.g. for the validation of Control-FREEC and ReadDepth.

Further, the comparison between true and predicted copy number has either been performed for segments, e.g. validation of Control-FREEC and ReadDepth, or for genes [24]. Different accurateness measures have been applied, including accuracy, false discovery rate, F1-score [25], receiver operating characteristic with true positive and false positive rates [26], Spearman correlation, and root mean squared error [27].

Finally, absolute copy numbers, $\log_2$ ratio values or the class of a copy number [25], which is either a deletion or an amplification, were considered. Absolute copy numbers are decimal or integer values mirroring the number of repeats of the sequence [7] whereas the $\log_2$ ratios stand for log transformed ratios to a matched sample or the ploidy [6].

The prediction accurateness of a regression model, which is the case if comparing copy numbers, can be evaluated using the mean squared error (MSE, equation 5) and mean absolute error (MAE). In general, an error based measure calculates the difference between a true and a predicted value, which in this case is the difference between true and predicted copy number per gene. The mean absolute error presents the average error, whereas the mean squared error combines systematic and random error into one value [28]. It also penalizes outliers: each distance is squared, and larger distances thus get more weight.

The true copy number per gene, x_i, was compared with the predicted one, $\hat{x}_i$, by calculating the residuals x_i - $\hat{x}_i$ for all genes N.

$$\text{Mean sqared error } MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{x}) \tag{5}$$

To avoid outlier penalty we compare additionally the prediction accurateness with the mean log ratio error:

$$\text{Mean log ratio error } MLRE = \frac{1}{N}\sum_{i=1}^{N} LR\ ,\ LR = abs\left(ln\frac{x_i+1}{\widehat{x_i}+1}\right) \tag{6}$$

If a caller predicts highly different copy numbers for genes of a genome having the same true copy number but also predicts similar copy numbers for genes with different true copy numbers, then the analysis of copy number variations can become difficult. Therefore, copy number callers were also evaluated by the variance of the residuals, which mirrors how close the predicted values surround

each unique true copy number (equation 7. For each unique true copy number (integer) value, the variance of the predicted values was calculated, however, default elements were replaced with the residuals z_i (the difference between actual x_i and predicted value $\hat{x}_i$). The variances are then averaged for the total number of unique true copy numbers M.

$$\text{Variance of residuals } \sigma^2_{Res} = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2\right) \quad (7)$$

Matthew's correlation coefficient (MCC) [29] can be applied for classification models and is especially applicable for unbalanced ratios of the four confusion matrix categories, which is the case for the cancer cell lines used in this paper (ratio of amplified/deleted genes from 0.8 to 5, Table 1). To define the classes of deletions and amplifications for ploidy $P = 2$, copy numbers > 2.75 were set as amplifications and copy numbers < 1.75 were set as deletions, while values in between were set as normal. In order to deal with these three classes (deletion, normal, amplification), the values for true positives, true negatives, false positives and false negatives were micro-averaged. This means that for example the true positive value TP is equal to $TP_{class1} + TP_{class2} + TP_{class3}$.

The prediction accurateness of several copy number callers and MetaCNV was evaluated using known amplified or deleted genes that were publicly available along with well-studied cancer cell lines. MetaCNV was developed for low coverage data. Therefore, MetaCNV's accurateness was verified on single sequenced cells of a cancer cell line (SKBR3) having 1x to 6x read coverage. Additionally, the accurateness was assessed on four cancer cell lines with normal and high coverage (62x to 104x coverage, Table 1). The range of copy numbers among the cancer cell lines was limited. To validate MetaCNV on a wider range of copy numbers and also genome-wide, mutated genomes of different coverages were simulated. MetaCNV's accurateness was compared with other callers' accurateness by MLRE and MCC (MSE, MAE, and Spearman's correlation in Figures S11, S13, S14).

For each caller, we map the output (segments and corresponding copy numbers) per cancer cell line and simulated genome to the human assembly GRCh38 Ensembl (release 84) [30]. Due to different segment sizes, this mapping resulted in one or several predicted copy numbers per gene. In such cases, the total copy number per gene was the sum of the weighted copy numbers, depending on their segment length within the gene (Figure S9). There were gaps in the prediction of SVDetect, CNVnator, and Control-FREEC, that is, no copy number for a segment was called. Such gaps within a gene were filled using the ploidy value. Some segments of the cancer cell lines given by COSMIC [31] (Cosmic, Suppl.) did not cover a gene completely. In these rare cases, the gene was reduced to the covered segment length within this gene.

Cancer cell lines

Cancer cell lines, like the HeLa cancer cell line, which can theoretically be divided and replicated indefinitely, contain cells taken from e.g. naturally-occurring cancer tissues [32]. They are publicly available and well-studied objects; mutations like CNVs and other structural variations are experimentally confirmed [33].

The use of cancer cell lines is advantageous due to the fact that the reviewed callers have to perform on real sequencing data, and the result can be compared to known CNVs. Using cancer cell lines can also be a disadvantage because of the lower heterogeneity found in them; high heterogeneity is a common characteristic of cancerous clinical samples. Further, variance and grades of deletions and amplifications are lowered, since a cancer cell line contains only a limited number of cells, compared to a bulk-sequenced clinical sample containing million of cells. Despite the disadvantages, testing a copy number caller's accurateness on cell lines is easy, transparent, and replicable.

The cancer cell lines HCC1187 (with matched blood sample), HCC2218 (with matched blood sample), MCF7, PC3, and SKBR3 (Table 1) were chosen to compare MetaCNV's accurateness with other

callers' accurateness. The sequenced DNA of a single cell of SKBR3 (1x per cell) by using a novel method to sequence both, genome and transcriptome of the same single cell [4], was used as low coverage data. The sequenced and aligned single cell genomes were stepwise merged to present increasing coverages.

Simulated mutated genomes

For each of four coverages (1x, 2x, 5x, and 10x), three mutated human genomes were simulated (gw1, gw2, lcd). The paired-end reads for the chromosomes 1-22 were generated using CNVsim v0.9.2, aligned using Bowtie2 v2.2.9 (Langmead *et al.*, 2009), and converted, sorted and indexed using Samtools v1.2 (Li, 2011). For each of the three genomes, 30 to 50 segments per chromosome and each segment with a copy number unequal to 2 and a length of 10 kbp to 100 kbp were generated. The simulated genomes were mapped to GRCh38 (Table 2). Two simulated genomes (gw1, gw2) were used to compare the callers' results for a genome wide copy number prediction. CNVnator and Control-FREEC only gave partial predictions, see Table S4. The third simulated genome (lcd) was therefore used to compare the callers' results on a reduced data set comprising segments for which a prediction of all callers (MetaCNV, SVDetect, CNVnator, ReadDepth, Control-FREEC) was available. This, however, limited the range of copy numbers from 0 to 4.

Results

The consensus copy number caller presented here, MetaCNV, combines a number of primary callers by rules that optimally harness the strengths of each method. MetaCNV's and other callers' accurateness were evaluated by calling copy numbers for cancer cell lines and comparing the results with experimentally confirmed deleted or amplified genes in the COSMIC database. To consider a higher number of mutated genes and a wider range of copy numbers, MetaCNV's and the other callers' accurateness was also evaluated on three simulated mutated human genomes for each of the coverages 1x, 2x, 5x, and 10x.

True and predicted absolute copy numbers were compared by the mean log ratio error (MLRE) across all genes. We also calculated Matthew's correlation coefficient, MCC, on the predicted classes of deleted and amplified genes. This measure evaluates how well callers can differentiate between amplifications and deletions in general.

Prediction accurateness on low coverage data

We assessed copy number prediction accurateness on a single-cell sequenced cancer cell line (SKBR3) with stepwise merged alignments of one additional cell, and stepwise increased coverage from 1x to 6x. MetaCNV was compared to the popular copy number callers CNVnator, Control-FREEC, ReadDepth, and SVDetect. In all benchmarks, MetaCNV was the most accurate method. MetaCNV outperformed the other methods in the majority of coverages in the MLRE benchmark. In all MCC benchmarks, MetaCNV was the most accurate method. MetaCNV was the most robust method as it was the top performer for MLRE, MSE, MAE, and MCC in the majority of the coverage levels (Figures 4, S13- S15, Tables S8-S10).

CNVnator produced only a reliable classification of segments into deletions and amplifications; absolute copy numbers were not usable. For example for the SKBR3 cell line, sequenced single cell nr. 1, copy numbers for regions of deletions ranged from 0 to 2,664 and copy numbers for regions of amplifications ranged from 0 to 44,084 (Figure S8). Increasing coverage from 1x to 6x improved the accurateness only for Control-FREEC, a coverage based approach, although it never reached the accurateness of its competitors.

Prediction accurateness on high coverage data

MetaCNV's accurateness on cancer cell lines with high read coverage was compared to the accurateness of other callers which are CNVnator, CopyCat, Control-FREEC, ReadDepth, and SVDetect.

Although MetaCNV was designed for low coverage data, it also performed well for high coverage data. In three of four tested cancer cell lines, it was the best performer (Figure 5). Several predictors reached an MCC near 1.0 in two cell lines, hence there was no clear winner there. To present an overall accurateness, MLRE and MCC were averaged over the four tested cancer cell lines per caller (Figures 5c & 5d). MetaCNV showed the best overall accurateness with MLRE and second best with MCC, only 0,004 behind CNVnator.

Prediction accurateness on simulated mutated genomes

MetaCNV's predictive accurateness on simulated mutated human genomes was compared to the predictive accurateness of CNVnator, SVDetect, ReadDepth, and Control-FREEC. The simulated datasets comprise the predictions genome-wide and reduced to genes for which an output from all callers was available. Each dataset was simulated in different coverages (1x, 2x, 5x, 10x). MetaCNV performed best in all tested simulated datasets using MLRE and MCC (Figures S16A, S17, and 18A); in two datasets (2x-gw2 and 5x-lcd) MetaCNV and ReadDepth performed equally well using MLRE.

The novel simNull alignment as matched sample

SVDetect, one of the current input callers for MetaCNV, requires a matched sample for copy number calling. For the cancer cell lines HCC1187 and HCC2218, a matched normal blood sample was used (HCC1187BL, HCC2218BL). Additionally, a normal sample was simulated with constant coverage. However, a simulated normal sample also contains noise and variance in coverage which negatively influences copy number calling for low coverage data. Further, comparing a low coverage sample (e.g. 3x) with a matched sample having higher coverage (e.g. 30x) will lead to an insensitive result. Therefore, we developed the novel idea of a simulated null alignment containing zero coverage, and thus no noise, to improve the sensitivity of copy number calling.

SVDetect using a simulated null alignment creates one of the inputs for MetaCNV. For comparison, MetaCNV was also run with input from SVDetect using a simulated normal sample. MetaCNV with a simulated null alignment achieved better overall results than with the simulated normal alignment when evaluated using MSE and MLRE (Figure S19). This was true for both low and high coverage data, although the effect was much stronger for low coverage data.

However, just replacing the matched simulated normal sample with a simulated null alignment in SVDetect did not always improve the prediction for high coverage data, see Figure S19.

Taken together, the superior accurateness of MetaCNV stems from: (i) combining the prediction of multiple callers to form a consensus, (ii) considering that the reliability of a copy number depends on the approach the caller was based on, and (iii) for low coverage data, using a matched sample with zero instead of normal coverage.

Discussion and conclusions

To investigate smaller areas within a tissue section or even sequenced single cells, callers are needed that are able to detect CNVs in low coverage alignments. We present MetaCNV a copy number caller specialised in low coverage data. MetaCNV is based on a consensus approach combining different calling approaches and achieved a better prediction accurateness on low coverage data than other reviewed callers. MetaCNV also performed well on high coverage data. (Figures 4 & 5)

In low coverage data, we observed that a read depth approach is better in predicting deleted segments, whereas amplifications were more accurately predicted by representatives of the paired-end mapping or hybrid approach. However, current callers, including those that employ a hybrid model, apply the developed calling model to each type of variation (deletions and amplifications). MetaCNV considers several calling approaches and builds a consensus based on rules to avoid overfitting, which results in predicting CN more accurately. Due to the different approaches of the current input callers

(ReadDepth, SVDetect, and CNVnator), MetaCNV is biased towards mid-sized and large deletions and short and mid-sized amplifications.

To identify somatic structural variants, callers such as SVDetect require a matched normal sample. In order to increase sensitivity for low coverage data, we developed the novel idea of a simulated null alignment used as matched sample. MetaCNV requires the calling result from SVDetect using this simNull as matched sample. For demonstration, MetaCNV was also tested with the input from SVDetect using a simulated normal, however, with a simNull performing a better prediction. The effect of a simulated null alignment compared with a simulated normal or matched normal is low if applied on high coverage data. In contrast, the impact increased immensely for low coverage data where, on the one hand, additional noise disturbs the prediction accurateness, and on the other hand, high sensitivity is required to predict reliable copy numbers. The simNull alignment leads to a distortion of predicted copy numbers that increases with the copy number. This distortion is corrected in a simple way, but with more alignments having different read coverages, a more sophisticated approach could lead to a further improvement in prediction.

The evaluation in this study was done using an error benchmark comparing true and predicted value per instance (gene), the mean log ratio error (MLRE), and Matthew's correlation coefficient (MCC) to assess how well a caller can distinguish between deletions and amplifications. The benchmarks mirror that coverage-based approaches are highly dependent on read coverage, and show that MetaCNV outperforms other callers on low coverage data and performs well on high coverage data.

It was not possible to include specialized low coverage scDNA callers in the benchmark. Ginkgo and SCNV can not be run for single samples and Lumpy does not output quantitative CNV values. CNV-seq v0.2-8 gave ambiguous copy numbers for overlapping bins. We could however run Ginkgo using all SKBR3 single cell samples, which gives it considerably more information than the callers

presented in the benchmark have. Despite this, Ginkgo achieved worse accurateness than MetaCNV for all six SKBR3 cells when evaluated with MCC, and worse or equal accurateness for four cells when evaluated with MLRE.

Availability and requirements

Project name: MetaCNV

Project home page: https://bitbucket.org/sonnhammergroup/metacnv

Operating system(s): unix

Programming language: C++

Other requirements: GTK+-2.0 or higher

Licence: GNU LGPL

Any restrictions to use by non-academics: none

List of abbreviations

CNV: Copy number variation

MCC: Matthew's correlation coefficient

MSE: Mean squared error

MAE: Mean absolute error

MLRE: Mean log ratio error

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: All data generated or analysed during this study are included in this published article and its supplementary information files.

Competing interests: The authors declare that they have no competing interests.

Funding: This work was supported by AstraZeneca AB project N1372.

Authors' contributions: SF and ES conceptualised and designed the project, performed the data analysis and wrote the manuscript. RB and SF developed the code. All authors read and approved the final manuscript.

Acknowledgements: Not applicable

References

1. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16:172–83. doi:10.1038/nrg3871.

2. Illumina. TruSeq DNA PCR-Free. 2017. https://www.illumina.com/library-prep-array-kit-selector.html%0A.

3. Bock C, Farlik M, Sheffield NC. Multi-Omics of Single Cells: Strategies and Applications. Trends Biotechnol. 2016;34:605–8. doi:10.1016/j.tibtech.2016.04.004.

4. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. 2015;33:285–9. doi:10.1038/nbt.3129.

5. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011;470 July 2010:59–65. doi:10.1038/nature09708.Mapping.

6. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. BMC Bioinformatics. 2013;14:S1. doi:10.1186/1471-2105-14-S11-S1.

7. Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: A parallel R package for detecting copy number alterations from short sequencing reads. PLoS One. 2011;6:e16327. doi:10.1371/journal.pone.0016327.

8. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 2009;10:1–9.

9. Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, et al. Cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation

sequencing data with a low false discovery rate. Nucleic Acids Res. 2012;40:e69. doi:10.1093/nar/gks003.

10. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28:423–5. doi:10.1093/bioinformatics/btr670.

11. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011.

12. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009.

13. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nature Reviews Genetics. 2011.

14. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6:677–81. doi:10.1038/nmeth.1363.

15. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25:2865–71.

16. Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, et al. SVDetect: A tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics. 2010;26:1895–6. doi:10.1093/bioinformatics/btq293.

17. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. Detecting copy number variation with mated short reads. Genome Res. 2010.

18. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, et al. Interactive analysis and assessment of single-cell copy-number variations. Nature Methods. 2015;12:1058–60.

19. Wang X, Chen H, Zhang NR. DNA copy number profiling using single-cell sequencing. Brief Bioinform. 2018.

20. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for structural variant discovery. Genome Biol. 2014.

21. Meier-Kolthoff JP, Auch AF, Huson DH, Goker M. COPYCAT : cophylogenetic analysis tool. Bioinformatics. 2007;23:898–900. doi:10.1093/bioinformatics/btm027.

22. Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, et al. A global reference for human genetic variation. Nature. 2015;526:68–74. doi:10.1038/nature15393.

23. Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, et al. pIRS: Profile-based illumina pair-end reads simulator. Bioinformatics. 2012;28:1533–5. doi:10.1093/bioinformatics/bts187.

24. Eisfeldt J, Nilsson D, Andersson-Assarsson JC, Lindstrand A. AMYCNE: Confident copy number assessment using whole genome sequencing data. PLoS One. 2018;13:e0189710. doi:10.1371/journal.pone.0189710.

25. Legault MA, Girard S, Perreault LPL, Rouleau GA, Dubé MP. Comparison of sequencing based CNV discovery methods using monozygotic twin quartets. PLoS One. 2015.

26. Duan J, Zhang J-GJ-G, Deng H-WH-W, Wang Y-PY-P. Comparative Studies of Copy Number Variation Detection Methods for Next-Generation Sequencing Technologies. PLoS One. 2013;8:59128. doi:10.1371/journal.pone.0059128.

27. Wang R, Lin D, Jiang Y. SCOPE: A normalization and copy number estimation method for single-cell DNA sequencing. 2019.

28. Wackerly D, Mendenhall W, Scheaffer RL. Mathematical Statistics with Applications. 2008. Accessed 24 Aug 2019.

29. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA - Protein Struct. 1975.

30. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018.

543 31. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: Somatic cancer
544 genetics at high-resolution. Nucleic Acids Res. 2017;45:D777–83. doi:10.1093/nar/gkw1121.
545 32. Landry JJM, Pyl PT, Rausch T, Zichner T, Tekkedil MM, Stütz AM, et al. The Genomic and
546 Transcriptomic Landscape of a HeLa Cell Line supplement. G3 Genes|Genomes|Genetics.
547 2013;3:1213–24. doi:10.1534/g3.113.005777.
548 33. Masters JRW. Human cancer cell lines: fact and fantasy. Nat Rev Mol Cell Biol. 2000;1:233–6.
549 doi:10.1038/35043102.
550 34. Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, Beare D, et al. PICNIC: An algorithm to
551 predict absolute allelic copy number variation with microarray cancer data. Biostatistics.
552 2010;11:164–75. doi:10.1093/biostatistics/kxp045.
553 35. Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstråhle J, et al. Spatial maps
554 of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nat Commun.
555 2018;9.
556 36. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short
557 DNA sequences to the human genome. Genome Biol. 2009;10:R25. doi:10.1186/gb-2009-10-3-
558 r25.
559 37. Broad Institute. Picard tools. https://broadinstitute.github.io/picard/. 2016.
560 38. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and
561 population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–
562 93. doi:10.1093/bioinformatics/btr509.

563

564 **Table 1.** Cancer cell lines used for accurateness testing. Coverage and number of deleted and
565 amplified genes extracted from COSMIC are presented. Copy numbers were called from sequenced
566 genomes and experimentally confirmed with PICNIC [34].

Cancer	Coverage	# of amplified	# of deleted	Total # of	Ratio

cell line		genes	genes	genes with CNVs	amplified/ deleted genes
HCC1187	104x	154	46	200	3.48
HCC2218	93x	200	36	236	5.55
MCF7	62x	49	30	79	1.63
PC3	76x	179	222	401	0.81
Single-cell SKBR3	6 stepwise merged single cells with 1x to 6x	161	29	190	5.55

Table 2. Simulated mutated genomes used for accurateness testing. The simulated data sets gw1 and gw2 were used to compare the prediction results for a genome-wide prediction. The simulated data sets lcd was used to compare the prediction results for genes for which a prediction of all tested caller was available.

Simulation	Depth of coverage	# of amplified genes	# of deleted genes	Total # of genes	Ratio amplified/ deleted genes
sim 1x gw1	1.1x	283	559	13906	0.51
sim 1x gw2	1.1x	236	573	13920	0.41
sim 1x lcd	1.5x	25	30	330	0.83
sim 2x gw1	2.2x	247	560	13937	0.44
sim 2x gw2	2.2x	254	588	13922	0.43
sim 2x lcd	2.9x	42	44	16	0.95

sim 5x gw1	5.7x	230	580	13987	0.40
sim 5x gw2	5.7x	251	593	13949	0.42
sim 5x lcd	7.3x	62	48	14	1.29
sim 10x gw1	11.3x	263	560	13882	0.47
sim 10x gw2	11.3x	240	551	13865	0.44
sim 10x lcd	14.9x	54	49	7	1.10

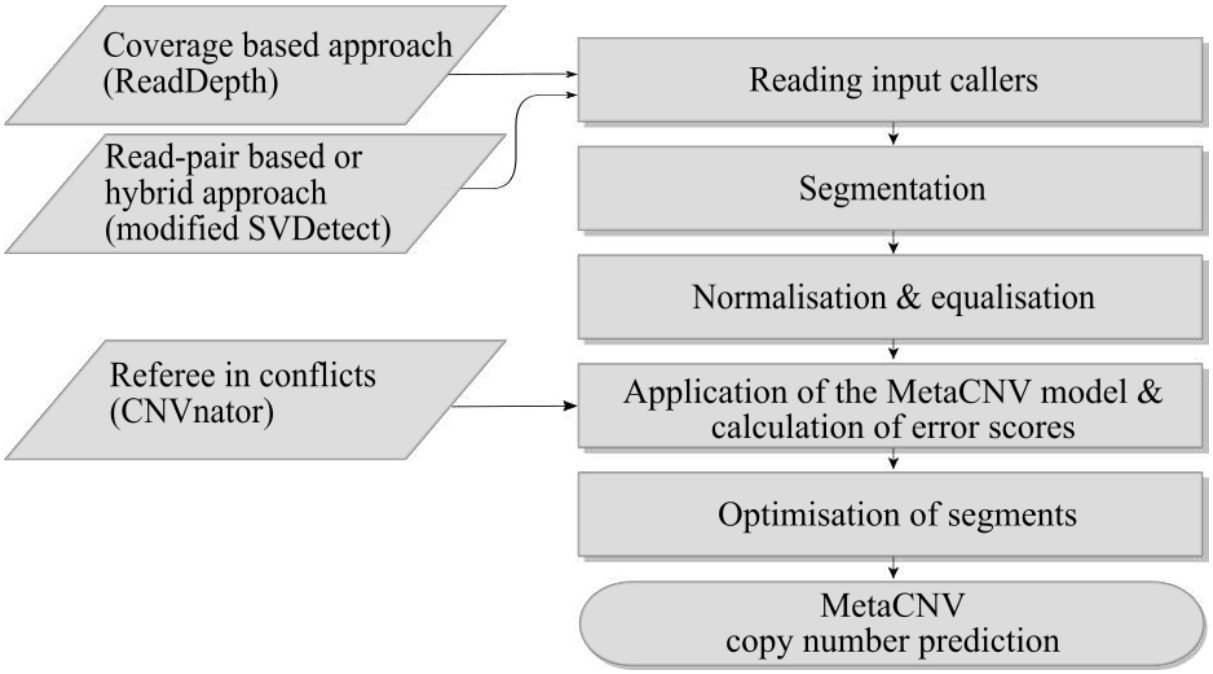
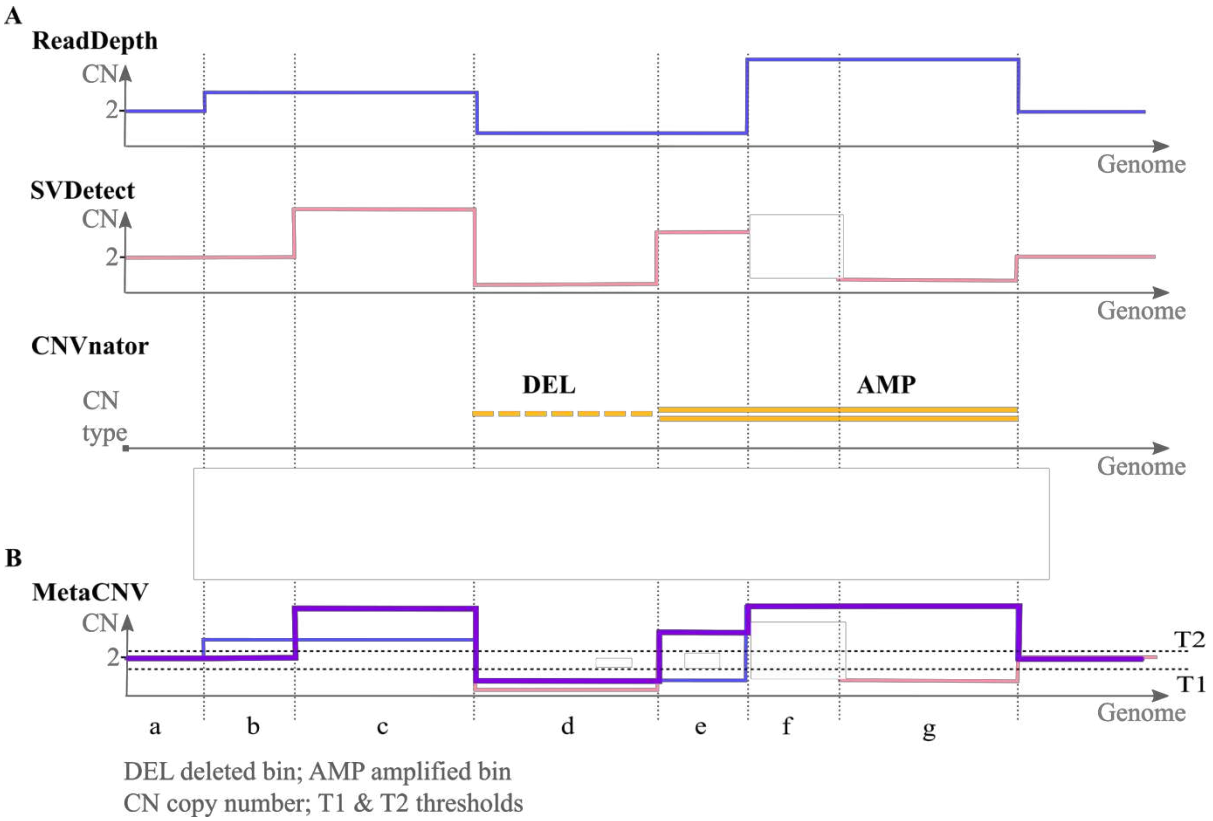


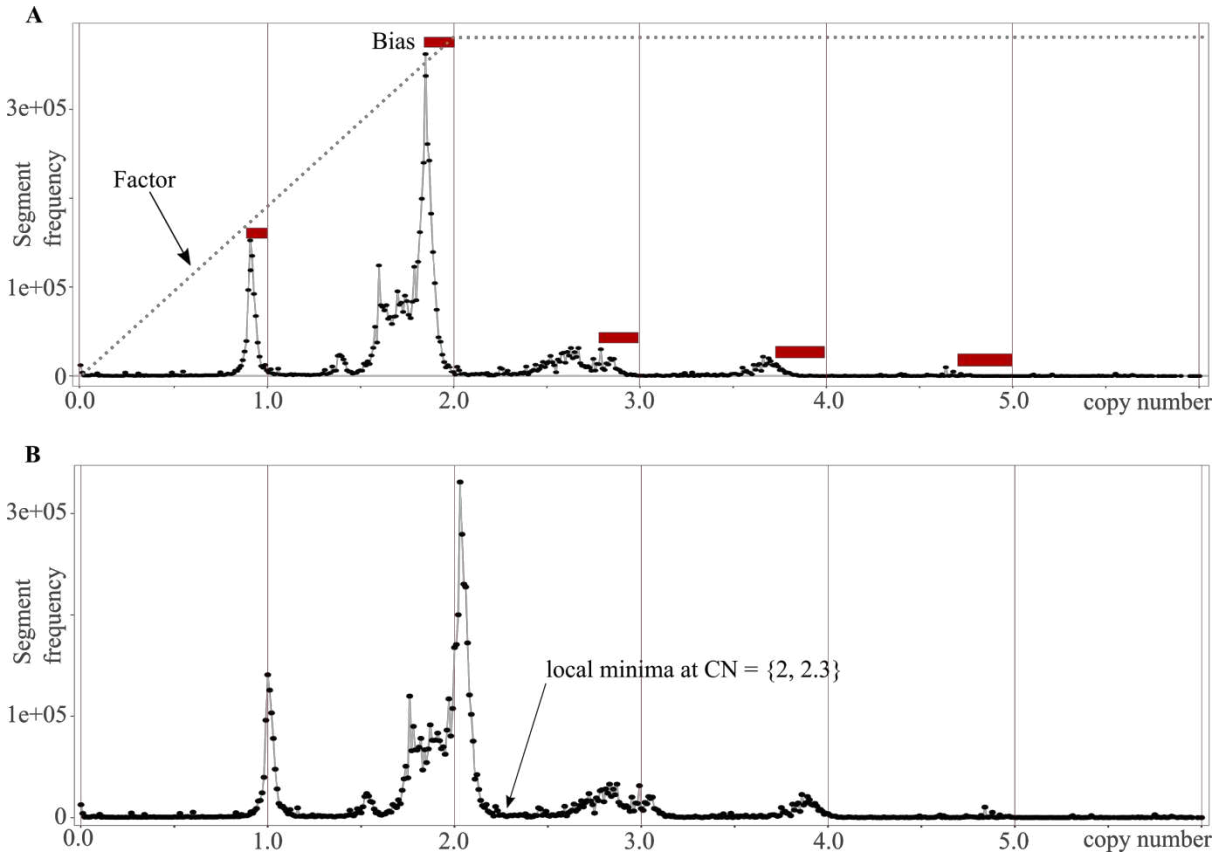
Figure 1. MetaCNV workflow to call copy numbers. MetaCNV requires as input an SVDetect prediction using a simulated null alignment (simNull) as matched sample (Figure S19).



578

579 **Figure 2.** Genome segmentation by MetaCNV. **A.** Segmentation of the genome according to the bins
580 and breakpoints from the input callers. **B.** Consensus segment prediction by MetaCNV is marked as a
581 thick line. Bins c is predicted as amplified and bin d as deleted. Here there is no conflict between the
582 input callers, hence the consensus is that c is amplified and d is deleted. Bin f is not predicted by
583 SVDetect, but because ReadDepth predicts it as an amplification, this becomes the consensus. Bins b,
584 e and g have conflicting ReadDepth and SVDetect predictions. CNVnator judges that e and g are
585 amplifications, hence this becomes the consensus. Bin b is set to CN 2 because only ReadDepth
586 predicts it as an amplification and CNVnator makes no prediction (Table S18).

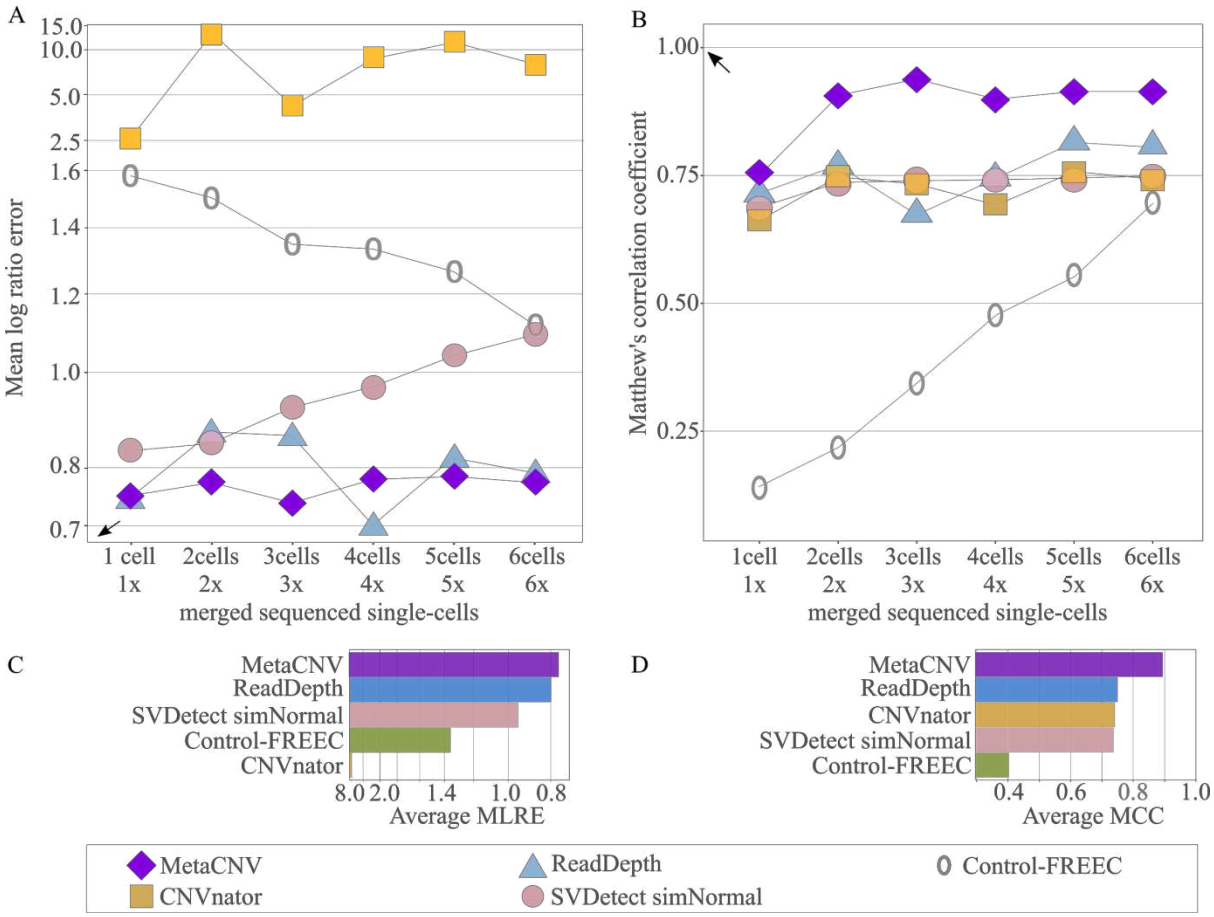
587



588

589 **Figure 3.** Distribution of absolute copy numbers. Copy numbers were called with ReadDepth of
590 cancer cell line HCC2218 after segmentation (6,867,679 segments with an average segment length of
591 450 bp). **A.** The bias occurs around each integer copy number and is lower for CN_{RD} ~1, but increases
592 for CN_{RD} ~2. The factor for normalisation is a linear function (equation 1b) which serves to adjust the
593 bias for different CN_{RD}. **B.** Distribution of absolute copy numbers called with ReadDepth after
594 normalisation

595



596

597 **Figure 4.** Prediction accurateness of MetaCNV on low coverage data, compared to CNVnator,
598 Control-FREEC, ReadDepth, and SVDetect. **A** and **B** present the benchmark results with MLRE and
599 MCC per merged alignment. **C** and **D** show the overall results averaged across all benchmarked
600 alignments.

601

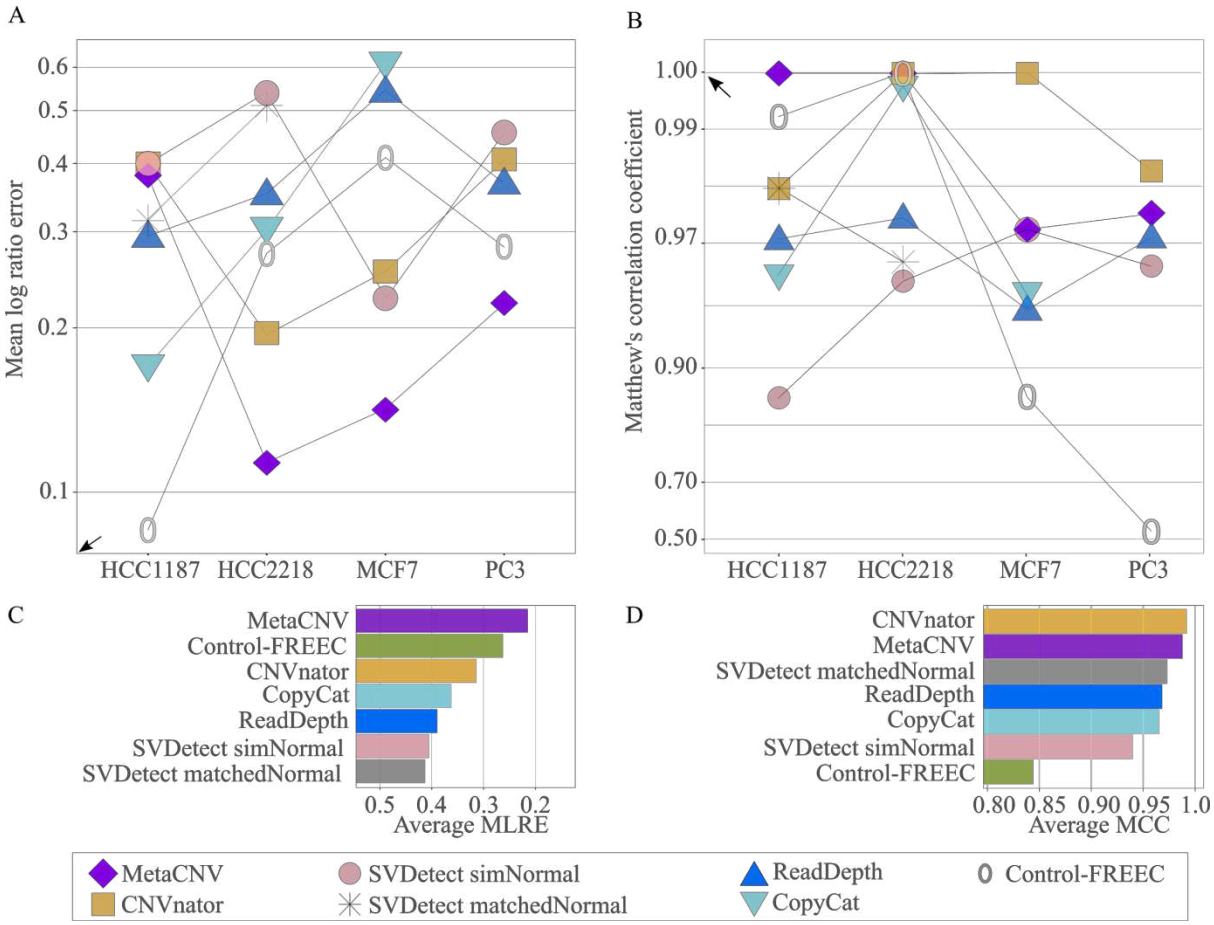


Figure 5. Prediction accurateness of MetaCNV and other copy number callers for high coverage data. **A** and **B** present the average benchmark results per cancer cell line. SVDetect using matchedNormal could only be performed for HCC1187 and HCC2218, for which matched blood samples were available. **C** and **D** show the overall results averaged across all benchmarked cell lines.

Electronic supplementary material

File name	Description of data
Additional file 1	Supplementary methods, figures and tables (DOCX)
Additional file 2	MetaCNV Readme (manual to run MetaCNV, TXT)
Additional file 3	Supplementary table containing genes per cancer cell line, incl. Copy number, segment length and evaluation length (XLS).

Figures

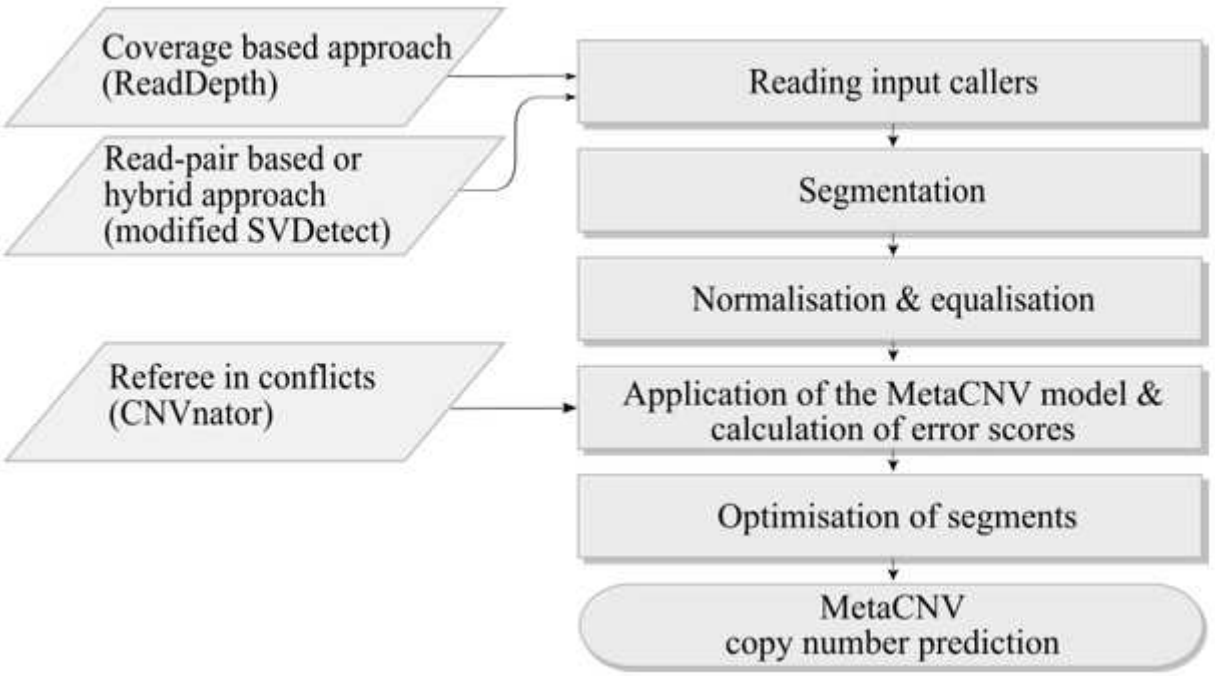


Figure 1

MetaCNV workflow to call copy numbers. MetaCNV requires as input an SVDetect prediction using a simulated null alignment (simNull) as matched sample (Figure S19).

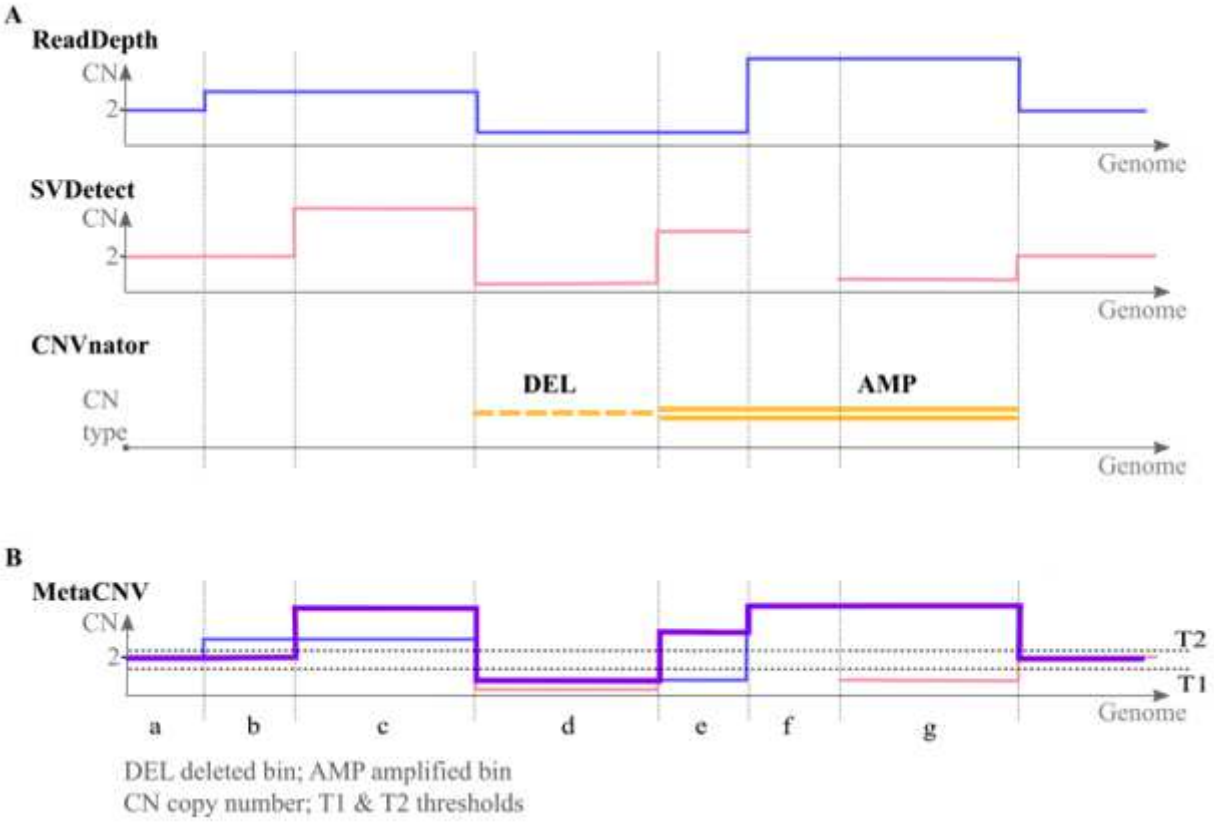


Figure 2

Genome segmentation by MetaCNV. A. Segmentation of the genome according to the bins and breakpoints from the input callers. B. Consensus segment prediction by MetaCNV is marked as a thick line. Bins c is predicted as amplified and bin d as deleted. Here there is no conflict between the input callers, hence the consensus is that c is amplified and d is deleted. Bin f is not predicted by SVDetect, but because ReadDepth predicts it as an amplification, this becomes the consensus. Bins b, e and g have conflicting ReadDepth and SVDetect predictions. CNVnator judges that e and g are amplifications, hence this becomes the consensus. Bin b is set to CN 2 because only ReadDepth predicts it as an amplification and CNVnator makes no prediction (Table S18).

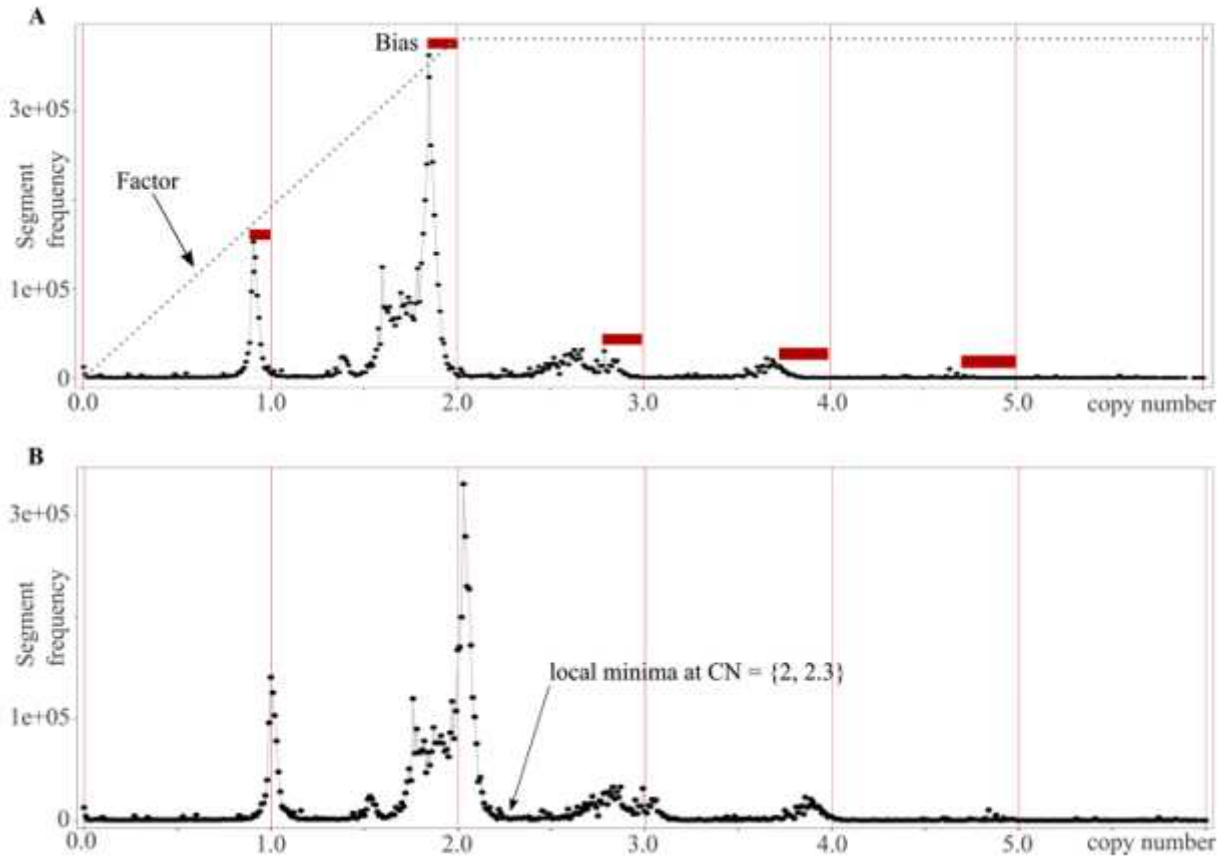


Figure 3

Distribution of absolute copy numbers. Copy numbers were called with ReadDepth of cancer cell line HCC2218 after segmentation (6,867,679 segments with an average segment length of 450 bp). A. The bias occurs around each integer copy number and is lower for CNRD ~1, but increases for CNRD ~2. The factor for normalisation is a linear function (equation 1b) which serves to adjust the bias for different CNRD. B. Distribution of absolute copy numbers called with ReadDepth after normalisation

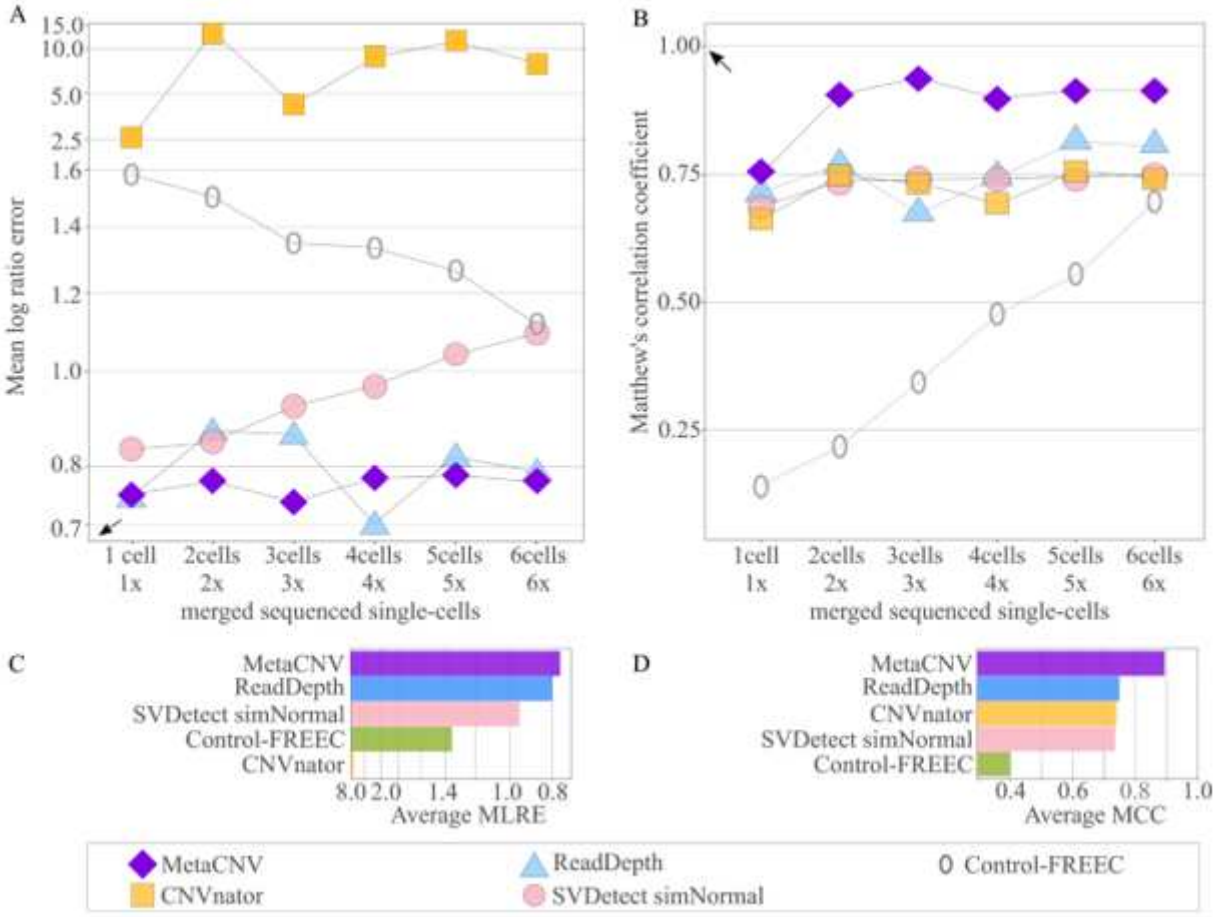


Figure 4

Prediction accurateness of MetaCNV on low coverage data, compared to CNVnator, Control-FREEC, ReadDepth, and SVDetect. A and B present the benchmark results with MLRE and MCC per merged alignment. C and D show the overall results averaged across all benchmarked alignments.

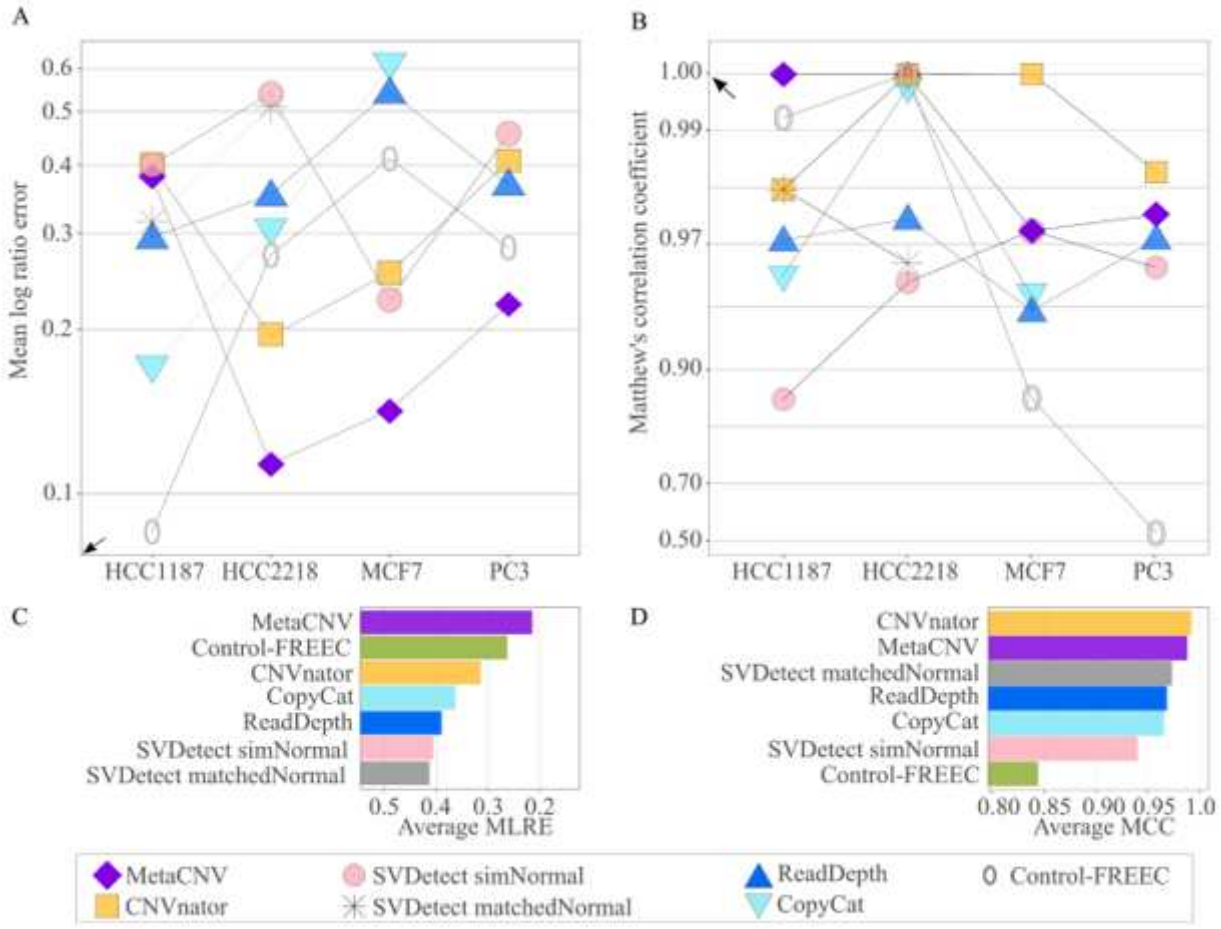


Figure 5

Prediction accurateness of MetaCNV and other copy number callers for high coverage data. A and B present the average benchmark results per cancer cell line. SVDetect using matchedNormal could only be performed for HCC1187 and HCC2218, for which matched blood samples were available. C and D show the overall results averaged across all benchmarked cell lines.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- metaCNVsupplementv2reSubmission.docx
- CCLknownCNVs.xlsx
- README.txt