BIOINFORMATICS

Vol. 28 ISMB 2012, pages i115–i120
doi:10.1093/bioinformatics/bts203

Statistical model-based testing to evaluate the recurrence of genomic aberrations

Atushi Niida[1,*], Seiya Imoto[1], Teppei Shimamura[1] and Satoru Miyano[1,2]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639 and [2]Computational Science Research Program, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

ABSTRACT

Motivation: In cancer genomes, chromosomal regions harboring cancer genes are often subjected to genomic aberrations like copy number alteration and loss of heterozygosity. Given this, finding recurrent genomic aberrations is considered an apt approach for screening cancer genes. Although several permutation-based tests have been proposed for this purpose, none of them are designed to find recurrent aberrations from the genomic dataset without paired normal sample controls. Their application to unpaired genomic data may lead to false discoveries, because they retrieve pseudo-aberrations that exist in normal genomes as polymorphisms.

Results: We develop a new parametric method named parametric aberration recurrence test (PART) to test for the recurrence of genomic aberrations. The introduction of Poisson-binomial statistics allow us to compute small *P*-values more efficiently and precisely than the previously proposed permutation-based approach. Moreover, we extended PART to cover unpaired data (PART-up) so that there is a statistical basis for analyzing unpaired genomic data. PART-up uses information from unpaired normal sample controls to remove pseudo-aberrations in unpaired genomic data. Using PART-up, we successfully predict recurrent genomic aberrations in cancer cell line samples whose paired normal sample controls are unavailable. This article thus proposes a powerful statistical framework for the identification of driver aberrations, which would be applicable to ever-increasing amounts of cancer genomic data seen in the era of next generation sequencing.

Availability: Our implementations of PART and PART-up are available from http://www.hgc.jp/~niiyan/PART/manual.html.

Contact: aniida@ims.u-tokyo.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Cancer genomes often exhibit chromosomal aberrations like copy number alteration and loss of heterozygosity (LOH). A chromosomal aberration potentially leads to the functional alteration of cancer genes and could be a driver for oncogenesis. For example, if the copy number of some locus is amplified, residing oncogenes would be functionally activated. Conversely the presence of tumor suppressor genes are associated with chromosomal deletion and LOH. However, most aberrations are so-called passengers, which accompany driver aberrations by chance and do not have any causal relationship with oncogenesis. Therefore, it is important problem to discriminate the driver aberrations from the passenger ones. Given that driver aberrations recurrently occur around driver cancer genes whereas passenger aberrations randomly exist across chromosomes, finding recurrent chromosomal aberrations is deemed a powerful approach for discovering driver aberrations and associated driver genes.

In the past decade, microarray technology has enabled genome-wide profiling of copy number and homozygosity (Michels *et al.*, 2007). The application of microarrays to cancer genomes has revealed prevalent aberrations in cancer cells, and produced a large amount of genomic data, which are rich resources for the identification of potential driver loci (Beroukhim *et al.*, 2010). By examining the presence of aberrations across all chromosomal positions in multiple samples, we have a binary aberration profile matrix whose rows and columns correspond to chromosomal positions and samples, respectively. We wish to find chromosomal positions where a significantly large fraction of samples are subjected to aberration. There are a number of computational methods to statistically screen for recurrent genomic aberrations, most of which are based on permutation tests (Morganella *et al.*, 2011). For example, GISTIC calculates the value of statistic scoring recurrence for each genomic position while the null distribution of the statistic is obtained using null aberration profile matrices, which are generated by permuting positions of the binary aberration profile matrix for each sample. Finally, GISTIC reports the significance of recurrence at each position and predicts driver loci by detecting the peaks of the significance plot. (Beroukhim *et al.*, 2007). Although the permutation approach is successful in finding driver aberrations, it is computationally intensive, especially when we need to calculate small *P*-values precisely.

Usually, aberration profiling of a cancer genome is performed by comparing a tumor sample with the paired normal sample. For example, LOH is called for a position whose genotype changes from a heterozygous state in a normal genome to a homozygous state in the paired-tumor genome. When the paired normal sample is not available, aberration profiling is also possible: LOH could be called for a chromosomal segment that has successive homozygous state in the tumor genome (Beroukhim *et al.*, 2006). However, it has been reported that we would confront false positive calls in such unpaired experimental designs: an obtained LOH might be only a polymorphic homozygous segment that exists in the paired normal genome (Heinrichs *et al.*, 2010).

To find recurrent aberrations, we developed a novel parametric test, parametric aberration recurrence test (PART). Using Poisson-binomial statistics (Wang, 1993), PART can be used for efficient and precise calculation of small *P*-values, as compared with the permutation approach. Moreover, we extend PART to cover unpaired data (PART-up) to find recurrent aberrations in unpaired experimental designs. PART-up computes the significance of aberration recurrence by taking into consideration the false positive rate, which is calculated from the unpaired normal sample data. By applying PART-up to simulated and real data, we demonstrate that

*To whom correspondence should be addressed.

our approach can identify recurrent genomic aberrations even in unpaired datasets.

2 METHODS

2.1 PART

First, we introduce a simple parametric test to find the recurrence of genomic aberrations (e.g. LOH) given an aberration profile matrix. Let $\mathbf{R}$ denote an $n \times m$ input binary matrix, whose rows and columns index chromosomal n positions and m samples. If the sample j has a genomic aberration at the position i, we set $R_{ij}=1$; otherwise, $R_{ij}=0$. The problem to be addressed is to statistically test whether a genomic aberration recurrently appears at each position. Namely, we need to calculate a P-value for a test statistic, which is defined as the count of aberrant samples in each position. For the position i, it is given by

$$s_i = \sum_{j=1}^{m} R_{ij}.$$

To calculate the P-value, we assume a null model where the aberration at each position appears with an equal probability within each sample, $\Pr(R_{ij}=1)=p_j$, and the probability can be estimated by the aberration ratio of the sample j:

$$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} R_{ij}.$$

Note that this null model is equivalent to the one generated by position permutation in each sample, which is the approach taken by GISTIC (Beroukhim *et al.*, 2007). Under the null model, if the aberration rate is constant across samples, that is $p_1=p_2=\ldots=p_m$, testing for aberration in each sample can be done by using independent Bernoulli trials with the equal success probability; in such a case, the test statistic s_i, which is the sum of the independent Bernoulli trials, follows the binomial distribution. However, in real data, the aberration varies across samples: some tumors have more genomic aberrations during oncogenesis than others. Therefore, for unequal p_js, testing for aberration at a position in each sample must consider independent Bernoulli trials with unequal success probabilities. In this case, the test statistic s_i follows a general case of the binomial distribution, the Poisson-binomial distribution (Wang, 1993) with the probabilistic function

$$\mathrm{PB}(k=s_i;p_1,p_2,\ldots,p_m)=\sum_{G\in F_k}\prod_{j\in G} p_j \prod_{j\in G^c}(1-p_j),$$

where F_k is the set of all subsets of k integers that can be selected from $1,2,\ldots,m$. For example, for $m=3$ and $k=2$, we have $F_2=\{\{1,2\},\{1,3\},\{2,3\}\}$. G^c is the complement of G, that is $G^c=\{1,2,3,\ldots,n\}\backslash G$. F_k will contain $m!/(m-k)!k!$ elements, over which summation is infeasible in practice unless m is small. However, efficient calculations using a discrete Fourier transformation or recursive formulae have been proposed (Hong, 2011). Using the Poisson-binomial distribution, the P-value $p(s_i)$ is calculated by

$$p(s_i)=1-\sum_{k=1}^{s_i}\mathrm{PB}(k;\hat{p}_1,\hat{p}_2,\ldots,\hat{p}_m).$$

To calculate the cumulative distribution function, we use the DFT-CF (the Discrete Fourier Transform of the Characteristic Function) method implemented in the R poibin package (Hong, 2011).

2.2 PART-up

In unpaired experimental designs, the aberration profile matrix of tumor samples, $\mathbf{R}$, may contain false aberration calls. Here, we address the problem of how to test for the recurrence of genomic aberration in the presence of false positive aberration calls. When the aberration profile matrix of unpaired normal samples from the same cohort is available, the Poisson-binomial approach enables us to test for the recurrence of aberration by denoising the false positive aberration calls.

Let $\mathbf{S}$ and $\mathbf{T}$ denote $n \times l$ and $n \times m$ binary matrices, and be referred to as the false aberration profile matrix and the true aberration profile matrix, respectively. $\mathbf{S}$ was prepared for l unpaired normal samples with the same procedure as that used for $\mathbf{R}$, and $\mathbf{T}$ is unobserved data that we can obtain by removing false positive calls from $\mathbf{R}$. We define the probability of aberration at the position i of the sample j as $p_{ij}=\Pr(R_{ij}=1)$, and use $\{p_{i1},p_{i2},\ldots,p_{im}\}$ as the Poisson-binomial parameters. Note that we must compute the Poisson-binomial parameters for each position by considering the existence of the false positive aberration calls from normal samples.

Under the assumption that the false positive rate of aberration calls at each position is constant across samples, the false positive rate p_i^{F} can be estimated from $\mathbf{S}$: $\hat{p}_i^{\mathrm{F}}=\sum_{j=1}^{l} S_{ij}/l$. We also define the probability that the aberration observed at the position i of the sample j is true: $p_{ij}^{\mathrm{T}}=\Pr(T_{ij}=1)$. As an observed aberration must be either a true aberration or a false positive aberration call, the following equation holds among these probabilities:

$$p_{ij}=p_{ij}^{\mathrm{T}}+(1-p_{ij}^{\mathrm{T}})\cdot p_i^{\mathrm{F}}. \quad (1)$$

Under the null model, the probability of true aberrations should be constant within each sample, that is, $p_{ij}^{\mathrm{T}}=p_j^{\mathrm{T}}$. By noting this, we can take the average of Equation (1) over positions:

$$p_j=p_j^{\mathrm{T}}+(1-p_j^{\mathrm{T}})\cdot p^{\mathrm{F}}, \quad (2)$$

where $p_j=\frac{1}{n}\sum_{i=1}^{n} p_{ij}$ and $p^{\mathrm{F}}=\frac{1}{n}\sum_{i=1}^{n} p_i^{\mathrm{F}}$. As $\hat{p}_j$, $\hat{p}^{\mathrm{F}}$ and $\hat{p}_i^{\mathrm{F}}$ are available from the data, we can calculate $\hat{p}_{ij}^{\mathrm{T}}$ from Equation (2):

$$\hat{p}_{ij}^{\mathrm{T}}=\hat{p}_j^{\mathrm{T}}=\frac{\hat{p}_j-\hat{p}^{\mathrm{F}}}{1-\hat{p}^{\mathrm{F}}}. \quad (3)$$

By substituting Equation (3) for Equation (1), we obtain:

$$\hat{p}_{ij}=\frac{\hat{p}_j\cdot(1-\hat{p}_i^{\mathrm{F}})-\hat{p}^{\mathrm{F}}+\hat{p}_i^{\mathrm{F}}}{1-\hat{p}^{\mathrm{F}}}.$$

Now, we have $\{\hat{p}_{i1},\hat{p}_{i2},\ldots,\hat{p}_{im}\}$, which are Poisson-binomial parameter estimates adjusted for each position. Using these parameter estimates, the P-value $p(s_j)$ is calculated as described in the previous section.

2.3 Preparation of simulation data

Simulation data for benchmark tests were generated partially based on a study by Guttman *et al.* (2007). Assuming unpaired experiments, we simulate $n \times m$ and $n \times l$ binary matrices, $\mathbf{R}$ and $\mathbf{S}$. In our simulation, we assumed three types of aberrations: concordant true, non-concordant true and concordant false positive aberrations. All types of aberrations exist in $\mathbf{R}$; therefore, we independently generated $n \times m$ binary matrices, $\mathbf{R}^{\mathrm{c}}$, $\mathbf{R}^{\mathrm{n}}$ and $\mathbf{R}^{\mathrm{f}}$. To obtain $\mathbf{R}$, we combined them using $R_{ij}=\max\{R_{ij}^{\mathrm{c}},R_{ij}^{\mathrm{n}},R_{ij}^{\mathrm{f}}\}$. These matrices are illustrated in Figure 1. Only concordant false positive aberrations exist in $\mathbf{S}$.

- Concordant true aberrations. $\mathbf{R}^{\mathrm{c}}$ contains concordant true aberrations of width w^{c} at row intervals specified by an integer set C, which contains the row indices of aberration centers. For each interval, aberrations recurrently appear in columns randomly sampled with rate r^{c}. Note that this type of aberrations should be a target of PART and the positions in the specified concordant intervals are used as actual positives in benchmark tests.
- Non-concordant true aberrations. Each column of $\mathbf{R}^{\mathrm{n}}$ has $k_j'^{\mathrm{n}}$ non-concordant true aberrations of length w^{n}. For each aberration, we sampled the interval length $w'^{\mathrm{n}} \sim$ Geometric $(1/w^{\mathrm{n}})$, and the interval position randomly. The number of aberrations $k_j'^{\mathrm{n}}$ was sampled for each column so that $k_j'^{\mathrm{n}} \sim \mathrm{Poisson}(k^{\mathrm{n}})$.
- Concordant false positive aberrations. Intervals of concordant false positive aberrations have width w^{f} and are specified by the concordant row interval set of size k^{f}. The k^{f} elements were randomly sampled from 1 to n. For each of the concordant intervals, $\mathbf{R}^{\mathrm{f}}$ and $\mathbf{S}$ have aberrations in columns randomly sampled with rate r^{f}.

In our simulation, we fix the parameters as $n=1000$, $m=100$, $l=300$, $w^{c}=w^{n}=w^{f}=5$ and $C=\{200,400,600,800\}$. For the other parameters, k^{n}, k^{f}, r^{c} and r^{f}, several combinations of parameter values were examined, as described later.

2.4 Preparation of real data

We obtain paired and unpaired LOH profile matrices for 294 pairs of colorectal cancer and normal samples. As a data source, TCGA (The Cancer Genome Atlas) Level 3 SNP data profiled by Affymetrix SNP Array 6.0 were downloaded from the TCGA data portal site (http://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). To obtain a paired LOH profile matrix, we performed LOH detection for each pair of the samples-based allelic imbalance (Staaf *et al.*, 2008). For SNPs whose genotypes are called heterozygous in the normal genome, the ratio of allelic copy intensities was calculated as the B allele frequency (BAF) score: $B=b/(a+b)$, where a and b are the copy number intensity for A and B alleles in the cancer genome. The BAF score should be 0.5 if the position has no allelic imbalance. We then computed the absolute deviation of the BAF score from 0.5 as the BAF′: $B'=|B-0.5|$. BAF′ was plotted along chromosomes and segmented using the circular binary segmentation algorithm with parameter $\alpha=0.01$ (Venkatraman and Olshen, 2007). We assumed that chromosomal segments with $B'>0.1$ are subjected to LOH. To obtain the unpaired LOH profile matrices, we applied the basic Hidden Markov Model method proposed by Beroukhim *et al.* (2006) to the cancer genome data. His method uses a hidden Markov model to detect successive LOH while taking into account SNP intermarker distances. LOH in unpaired samples was also detected by the same procedure.

We also prepared the unpaired aberration profile matrices for the Sanger cell line data. We obtained Affymetrix SNP Array 6.0 data containing 764 cell lines and 466 unpaired normal samples from the Cancer Genome Project site (http://www.sanger.ac.uk/genetics/CGP/Archive/) (Bignell *et al.*, 2010). The unpaired LOH profile matrices were obtained in the same way as the TCGA data. The copy number amplification and deletion profile matrices were obtained by binarizing the copy number profiles predicted by PICNIC (Greenman *et al.*, 2010).

3 RESULTS

3.1 Simulation data test

First, we numerically show that our Poisson-binomial approach is statistically equivalent to the permutation approach adopted by other methods. We prepared an LOH profile matrix from the Sanger cell line data, and compare PART P-values with those based on 10 000 permutations of chromosomal positions for each sample. The minus log scale P-value plot in Figure 2 shows that they correspond with each other, although the permutation approach has a limitation for the calculation of small P-values. We also permutated chromosomal positions for each sample, apply PART to the permutated data and plot a histogram of P-values. Supplementary Materials Figure S1 shows that the histogram is close to uniform distributions, demonstrating that our method successfully controls the rate of false positives.

Next, we numerically compare the performance of PART and PART-up. We simulated aberration profile matrices; in each simulation, we obtain a pair of true and false positive aberration matrices. To generate matrices, we assumed three types of aberrations: concordant true non-concordant true and concordant false positive aberrations. In the true aberration matrix, concordant true aberrations appear recurrently at specific positions whereas non-concordant true aberrations appear randomly. Namely, concordant true aberrations mimic drivers targeted by PART whereas non-concordant true aberrations mimic passengers. As concordant false positive aberrations mimic polymorphisms which exist in both cancer and normal genomes, they appear recurrently at the same position and frequency in both the true and false positive aberration matrices. The simulation data were generated from a simulator with four free parameters: k^{n} for the number of non-concordant true aberrations, k^{f} for the number of concordant false positive aberrations, r^{c} for the rate of samples subjected to concordant true aberration, and r^{f} for the rate of samples subjected to concordant false positive aberrations. A simulation instance is illustrated in Figure 1.

We prepared 16 types of simulators with different parameter settings and obtained 100 matrix pairs from 100 Monte Carlo trials for each simulator. For each Monte Carlo matrix pair, we applied PART to the true aberration matrices and PART-up to both the matrices. The result for a Monte Carlo matrix pair is presented in Figure 3. From the results pooled across the 100 Monte Carlo trials, we calculated precision and recall for each method over the whole range of significance cutoffs to depict precision-recall (PR) curves. Precision is defined as the proportion of actual in predicted positives, whereas recall is defined as the proportion of predicted in actual positives. We assumed positions within concordant intervals

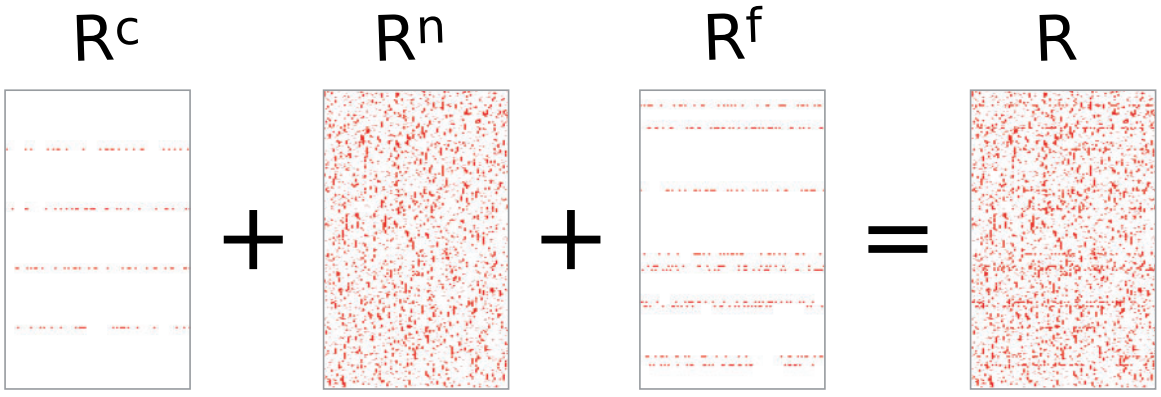

Fig. 1. Simulation of aberration matrices. In a Monte Carlo trial, three matrices, $\mathbf{R}^{c}$, $\mathbf{R}^{n}$ and $\mathbf{R}^{f}$, were simulated with $k^{n}=10$, $k^{f}=10$, $r^{c}=0.3$ and $r^{f}=0.3$. $\mathbf{R}^{c}$, $\mathbf{R}^{n}$ and $\mathbf{R}^{f}$ contain concordant true, non-concordant true, and concordant false positive aberrations, respectively. The aberration matrix $\mathbf{R}$ is obtained by overlapping the three matrices. Although the false positive aberration matrix $\mathbf{S}$ is not shown here, its simulation is similar to that of $\mathbf{R}^{f}$

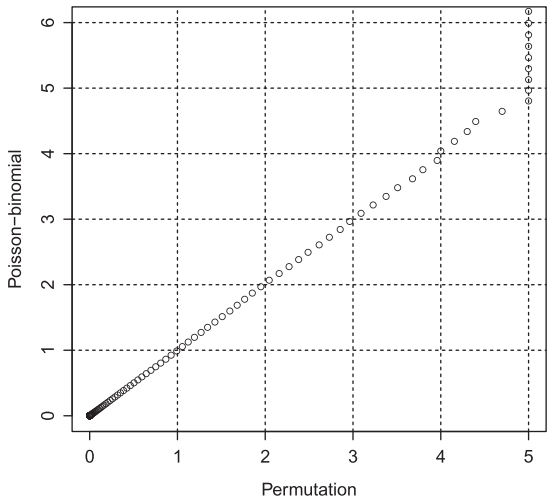

Fig. 2. Comparison of *P*-values between the Poisson-binomial and permutation approaches. The *P*-values for recurrent genomic aberration were obtained by the Poisson-binomial statistics and permutations, and plotted in minus log scale

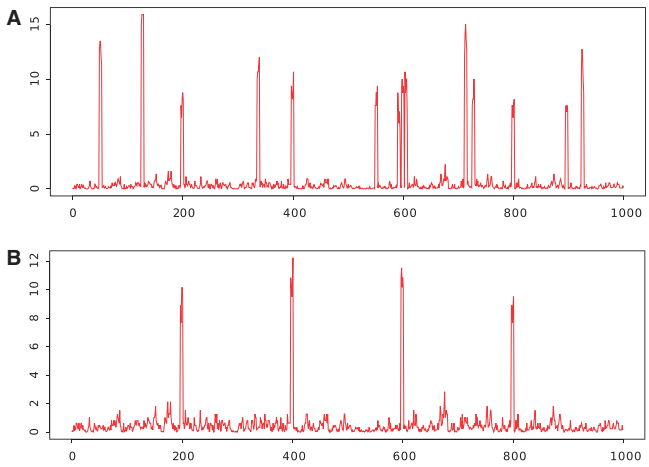

Fig. 3. An example of significance plots for simulation data. PART and PART-up were applied to the simulated data presented in Figure 1. The *P*-values from PART **(A)** and PART-up **(B)** were plotted in minus log scales across chromosomal positions

as actual positives and positions determined by a significance cutoff as predicted positives. The PR curve shows the discriminative ability of each method to find positions associated with true concordant aberrations. For this type of benchmark tests, the receiver operating characteristic curve is also popular; however, we chose the PR curve because it is preferred for our case where the number of actual positives is relatively small (Davis and Goadrich, 2006). The heatmap in Figure 4 shows the PR curves for the 16 different parameter settings. PART-up performs better in the presence of more concordant false positive aberrations (i.e. when the parameters controlling the frequency of concordant false positive aberrations, k^{f} and r^{f}, are larger), while the performance of both methods are attenuated by the presence of non-concordant true aberrations (i.e. when the parameter controlling the frequency of non-concordant

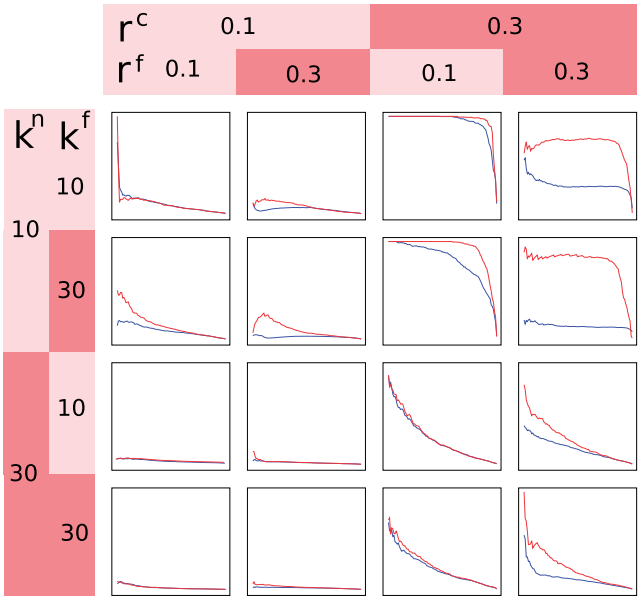

Fig. 4. PR curves of PART and PART-up. PR curves were obtained by applying PART (blue lines) and PART-up (red lines) to simulation data from 16 different parameter settings. The horizontal axis indicates recall whereas the vertical axis indicates precision

true aberrations, k^{n}, is larger). These results demonstrate that PART-up can successfully discriminate concordant true aberrations from concordant false positive aberrations, suggesting its applicability to genomic data obtained in unpaired experimental designs.

3.2 Real data test

In this section, we compare PART and PART-up using real experimental data. First, we focus on TCGA colorectal cancer SNP data that were obtained in a paired experimental design. To examine the performance of PART-up, we prepared two types of LOH profile matrices in two different ways: paired and unpaired LOH profile matrices. The former was authentically obtained based on paired genomes: LOHs were called by comparing genotypes between tumor and paired normal genomes. The latter was approximately obtained based only on tumor genomes: LOHs were calls for segments that harbor successive homozygous calls in tumor genomes. An unpaired LOH profile matrix for normal samples was also prepared for PART-up input.

We applied PART to the paired and unpaired matrices, and plot minus log *P*-value across chromosomes. The paired LOH profile matrix produces a clear significance plot, whereas the unpaired LOH profile matrix yields a noisy plot with many spikes, as shown in Figure 5A and B. The spikes would reflect false positive LOH calls for polymorphic homozygous segments that exist in normal genomes. We also applied PART-up to the same unpaired LOH profile matrix from tumor samples combined with that from normal samples. Figure 5C shows that PART-up successfully removes most of the spikes and the result corresponds well to that from PART applied to the paired matrix. This observation demonstrates that, if the aberration profiles for unpaired normal samples are available, PART-up performs as well for unpaired data as PART does for paired data.

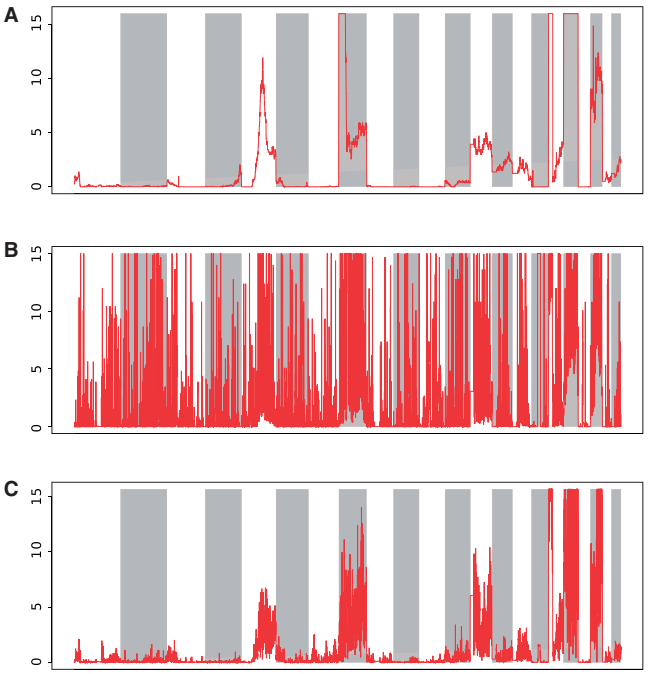

Fig. 5. Significance plots for recurrent LOH in the TCGA colorectal cancer data. Minus log scaled *P*-values for the recurrence of LOH were plotted across chromosomes, each of which is indicated by vertical stripes. **(A)** PART was applied to the paired LOH profile matrix. **(B)** PART was applied to the unpaired LOH profile matrix from the tumor samples. **(C)** PART-up was applied to the unpaired LOH profile matrices from the tumor and normal samples

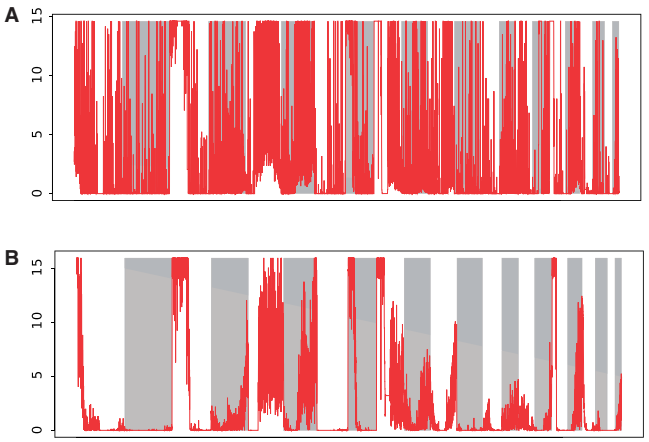

Fig. 6. Significance plots for recurrent LOH in the Sanger cell line cancer data. Minus log scaled *P*-values for the recurrence of LOH were plotted across chromosomes, each of which is indicated by vertical stripes. **(A)** PART was applied to the unpaired LOH profile matrix from the tumor samples. **(B)** PART-up was applied to the unpaired LOH profile matrices from the tumor and unpaired normal samples

Next, we obtained unpaired LOH profile matrices from a cancer cell line dataset published by the Sanger institute (Bignell *et al.*, 2010). As paired normal controls are usually unavailable for cell lines, it has been difficult to find recurrent aberrations in cell line data. However, as the Sanger dataset is accompanied by hundreds of unpaired normal sample data items, it can be subject to PART-up. We apply PART and PART-up to the unpaired LOH profile matrices and the significance plots are shown in Figure 6A and B. As in the TCGA case, although PART produces a noisy plot, PART-up produces a much clearer plot, revealing recurrent LOH region in the cancer cell lines. We also applied PART and PART-up to the unpaired copy number amplification and deletion profile matrices, and these results are shown in Supplementary Materials Figures S2 and S3. The comparison between the significance plots demonstrates that PART-up removes some spikes, which would be from copy number polymorphisms in normal genomes. The differences are less dramatic than in the LOH case, reflecting the low rate of pseudoaberrations from normal genomes for copy number aberrations as compared with LOH (see Supplementary Materials Table S1). As such, we conclude that PART-up successfully identifies recurrently aberrant regions from unpaired genomic data.

4 DISCUSSION

In this study, we presented a novel statistical method, PART, to test the recurrence of genomic aberration. Although a number of methods have been developed for similar purposes, most of them take permutation approaches to assess statistical significance. Conversely, our method takes a novel parametric approach by employing Poisson-binomial statistics. There are pros and cons between the two approaches. Our parametric approach needs less computational time and can calculate small *P*-values more accurately than the permutation approach. This property is important for genomic analysis, because we usually need to calculate small *P*-values to correct large-scale multiple hypothesis testing. Conversely although our approach must take the count of aberrant samples as a test statistic, the permutation approach is flexible for the type of statistic and can enable more biologically plausible tests. For example, the GISTIC statistic takes into consideration aberration strength in addition to recurrence (Beroukhim *et al.*, 2007). However, in spite of these differences, we found that PART and GISTIC produce consistent results on the copy number data (See Supplementary Note).

The most notable advantage of our parametric approach is highlighted by the extension of PART to PART-up. Although it is preferable that genomic aberration profilings are performed in paired experimental designs, it is not always possible to obtain paired normal samples. However, PART-up is applicable only if data are accompanied with unpaired normal samples from the same cohort, which are generally easier to obtain. Although previously proposed methods are not able to deal with such unpaired data, the introduction of Poisson-binomial statistics enables us to test recurrent aberrations in unpaired data while considering the rate of pseudo-aberrations that originate from normal genomes. Although a method has been proposed to call aberrations in individual tumor samples using the pooled unpaired normal sample as a reference (Yamamoto *et al.*, 2007), testing for the recurrence of aberrations in a group of unpaired tumor samples is a novel approach. It is expected that combining these complementary approaches will reduce false positives and lead to higher performance.

In this study, we applied our method to genomic aberration profiles obtained by the microarray technology. The next-generation sequence technology has result in a torrent of cancer genome data (Meyerson *et al.*, 2010). The current dataset is not only large but also complex, in that the new technology can profile various types of genomic aberrations that cannot be captured by the old technology: point mutations, short indels, translocations, and so forth. A population-scale sequencing project targeting thousands of normal genomes is also ongoing (1000 Genomes Project Consortium, 2010); data produced by the project would be used as unpaired normal sample controls for PART-up. Based on this study, future studies would address the application of our Poisson-binomial approach to a wider spectrum of genomic aberrations revealed by the next-generation sequence technology.

ACKNOWLEDGEMENTS

Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo.

Funding: This work was supported by Research Fellowship for Young Scientists from Japan Society for the Promotion of Science, and 'Systems Cancer' (Project No 4201), a Grant-in-Aid for Scientific Research on Innovative Areas by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

Conflict of Interest: none declared.

REFERENCES

Beroukhim,R. *et al.* (2006) Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. *PLoS Comput. Biol.*, **2**, e41.

Beroukhim,R. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA*, **104**, 20007–1200.

Beroukhim,R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Bignell,G. *et al.* (2010) Signatures of mutation and selection in the cancer genome. *Nature*, **463**, 893–898.

Davis,J. and Goadrich,M. (2006) The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, pp. 233–240.

Greenman,C. *et al.* (2010) PICNIC: An algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, **11**, 164–175.

Guttman,M. *et al.* (2007) Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genet.*, **3**, e143.

Heinrichs,S. *et al.* (2010) SNP array analysis in hematologic malignancies: avoiding false discoveries. *Blood*, **115**, 4157–4161.

Hong,Y. (2011) On computing the distribution function for the sum of independent and nonidentical random indicators. *Technical Report No. 11–3*, Department of Statitics, Virginia Tech.

Meyerson,M. *et al.* (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.*, **11**, 685–696.

Michels,E. *et al.* (2007) Detection of dna copy number alterations in cancer by array comparative genomic hybridization. *Genet. Med.*, **9**, 574–584.

Morganella,S. *et al.* (2011) Finding recurrent copy number alterations preserving within-sample homogeneity. *Bioinformatics*, **27**, 2949–2956.

Staaf,J. *et al.* (2008) Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol.*, **9**, R136.

Venkatraman,E. and Olshen,A. (2007) A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, **23**, 657–663.

Wang,H. (1993) On the number of successes in independent trials. *Statistica Sin.*, **3**, 295–312.

Yamamoto,G. *et al.* (2007) Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. *Am. J. Hum. Genet.*, **81**, 114–126.

1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.