Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

14-11

The Algorithm of Equal Acceptance Region for Detecting Copy Number Alterations: Applications to Next-Generation Sequencing Data

Yan-Xia Lin

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

The Algorithm of Equal Acceptance Region for Detecting Copy Number Alterations: Applications to Next-Generation Sequencing Data

Yan-Xia Lin*[1]

[1]Centre for Statistical and Survey Methodology
School of Mathematics and Applied Statistics, University of Wollongong
Wollongong, Australia

Email: Yan-Xia Lin*- yanxia@uow.edu.au;

*Corresponding author

Abstract

The information of copy number alterations (gains and losses) in tumour genomes can be used to discovery cancer-causing genes. The estimate of copy number can be obtained from the estimate copy number ratio. The higher the depth of underlying sequencing data, the more accurate the estimate of copy number ratio. At the same time, the higher depth of a sequencing data used in copy number analysis, the more cost of data analysis. To develop a method for identifying a necessary depth of sequencing data for copy number analysis before test data are produced is of interest. In this paper, we proposed an algorithm of equal acceptance regions for detecting copy number ratios. This algorithm can be used to determine the depth of sequencing data required for copy number analysis.

1 Introduction, basic assumptions and notation

The information of copy number alterations (gains and losses) in tumour genomes can be useful for discovering cancer-causing genes. Recently, Chiang *et al.* (2009) and Alkan *et al.* (2009) have studied how to estimate copy number alterations using massively parallel sequence data. Chiang *et al.* (2009) suggested using log normal distribution to approach the distribution of log ratio of copy-number and use the approximated distribution to work out the number of aligned reads required to ensure the power of the detection of copy

number alterations in tumour genomes. Other different approaches for detecting copy number variation using next generation sequencing data can be found from Xie and Tammi (2009) and Kim *et al.* (2010).

SegSeq, a recently proposed sequenced-based algorithm, utilizes windows defined by a predefined number of normal reads to detect breakpoints between copy number alternation's (CNA) (Chiang *et al.*, 2009). Kim *et al.* (2010) commented that major disadvantage of window-based approaches is that the window size must be determined a priori and that the overall performance of the algorithm is influenced strongly by the value.

Xie and Tammi (2009) discussed the impact of window size on the confidential level of CNAs. They identified and focused on the confidential level for testing true copy number ratio 1 only. However, they made mistakes by applying the p-value for test $H_0 : r = 1$ to test $H_0 : r = r_0$, where $r_0 \neq 1$.

The definition of tumour-normal copy-number ratio in terms of next generation sequencing data can be found from Chiang *et al.* (2009). For reading convenience, the definition is briefly introduced below.

Consider a genomic window of length L within the alignable portion of a reference genome with length A. Let a_N and a_T the total number of aligned sequence reads from the normal and tumour sample, respectively. Let N and T the number of aligned sequence reads from the normal and tumour samples within the genomic window L, respectively. N and T are two independent random variables.

We adopt Assumptions: (i) there are no copy-number alterations within a genomic window of length L; (ii) N and T follow Poisson distributions with parameters $\lambda_N = a_N L/A$ and $\lambda_T = r a_T L/A = rm\lambda_N$, respectively, where $m = a_T/a_N$ and r is the true copy-number ratio given by the genomic window.

The tumour-normal copy-number ratio, R, is defined as

$$R = \frac{T/a_T}{N/a_N}, \qquad \text{if } N > 0$$

else R is undefined. R is a random variable and r needs to be estimated. Random variable R is the ratio of two independent variables. The probability distribution of R does not follow any well known probability distributions.

In this paper, we derive formulas for evaluating Type I and Type II errors in detecting copy number ratios and propose an algorithm of equal acceptance regions for detecting copy number ratios. The algorithm can be used to determine the necessary depth of sequencing data required for copy number analysis. By using this opportunity, we also give a detail explanation and discussion on relevant inference concepts used in copy number analysis.

This paper consists of four sections. The relationship between the value of λ_N and the estimate of copy number ratio is discussed in Section 2. Section 3 discusses hypothesis tests on copy number ratios and the

power of the test. The algorithm of equal acceptance regions is introduced in Section 3 and its application is presented in Section 4. Some R code used in this paper, DNACopy outputs and big tables are listed in the Appendix.

2 Impacts of λ_N and λ_T on the estimate of copy number ratio r

In literature, the true copy number r is usually estimated through the tumour-normal ratio R. It can be shown that R is not an unbiased estimator of r. It will be interesting to know if R is an appropriate estimator of r. Based on the definition of R and Assumptions in Section 1, we can prove the following results:

$$E(R|N \neq 0) = m\lambda_T/\lambda_N + o(1) = r + o(1) \tag{1}$$

and

$$Var(R|N \neq 0) = E(R^2|N \neq 0) - (E(R|N \neq 0))^2 = r^2 + o(1) - (r + o(1))^2 = o(1), \tag{2}$$

respectively, as $\lambda_N \to \infty$, where $a_N/a_T = m$. The proof of (**??**) is presented in the Appendix A. Equation (**??**) can be proved in a similar way and the proof is omitted. Therefore, both $E(R) - r$ and $Var(R)$ tend to 0 as $\lambda_N \to \infty$. Although R is not an unbiased estimator of r, but the above results guarantee that R converges to r in probability as $\lambda_N \to \infty$ (recall $\lambda_T/\lambda_N = r/m$), i.e R is an asymptotically unbiased estimator of r. Equation (**??**) ensures that, if λ_N or λ_T is reasonably large, one can confidently claim that the value of R will always appear in a close neighbourhood of true copy number ratio r.

3 Detection of copy-number alterations and the power of the detection

Based on the number of aligned sequence reads from "test" and "reference" samples, to detect whether the true copy number ratio is r_0 or not is equivalent to make a decision between null hypothesis $H_0 : r = r_0$ and alternative hypothesis $H_1 : r \neq r_0$. A test is desirable if the test has small Type I error as well as small Type II error. The smaller the Type II error is, the more powerful the test will be.

Since the tumour-normal ratio R converges to the true copy number ratio r in probability as $\lambda_N \to \infty$, it is reasonable to employ R as a test statistics for the test $H_0 : r = r_0$ vs $H_1 : r \neq r_0$. To carry out the test, knowing the probability distribution of R or understanding how to evaluate $P(a \leq R \leq b)$ for any real numbers $0 < a < b$ is necessary.

A Poisson distribution $Poi(\lambda)$ can be excellently approximated by a normal distribution $N(\lambda, \sqrt{\lambda})$ if $\lambda > 1000$ and can be well approximated by a normal distribution $N(\lambda, \sqrt{\lambda})$ if $\lambda > 10$. For next generation sequencing data, both λ_N and λ_T are more likely to be greater than 10. Therefore, we adopt that $T \sim$

$N(ra_TL/A, \sqrt{ra_TL/A})$ and $N \sim N(a_NL/A, \sqrt{a_NL/A})$. Thus, the probability distribution of R can be approximated by the distribution of the ratio of two independent normally distributed random variables.

The probability distribution of the ratio of two independent normally distributed random variables was investigated by Hayya *et al.* (1975). Their results are briefly introduced below.

Denote Y and X two independent normally distributed random variables and $W = Y/X$. Then W has the following properties

(i) If $c.v.(X) = \sigma_X/\mu_X \leq 0.09$ and $c.v.(Y) > 0.19$, $W = Y/X$ is approximately normally distributed at 5% significant level.

(ii) If $c.v.(X) < 0.39$ and $c.v.(Y) > 0.005$,

$$Z = \frac{\mu_X W - \mu_Y}{\sqrt{\sigma_X^2 W^2 + \sigma_Y^2}}$$

is approximately $N(0,1)$ at 5% significant level.

We apply Hayya *et al.* (1975) results to R and re-express R in the following way

$$R = \frac{T/a_T}{N/a_N} = W\frac{a_N}{a_T},$$

where $W = T/N$. Denote

$$z(r) = \frac{\mu_N W - \mu_T}{\sqrt{\sigma_N^2 W^2 + \sigma_T^2}}, \tag{3}$$

where $\mu_T = ra_TL/A$, $\mu_N = a_NL/A$, $\sigma_T^2 = ra_TL/A$ and $\sigma_N^2 = a_NL/A$. Then, if

$$c.v.(N) = \sqrt{Var(N)}/E(N) = 1/\sqrt{a_NL/A} < 0.39 \tag{4}$$

and

$$c.v.(T) = \sqrt{Var(T)}/E(T) = \frac{1}{\sqrt{ra_TL/A}} > 0.005, \tag{5}$$

that is, $\lambda_T = ra_TL/A < 40,000$ and $\lambda_N = a_NL/A > 6.5747$,

$$z(r) = \frac{\mu_N W - \mu_T}{\sqrt{\sigma_N^2 W^2 + \sigma_T^2}}$$

has standard normal distribution at significance level 0.05. In this paper, we always assume that

$$z(r) = \frac{\mu_N W - \mu_T}{\sqrt{\sigma_N^2 W^2 + \sigma_T^2}} \sim N(0,1),$$

because (**??**) and (**??**) are always held for next generation sequencing data.

Xie and Tammi (2009) proposed a method for detecting copy number variation using next generation sequencing data. They mainly focused on detection of copy number ratio $r = 1$ and gave an interesting discussion on the impact the length of genomic window L on the p-value of the test of $H_0 : r = 1$ vs $H_1 : r \neq 1$. However, they wrongly apply the p-value determined by the test of $H_0 : r = 1$ vs $H_1 : r \neq 1$ to the test $H_0 : r = r'$ vs $H_1 : r \neq r'$ where $r' \neq 1$.

In this section, we extend Xie and Tammi's work to general situations including detections of CNA, determination of p-value of the detections and evaluation of the power of the detections. Unless further notice, we always consider hypothesis test

$$H_0 : r = r_0 \qquad \text{vs} \qquad H_1 : r \neq r_0, \tag{6}$$

where r_0 is the copy number ratio to be tested. It might take value 0 or 0.5 or 1 or 1.5 or $\cdots$ with increment 0.5.

1. Rejection Region for $H_0 : r = r_0$ vs $H_1 : r \neq r_0$

The test statistics used for the test (**??**) is defined as

$$z(r) = \frac{\mu_N W - \mu_T}{\sqrt{\sigma_N^2 W^2 + \sigma_T^2}} \sim N(0, 1).$$

Under H_0,

$$\begin{aligned} p(z_0) &= P(|z(r_0)| > z_0) = P((\mu_N W - \mu_T)^2 > z_0^2(\sigma_T^2 + W^2\sigma_N^2)) \\ &= P((1 - \frac{Az_0^2}{a_N L})R^2 - 2r_0 R + (r_0^2 - r_0\frac{z_0^2 A}{a_T L}) > 0). \end{aligned}$$

Thus,

- If $(1 - \frac{Az_0^2}{a_N L}) > 0$, in terms of the value of R, the rejection region for H_0 at significance level $p(z_0)$ is

$$(R > r_0\frac{1 + \sqrt{1 - N_A T_A}}{N_A}) \bigcup (R < r_0\frac{1 - \sqrt{1 - N_A T_A}}{N_A}); \tag{7}$$

- If $(1 - \frac{Az_0^2}{a_N L}) < 0$, in terms of the value of R, the rejection region for H_0 at significance level $p(z_0)$ is

$$r_0\frac{1 - \sqrt{1 - N_A T_A}}{N_A} < R < r_0\frac{1 + \sqrt{1 - N_A T_A}}{N_A}, \tag{8}$$

where $T_A = 1 - (z_0^2 A)/(r_0 a_T L)$ and $N_A = 1 - (z_0^2 A)/(a_N L)$.

By using simulation data it can be clearly demonstrated that, for any a pre-set Type I error and fixed a_T and a_N, the size of acceptance region increases as the the values of r_0 increases and the overlapped area

between acceptance regions for r_0 and $r_0 + 0.5$ increases as the value of r_0 increases. This means that power of test will be reduced as the value of r_0 increases.

2. Evaluation of the power of tests when $(1 - \frac{Az_0^2}{a_N L}) > 0$

In practice, next generation sequencing data tends to give $(1 - \frac{Az_0^2}{a_N L}) > 0$. Therefore, the scenario of $(1 - \frac{Az_0^2}{a_N L}) > 0$ is of interested.

In this subsection, we evaluate the powers of two types of tests below,

(1) $H_0 : r = r_0$ versus $H_1 : r = r_0 + \triangle r^+$ where $\triangle r^+ > 0$.

(2) $H_0 : r = r_0$ versus $H_1 : r = r_0 + \triangle r^-$ where $\triangle r^- < 0$.

In the following discussion, the Type I errors for both tests are set to be the same $\alpha = (1 - \Phi(z_0))$ with $z_0 > 0$.

Under $H_0 : r = r_0$ and given significance level $\alpha = (1 - \Phi(z_0))$ with $z_0 > 0$, the acceptance regions for the above tests are determined by (**??**).

(i) Consider test $H_0 : r = r_0$ versus $H_1 : r = r_0 + \triangle r^+$.

Denote

$$U = r_0 \frac{1 + \sqrt{1 - N_A T_A}}{N_A} = r_0 \left[1 + \sqrt{1 - (1 - \frac{z_0^2 A}{a_N L})(1 - \frac{z_0^2 A}{r_0 a_T L})} \right] / (1 - \frac{z_0^2 A}{a_N L}).$$

Based on (**??**), the condition for accepting H_0 at significance level $\alpha = (1 - \Phi(z_0))$ is $R \leq U$. Therefore, type II error for the test $H_0 : r = r_0$ vs $H_1 : r = r_0 + \triangle r^+$ is

$$\beta = P(\text{accept } H_0 \text{ under } \alpha = (1 - \Phi(z_0)) | H_1 : r = r_0 + \triangle r^+) = P(R \leq U | H_1 \text{ is true}).$$

The value of β can be evaluated by using the Monte Carlo method. An R function - typeIIerrorPU (in the Appendix C) is developed for the purpose.

(ii) Consider test $H_0 : r = r_0$ versus $H_1 : r = r_0 + \triangle r^-$.

Denote

$$L = r_0 \frac{1 - \sqrt{1 - N_A T_A}}{N_A} = r_0 \left[1 - \sqrt{1 - (1 - \frac{z_0^2 A}{a_N L})(1 - \frac{z_0^2 A}{r_0 a_T L})} \right] / (1 - \frac{z_0^2 A}{a_N L}).$$

Type II error for the test $H_0 : r = r_0$ vs $H_1 : r = r_0 + \triangle r^-$ is

$$\beta = P(\text{accept } H_0 \text{ under } \alpha = (1 - \Phi(z_0)) | H_1 : r = r_0 + \triangle r^-) = P(R \geq L | H_1 \text{ is true}),$$

Table 1: The formulas used to calculate p-values

	$R > r_0$	$R < r_0$
H_1	p-value	p-value
$r \neq r_0$	$2 \times (1 - \Phi(z(r_0))$	$2\Phi(z(r_0))$
$r < r_0$	$\Phi(z(r_0))$	$\Phi(z(r_0))$
$r > r_0$	$1 - \Phi(z(r_0))$	$1 - \Phi(z(r_0))$

The value of $z(r_0)$ is calculated by (**??**). If $R > r_0$, then $z(r_0) > 0$; If $R < r_0$, then $z(r_0) < 0$.

which can be evaluated by using Monte Carlo method. A R function - typeIIerrorPL (in the Appendix C) can be used for the purpose.

Using simulation data one is able to show the power of the test $H_0 : r = r_0$ vs $H_1 : r = r_1$ will decrease as the value of r_0 increases. In addition, the power of the test will increase as the difference between r_0 and r_1 increases. To ensure the power of test $H_0 : r = r_0$ vs $H_1 : r = r_1$, as the value of r_0 increases, a larger value of $\lambda_N = a_N L/A$ is required. Since the size of genomic window L and the length of reference genome A are naturally determined by underlying genome sequence, the power of test is determined by a_N (or a_T as $a_T/a_N = m$), the depth of underlying sequencing data. The higher the depth sequencing data is, the more powerful of the test will be.

3. Calculation of p-values for test statistics

In a test of statistical significance, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed such that H_0 could be rejected. Consider test statistics

$$z(r_0) = \frac{\mu_N W - \mu_T}{\sqrt{\sigma_N^2 W^2 + \sigma_T^2}} = \sqrt{\frac{a_T L}{A}} \frac{R - r_0}{\sqrt{\frac{a_T}{a_N} R^2 + r_0}} \sim N(0, 1)$$

for $H_0 : r = r_0$. The formula used to calculate p-value depends on alternative hypothesis H_1 and the value of R. Different formulas for the purpose are listed in Table 1. Different alternative hypotheses are listed in the first column. If the p-value of $z(r_0)$ is less than significance level $\alpha > 0$, the probability of wrongly rejecting $H_0 : r = r_0$ is at most α.

From the formulas in Table 1, one cannot directly make a decision, rejecting or accepting H_0, based on the value of R. In Theorem **??** below, we derive a sufficient condition in terms of the values of $a_T L/A$ and R for determining whether there are sufficient evidences to reject $H_0 : r = r_0$ at significance level of α.

Theorem 1 Consider a genome sequence with length A. Denote a_T and a_N the number aligned reads from "Test" and "Reference" samples, respectively. Assume $a_T/a_N = m$ and the length of genomic window is L.

Table 2: The values for W_1 and W_2 determined by different value of r_0, where Type I error is α, $d_{r_0,1} = d_{r_0,2} = 0.25$ and $a_T/a_N = 1$.

r	0.5	1	1.5	2
W_1	112.2833	270.8009	482.1577	746.3537
W_2	59.4441	165.1225	323.6401	534.9969
r	2.5	3	3.5	4
W_1	1063.3889	1433.2633	1855.9769	2331.5297
W_2	799.1929	1116.2281	1486.1025	1908.8161

$z_0 = 2.57$ and $\alpha = 0.01$

Denote

$$W_1(z_0, r_0, d_{r_0,1}, m) = \frac{z_0^2\left[m(r_0 + d_{r_0,1})^2 + r_0\right]}{d_{r_0,1}^2} \tag{9}$$

and

$$W_2(z_0, r_0, d_{r_0,2}, m) = \frac{z_0^2\left[m(r_0 - d_{r_0,2})^2 + r_0\right]}{d_{r_0,2}^2} \tag{10}$$

where z_0 is a real number, $r_0 > 0$ and both $d_{r_0,1}$ and $d_{r_0,2}$ are positive real numbers.

Consider hypothesis test $H_0 : r = r_0$ versus $H_1 : r \neq r_0$. If

$$\frac{a_T L}{A} > W_1 \text{ and } R > r_0 + d_{r_0,1} \tag{11}$$

or

$$\frac{a_T L}{A} > W_2 \text{ and } R < r_0 - d_{r_0,2} \tag{12}$$

then, there are sufficient events to reject H_0 with Type I error at most $2\left[1 - \Phi(|z_0|)\right]$.

The proof of Theorem **??** is presented in the Appendix B.

Remark: (i) For hypothesis test $H_0 : r = r_0$ versus $H_1 : r \neq r_0$ with preset significance level $2(1-\Phi(|z_0|)$ and pre-set interval $(r_0 - d_{r_0,2}, r_0 - d_{r_0,1})$, if $a_T L/A > \max\{W_1, W_2\}$, a decision on whether H_0 is held or not can be simply made based on the value of R. If $R \notin (r_0 - d_{r_0,2}, r_0 - d_{r_0,1})$, H_0 can be rejected with at most Type I error $2(1 - \Phi(|z_0|)$. An example of the values of W_1 and W_2 based on $z_0 = 2.57$, i.e. Type I error $= 0.01$, are reported in Table 2, where $d_{r_0,1} = d_{r_0,2} = 0.25$ and $a_T/a_N = 1$. (ii) In addition, Theorem **??** shows a manner for determining a sufficient depth of testing sequencing data for test $H_0 : r = r_0$ versus $H_1 : r \neq r_0$ with preset level of Type I error. This is demonstrated in Example 1 below. (iii) From (**??**) and (**??**), given fixed A, W_1 and W_2, the shorter the window length L, the higher the depth of underlying sequencing data is required. The window length L may be empirically estimated through prior tests or determined based on the knowledge of analysts.

Example 1 Assume that one wishes to detect a 50kb region of a single-copy loss in the alignable portion of human genome ($A = 2.2 \times 10^9$ for 36-bp reads) with Type I error 0.01. To achieve the result and at the same time not to over supply data, the depth of sequencing data, i.e. the values of a_T and a_N, needs to be determined before the data are produced. In this study, we let $a_T = a_N$ and their values are determined below.

For detecting a single-copy loss, we consider test $H_0 : r = 0.5$ versus $H_1 : r \neq 0.5$ and apply the result of Theorem **??** to this example. We use the value $W_1 = 112.2833 \approx 113$ in Table 2 to calculate the number of aligned reads a_T. Therefore, $a_T = W_1 A/L \approx 113 \times 2.2 \times 10^9/50000 = 4.972 \times 10^6$, i.e. 4.972 million. Based on Theorem **??**, if the depths of the sequencing data for normal and tumour samples are the same 4.972 million, then, if we use rule set by Theorem **??** to carry out hypothesis $H_0 : r = 0.5$ vs $H_0 : r \neq 0.5$, the Type I error of the test is at most 0.01.

By noting that, to ensure the Type I error for the test $H_0 : r = r_0$ vs $H_1 : r \neq r_0$ is at most 0.01, the larger the value of r_0 is, the higher the depth of underlying sequencing data is required. Given L and A fixed, if an a_T is able to ensure the Type I error for the test $H_0 : r = r_0$ vs $H_1 : r \neq r_0$ is at most 0.01, then this a_T will be also able to ensure the Type I error for the test $H_0 : r = r^*$ vs $H_1 : r \neq r^*$ is at most 0.01 for all $r^* < r_0$.

Based on the above fact, we suggest an algorithm below for determining the depth of sequencing data when we wish to detect copy number ratios $0.5, 1, \cdots, r^*$ simultaneously and ensure Type I and Type II errors are at most α.

Algorithm of Equal Acceptance Region (EAR): Denote a genomic window of length by L and reference genome length by A.

(i) Use (**??**) to find the W_2 for $r = a + 0.5$, denoting by $W_2(a + 0.5)$.

(ii) Solve a_T from $a_T L/A > W_2(a + 0.5)$ and let $a_N = a_T/m$.

(iii) Following the rule below to make a decision on the estimate of r. If $R < 0.75$, the estimate of r is 0.5; if $0.75 < R < 1.25$, the estimate of r is 1; $\cdots$, if $a - 0.25 < R < a + 0.25$, the estimate of r is a. Otherwise, the estimate of r is greater than a.

If the probability of having copy number ratio greater than a is less that 0.01 and the a_T is determined based on Type I error $\alpha = 0.01$, the a_T will ensure all underlying tests have Type I and Type II errors at most 0.01. This can be briefly explained as follows. The value of a_T is determined by a, therefore, the

EAR algorithm makes sure that $P(R < r_0 - 0.25 \text{ or } R > r_0 + 0.25|\text{the true copy number ratio is } r_0) \leq 0.01$ for all $r_0 = 0, 1, \cdots, a$, i.e. the probability of wrongly identifying copy number given the true value of copy number "$\leq a$" is less than 0.01. If the probability of having copy number ratio greater than a is less that 0.01, it will mean that the maximum probability of wrongly detecting a true copy number ratio with value $> a$ is less than 0.01.

Comparing to the process of carrying out hypothesis for each copy number individually, the advantage of using the EAR algorithm to determine the estimate of r is that the process is simple and all the interested tests can be processed simultaneously. It means that the tedious and time consuming data analysis process can be avoided. Since the a_T is used for testing copy number ratios up to $r = a$ simultaneously, the depth of underlying sequencing determined by the value of a_T may be more than those required for the tests of lower value copy number ratios. It means that following the AER algorithm to determine the depth of tested sequencing data might lead to slightly over supplying data. To balance between simplicity and costs in data analysis the algorithm provides a option.

4 Application

An application of the EAR algorithm is presented in Example 2. We use it to demonstrate that the algorithm can benefit in providing a guideline on the level of the depth of sequencing data for copy number ratio analysis and, consequently, benefit in reducing data analysis cost sometimes.

Example 2. A sequencing dataset of colon tumour is used in this example The data files, chom-1.tsv to chom-Y.tsv, can be download from www.uow.edu.au/∼yanxia/bioinformatics /sequencing_data/. Each file has 6 columns with heading "chrom", "start", "end", "sampleRead", "refRead" and "ratio". The values under "Ratio" are the ratio ("sampleRead") to ("sampleRead" +"refRead") rather than the observations of copy number ratios. The total number of reads for tumour sample and reference sample are $a_T = 6460100$ and $a_N = 59999826$ respectively and the length of each Mark in the data is 3000bp.

In this example, we use the EAR algorithm to determine a_T^* and a_N^*. If $a_T^* << a_T$ and $a_N^* << a_N$, we want to further check whether the sequencing data with depth (a_T^*, a_N^*) enables to provide as much information on copy number ratios as the sequencing data with depths (a_T, a_N) does.

In this example we are interested to detect $r = 0.5, 1, 1.5$ and $r \geq 2$ and require each test has Type I error at most $\alpha = 0.01$.

Firstly, a_T^* and a_N^* are determined in the following steps.

Table 3: The values of W_1 and W_2 with $m = 1.076686$, $d_{r_0,1} = d_{r_0,2} = 0.25$ and $2[1 - \Phi(|z_0|)] = 0.01$. $\alpha = 0.01$

	$r = 0.5$	$r = 1$	$r = 1.5$	$r = 2$
W_1	116.7310	283.1555	506.3727	787.3826
W_2	59.93828	169.57016	335.99470	559.21192
	$r = 2.5$	$r = 3$	$r = 3.5$	$r = 4$
W_1	1123.1852	1516.7804	1967.1683	2474.3489
W_2	839.22181	1176.02438	1569.61961	2020.00752

Step i Preset Type I error $\alpha = 0.01$ and let $m = a_T/a_N = 64601000/59999826 = 1.076686$. For comparison purpose, we require that $a_T^*/a_N^* = m = 1.076686$.

Step ii Let the range of copy number of ratios be 0.5 to 1.5.

Step iii Determine the length of genomic window L. In practice, L might be empirically determined. But in this example, we use the colon tumour data to determine L. We apply R package DNAcopy to the colon tumour data and identify segments of copy number ratios. We found that the lengths for majority of segments are greater than $L = 7 \times 3000$bp. Therefore, we choose $L = 7 \times 3000$bp in this study. For human genome, we use $A = 2.2 \times 10^9$ (Chiang *et al.*, 2008).

Step iv Use formulae (**??**) and (**??**) to calculate W_1 and W_2 for different values of r with Type I error $\alpha = 0.01$. The values of W_1 and W_2 are reported in Table 3. From Table 3, $W_2^* = 559.21192 \approx 560$ is identified, which is given by $r = 2$.

Step v Solve a_T^* for $a_T^* L/A \geq W_2^*$ and give $a_T^* = 58666667$. Therefore, $a_N^* = a_T^*/m = 54488163$. Both the values of a_T^* and a_N^* are much less than a_T and a_N, respectively.

Secondly, we re-sample data with depth $(a_T^*, a_N^*) = (58666667, 54488163)$ from the original colon tumour data, i.e. the data with $(a_T, a_N) = (64601000, 59999826)$.

Finally, we compare the copy number ratio data analysis given by the original data (with (a_T, a_N)) and the subset data (with (a_T^*, a_N^*)). To save the time in evaluating information loss, as an example, we only compare copy number ratio analysis for the data given by Chromosome 22.

We apply R package DNAcopy to both original data set and subset data. The segmental analysis reports for both datasets are reported in the Appendix C, Tables 4 and 5. From the tables, the length of each segment is calculated by num.mark $\times 3000$ and the tumour-normal copy number ratio R on relevant segment are listed under Column "seg.mean".

Then we applied the EAR algorithm to Column "seg.mean" in Tables 4 and 5, respectively and report the estimates of copy number ratios for both datasets in Table 6.

Table 6 shows that the total number of mark positions where both original and subset data gave the same estimate on copy number ratio is 11418, out of 11583. This means that both original and subset datasets provide the same information on copy number ratio on 98.6% of the total mark positions. The total information loss per make position, due to using subset data, is $(11583 - 11418)/11583 = 0.0142 = 1.4\%$. In the meanwhile, the depths of tumour sample and normal sample in the subset dataset are 10% less than those in the original dataset. The total costs in data might be saved by 10% if data analysis is based on subset data. This study shows that it will be possible to reduce the cost of data analysis if the depth of testing sequencing data are determined by the EAR algorithm .

Acknowledgements

The author would like to acknowledge Dr Jianghua Zhang for providing a sequencing dataset of colon tumour for this study and his helpful discussion in the preparation of this paper. The author would like to acknowledge Prof. John Rayner for his carefully reading on earlier version of this paper.

References

Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O., Baker, C., Malig, M., Mutlu, O., *et al.* (2009) 'Personalized copy number and segmental duplication maps using next-generation sequencing', *Nature Genetics*, Vol. 41, pp1061-1067.

Chiang, D.Y., Getz, G., Jaffe, D.B., O'Kelly, M. J.T., Zhao, X., Carter, S. L. , Riss, C., Nusbaum, C., Meyerson, M. and Lander, E.S. (2009) 'High-resolution mapping of copy-number alterations with massively parallel sequencing', *Nature Methods*, Vol. 6, pp99-103.

Hayya,J., Armstrong, D. and Gressis, N. (1975) 'A note on the ratio of two normally distributed variables', *Management Science*, Vol. 21, pp1338-1341.

Kim, T.-M., Luquette, L.J. , Xi, R. and Park, P.J. (2010) 'rSW-seq: Algorithm for detection of copy number alterations in deep sequencing data', *BMC Bioinformatics* Vol. 11, No. 1, p432. doi:10.1186/1471-2105-11-432

Xie, C. and Tammi, M.T. (2009) 'CNV-seq, a new method to detect copy number variation using high-throughput sequencing', *BMC Bioinformatics*, 10:80, doi:10.1186/1471-2105-10-80.

Appendix

Appendix A: Proof of Equations ??

$$E(R|N\neq 0)=\frac{1}{1-e^{-\lambda_N}}\sum_{n=1}^{\infty}\sum_{t=0}^{\infty}\frac{t/a_T}{n/a_N}P(T=t)P(N=n)$$

$$=\frac{1}{1-e^{-\lambda_N}}\sum_{n=1}^{\infty}\sum_{t=0}^{\infty}m\frac{t}{n}\frac{\lambda_T^t}{t!}e^{-\lambda_T}\frac{\lambda_N^n}{n!}e^{-\lambda_N}=\frac{m}{1-e^{-\lambda_N}}\lambda_T\sum_{t=0}^{\infty}\frac{\lambda_T^t}{t!}e^{-\lambda_T}\sum_{n=1}^{\infty}\frac{\lambda_N^n}{nn!}e^{-\lambda_N}$$

$$=\frac{m\lambda_T}{1-e^{-\lambda_N}}\sum_{n=1}^{\infty}\frac{\lambda_N^n}{(n+1)!}\frac{n+1}{n}e^{-\lambda_N}=\frac{r}{1-e^{-\lambda_N}}(1+\sum_{n=1}^{\infty}\frac{\lambda_N^n}{(n+1)!}\frac{1}{n}e^{-\lambda_N}).$$

Given the facts that $\lim_{\lambda_N\to\infty}\lambda_N^k e^{-\lambda_N}=0$ for any constant $k\geq 0$ and $\sum_{n=k+1}^{\infty}\frac{\lambda_N^{n+1}}{(n+1)!}\frac{1}{n}e^{-\lambda_N}\leq 1/k$ for $k>1$, it can be shown that

$$\lim_{\lambda_N\to\infty}E(R|N\neq 0)=\lim_{\lambda_N\to\infty}\frac{r}{1-e^{-\lambda_N}}(1+\sum_{n=1}^{\infty}\frac{\lambda_N^n}{(n+1)!}\frac{1}{n}e^{-\lambda_N})=r.$$

Appendix B: The proof of Theorem ??

Under H_0,

$$z=\sqrt{\frac{a_TL}{A}}\frac{R-r_0}{\sqrt{\frac{a_T}{a_N}R^2+r_0}}=\sqrt{\frac{a_TL}{A}}\frac{R-r_0}{\sqrt{mR^2+r_0}}\sim N(0,1)$$

Therefore,

$$P(\sqrt{\frac{a_T L}{A}}\frac{R-r_0}{\sqrt{mR^2+r_0}} < z_0) = \Phi(z_0).$$

Given the alternative hypothesis $H_1 : r \neq r_0$, $H_0 : r = r_0$ will be rejected at significance level $2(1-\Phi(|z_0|))$ if $\sqrt{\frac{a_T L}{A}}\frac{R-r_0}{\sqrt{mR^2+r_0}} < -|z_0|$ or $\sqrt{\frac{a_T L}{A}}\frac{R-r_0}{\sqrt{mR^2+r_0}} > |z_0|$. In the following, we need to show that these two conditions will be held respectively if corresponding condition (**??**) or (**??**) is held.

Since $\frac{R-r_0}{\sqrt{mR^2+r_0}}$ is an increasing function of R, if (**??**) is held, then,

$$\sqrt{\frac{a_T L}{A}}\frac{R-r_0}{\sqrt{mR^2+r_0}} > \sqrt{\frac{z_0^2(m(r_0+d_{r_0,1})^2+r_0)}{d_{r_0,1}^2}}\frac{d_{r_0,1}}{\sqrt{m(r_0+d_{r_0,1})^2+r_0}} = |z_0|;$$

If (**??**) is held,

$$\sqrt{\frac{a_T L}{A}}\frac{R-r_0}{\sqrt{mR^2+r_0}} < \sqrt{\frac{z_0^2(m(r_0-d_{r_0,2})^2+r_0)}{d_{r_0,2}^2}}\frac{-d_{r_0,2}}{\sqrt{m(r_0-d_{r_0,2})^2+r_0}} = -|z_0|,$$

as required.

Appendix C: R Functions 1. R function - typeIIerrorPU

```
# m=aT/aN

# LT=aT*L/A

# LN=aN*L/A

# r1 = \triangle r must be > 0

typeIIerrorPU= function(m, r0, r1, LT, LN, z0, n){

# n denotes the size of Monte Carlo

# simulation sample

# and 1-\Phi(z0) is the type I error.

set.seed(123)

U= r0*(1+sqrt(1-(1-z0^2/LN)

  *(1-z0^2/(r0*LT))))/(1-z0^2/LN)

N=c()

T=c()

R=c()

count1=0

while (count1 <n){

a =rpois(1,LN)
```

```
if ( a>0) {
count1=count1+1
N[count1]=a
}
}
T=rpois(n, (r0+r1)*LT)
R=T/(N*m)
count=0
for(i in 1:n){
if(R[i]<= U)
count=count+1
}
q=count/n
q
}
```

2. R function - typeIIerrorPL

```
# m=aT/aN
# LT=aT*L/A
# LN=aN*L/A
# r2 = \triangle r must be < 0
typeIIerrorPL= function(m, r0, r2, LT, LN, z0, n){
set.seed(123)
L= r0*(1-sqrt(1-(1-z0^2/LN)
   *(1-z0^2/(r0*LT))))/(1-z0^2/LN)
N=c()
T=c()
R=c()
count1=0
while (count1 <n){
```

```
a =rpois(1,LN)
if ( a>0) {
count1=count1+1
N[count1]=a
}
}
T=rpois(n, (r0+r2)*LT)
R=T/(N*m)
count=0
for(i in 1:n){
if(R[i]>= L)
count=count+1
}
q=count/n
q
}
```

Appendix D: Analysis outputs

The DNAcopy data analysis for subset data and original data are reported in Tables 4 and 5. The final data analysis for Example 2 is in Table 6.

Table 4: The DNAcopy data analysis on subset data of Chromosome 22

ID	chrom	loc.start	loc.end	num.mark	seg.mean
1	22	1	124	124	0.7496
2	22	125	343	219	1.0094
3	22	344	354	11	0.1581
4	22	355	836	482	1.0589
5	22	837	840	4	2.6716
6	22	841	1391	551	1.0494
7	22	1392	1404	12	0.6111
8	22	1405	1406	2	20.0000
9	22	1407	1761	355	1.1443
10	22	1762	1765	4	15.1396
11	22	1766	1829	64	0.9762
12	22	1830	1831	2	3.0685
13	22	1832	1837	6	1.5122
14	22	1838	2626	789	1.0324
15	22	2627	2629	3	0.0000
16	22	2630	3002	373	1.0332
17	22	3003	3009	7	0.2120
18	22	3010	3011	2	2.7918
19	22	3012	3934	923	1.0072
20	22	3935	3937	3	0.0000
21	22	3938	5541	1604	1.0582
22	22	5542	5543	2	20.0000
23	22	5544	11583	6040	1.0372

Table 5: The DNAcopy data analysis on original data of Chromosome 22

ID	chrom	loc.start	loc.end	num.mark	seg.mean
1	22	1	145	145	0.7631
2	22	146	254	109	1.0548
3	22	255	266	12	0.3132
4	22	267	320	54	1.0737
5	22	321	325	5	0.2941
6	22	326	343	18	1.0631
7	22	344	354	11	0.1754
8	22	355	1391	1037	1.0508
9	22	1392	1404	13	0.6685
10	22	1405	1406	2	20.0000
11	22	1407	1438	32	1.2014
12	22	1439	1441	3	13.3829
13	22	1442	1748	307	1.0010
14	22	1749	1751	3	3.2737
15	22	1752	1762	11	1.2997
16	22	1763	1765	3	13.5191
17	22	1766	1829	64	0.9262
18	22	1830	1832	3	3.7151
19	22	1833	1837	5	1.6766
20	22	1838	2626	789	1.0267
21	22	2627	2629	3	0.0000
22	22	2630	3002	373	1.0254
23	22	3003	3009	7	0.2752
24	22	3010	3011	2	2.4767
25	22	3012	5541	2530	1.0332
26	22	5542	5543	2	20.0000
27	22	5544	11583	6040	1.0344

Table 6: The final data analysis for Example 2, where R_{old} and R_{new} are the observations of tumour-normal copy-number ratios obtained from the original data and subset data respectively; $\hat{r}_{old}$ and $\hat{r}_{new}$ are the estimations of the true copy number rations based on the original data and subset data respectively.

start(bp)	end (bp)	R_{old}	R_{new}	$\hat{r}_{old}$	$\hat{r}_{new}$	num. marks matched
3000	37500	0.7631	0.7496	1	0.5	0
375000	438000	0.7631	1.0094	1	1	21
438000	765000	1.0548	1.0094	1	1	109
765000	801000	0.3132	1.0094	0.5	1	0
801000	963000	1.0737	1.0094	1	1	54
963000	978000	0.2941	1.0094	0.5	1	0
978000	1032000	1.0631	1.0094	1	1	18
1032000	1065000	0.1754	0.1581	0.5	0.5	11
1065000	2511000	1.0508	1.0589	1	1	482
2511000	2523000	1.0508	2.6716	1	na	0
2523000	4176000	1.0508	1.0494	1	1	551
4176000	4215000	0.6685	0.6111	0.5	0.5	13
4215000	4221000	20	20	na	na	2
4221000	4317000	1.2014	1.1443	1	1	32
4317000	4326000	13.389	1.1443	na	1	0
4326000	5247000	1.001	1.1443	1	1	307
5247000	5256000	3.2732	1.1443	na	1	0
5256000	5286000	1.2997	1.1443	1.5	1	0
5286000	5289000	1.2997	15.1396	1.5	na	0
5289000	5298000	13.5191	15.1396	na	na	3
5298000	5490000	0.9262	0.9762	1	1	64
5490000	5496000	3.7151	3.0685	na	na	2
5496000	5499000	3.7151	1.5122	na	na	1
5499000	5514000	1.6766	1.5122	na	na	5
5514000	7881000	1.0267	1.0324	1	1	789
7881000	7890000	0	0	0.5	0.5	3
7890000	9009000	1.0254	1.0332	1	1	373
9009000	9030000	0.2752	0.212	0.5	0.5	7
9030000	9036000	2.4767	2.7918	na	na	2
9036000	11805000	1.0332	1.0072	1	1	923
11805000	11814000	1.0332	0	1	0.5	0
11814000	16626000	1.0332	1.0582	1	1	1604
16626000	16632000	20	20	na	na	2
16632000	34752000	1.0344	1.0372	1	1	6040

Notation "na" means that there is no sufficient samples from normal tissue and tumour tissue in relevant genomic window. The estimation of copy number ratio given by the EAR algorithm for the genomic window is not reliable.