# Minimum Error Calibration and Normalization for Genomic Copy Number Analysis

Bo Gao[1,2] and Michael Baudis[1,2]

[1]Institute of Molecular Life Sciences, University of Zurich
[2]Swiss Institute of Bioinformatics

2020-04-30

## Abstract

### Background

Copy number variations (CNV) are regional deviations from the normal autosomal bi-allelic DNA content. While germline CNVs are a major contributor to genomic syndromes and inherited diseases, the majority of cancers accumulate extensive "somatic" CNV (sCNV or CNA) during the process of oncogenetic transformation and progression. While specific sCNV have closely been associated with tumorigenesis, intriguingly many neoplasias exhibit recurrent sCNV patterns beyond the involvement of a few cancer driver genes. Currently, CNV profiles of tumor samples are generated using genomic micro-arrays or high-throughput DNA sequencing. Regardless of the underlying technology, genomic copy number data is derived from the relative assessment and integration of multiple signals, with the data generation process being prone to contamination from several sources. Estimated copy number values have no absolute or strictly linear correlation to their corresponding DNA levels, and the extent of deviation differs between sample profiles, which poses a great challenge for data integration and comparison in large scale genome analysis.

### Results

In this study, we present a novel method named "Minimum Error Calibration and Normalization for Copy Numbers Analysis" (*Mecan4CNA*). It only requires CNV segmentation files as input, is platform independent, and has a high performance with limited hardware requirements. For a given multi-sample copy number dataset, *Mecan4CNA* can batch-normalize all samples to the corresponding true copy number levels of the main tumor clones. Experiments of *Mecan4CNA* on simulated data showed an overall accuracy of 93% and 91% in determining the normal level and single copy alteration (i.e. duplication or loss of one allele), respectively. Comparison of estimated normal levels and single copy alternations with existing methods and karyotyping data on the NCI-60 tumor cell line produced coherent results. To estimate the method's impact on downstream analyses, we performed GISTIC analyses on the original

and *Mecan4CNA* normalized data from the Cancer Genome Atlas (TCGA) where the normalized data showed prominent improvements of both sensitivity and specificity in detecting focal regions.

**Conclusions**

*Mecan4CNA* provides an advanced method for CNA data normalization, especially in meta-analyses involving large profile numbers and heterogeneous source data quality. With its informative output and visualization options, Mecan4CNA also can improve the interpretation of individual CNA profiles. *Mecan4CNA* is freely available as a Python package and through its code repository on Github.

# Introduction

Copy number aberrations (CNA[1]) represent the gain and loss of DNA compared to the normal bi-allelic status, for genomic regions of varying sizes. Somatic copy number aberrations are a typical hallmark of cancer with complex and intrinsic connections to the development of many malignant diseases [1, 2]. The determination of CNA profiles is a core element of cancer genome analyses in research projects and also used in clinical practice for different types of cancer. While disease-related mutational patterns and genes associated with CNAs have been delineated in various cancer types [3], it is expected that rapid advances in genomic high-throughput technologies and the accumulation of massive amounts of data from cancer studies will reveal more comprehensive patterns in the near future [4]. However, although CNA data has become a principal element in cancer genome analysis, the reliable processing and interpretation of such data still pose challenges due to heterogeneous technologies, data formats and data quality.

For the last two decades, the predominant methods for the generation of genome-wide CNA profiles have been based on genomic micro-array technologies [5, 6, 7]. More recently, high-throughput sequencing technologies have been co-opted to generate DNA copy number profiles in rare disease diagnostics and cancer genome analyses [8, 9, 10]. Independent of technology or specific platform, CNA data often suffers from signal deviations (Figure 1), which can be attributed to three main sources. The main component influencing the signal levels is the varying clonal purity of the tumor sample. Ideally, a derived DNA copy number should be an integer value representing the DNA (allele) count at a given genome position, as it exists in the cells of a clonal tumor cell population. In practice, a given tumor sample frequently represents a mixture of normal cells (e.g. stroma), cells from the predominant malignant clone and cells from minor tumor clones (sub-clones) representing different branches in the malignant evolution process. While stromal admixture leads to a uniform attenuation of the CNA signal, sub-clonal components can result in regional and fractional divergences from the copy number levels expected from a homogeneous tumor sample. Additionally, CN values are derived from the relative measurement of DNA content, with a possibly non-linear correlation between the true regional CN and the measured - intensity or count based - signal. Also, systematic errors may accumulate in each experimental step throughout the pipeline, e.g. in sample preparation, DNA labelling or hybridization procedures. In theory, the sources of experimental noise can be alleviated by performing additional

[1] Alternatively to "Copy Number Aberrations" (CNA) the term "Somatic Copy Number Variations" (sCNV) is being used in the literature. We prefer the CNA term due to precedence and epistemic relation to the oncogenetic process.

experiments, increasing the amount of source material used or the depth of sequencing, or through fine tuning of experimental and analytical protocols. However, such efforts are negatively impacted by scarce or qualitatively limited source material (e.g. archival tissue), associated costs (e.g. for technical replicates or increased read depth) or analytical overhead. Most importantly, *experimental* improvements cannot be pursued in meta-analyses on pre-existing data, which possibly was derived from a variety of sources.

Currently, no "Gold Standard" solution addresses the problem of copy number calibration and the generation of integer CN counts, in a consistent and universal manner. In primary research and clinical studies, the most common approach lies in a supervised assessment under the potential inclusion of contextual information. This methodology can work well for limited numbers of profiles with simple aberration patterns; however, it is dependent on individual observer experience, shows limited performance for highly complex CNA profiles and is incompatible with the consistent assessment of CNA profiles for meta-analyses which potentially include tens of thousands of genome profiles.

To allow a basic comparison of CNA profiles from different experiments, a widely used approach has been to apply median centering of probe values or derived CNA segments. However, while this statistical methodology is intuitive and can be easily implemented, it also has been shown to struggle when processing complex CNA profiles [11, 12]. In the last decade, several computational methods have been developed to address the detection of tumor purity or actual DNA levels [13, 14, 15, 16, 17, 11, 18, 19, 12]. The *ABSOLUTE* method [11] uses allele specific copy number ratios and pre-computed models to estimate purity and ploidy, and applies this information to computes potential models of absolute copy numbers for individual genome regions. *BACOM* [12] exploits allele specific copy number signals with a Bayesian model to differentiate homozygous and heterozygous deletions and to estimate normal cell fractions. *AbsCN-seq* [19] uses a statistical method to infer purity, ploidy and absolute copy number. These computational methods can provide a reliable estimation of regional copy number levels, for samples with sufficient, homogeneous data quality and appropriate access to source data (e.g. allelic SNP information). Additionally to the calibration of individual CNA profiles, other methods rely on the multi-sample statistics of probe signals or segmentation data without resolving each individual sample by itself [20] and usually are being applied to homogeneous series of cancer profiling experiments or otherwise homogeneous input data [21, 22, 23].

However, in many scenarios, the application of the aforementioned computational methods may not be feasible for one or more reasons. First, comparative analysis across multiple original studies typically has to rely on segmented CNA data instead of raw array- or sequencing data, since a) the original data may not be accessible due to privacy concerns, administrative unspecified reasons; and b) the re-analysis of data from legacy platforms may not be practically feasible. Ploidy estimation methods frequently rely on specific information from raw source files and are limited to selected platforms which even with available source data may limit their application to subsets represented by supported platforms. Additionally, raw data based ploidy estimation methods can have extensive demand with respect to computational hardware and processing time, especially when processing high-volume datasets. To our knowledge, currently no generic method addresses the normalization of copy number data for aggregated analysis, in a manner compatible with the limitations described.

In this study, we present Minimum Error Calibration and Normalization for Copy Numbers Analysis (Mecan4CNA), a generic computational method aimed to address issues associated with the meta-

analysis of very large, heterogeneous copy number profiling datasets in cancer genomics. The method does not require access to raw data; it is platform agnostic; and its performance enables the processing of very large CNA datasets. As input, *Mecan4CNA* solely uses copy number segmentation files, which can be generated as a standard output by most platform-specific or generic CNA software packages or processing pipelines. By primarily estimating normal signal levels and determining abnormal single-count deviations, the method avoids the intricacies of deterministic ploidy identification. In this fashion, *Mecan4CNA* can calibrate and normalize large sets of copy number data both accurately and efficiently. After applying *Mecan4CNA*, all copy number values will be aligned to the corresponding true copy number levels of the main tumor clones (i.e. "3" corresponding to 3 alleles at any given region, across all samples) and be ready for follow-up analysis of regional CNA involvement as well as comparisons of whole-genome CNA patterns between experiments. Benchmarking analyses on simulated and real data showed the consistency of *Mecan4CNA* over a wide range of input data qualities.

# Method

A generic CNA segmentation file stores information about regional copy number values as structured text. Usually, each line represents a chromosome segment with its position, size and signal intensity, with possible addition of optional data (e.g. probe count in segment, categorical CNA status). As explained in the introduction, the signal intensity reflects a value derived from a possible admixture of cell types (normal cells, main tumor cells and sub-clonal tumor cells). This preservation of information is crucial for studying the details of tumor ploidy, but it also makes the copy number data difficult to interpret and compare.

Let $X$ denotes a CNA profile and $x_i$ denotes the signal value of one record in $X$ (i.e. one segment), then a CNA profile can be modelled as the following:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{bmatrix}, \; x_i = (aN_i + bT_i + \sum_n c_i^n S_i^n + \sum_m E_i^m) \prod_j (1 + e_j), \quad for \; x_i \in X, \; a + b + \sum_n c_i^n = 1$$

where $N_i, T_i \; and \; S_i$ denote the actual copy numbers of normal cells, primary clone tumor cells and sub-clonal cells, respectively. Likewise, $a$, $b$ and $c$ are the composition ratio of normal cells, primary clone tumor cells and sub-clonal cells of this segment, respectively. $E_i^m$ represents an independent error of the segment and $e_j$ represents a systematic error of the sample in each processing step.

Normally, for a copy number profile $X$, one needs to solve $a$, $b$ and $c_i$ to reveal the purity of the sample, and to solve at least $T_i$ to know the ploidy of the main tumor. Both tasks are very challenging because the only known value from a segmentation file is $x_i$. For this reason, computational methods usually need to rely on additional information besides the segmentation profile (e.g. raw data). In this study, our goal is to calibrate and normalize $x_i$ from different profiles. This change of aim allows us to circumvent the difficulty of evaluating the exact purity and ploidy. Instead, we focus on estimating the deviation of the normal level and the corresponding value of one copy alternation in each profile. By pursuing a sub-optimal partial solution, we can dramatically reduce the computation complexity.

Although we are solving for a partial solution, we still have to deal with the problem of insufficient known values. Specifically, we eliminate variables and compute determined values by approximations and transformations. First, we reduce the complexity of the representation of $x_i$. For each $x_i$, the two terms $\sum_m E_i^m$ and $\prod_j (1+e_j)$ are always constants. If we also consider the summation of sub-clones as a single pseudo-sub-clone, $x_i$ can be simplified as the following:

$$\begin{aligned} x_i &= (aN_i + bT_i + cS_i + E_i')(1+e) \\ &= aN_i + bT_i + cS_i + E_i \end{aligned}$$

$$1 = a + b + c$$

where $a$, $b$, $c$ and $e$ are constants in each sample. $N_i, T_i, S_i$ and $E_i$ are variables of each $x_i$. Here, $E_i$ represents the integral of all errors.

Next, we introduce two new values: the distance and the ratio of the distance. Let $D(i,j)$ be the distance of $x_i$ and $x_j$:

$$\begin{aligned} D(i,j) &= |x_i - x_j| \\ &= |(aN_i + bT_i + cS_i + E_i) - (aN_j + bT_j + cS_j + E_j)| \end{aligned}$$

Because germline copy number changes are rare events and are usually filtered by most processing pipelines, we can assume $N_i$ to be a constant of 2. Then, $D(i,j)$ can be simplified as:

$$D(i,j) = |b(T_i - T_j) + c(S_i - S_j) + E_{i,j}|$$

Let $R(i,j,k)$ represents the ratio of two distance $D(i,j)$ and $D(i,k)$

$$R(i,j,k) = \begin{cases} \frac{D(i,k)}{D(i,j)}, & \text{if } D(i,k) \geq D(i,j) \\ \frac{D(i,j)}{D(i,k)}, & \text{otherwise} \end{cases}$$

Here, we illustrate the deduction of one case, and the other scenario is the same.

$$\begin{aligned} R(i,j,k) &= \frac{D(i,k)}{D(i,j)} \\ &= |\frac{b(T_i - T_k) + c(S_i - S_k) + E_{i,k}}{b(T_i - T_j) + c(S_i - S_j) + E_{i,j}}| \\ &= |\frac{T_i - T_k}{T_i - T_j}|(1 + |\frac{c(S_i - S_j) + E_{i,j}}{b(T_i - T_j) + c(S_i - S_j) + E_{i,j}}|) + |\frac{c(S_i - S_k) + E_{i,k}}{b(T_i - T_j) + c(S_i - S_j) + E_{i,j}}| \end{aligned}$$

The term $|\frac{T_i - T_k}{T_i - T_j}|$ is the ratio of the distances between actual copy number levels. When $b > c$, which means a biopsy where the proportion of the main tumor is higher than the proportion of sub-clones, we can have:

$$\frac{c(S_i - S_j) + E_{i,j}}{b(T_i - T_j) + c(S_i - S_j) + E_{i,j}} < 1 \ , \ \frac{c(S_i - S_k) + E_{i,k}}{b(T_i - T_j) + c(S_i - S_j) + E_{i,j}} < 1$$

$$R(i,j,k) \approx |\frac{T_i - T_k}{T_i - T_j}| + e_{i,j,k}$$
$$= r_{i,j,k} + e_{i,j,k}$$

The new term $e_{i,j,k}$ reflects the deviation from the actual copy number, it reaches zero when $x_i, x_j$ and $x_k$ are the same value as the actual copy number level. Its value is influenced by the purity of the sample. If the proportion of the primary tumor falls below 50% in the sample, the error term may become the dominant value. However, in most experiments, the purity of the primary tumor is usually at a reasonable level. If we can find a tuple of $x_i$ and $x_j$ that gives the smallest $e_{i,j,k}$, we can say that this tuple represent two copy number levels in a solution of $X$. Moreover, that is to calculate the following equation:

$$argmin\ f(i,j) = \{\sum_k e_{i,j,k} | i,j,k \in X\}$$

Furthermore, if the copy number difference of $T_i$ and $T_j$ is one, which means $|T_i - T_j| = 1$, then $r_{i,j,k}$ becomes an integer. Usually, the segmentation data is generated from pipelines, where the systematic error is properly reduced to a level that is smaller than the deviation. Therefore, we can evaluate $e_{i,j,k}$ as the following:

$$e_{i,j,k} = I_{i,j,k} - R_{i,j,k},\ where\ I_{i,j,k} \text{ is the nearest integer to } R_{i,j,k}$$

Theoretically, we have to evaluate $f(i,j)$ for all pairwise combinations in $X$, which is a nightmare of computation as it involves a tensor *e(i,j,k)*. In practise, there are two factors that can help reduce the daunting searching space to a constant. First, we know the actual copy number values are integers, therefore, the copy number values in a segmentation file are normally distributed around the deviations of these actual values. By computing the mean of each local normal distribution, we can reduce the number of candidate values of *i, j and k* to a minimal set of size *n*, where $n$ is usually no more than 20 (7 possible integer copy number values from 0 to 6, signals from normal cells, the main tumor and a few sub-clones). Second, because we want to find the value of the normal copy number level, we always assume one of *i and j* is the normal level. Therefore, we only need to evaluate a maximum of $n$ combination for each sample. Figure 2 illustrates the general problem and why we can reduce the search space.

Now, we have acquired a set of potential solutions: a pair of $x_i$ and $x_j$, where $D(i,j)$ is 1 and one of them is the normal level. In the last step, we need to determine which one is actually the normal level and which one is the one copy alternation. By using a weighted function, we consider the strength of the signal, the distance to the center and the validness of models comprehensively. One with a higher $B_{score}$ is determined as the normal level:

$$Bscore = \frac{S_{signal} + S_{distance} + S_{model}}{3}$$

Finally, let *base* and *alt* represent the values that corresponds to the normal copy number level and one copy alternation in a segmentation profile, respectively; let $x'_i$ denotes the normalized value. We calibrate and normalize the value of each segment through the following linear transformations:

$$x'_i = \frac{x_i - base}{|base - alt|} + 2$$

After normalization, the copy number value of each segment will be aligned to the actual integer copy number level of the main tumor clone. In a segmentation file, the values come as log2 ratios. They will be converted to integer copy number values for the computation and converted back to log2 ratios after normalization.

# Results

In order to validate the performance of *Mecan4CNA*, first we compared the estimation results of our method with results from karyotyping and ABSOLUTE on cell line data. Next, we applied the method on a series of simulated data to evaluate its consistency and generalization ability. Additionally, we explored the baseline deviation situation of copy number data from TCGA. Finally, we used *Mecan4CNA*-normalized TCGA data as input for a GISTIC [20] analysis, to example a possible improvement compared to using the unprocessed original data.

## Performance on real data

The NCI-60 tumor cell lines were selected by the National Cancer Institute of the United States (NCI) as a reference panel for drug screening experiments. It comprises 60 human cancer cell lines from 9 different cancer types. Their molecular characteristics have been well studied in the past two decades. Particularly, 58 cell lines have been karyotyped using the spectral karyotyping protocol [24], and the copy number variations of all cell lines have been profiled using microarray experiments [25]. DNA copy number changes on chromosome 13 were explored in detail in the karyotyping study. Among all karyotyped cell lines, 30 have either no changes or only chromosome level changes on chromosome 13. They represent ideal examples to validate the performance of our method, because we can match the karyotyping data with the microarray data to identify the copy number values (from microarrays) that correspond to the actual normal and abnormal copy number levels (from karyotyping). If we compare these values with the estimation results of *Mecan4CNA* on chromosome 13, we can evaluate the performance of the method.

When comparing data from microarrays with the karyotyped standard, 6 cell lines showed contradicting results and were excluded from further analysis (5 karyotyped as one copy loss but microarray showed normal; 1 karyotyped as one copy loss but microarray showed copy gain). These differences may be explained by a possible clonal evolution between the reference analysis and the cell line batches from which the microarray data was prepared. The remaining 24 cell lines were used for further analysis. In order to compare the performance of *Mecan4CNA* with existing methods, we also used ABSOLUTE, a widely used method to estimate purity and ploidy from copy number data, to infer the corresponding copy number levels of each cell line. Figure 3 shows the comparison results of *Mecan4CNA* and ABSOLUTE, respectively. On both graphs, the estimated values were plotted against the actual value of a copy number level. *Mecan4CNA* achieved 0.987 for spearman correlation coefficient and 0.033 for root mean square error (RMSE). Both scores indicate a strong and confident correlation between the estimation and actual values. ABSOLUTE achieved 0.988 for spearman correlation coefficient and 0.049 for root mean square error (RMSE). This high concordance confirmed the solid performance of both methods on estimating copy number levels. Data is available in Supplementary S1.

## Performance on simulated data

To further evaluate the performance of *Mecan4CNA*, we applied the method to 10 simulated datasets of different cell composition and noise levels. Every dataset comprises 100 samples, which are copy number segment profiles generated based on our modeling equation introduced in the method section. The cell composition was generated using a Dirichlet distribution. In samples where the main tumor is dominant in the composition (later referred to as: high tumor), the composition ratio of normal and sub-clones are capped at 0.2. Otherwise, the sample is considered to have a low main tumor proportion (later referred to as: low tumor). For both high tumor datasets and low tumor datasets, the independent errors were generated using a normal distribution with the standard deviation increasing from 0.001 to 0.009 with a step size of 0.002; the global errors were generated using a normal distribution with the standard deviation increasing from 0.1 to 0.18 with a step size of 0.02. The ploidy of normal cells was assumed to always be two. A total of 1000 samples were generated in this manner, and both the actual cell composition and tumor ploidy were known.

The five datasets containing samples with a high contribution of the main (virtual) tumor clone show a gradual increase of noise levels from dataset to dataset. A similar observation can be made in the other five datasets with low tumor composition (high sub-clone or normal). We used *Mecan4CNA* to estimate the values of the normal level and one copy alternation of each sample, then compare the estimation accuracy of the normal level (later referred as baseline) and the distance between the normal level and one copy alternation level (later referred as level distance).

As shown in Figure 4, the estimation accuracy of *Mecan4CNA* decreases when the noise level increases or when the tumor composition decreases. Specifically, when under the same noise level, for both the baseline and level distance, the estimation generates more outliers in low tumor datasets. This is often an indication of a low quality sample or a very complex sample. When under the same tumor composition, the estimation accuracy only shows very slight declines as the noise level increases. It shows that the quality of the sample has a greater impact on the accuracy of segmentation than the level of noise. When comparing the estimation accuracy of the baseline and level distance, the baseline estimation shows better consistency, even when facing the low tumor and high noise data. This is because even in low tumor data, as long as the genome is not in complete chaos, which occasionally happens in cancer, the signal of normal DNA level (baseline) is often still detectable. However, when the biopsy has a high proportion of normal and sub-clone cells, the signal of the main tumor will be significantly reduced, and the signal of sub-clones becomes more influential at the same time. The combination of these two effects makes it much more difficult to have an accurate estimation of the main tumor's abnormal copy number levels in samples with low tumor composition. *Mecan4CNA* provides conservative estimates of the level distance since overestimation can lead to undesired information loss during normalization, and underestimation only introduces false-positive information.

Overall, *Mecan4CNA* performed with high accuracy and consistency in all scenarios. Specifically, for the estimation of baseline, *Mecan4CNA* achieved 97% average accuracy in the combination of all high tumor composition datasets; 89% average accuracy in the combination of all low tumor composition datasets; and 93% overall accuracy of all datasets. For the estimation of level distance, *Mecan4CNA* achieved 96% average accuracy in the combination of all high tumor composition datasets; 86% average accuracy in the combination of all low tumor composition datasets; and 91% overall accuracy of all

datasets. In the method section, we discussed that the proportion of the main tumor should be much higher than the sub-clones, and the noise level should have been calibrated and reduced by the pipeline. However, *Mecan4CNA* showed good and consistent performance when dealing with samples of high heterogeneity and noise level. It further illustrates the consistency and generalization capability of *Mecan4CNA*.

**Performance on low cellularity data**

To test *Mecan4CNA's* capacity in extreme cases, we further simulated two datasets of very low cellularity. One dataset is comprised of 20˜40% primary tumor and 40˜50% normal cells, and the other dataset is comprised of 20˜40% primary tumor and 40˜50% sub-clonal cells. The identification of the main tumor's copy number status is extremely challenging in both datasets, because the signal is heavily clouded and distorted by the excessive normal and sub-clonal cells. For the first dataset, *Mecan4CNA* archived 72% accuracy in estimating the baseline and 75% accuracy in estimating the level distance. When the tumor cells are admixed with a high proportion of normal cells, the signal intensity of abnormal copies is significantly reduced [Fig 9(a)], thus rendered in similarity to a noisy signal of the "normal" level. We used a strategy of increasing the sensitivity (at the cost of decreased specificity) to capture subtle changes. However, this is only applicable if the admixture composition is already known (in practice e.g. from histopathology). For the second dataset, estimation accuracy dropped considerably to 60% accuracy for the baseline and 54% for the level distance. When the primary tumor cells are accompanied with sub-clonal cells of high heterogeneity and proportion, their signal is rivaled by the noise of equal strength [Fig 9(b)] and therefore extremely difficult to identify. Without additional information, even manual interpretation failed to give definitive solutions in most cases. While we reduced the resolution of signal binning to tolerate more noise, this also is related to prior knowledge about the cellularity and only provided marginal improvement. Without fine tuning of the *Mecan4CNA* parameters based on known cellular composition, the performance in both scenarios was below 50%. The accurate interpretation of low cellularity samples represents a significant challenge and our testing indicates that the pure computational approach frequently may be insufficient to solve cases with low cellularity. In practice, the assessment can be improved when combined with auxiliary data and domain expertise.

**Performance on difference segmentation methods**

Accurate segmentation is an important and challenging step in the analysis of copy number profiles with a number of computational methods having been proposed in the last two decades. However, there is still a lack of a generic methods to fulfill all practical scenarios and the performance of different methods is usually context dependent [26]. In order to understand the influence of different segmentation methods on *Mecan4CNA's* performance, we compared the estimation results on three tumor datasets (Table 4) using two different segmentation methods. In order to achieve broad coverage and diversity, first, we chose three tumor data datasets of different diseases and platforms. Then, each dataset was being segmented using two methods: *DNAcopy*, which is based on Circular Binary Segmentation (CBS); and *copynumber*, which is based on Piecewise Constant Fitting (PCF). Figure 8 shows the estimation results of both the baseline and the level distance, where the comparison of the two methods shows

high consensus. Specifically, 95% of baseline comparisons and 92% of level distance comparisons show identical or close (less than 10% difference in value) results. Next, we inspected the samples with inconsistent estimations and summarized those into two causes: low quality of the raw data (e.g. high noise data in GSM889595) and ambiguous interpretation (e.g. strong signals from high-level duplications and deletions in GSM1098760). In all inconsistent cases, at least one of the segmentation method generated the correct estimate. By tuning parameters of the segmentation method, all inconsistent estimations could be corrected. *Mecan4CNA* relies on the distribution of signals instead of the relations of segment levels and is robust against regional inconsistency or high level of fragments. The benchmark on two different segmentation methods showed *Mecan4CNA's* capacity to tolerate a moderate discrepancy of different segmentation methods in producing coherent results.

## Applications of *Mecan4CNA*

### Baseline deviation of TCGA data

The Cancer Genome Atlas (TCGA) provides a large collection of copy number data from 33 cancer projects. The copy number data was generated using Affymetrix SNP6 microarrays and a uniform processing pipeline. We used *Mecan4CNA* to estimate the baseline value of all TCGA copy number data, summarized by the project. As shown in Figure 5, baseline (the normal copy number level) deviation is widespread with varying contributions related to individual projects. In general, the deviation level is strongly correlated with the difficulty of acquiring pure biopsies. In projects targeting diseases such as acute myeloid leukemias (e.g. TCGA's LAML), where biopsies are usually pure, or cervical squamous cell carcinomas and endocervical adenocarcinomas, with predominant low amount of intra-tumor heterogeneity [27, 28], the original baseline levels are often accurate or have a low deviation. In entities such a lung squamous cell carcinomas, where biopsies frequently are difficult to obtain and different biopsy methods may lead to a high heterogeneity [29], or serous ovarian (cyst-)adenocarcinomas, where tumor cells are frequently in mixture with immune cells [30], the original baseline estimates frequently deviate by a large amount from the optimal value.

Although great efforts went into limiting errors and noise in TCGA copy number data, they originated from projects addressing diverse diseases and where data generation were different in time, space and experimental conditions. The variation in the results does not directly reflect on the accuracy of the results since the determination of an underlying "Gold Standard" is beyond the scope of our methodology. The generally good quality of these datasets allowed us to show that baseline variation is a common and recurring problem among copy number data.

### Normalized data as GISTIC input

In order to demonstrate the effect of *Mecan4CNA* in copy number data analysis, we applied GISTIC with both the original and the normalized copy number data from the TCGA project. GISTIC is a widely used multi-sample tool to detect and score focal regions in copy number datasets. We chose data from 3 TCGA projects with different baseline deviation levels to look for focal regions and potential driver genes using GISTIC. GISTIC was run in *gene-gistic* mode with default parameters (Supplementary S2).

Table 1 illustrates the comparisons of called regions from GISTIC using the original and normalized data. Overall, when using the normalized data, the number of detected regions is significantly reduced

(Figure 6). At the same time, when mapping focal regions to Cancer Census Genes [31], the number the detected drivers actually increases as shown in Figure 7. Some driver genes are only captured using the normalized data, such as PDGFRA in GBM, XPC in SKCM and CASP3 in OV. For focal regions detected by both settings, the size of the region is usually reduced by the normalized data. A few new focal regions containing driver genes, which were too weak in signal when using the original data, became significant and were detected using the normalized data. For example, in the SKCM datasets, region *chr6:9090525-9120556*[2] was detected by the original data with high significance but harbored no protein coding genes. When using the normalized data, the signal strength of this region was reduced dramatically and it was not called as a focal region. Region *chr12:68834906-68893964* was called only when using the normalized data and harbored two known oncogenes (CMP and MDM2). Region *chr1:204403604-204437602* was narrowed and shifted to *chr1:204477206-204586780* by using the normalized data; this subtle change removed a passenger gene PPP1R15B and included the oncogene MDM4.

The Molecular Signatures Database (MSigDB) [32] is a collection of annotated gene sets for use with GSEA software. The "C6: oncogenic signatures" represent signatures of cellular pathways that are often dis-regulated in cancer. We mapped the genes called by GISTIC using the original and the normalized data to each pathway in C6. Table 2 shows pathways that are covered by a high number of genes in both data settings. Pathways in bold are with known associations to melanoma, for example, KRAS and P53. Although these pathways stood out in both data settings, most of these pathways were covered with more genes when using the normalized data. Table 3 shows pathways with significant coverage difference between the original and the normalized data. The coverage of several pathways (KRAS, PTEN, MYC), which have strong associations with tumorigenesis in melanoma, were significantly increased when using the normalized data. In summary, when using the normalized data instead of the original data, the analysis result of GISTIC showed a prominent improvement in both sensitivity and specificity.

## Conclusions

In this study, we presented a novel method dubbed "Minimum Error Normalization for Copy Numbers" (*Mecan4CNA*), aimed at supporting calibrated meta-analyses on heterogeneous genomic copy number profiling datasets. We tested the method on simulated and cell line data and showed promising results in estimating baseline and copy number levels from segmented copy number profiles, without the need for probe intensities or SNP status information. When applied to the copy number profiling data from the TCGA project, we were able to detect frequent baseline deviations, which could be primarily attributed to the purity of the corresponding samples. Finally, the comparison of results from GISTIC analyses, when using the original and *Mecan4CNA* normalized data as input, showed that the normalized data could provide a significant improvement in both sensitivity and specificity for the detection of focal alterations of cancer-related genes. As a generic tool for analyzing copy number datasets, *Mecan4CNA* does not rely on any additional data types beyond the segmented data from copy number analysis pipelines, and is also efficient in speed and resource utilization. While the primary utility of the method

[2] base positions according to reference genome GRCh38

can be found in the meta-analysis of large CNA datasets, *Mecan4CNA* also can facilitate single sample CNA interpretation.

## Availability and requirements

Home page: https://github.com/baudisgroup/mecan4cna
Operating system(s): Platform independent
Programming language: Python
Requirements: Python3.6 or higher
License: MIT
Restrictions: None

## List of abbreviations

- CNV: Copy Number Variations
- sCNV: Somatic Copy Number Variations
- CNA: Copy Number Aberrations
- Mecan4CNA: Minimum Error Calibration and Normalization for Copy Numbers Analysis
- SNP: Single Nucleotide Polymorphism

## Declarations

### Ethics approval and consent to participate

Not applicable

### Consent for publication

Not applicable

### Competing interests

The authors declare that they have no competing interests.

### Author's contributions

BG developed the method and implemented the software tool. MB conceived the original concept and supervised the project. Both authors contributed to the writing and revisions of the manuscript.

## Availability of data and material

- The NCI-60 karyotyping data used during the current study is available in the Gene Expression Omnibus repository (GSE32264: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32264`)
- The TCGA datasets used during the current study are available in the TCGA Research Network: `https://www.cancer.gov/tcga`
- The C6 pathway data used during the current study is available in the Molecular Signatures Database, `http://software.broadinstitute.org/gsea/msigdb/genesets.jsp?collection=C6`


## Funding

BG is recipient of a grant from the China Scholarship Council. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Acknowledgements

Not applicable


## Supplementary material

S1. NCI-60 experiment results: NCI60_results.csv

S2. GISTIC score plots: gistic_peaks.pdf

# References


[1] Baudis M. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. BMC Cancer. 2007;7(1):226.

[2] Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463:899–899.

[3] Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nature Genetics. 2013;45:1134–1134.

[4] Nakagawa H, Wardell C, Furuta M, Taniguchi H, Fujimoto A. Cancer whole-genome sequencing: present and future. Oncogene. 2015;34(49):5943–5950.

[5] Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer. 1997;4(20):399–407.

[6] Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet. 1998;2(20):207–211.

[7] Zhao X, Li C, Paez J, Chin K, Jänne P, Chen T, et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res. 2004;64(9):3060–3071.

[8] Campbell P, Stephens P, Pleasance E, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008;40(6):722–729.

[9] Ley T, Mardis E, Ding L, Fulton B, McLellan M, Chen K, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008;456(7218):66–72.

[10] Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Research. 2009;19(9):1586–1592.

[11] Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nature Biotechnology. 2012;30:413–413.

[12] Fu Y, Yu G, Levine DA, Wang N, Shih IM, Zhang Z, et al. BACOM2.0 facilitates absolute normalization and quantification of somatic copy number alterations in heterogeneous tumor. Scientific Reports. 2015;5:13955–13955.

[13] Attiyeh EF, Diskin SJ, Attiyeh MA, Mossé YP, Hou C, Jackson EM, et al. Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. Genome Research. 2009;19(2):276–283.

[14] Butler A, Greenman CD, Beare D, Bignell G, Hinton J, Chen L, et al. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics. 2009;11(1):164–175.

[15] Bengtsson H, Neuvial P, Speed TP. TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. BMC Bioinformatics. 2010;11(1):245.

[16] Yau C, Mouradov D, Jorissen RN, Colella S, Mirza G, Steers G, et al. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome Biology. 2010;11(9):R92.

[17] Gusnanto A, Wood HM, Rabbitts P, Berri S, Pawitan Y. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics. 2011;28(1):40–47.

[18] Oesper L, Mahmoody A, Raphael BJ. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biology. 2013;14(7):R80.

[19] Messer K, Bao L, Pu M. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. Bioinformatics. 2014;30(8):1056–1063.

[20] Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology. 2011;12(4):R41.

[21] Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, et al. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics. 2010;27(2):268–269.

[22] Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Research nar. 2015;43(6):e39.

[23] Rigaill G, Hupé P, Almeida A, La Rosa P, Meyniel JP, Decraene C, et al. ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. Bioinformatics. 2008;24(6):768–774.

[24] Roschke AV, Tonon G, Gehlhaus KS, McTyre N, Bussey KJ, Lababidi S, et al. Karyotypic Complexity of the NCI-60 Drug-Screening Panel. Cancer Research Cancer Res. 2003;63(24):8634.

[25] Ruan X, Kocher JPA, Pommier Y, Liu H, Reinhold WC. Mass Homozygotes Accumulation in the NCI-60 Cancer Cell Lines As Compared to HapMap Trios, and Relation to Fragile Site Location. PLOS ONE. 2012;7(2):e31628.

[26] Ruan J, Liu Z, Sun M, Wang Y, Yue J, Yu G. DBS: a fast and informative segmentation algorithm for DNA copy number analysis. BMC Bioinformatics. 2019;20(1):1.

[27] Guo Z, Pontén F, Wilander E, Pontén J. Clonality of Precursors of Cervical Cancer and Their Genetical Links to Invasive Cancer. Modern Pathology. 2000;13(6):606–613.

[28] Ueda Y, Enomoto T, Miyatake T, Ozaki K, Yoshizaki T, Kanao H, et al. Monoclonal Expansion with Integration of High-Risk Type Human Papillomaviruses Is an Initial Step for Cervical Carcinogenesis: Association of Clonal Status and Human Papillomavirus Infection with Clinical Outcome in Cervical Intraepithelial Neoplasia. Laboratory Investigation. 2003;83(10):1517–1527.

[29] Vanderlaan PA, Yamaguchi N, Folch E, Boucher DH, Kent MS, Gangadharan SP, et al. Success and failure rates of tumor genotyping techniques in routine pathological samples with non-small-cell lung cancer. Lung Cancer Lung cancer (Amsterdam, Netherlands). 2014;84(1):39–44.

[30] Zhang L, Conejo-Garcia JR, Katsaros D, Gimotty PA, Massobrio M, Regnani G, et al. Intratumoral T Cells, Recurrence, and Survival in Epithelial Ovarian Cancer. New England Journal of Medicine N Engl J Med. 2003;348(3):203–213.

[31] Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer. 2018;18(11):696–705.

[32] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov J, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems. 2015;1(6):417–425.

# Appendix

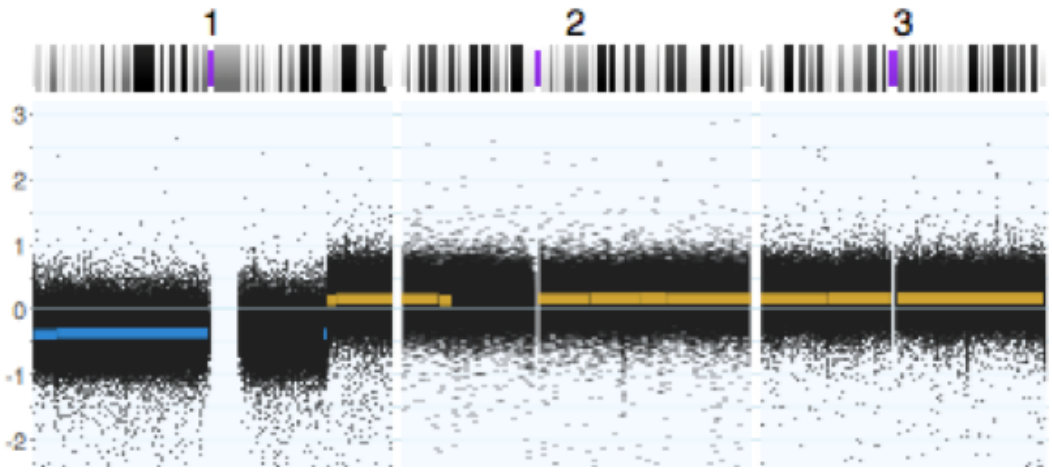


Figure 1: The detected SNPs (black dots) and called segments (colored lines) of the copy number profile of a gastrointestinal stromal tumor. The labels of y-axis are the expected copy numbers. Only showing chromosome 1, 2, and 3.

| | Data | No. regions | No. genes | No. Census genes | Gene density | Driver density |
|---|---|---|---|---|---|---|
| GBM | Original | 29/46 | 143/1722 | 19/45 | 4.93/37.43 | 0.13/0.03 |
| | Normalized | 26/40 | 110/1953 | 19/50 | 4.23/48.83 | 0.17/0.03 |
| SKCM | Original | 33/37 | 1494/1721 | 58/43 | 45.27/46.51 | 0.04/0.02 |
| | Normalized | 26/21 | 1860/2099 | 47/66 | 71.54/99.95 | 0.03/0.03 |
| OV | Original | 38/51 | 423/2515 | 13/64 | 11.13/49.31 | 0.03/0.03 |
| | Normalized | 35/36 | 382/2240 | 13/54 | 10.91/62.22 | 0.03/0.02 |

Table 1: Summary of detected regions of GISTIC using original and normalized data. Numbers in column represents the count of gain/loss.

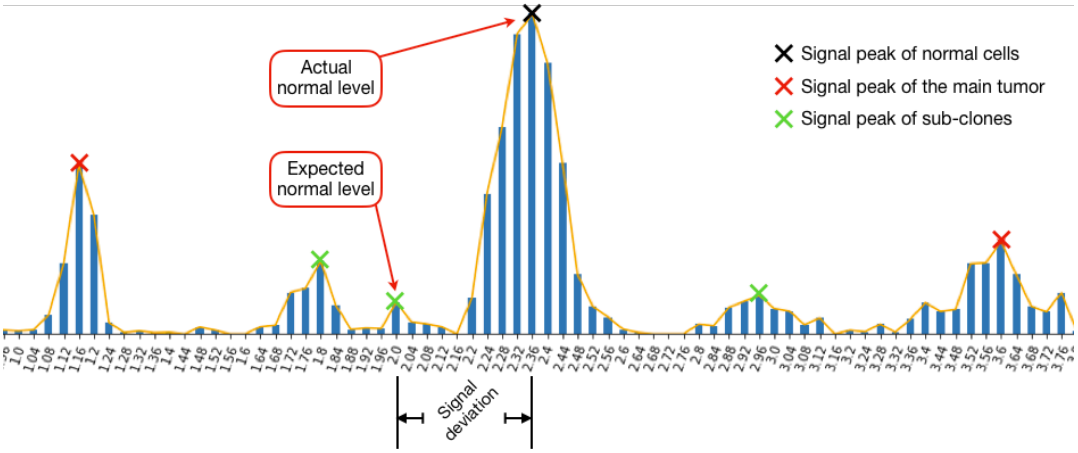


Figure 2: The central part of an example histogram from a copy number profiling experiment, showing the distribution of values for given - binned - intensities. Intensity values form approximately local normal distributions, and - as in this example - discrepancies between frequency peaks and expected values may illustrate signal deviations from actual copy number levels.

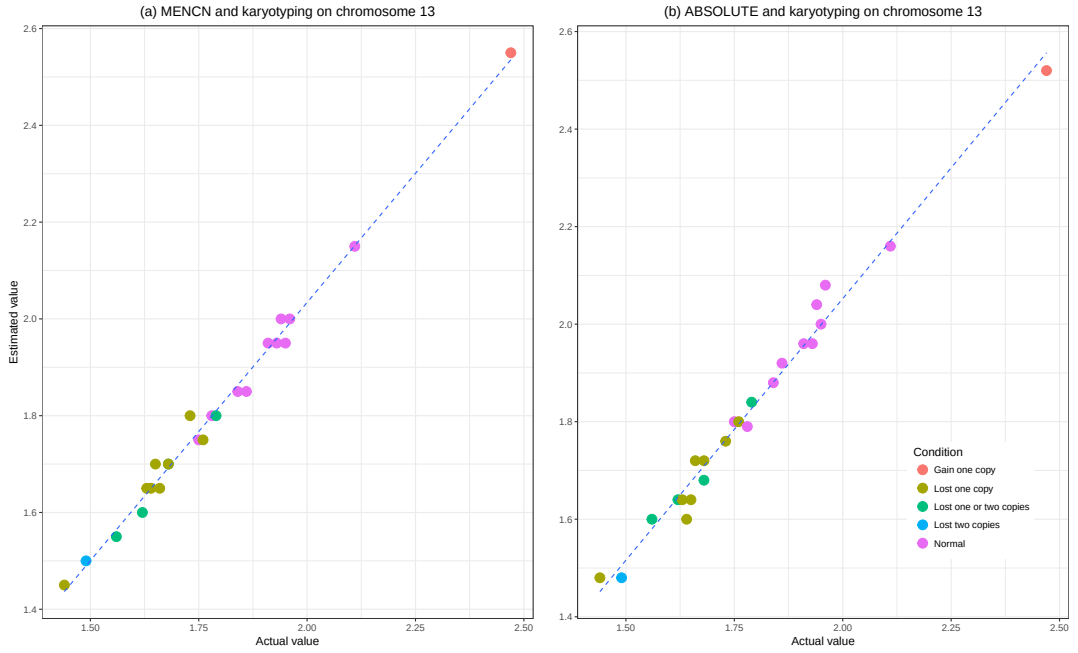


Figure 3: The comparison of estimated values and actual values on chromosome 13. The estimated values were calculated from the CNA file; the actual values were derived from karyotyping results. (a) estimation by Mecan4CNA; (b) estimation by ABSOLUTE. The estimation results of two methods are highly coherent.

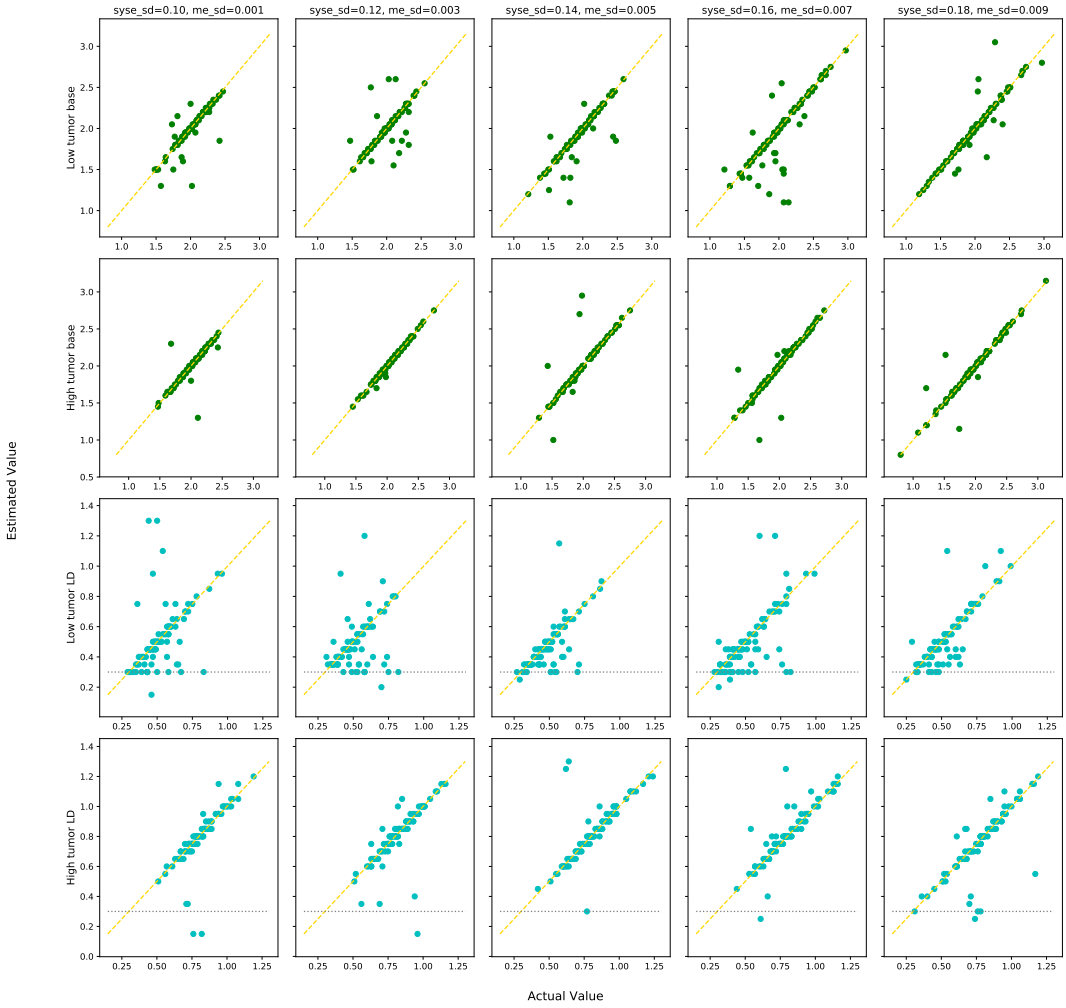


Figure 4: The comparisons of estimated and actual values for CN baseline and CN level distances on simulated data with different settings. Rows 1 & 2: Baseline estimation of low & and high tumor content data, respectively; Rows 3 & 4: Level distance estimation of low & and high tumor content data, respectively. The noise levels increase gradually in each column from left to right. *syse_sd* is the standard deviation used to simulate systematic errors, and *me_sd* is the standard deviation used to simulate measurement (individual) errors. The dotted lines indicate a threshold for filtering problematic results.

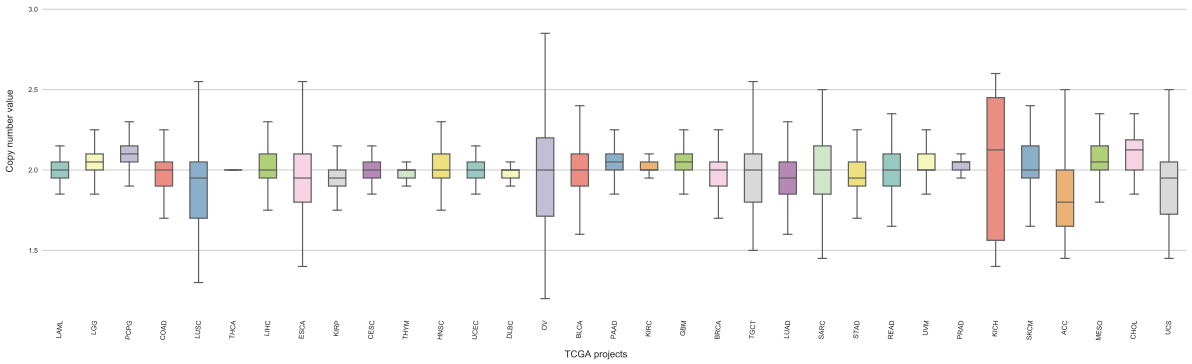


Figure 5: The distribution of baseline values estimated with *Mecan4CNA* in copy number profiling experiments from different TCGA projects (outliers not shown).

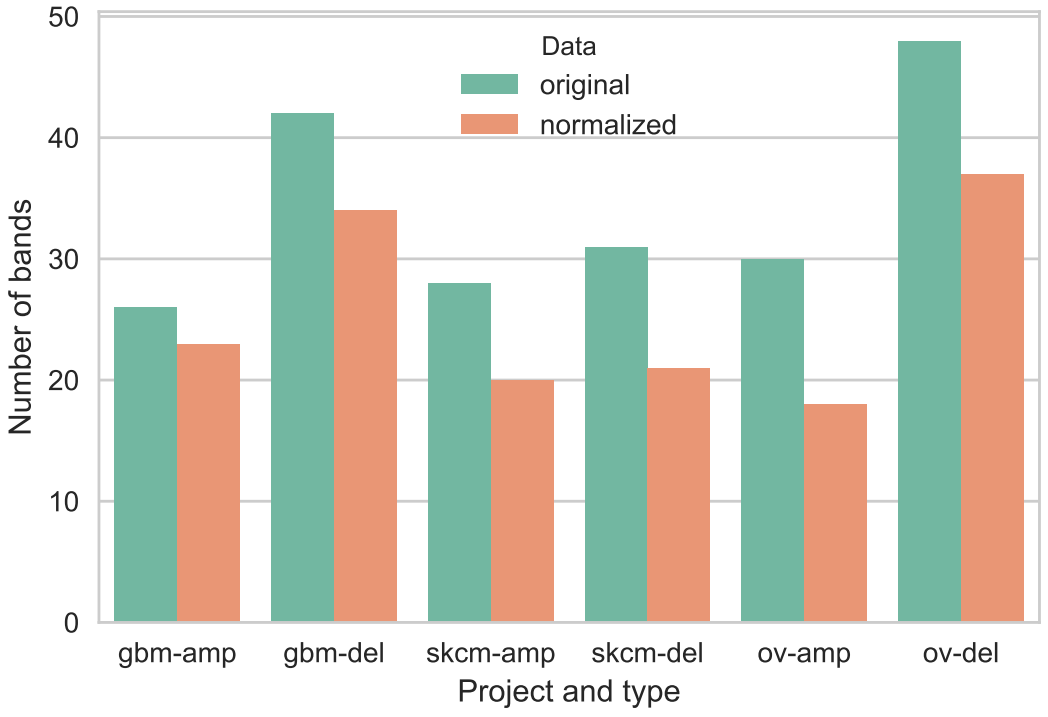


Figure 6: The number of significant bands detected by GISTIC using different datasets.

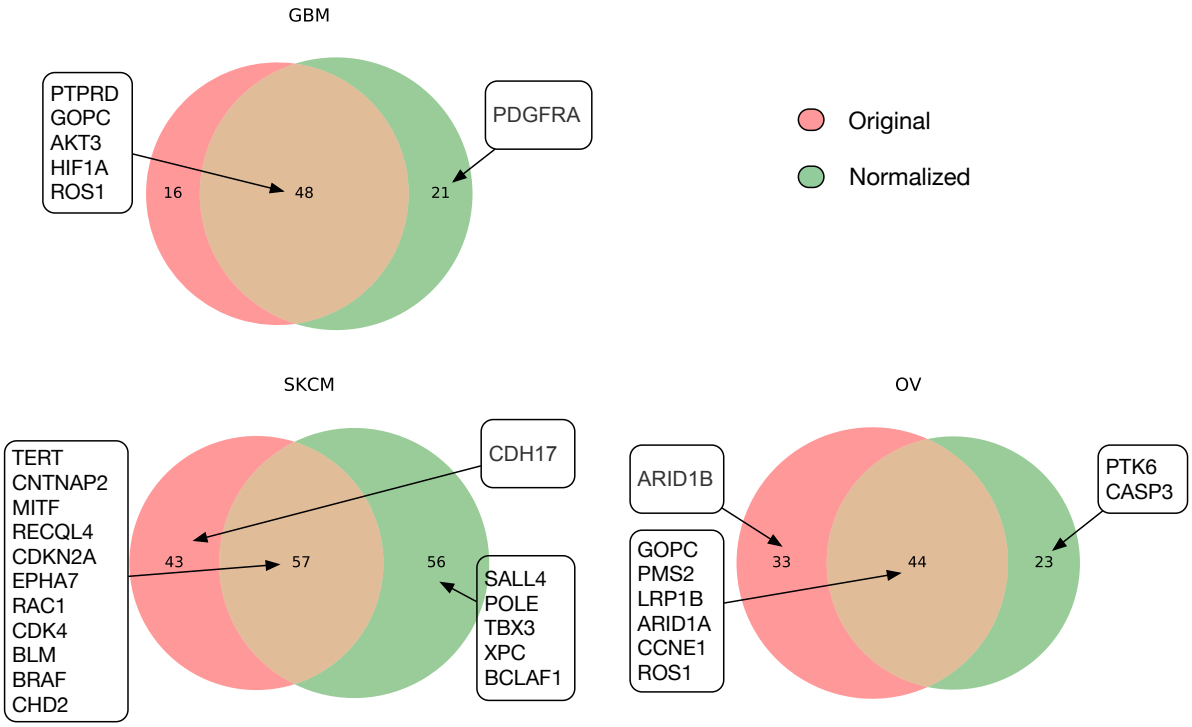


Figure 7: Detected cancer census genes using the original and normalized data in the GBM, SKCM and OV datasets from TCGA. Numbers in circles represent the total counts of detected census genes. Gene symbols in boxes indicate known cancer drivers among the detected genes for the respective diseases.

| **Pathway** | **Original Hits** | **Norm Hits** | **Diff** |
|---|---|---|---|
| **NFE2L2.V2** | 13 | 19 | 6 |
| **KRAS.600_UP.V1_DN** | 10 | 19 | 9 |
| IL21_UP.V1_UP | 11 | 16 | 5 |
| PRC2_EED_UP.V1_DN | 12 | 15 | 3 |
| ERB2_UP.V1_UP | 11 | 13 | 2 |
| **TBK1.DF_UP** | 12 | 12 | 0 |
| CSR_EARLY_UP.V1_UP | 10 | 12 | 2 |
| ESC_V6.5_UP_LATE.V1_UP | 10 | 12 | 2 |
| STK33_SKM_UP | 10 | 11 | 1 |
| STK33_UP | 13 | 10 | -3 |
| **CYCLIN_D1_KE_.V1_UP** | 10 | 10 | 0 |
| **CYCLIN_D1_UP.V1_UP** | 10 | 10 | 0 |
| **P53_DN.V1_DN** | 10 | 10 | 0 |
| **KRAS.600_UP.V1_UP** | 10 | 10 | 0 |

Table 2: MSigDB pathways with high coverage in both the original and normalized data of SKCM. Pathways in bold have known associations with melanoma.

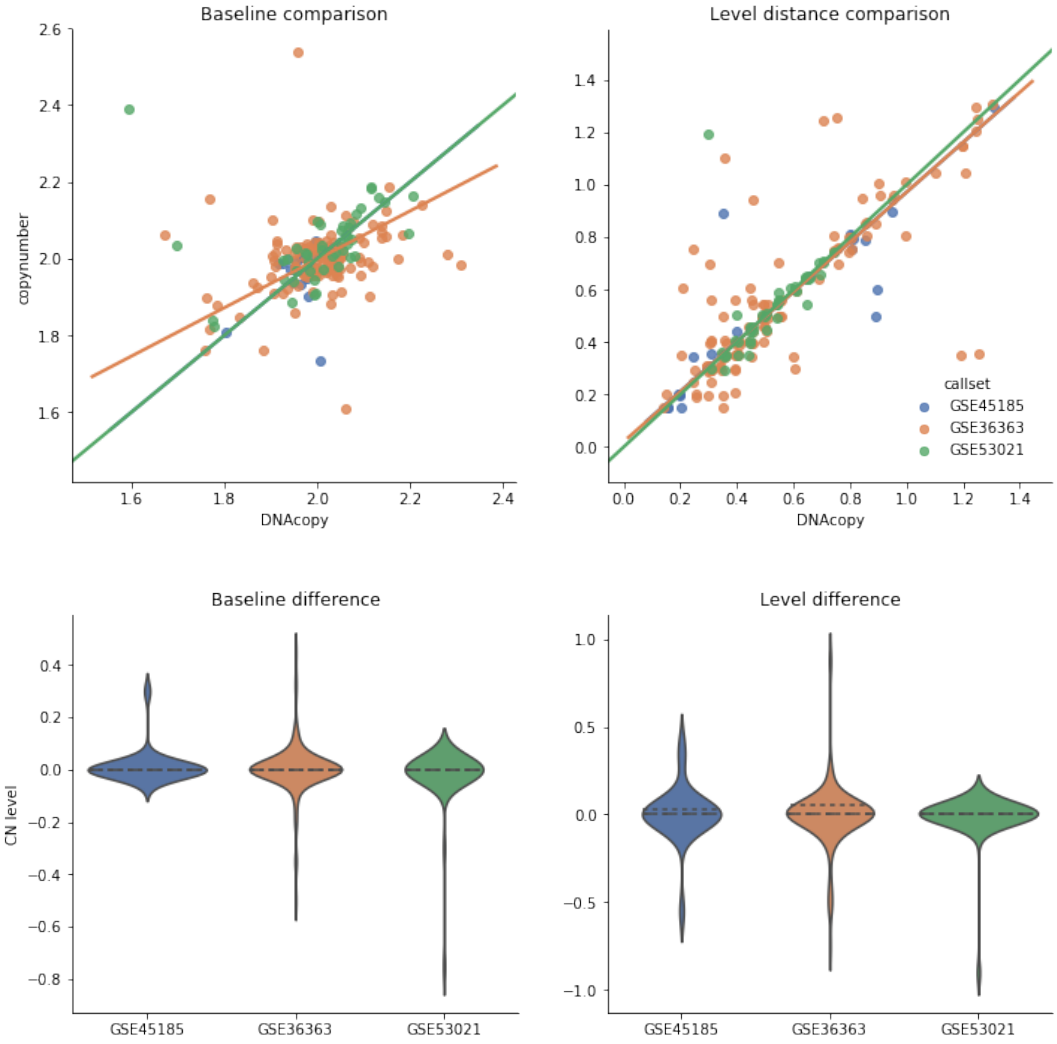


Figure 8: The comparison of Mecan4CNA's estimation results by using two different segmentation methods on three copy number datasets from GEO. Estimation from different methods showed high coherence. Points in scatter plots were jittered to improve visual comprehension.

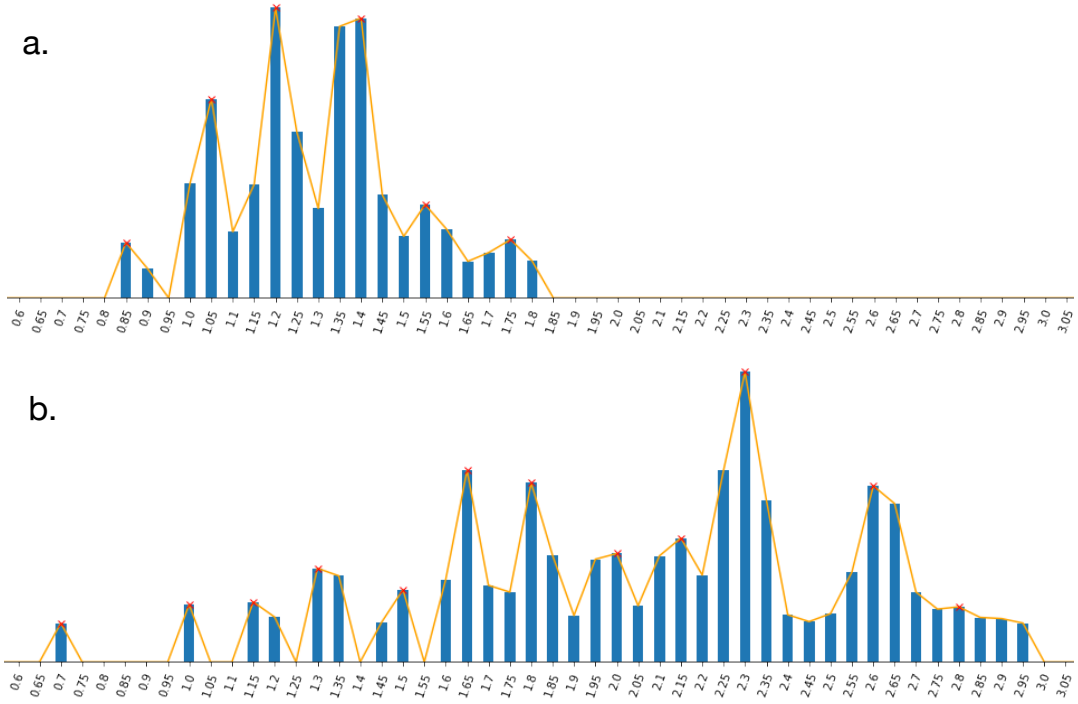


Figure 9: Two examples of low cellularity profiles shown in the signal histogram. In profile (a) with a high proportion of normal cells and high deviation, the signals tend to form a tight cluster. (b) represents a profile with a high proportion of sub-clonal cells, where the exceptional noise level makes estimation extremely challenging. Scales of the X-axis are uncalibrated copy number levels.

| **Pathway** | **Original Hits** | **Norm Hits** | **Diff** |
|---|---|---|---|
| PDGF_ERK_DN.V1_DN | 3 | 15 | 12 |
| NRL_DN.V1_UP | 1 | 11 | 10 |
| **KRAS.600_UP.V1_DN** | 10 | 19 | 9 |
| IL2_UP.V1_UP | 8 | 17 | 9 |
| VEGF_A_UP.V1_UP | 7 | 16 | 9 |
| STK33_DN | 3 | 12 | 9 |
| CRX_NRL_DN.V1_UP | 1 | 10 | 9 |
| MTOR_UP.V1_UP | 5 | 13 | 8 |
| PIGF_UP.V1_DN | 4 | 12 | 8 |
| STK33_NOMO_DN | 2 | 10 | 8 |
| TGFB_UP.V1_UP | 7 | 15 | 8 |
| RPS14_DN.V1_DN | 7 | 15 | 8 |
| **KRAS.300_UP.V1_DN** | 5 | 12 | 7 |
| **KRAS.AMP.LUNG_UP.V1_DN** | 2 | 9 | 7 |
| PDGF_UP.V1_UP | 6 | 13 | 7 |
| RAPA_EARLY_UP.V1_UP | 8 | 15 | 7 |
| IL15_UP.V1_UP | 4 | 11 | 7 |
| MTOR_UP.V1_DN | 3 | 10 | 7 |
| **PTEN_DN.V1_UP** | 9 | 16 | 7 |
| MEK_UP.V1_UP | 7 | 14 | 7 |
| **MYC_UP.V1_UP** | 5 | 12 | 7 |
| CAHOY_ASTROCYTIC | 1 | 8 | 7 |
| MEL18_DN.V1_DN | 9 | 4 | -5 |
| PRC2_EZH2_UP.V1_DN | 10 | 3 | -7 |

Table 3: MSigDB pathways with significant changes of coverage between the original and normalized data of SKCM. Pathways in bold have known associations with melanoma.

| **Callset** | **Cancer type** | **No. of samples** |
|---|---|---|
| GSE45185 | Glioblastoma | 24 |
| GSE36363 | Lung adenocarcinoma | 131 |
| GSE53021 | High-grade myogenic cancers | 47 |

Table 4: Public copy number datasets from Gene Expression Omnibus (GEO)