Published in final edited form as:

J Appl Stat. 2017 ; 44(1): 57–70. doi:10.1080/02664763.2016.1158798.

Beta approximation of ratio distribution and its application to next generation sequencing read counts

Shengping Yang¹ and Zhide Fang^{2,*}

¹Department of Pathology, School of Medicine, Texas Tech University Health Science Center, Lubbock, Texas, USA

²Biostatistics Program, School of Public Health, LSU Health Sciences Center, New Orleans, Louisiana, USA

Abstract

Paired sequencing data are commonly collected in genomic studies to control biological variation. However, existing data processing strategies suffer at low coverage regions, which are unavoidable due to the limitation of current sequencing technology. Furthermore, information contained in the absolute values of the read counts is commonly ignored. We propose a read count ratio processing/modification method, to not only incorporate information contained in the absolute values of paired counts into one variable, but also mitigate the discrete artifact, especially when both counts are small. Simulation shows that the processed variable fits well with a Beta distribution, thus providing an easy tool for down-stream inference analysis.

AMS 2010 SUBJECT CLASSIFICATIONS

Primary 62E17; secondary 62P10

Keywords

Beta distribution; Next-generation sequencing; Paired tumor and normal tissues; Ratio of counts; Statistical artifacts

1. Introduction

High throughput sequencing technologies have been widely used in interrogating genomic variations that cause certain diseases such as cancers. Comparing with those by microarray-based technologies, sequencing data are “digital”, and thus discrete distributions such as Poisson or Negative Binomial distributions are commonly used to model these data. In the meantime, due to the inherent variation in such data, a considerable proportion of genomic studies collect data on both samples with tumor and normal tissues from the same patients to remove many of the biological variations that are not of interest.

*Corresponding author: zfang@lsuhsc.edu.

As pointed out by Hawkins, Hon et al., [1], there are many challenges associated with generating and analyzing the enormous amount of Next Generation Sequencing (NGS) data, so that they can be appropriately translated into meaningful biological and clinical conclusions. To tackle these challenges, attempts from different perspectives have been made, including enhancing sequencing technique to obtain longer sequencing reads and more accurate readings; improving experimental design to remove both experimental and biological variations, as well as developing new analytical tools to achieve less false discovery and better resolution.

Comparing with microarray, one of the limitations of the current sequencing technique is that the sequencing coverage of short reads is not uniform across the genome, even among genomic regions that are close to each other. The possible causes of this non-uniform coverage include differential GC content, the existence of repeat sequences as well as platform specific biases, etc. (Holt, Losic et al., [2]). Meanwhile, the discrete nature of NGS data directly causes the distribution of the NGS data to be different from that of the microarray data, which have been showed to approximately follow a log normal distribution.

From the experimental design perspective, many studies chose to sequence both the tumor and paired normal control samples simultaneously, with the aim of using the paired normal samples as the references to remove/reduce the (experimental and biological) variability among the experimental units. This is a statistically valid strategy, and generally more meaningful results can be obtained from analyzing paired sample data if such biological variability is known and substantial.

Many statistical methodologies have been developed to utilize such paired data, for example, using the difference or ratio of the read counts for the tumor and the paired control samples as a constructed variable to build statistical models. However, to our knowledge, there is no clear answer with respect to what is the best approach to incorporate the paired data into a statistical model so that all the information contained in the data can be optimally used.

For example, in Copy Number Alteration (CNA) analysis of tumor sample, the changes in DNA copy number at certain genomic regions can be inferred by comparing the read counts of the sequences of paired samples, and using change-point detection method to determine the exact locations of the change points. Note that there is in general a trade-off between accuracy (detection resolution) and specificity, i.e., if better detection resolution is of primary concern, then a large number of small CNAs can be “detected”, but many of these “detected” CNAs may be false; on the other hand, if detection resolution is not critical, then many small real CNAs may be missed, but most of the detected CNAs are more likely real (Mosen-Ansorena, Aransay et al., [3]).

One of the earliest attempts to use sequencing data for detecting CNAs was made by Chiang and colleagues. They proposed using a log-normal distribution to approximate the logarithm

of the copy ratio $R(x_L, x_R) = \frac{T(x_L, x_R)/a_T}{N(x_L, x_R)/a_N}$, a ratio of the numbers of aligned reads from the tumor (“T”) and the paired normal (“N”) samples separately, where $\alpha_T(\alpha_N)$ denotes the total number of aligned sequence reads from the tumor (normal) sample, and $T(x_L, x_R)$ and

$N(x_L, x_R)$ denote the numbers of aligned sequence reads from the tumor and normal samples within a genomic window bounded by its left and right positions x_L and x_R , respectively. In other words, the copy ratio was considered to approximately follow a log-normal distribution. However, its simulation study shows that the approximation conforms well to the log-normal distribution only when on average $N(x_L, x_R) > 80$ (Chiang, Gatz et al., [4]).

In another attempt, Xie and Tammi [5] used the Gaussian distribution to approximate the Poisson distribution and concluded that the approximation, with continuity correction, is good when the Poisson mean is greater than 10. For prediction, the ratio of read counts in a sliding window was used. Then, with the assumption of the independence of the two read counts, by Geary-Hinkley transformation (Hayya, Armstrong et al., [6]), the read count ratio

was transformed to $t = \frac{\mu_Y z - \mu_X}{\sqrt{\sigma_Y^2 z^2 + \sigma_X^2}}$, where z is the ratio of read counts in the window, and μ_X, σ_X^2, μ_Y and σ_Y^2 are the means and the variances for X and Y respectively. The distribution of t is approximately normal when the estimated mean is greater than 6 for Y and less than 40,000 for X . The assumption for this approximation is that X is greater than 0 (Hinkley [7]).

More recently, in rSW-seq, weights are assigned to genomic locations according to the read counts of the tumor and paired normal samples. Specifically, at the locations where the read counts for the tumor and normal samples are the same, equal weights are assigned with opposite signs, that is, 1 and -1 for the tumor and normal samples, respectively. Otherwise, the weights were 2*tumor read count/total read count and 2*normal read count/total read count for the tumor and normal samples, respectively. After converting the sequencing counts to weighted data, a Smith-Waterman algorithm was then used to detect copy number changes (Kim, Luquette et al., [8]). However, similar to the methods used in previous two examples, there is an obvious limitation on this transformation. That is, the information contained in the absolute values of the tumor and normal read counts is ignored, because only the information contained in the ratio of the read counts is used in the downstream analysis. The consequence of such data processing/transformation is that the locations with different absolute read counts but the same ratio are treated equally. This is not consistent with a well-known fact that the locations with larger read counts generally provide more reliable information than those with smaller read counts.

To mitigate this limitation, a moving window approach, with fixed/adjustable size, was proposed to obtain large read counts (Chiang, Gatz et al., [4]; Xie and Tammi [5]). Basically, by adding the read counts within such a window, the absolute number of read counts increases substantially, and then by treating the sum from one window as one observation, most of the observations achieve “large” absolute count values, and thus improved reliability in estimation.

One downside of the moving window approach is that the pre-defined window has to be large enough so that all individual windows have large read counts. And as a trade-off, the resolution of variant/alteration detection is compromised.

In this paper, we propose a different type of read count data processing/modification method, with which the transformed variable can be satisfactorily approximated even when the paired sample read counts are small. We will use paired tumor and normal tissue sample data as an example to discuss the distribution of the ratio of the tumor and the total (the sum of tumor and the paired control) read counts, and evaluate the performance of the proposed data modification strategy, especially for the situations where both the tumor and control read counts are small. In fact, such read count ratios can be considered as sampled from a distribution over the rational numbers in the unit interval (Trifonov, Pasqualucci et al., [9]). The exact distribution of the ratio of two power-law or two geometric distributions were developed in Trifonov's paper; however, no closed form is available for the distribution of the ratio of two Poisson distributions. In fact, even if a closed form can be obtained, it is computationally expensive to apply the complex exact distribution to derive statistical inferences, due to its discrete and multi-modality nature, given the enormous size of sequencing data. On the other hand, as pointed out by Johnston, Shayla et al., [10], the discrete nature of the ratio of two random integers is in fact caused by some statistical artifacts, and is independent of the underlying distribution. Motivated by this observation, in this paper we propose a feasible solution to approximate the distribution of read count ratio, especially for NGS data analysis.

2. Method

We define X_{ij} as the read count of the i^{th} genetic location for the j^{th} tumor sample, where $X_{ij} \in \{0, 1, 2, 3, \dots\}$, and Y_{ij} as the read count of the i^{th} genetic location for the corresponding j^{th} paired normal sample, where $Y_{ij} \in \{0, 1, 2, 3, \dots\}$. Note that in practice, we will exclude the genetic locations from the analysis where both X_{ij} and Y_{ij} are 0, because these are locations that contain no information for data analysis purpose. Furthermore, we will assume that both X_{ij} and Y_{ij} follow Poisson distributions. Additionally, we define $B_{ij} = X_{ij} / (X_{ij} + Y_{ij})$, the ratio of X_{ij} and $(X_{ij} + Y_{ij})$, as the read count ratio of interest. We comment that, although the proposed data modification strategy (details next) was mainly developed for analyzing NGS data, especially, NGS data generated for CNA detection, it is, however, also applicable to any data representing the ratio of two Poisson distributions.

Note that we will primarily focus on the situations where the means of both X_{ij} and Y_{ij} are relative small since otherwise the distributions of X_{ij} and Y_{ij} can be conveniently approximated as normal distribution because of the Central Limit Theorem (CLT). Furthermore, the following lemma indicates that if X_{ij} and Y_{ij} are independent with large means, the ratio B_{ij} approximately follows a Beta distribution.

Lemma 1

If $X_{ij} \sim \text{Poisson}(\lambda_{ij}^X)$, $Y_{ij} \sim \text{Poisson}(\lambda_{ij}^Y)$ and they are independent, then B_{ij} approximately follows $\text{Beta}(\lambda_{ij}^X, \lambda_{ij}^Y)$ when $\lambda_{ij}^X, \lambda_{ij}^Y$ are large.

The proof of the lemma is very straightforward. In fact, for large λ , both $Poisson(\lambda)$ and $Gamma(\lambda, 1)$ can be approximated by a normal distribution with the same mean and variance λ . This implies that if the means, $E(X_{ij}) = \lambda_{ij}^X$, $E(Y_{ij}) = \lambda_{ij}^Y$ are large, the read counts X_{ij} , Y_{ij} approximately follow $Gamma(\lambda_{ij}^X, 1)$ and $Gamma(\lambda_{ij}^Y, 1)$ respectively. Hence, simple calculation of the distribution of transformation of random variables shows that B_{ij} approximately follows $Beta(\lambda_{ij}^X, \lambda_{ij}^Y)$.

In addition, the variance of this ratio B_{ij} can be estimated by $\frac{X_{ij}Y_{ij}}{(X_{ij}+Y_{ij})^2(X_{ij}+Y_{ij}+1)}$. And it is obvious that proportionally increasing the values of both X_{ij} and Y_{ij} will result in the corresponding decrease of this estimated variance, i.e.,

$$\frac{fX_{ij}fY_{ij}}{(fX_{ij}+fY_{ij})^2(fX_{ij}+fY_{ij}+1)} \leq \frac{X_{ij}Y_{ij}}{(X_{ij}+Y_{ij})^2(X_{ij}+Y_{ij}+1)},$$

where $f(>1)$ is a multiplication factor.

It has been demonstrated that, when X_{ij} and Y_{ij} are not large, the distribution of B_{ij} has a distribution similar to Thomae's function, which has a complex probability mass function. Even if a closed-form distribution function can be derived, it will be computationally expensive in making statistical inferences given the large number of genomic locations in sequencing data analysis. Thus, by taking a different approach, we propose to perform some meaningful data modification, rather than deriving the closed-form distribution function, to get a good approximation of B_{ij} . In particular, we propose to incorporate a very small random variation to the read counts of paired samples so that the count ratio can be approximated by a known continuous distribution even when both λ s are small. We start with constructing the following quantity,

$$M_{ij} = \frac{X_{ij} + u_1}{X_{ij} + u_1 + Y_{ij} + u_2}, \quad \text{if } x_{ij} + y_{ij} \neq 0, \text{ or}$$

$$M_{ij} \text{ is not defined,} \quad \text{if } x_{ij} + y_{ij} = 0,$$

where u_1 and u_2 are random variables with uniform distributions (0, 1). We will call M_{ij} the variable with initial modification (viM). Note that we do not define M_{ij} when $x_{ij} + y_{ij} = 0$ (or $x_{ij} = y_{ij} = 0$) because these locations (no coverage locations) do not contain any information for data analysis purpose.

Furthermore, to ensure that the approximation fits well for extremely small values of counts, for example, the expected values are both less than 5, we propose adding another small random variation to M_{ij} , and define,

$$mM_{ij} = \frac{X_{ij} + \frac{x_{ij}}{x_{ij} + y_{ij}}u_3 + b_1}{X_{ij} + \frac{x_{ij}}{x_{ij} + y_{ij}}u_3 + b_1 + Y_{ij} + \frac{y_{ij}}{x_{ij} + y_{ij}}u_4 + b_2} \quad \text{if } x_{ij} + Y_{ij} \neq 0, \text{ or}$$

$$M_{ij} \text{ is not defined,} \quad \text{if } x_{ij} + y_{ij} = 0,$$

where u_3 and u_4 are random variables with uniform distribution (0, 1), b_1 and b_2 are Beta distributions with parameters $(\frac{y_{ij}}{x_{ij} + y_{ij}}c, \frac{x_{ij}}{x_{ij} + y_{ij}}c)$ and $(\frac{x_{ij}}{x_{ij} + y_{ij}}c, \frac{y_{ij}}{x_{ij} + y_{ij}}c)$, respectively. Note that c is a fine-tuning constant that can take a wide range of values, and in this paper we set $c = 2$.

3. Results

3.1 Data simulation

To evaluate the performance of the proposed variable modification, we generated a number of simulations, and visually assessed the distributions of the modified count ratios.

First, to represent situations where both read counts are large, one million random numbers from a Poisson distribution with mean 35 were generated as the sample of the read count in the numerator. Similarly, an independent set of one million random numbers from a Poisson distribution with mean 70 were generated as the random sample of another read count, and the sum of the two corresponding Poisson random numbers was used as the denominator of the count ratio. Figure 1.A presents the histogram of these ratios. Based on visual assessment, the distribution fits quite well with a Beta distribution with parameters 36 and 71 (imposed curves in bold). To make the proposed modification, one random number generated from a uniform distribution (0, 1) was added to each of the Poisson distributions. Figure 1.B is the histogram of M_{ij} , and it also fits quite well with a Beta distribution with parameters 36 and 71. Next, we added randomly generated random numbers from a Beta distribution (36/107, 71/107) to the Poisson counts in the numerator, and the sums of random numbers from Beta distribution (36/107, 71/107) and Beta distribution (71/107, 36/107) to the denominator. The histogram of such further modified ratios also fits well with a Beta distribution B(36, 71) (Figure 1.C). For illustration purpose, we include the histogram of Beta distribution B(36, 71) in Figure 1.D. To mitigate the uncertainty in data simulation, each of the above simulations was repeated 100 times. Based on these repeated simulations, the averages of the means of the raw Poisson ratios, $v_i M_{ij}$, mM_{ij} , and a set of random numbers generated from Beta distribution B(36, 71) were 0.3364445, 0.3379777, 0.3394586, 0.3394443, respectively. The corresponding standard deviations were all close to 0.00004. The differences in mean between the raw ratio and both the initial and the further modified ratios are all less than 1%. Additionally, the corresponding average standard deviations of these simulated values were 0.04589555, 0.04550284, 0.04529678 and 0.04514396, respectively. The corresponding standard deviations of these standard deviations were all close to 0.00003. Note that, the differences in average standard

deviations between the raw ratio, viM_{ij} and mM_{ij} were both less than 0.1%. Similar results were observed when using 70 as the Poisson parameter in the numerator and 35 as the parameter of another Poisson count (plots not shown). These indicate that the Beta distribution fits the count ratios well when both λ s are large, regardless of whether data processing/modification is performed.

Next, we set the parameters of the two Poisson distributions to be relative smaller, that is, 8 and 10, respectively. There seems to be some obvious discreteness for the distribution of the raw ratio (Figure 2.A). However, after performing the initial and further modifications, both viM_{ij} and mM_{ij} can be modelled nicely by a Beta distribution with parameters (9, 11) (Figure 2.B, 2.C). Based on 100 repeated simulations, the averages of the means of the ratios were 0.4444364, 0.4474971, 0.4501244 and 0.4499963, for the raw, viM_{ij} , mM_{ij} , and the Beta distribution, respectively. The standard deviation s of the mean ratios based on the 100 simulations were all close to 0.0001. The differences in mean between the raw ratios, viM_{ij} and mM_{ij} were all less than 1%. On the other hand, the differences in standard deviation between the raw ratio, viM_{ij} and mM_{ij} increased around 5~10%. This larger difference is expected since the raw ratio is substantially more discrete.

Last, we used very small values as the parameters of the two Poisson distributions, that is, 1 and 3, respectively. The distribution of the raw ratios show high level of discreteness, and only a small number of values between 0 and 1 have non-zero probability, which is an artifact of dividing a small integer by another slightly larger integer (Johnson, Shayla et al, [10]). As a result, the frequencies of some values are very high (Figure 3.A). For example, more than one third of the raw count ratios take the value of 0 (the probability of observing 0 from a Poisson with mean 1 is 0.368). The viM_{ij} and mM_{ij} , comparing with the raw ratio, had substantially improved the continuity of the distribution, and roughly follows a Beta distribution with parameters (2, 4). However, we point out that, although adding a small value to each of two counts can turn all 36.8% of zero ratios into non-zero ratios, there are still a large number of ratios close to 0, comparing to the fitted Beta distribution (Figure 3.B). The histogram of mM_{ij} is presented in Figure 3.C and shows a better fit with the Beta distribution. The means of the raw Poisson ratios, viM_{ij} , mM_{ij} and random numbers generated from Beta distribution B(2,4) are 0.2499811, 0.3059688, 0.334449 and 0.3333187, respectively. Note that the differences in mean between the raw ratio and the initial and further modified ratios are large, comparing to those when both λ s are (relative) large.

In fact, the discreteness of the raw ratio distribution is largely a statistical artifact due to small λ s (Johnson, Shayla et al., [10]). Assuming that the total of the paired counts is fixed, then the ratio of one count, for example, tumor sample count, over the total count provides a more accurate estimate for the probability of being tumor if additional count(s) is obtained. However, it is clear that if the total of counts is small, then the estimate of this probability is more likely to be affected by a small change of the absolute count values. To address this issue, a uniform prior, which is a conjugate prior of the parameter of a Bernoulli process, is commonly used. Suppose that one out of four Bernoulli trials has a positive outcome, then the unbiased estimate of the probability of success is 0.25. By applying a uniform prior, the

posterior distribution has a mean of 0.33. Thus it can be seen that our proposed modification method is numerically similar to applying a uniform prior to the distribution of the count ratio. In addition, due to the fact that a closed form of the exact distribution is very hard to derive (or does not exist), the approximations we have obtained are highly desirable.

To demonstrate whether the artificially introduced data/noise is consistent with reality if the read level were allowed to be high, we performed the following simulations:

First, we generated a random number from a Poisson distribution with mean 200, and then generated an independent random number from a Poisson distribution with mean 400. Next, we generated a random number from a Poisson distribution with mean 2, and another random number from a Poisson distribution with mean 4. Then, we calculated the raw ratios of both the first and the second pairs of the random numbers, as well as viM of the second pair. We repeat the process 10,000 times, and the mean of the ratios obtained from the first pair counts was used as the “gold” standard. And we calculated the percentages of the raw count ratio and viM lying within the $\pm 10\%$ of the “true” ratio (the “gold” standard). We repeated the above process as a whole for 1,000 times, and the distribution of the percentages of the count ratios lying within 10% interval of the true ratio were displayed in Figure 4. About 43% (43%) of the viM (mM) values lie within 10% interval of the “true” ratio compared with about 39% for the raw count ratio. The corresponding percentages for a 5% interval were 22%, 22% and 20%, respectively, and 74%, 74% and 69% for a 20% interval. In general, due to the discrete nature of the count ratio when the counts are small, the raw count ratio values have wider spread compared to viM and mM , and thus have poorer consistency with the count ratio calculated using large counts, given the “true” ratio.

3.2 Application on CNA detection

To illustrate the effectiveness of the proposed modification, we evaluated how the proposed data modification affecting CNA detection using publicly available NGS data. These data were obtained from the Gene Expression Omnibus website (Accession number: GSM665129). Eight commonly used breast cell lines were sequenced using Illumina Genome Analyzer II, including 4 estrogen receptor (ER) positive and 3 ER negative, as well as a non-tumor epithelium cell lines. Short reads were aligned by Bowtie (Langmead, Trapnell et al., [11]) to human genome hg19. A maximum of two mismatches were allowed in each alignment, and reads mapped to multiple genomic locations were discarded. SAMtools was used to generate the BAM format output files (Li, Handsaker et al., [12]), and subsequently read count for each location was calculated. The same alignment process was performed on both the tumor and the non-tumor samples.

We comment that, in many studies, investigators obtained both tumor and the paired/unpaired non-tumor samples simultaneously, with the aim of using the aligned read counts as a surrogate to detect where the genomic structure changes in the tumor sample, while the paired/unpaired sample serving as a reference. For demonstration purposes, we randomly chose a cell line from the 7 tumor cell lines, and used the only non-tumor epithelium cell line as the reference since the non-tumor genome is supposed to have minimal CNAs. Note that the goal of this study is mainly to introduce read count ratio modification in situations where the paired read counts are both small, but not to develop a novel algorithm for CNA

detection. Thus, we use a modified version of a Hidden Markov Model (HMM) based CNA detection method that we developed in another paper (Yang, Pounds et al., [13]), and focus primarily on demonstrating how using ratio data modified by the proposed method can improve CNA detection comparing with using ratio data without modification. With the assumption that if the paired read counts are both very large, count ratio is an excellent surrogate measurement of the DNA copy number, we use CNAs detected using large counts, at the cost of compromised resolution, as the “gold” standard, and then artificially reduce the magnitude of the read counts and evaluate how the proposed count ratio modification improves the performance of CNA detection using the artificially reduced read counts.

To obtain large counts for both tumor and non-tumor samples, we treated each gene as a unit, and summed all the read counts within a gene to get one total read count. The summation was carried out for all genes in a genome, so that the individual read counts are predominately large values, i.e., more than 99.5% of read counts are greater than 100. Next, we performed CNA detection using raw count ratio data calculated from these read counts, and termed the regions with CNA detected the “gold” standard CNA regions. Then we performed the same CNA detection using count ratio data modified by our proposed method to evaluate whether the modification affects how CNAs are detected. In fact, since the raw read counts for individual genes are in general very large, adding a random value with a uniform distribution has very little effect on the nature of the count ratio. It turned out that the CNA regions detected were exactly the same regardless whether the uniform distributed noise was added (Figure 5A and 5B).

Next, we divided all the original read counts by 1,000 to obtain the artificially reduced read counts with all the reduced read counts being rounded to their nearest integers to mimic the situations where the read counts generated by the sequencer are small. After the read count reduction, about 50% of the read counts are less than or equal to 10. We performed the same CNA detection using both the raw reduced read count ratios and the modified read count ratios. With the CNA regions detected using the original large counts ratio being the “gold” standard, we found that about 93% of the DNA copy number states detected using the modified reduced read count ratios were correctly identified (Figure 5D), while only 68% were correct if no data modification was made (Figure 5C). Note that without performing the proposed data modification, there are a considerable amount of extreme count ratio values, which result in the incorrect assignment of the copy number states for those regions (Figure 5C). Considering the fact that after read count artificial reduction, about half of the read counts are less than or equal to 10, the proposed count ratio modification substantially improves the accuracy of the CNA detection, comparing with no modification. In fact, a 1,000 fold read count reduction can translate into substantially reduction in cost for generating the read counts, while a 7% reduction in accuracy seems to be quite acceptable. In particular, unlike the simulation, low read count regions always exist in reality due to the non-uniformity nature of synthesis efficiency, and the cost for increasing the counts in low read count regions can be inhibitive.

In order to better visualize how the proposed data modification smooths the data distribution, we present the histograms of the distributions before and after the data modification for the artificially reduced data in Figures 6A and 6B. It is quite clear that due

to the statistical artifact, the distribution of the raw reduced read count ratios shows substantial discreteness, and several values have very high frequencies (Figure 6A). As a comparison, the modified reduced read count ratios have much smoothed distribution, and the several modes are not due to data discreteness, rather reflect the existence of several distinct copy number states (Figure 6B). It is obvious that the raw reduced count ratios have a considerable number of 1s, which might have caused the enormous amount of change points detected in Figure 5C (observations in purple), which are clearly false discoveries that are hard to eliminate. After data modification, these artifacts are substantially mitigated (Figure 5D), and the CNAs detected largely agree with those detected based on ratios of very large read counts.

4. Discussions

Digitalization very often results in count data, which is commonly thought to be able to provide more precise measurements of the items of interest. This might be true if the counts are in general large. However, if the counts are small, and the ratios of counts are of interest, then statistical artifacts might occur, and it is difficult to derive the exact distribution of such count ratios. Trifonov, Pasqualucci et al, [9] provided a closed form distribution for such count ratios with the requirement that the counts follow a power-law or a geometric distribution. However, as far as we know, the exact distribution of the count ratios has not yet been available when the counts follow a Poisson or negative binomial distribution.

Deriving the distribution of count ratios has important applications, especially, in the area of NGS data analysis. As we know, short read counts are “digital” data, and very often assumed to follow a Poisson or negative binomial distribution. Since the genetic characteristics of the tumor and paired normal samples from the same patients are highly similar, paired data are commonly used to eliminate many of the systematic variations so that the real mutations/alterations can be more accurately detected. One way to incorporate both the tumor and paired sample data into the analysis is to calculate the ratio of the tumor short read count over the total (the sum of tumor and paired normal) read counts. This ratio approximately follows a Beta distribution if both read counts are large, as demonstrated by the example and proof given above. However, if both read counts are small, the count ratio does not follow a Beta distribution well due to statistical artifact caused by dividing an integer by another integer (Johnson, Shayla et al, [10]). In other words, although the true underlying distribution of such ratios might have a continuous distribution (as reflected when the counts are large), but it is inappropriately discretized when both counts are small. To appropriately address this issue, and to make it easy to approximate the distribution of count ratios, especially for situations when both counts are small, in this paper we propose a count data processing method, so that the above mentioned statistical artifact can be effectively removed/mitigated.

The initial modification of adding a very small random noise, i.e., uniform (0, 1) random number, to both the numerator and denominator of the raw count ratio, works quite well in smoothing the distribution of the ratios of moderately small counts. However, when both read counts are extremely small, like as small as 1, we observed moderate discrepancy between the modified distribution and the targeted Beta distribution. To reduce this

discrepancy, we proposed further modification by adding another random number from a Beta distribution, defined from the read counts. Note that we suggest to exclude the observations with both counts being zero since they would not provide any information, at least in sequencing data analysis.

The advantages of our proposed data modification strategy are three folds: 1) the modification is very straightforward and easy to implement; 2) the modified distribution has a well-behaved continuous distribution, which allows us to utilize many of the existing software and methodologies developed for microarray data analysis, with or without further modification – note that Beta distribution is very similar to a normal distribution when the mean is close to 0.5, and CLT can be applied even with the summation of a very small number of ratios; 3) The distribution of the modified ratios can be conveniently approximated by a known continuous distribution, that is, Beta distribution with known parameters, which makes the downstream analyses more efficient; 4) Should we prefer to use a Beta distribution to approximately model the modified ratio, it is feasible to utilize the information inherited in the absolute values of the read counts for modeling, because the variance of the Beta distribution directly related to the absolute values of the two read counts.

We also demonstrated how to use a “naïve” change point detection method to identify CNAs in cancer genome. We first summed all read counts for one gene so that the count corresponding to each gene was very large, and thus we could comfortably use CNAs detected using these large read counts as the “gold” standard. Then, by dividing these summed counts by 1000, we obtained the “reduced” read counts, of which about 50% were less than or equal to 10. Our results show that CNAs detected using ratios calculated from the modified “reduced” read counts are highly consistent with the “gold” standard. In contrast, CNAs detected using ratios calculated from the raw “reduced” read counts have much larger discrepancy with the “gold” standard. In fact, this example also, in a reverse manner, demonstrated that the proposed data processing method could achieve high consistency with what would be expected if the read counts were allowed to be high. In other words, we first had the large read counts, and then proportionally reduced the counts, instead of having small counts first and then allowing them to increase (for example, increase coverage depth), because the ratios from small counts are unlikely to be consistent if the experiments are repeated.

In summary, we propose an easy to implement approximation method for evaluating the ratio of discrete read counts. We showed that the approximation works well in “smoothing” the underlying discrete distribution. In the meantime, the distribution of the “smoothed” read count ratios is well known, with parameters that can be straightforwardly estimated. We comment that we do not make a firm recommendation on how NGS should be analyzed because the difference in coverage depth from different studies can be quite large, and thus other strategies, such as using a small sized moving window, might need to be considered in addition to our approach. Furthermore, in this paper, we propose to modify the ratios of the read counts that follow Poisson distribution. Should over-dispersion occur and a Negative Binomial distribution with mean μ and variance $\mu(1+\mu\phi)$ be used to model the counts, the

modification would still work well for small ϕ . How to modify the ratio when ϕ is large is the topic of our future research.

Acknowledgments

ZF's research was supported by 1 U54 GM104940 from the National Institute of General Medical Sciences of the National Institutes of Health which funds the Louisiana Clinical and Translational Science Center of Pennington Biomedical Research Center.

References

1. Hawkins RD, Hon GC, Ren B. Next-Generation Genomics: an Integrative Approach. *Nature Reviews Genetics*. 2010; 11(7):476–486.
2. Holt C, Losic B, Pai D, Zhao Z, Trinh Q, Syam S, Arshadi N, Jang GH, Ali J, Beck T, McPherson J, Muthuswamy LB. WaveCNV: Allele specific copy number alterations in primary tumors and xenograft models from next-generation sequencing. *Bioinformatics*. 2014; 30(6):768–774. [PubMed: 24192544]
3. Mosen-Ansorena D, Aransay AM, Rodriguez-Ezpeletz N. Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. *BMC Bioinformatics*. 2012; 13:192. [PubMed: 22870940]
4. Chiang DY, Gatz G, Jaffe DJ, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods*. 2009; 6(1):99–103. [PubMed: 19043412]
5. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*. 2009; 10:80. [PubMed: 19267900]
6. Hayya J, Armstrong D, Gressis N. A note on the ratio of two normally distributed variable. *Management Science*. 1975; 21(11):1338–1341.
7. Hinkley DV. On the ratio of two correlated normal random variables. *Biometrika*. 1969; 56(3):635–639.
8. Kim TM, Luquette LJ, Xi R, Park PJ. rSW-seq: Algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics*. 2010; 11:432. [PubMed: 20718989]
9. Trifonov V, Pasqualucci L, Dalla-Favera R, Rabadan R. Fractal-like Distributions over the Rational Numbers in High-throughput Biological and Clinical Data. *Scientific Reports*. 2011; 1:191. [PubMed: 22355706]
10. Johnson RG, Shayla SD, Mallawaarachy AR. Statistical Artifacts in the ratio of discrete quantities. *The American Statistician*. 1995; 49(3):285–291.
11. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 2009; 10(3):R25. [PubMed: 19261174]
12. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25(16):2078–2079. [PubMed: 19505943]
13. Yang S, Pounds S, Zhang K, Fang Z. PAIR: paired allelic log-intensity-ratio based normalization method for SNP-CGH arrays. *Bioinformatics*. 2013; 29(3):299–307. [PubMed: 23196989]

Figure 1.

A. Ratio of Poisson distributions; B. Ratio with uniform distributions added; C. Ratio with uniform and Beta distributions added; D. Beta distribution Beta(36, 71). The superimposed curve in each plot is the histogram profile of Beta(36, 71).

Figure 2.

A. Ratio of Poisson distributions; B. Ratio with uniform distributions added; C. Ratio with uniform and Beta distributions added; D. Beta distribution Beta(9, 11). The superimposed curve in each plot is the histogram profile of Beta(9, 11).

Figure 3.

A. Ratio of Poisson distributions; B. Ratio with uniform distributions added; C. Ratio with both uniform and Beta distributions added; D. Beta distribution Beta(2, 4). The superimposed curve in each plot is the histogram profile of Beta(2, 4).

Figure 4.

A. Histogram of the percentages of raw count ratios lying within 10% interval of the “true” ratio. B. Histogram of the percentages of viM values lying within 10% interval of the “true” ratio. C. Histogram of the percentages of mM values lying within 10% interval of the “true” ratio.

Figure 5.

A. CNA detection using the raw original count ratios. B. CNA detection using the modified original count ratios. C. CNA detection using the raw reduced count ratios. D. CNA detection using the modified reduced count ratios.

Figure 6.

A. Histogram of the raw ratios of reduced counts; B. Histogram of the modified ratios of reduced counts.