BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 4 2010, pages 464–469
doi:10.1093/bioinformatics/btp708

Genome analysis

CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data

Qunyuan Zhang[1,*], Li Ding[2], David E. Larson[2], Daniel C. Koboldt[2], Michael D. McLellan[2], Ken Chen[2], Xiaoqi Shi[2], Aldi Kraja[1], Elaine R. Mardis[2], Richard K. Wilson[2], Ingrid B. Borecki[1] and Michael A. Province[1]

[1]Division of Statistical Genomics and [2]The Genome Center, Washington University School of Medicine, St Louis, MO, USA

Received on October 16, 2009; revised on December 18, 2009; accepted on December 21, 2009

Advance Access publication December 23, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: DNA copy number aberration (CNA) is a hallmark of genomic abnormality in tumor cells. Recurrent CNA (RCNA) occurs in multiple cancer samples across the same chromosomal region and has greater implication in tumorigenesis. Current commonly used methods for RCNA identification require CNA calling for individual samples before cross-sample analysis. This two-step strategy may result in a heavy computational burden, as well as a loss of the overall statistical power due to segmentation and discretization of individual sample's data. We propose a population-based approach for RCNA detection with no need of single-sample analysis, which is statistically powerful, computationally efficient and particularly suitable for high-resolution and large-population studies.

Results: Our approach, correlation matrix diagonal segmentation (CMDS), identifies RCNAs based on a between-chromosomal-site correlation analysis. Directly using the raw intensity ratio data from all samples and adopting a diagonal transformation strategy, CMDS substantially reduces computational burden and can obtain results very quickly from large datasets. Our simulation indicates that the statistical power of CMDS is higher than that of single-sample CNA calling based two-step approaches. We applied CMDS to two real datasets of lung cancer and brain cancer from Affymetrix and Illumina array platforms, respectively, and successfully identified known regions of CNA associated with *EGFR*, *KRAS* and other important oncogenes. CMDS provides a fast, powerful and easily implemented tool for the RCNA analysis of large-scale data from cancer genomes.

Availability: The R and C programs implementing our method are available at https://dsgweb.wustl.edu/qunyuan/software/cmds.

Contact: qunyuan@wustl.edu

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

1 INTRODUCTION

DNA copy number aberration (CNA) is a significant phenomenon of amplification and deletion of chromosomal regions in tumor cells. Identification of CNAs may provide important insight into the molecular mechanism of oncogenesis, as well as useful information for the diagnosis and treatment of cancers. A recurrent CNA (RCNA) is a CNA that occurs in multiple patients across the same chromosomal region. In contrast to non-recurrent CNAs, RCNAs are often identified as more significant events with greater implication in tumorigenesis.

In recent years, array-based technologies, such as array comparative genomic hybridization (CGH), single nucleotide polymorphism (SNP) arrays and copy number (CN) arrays, have facilitated the genome-wide studies of CNA. Numerous mathematical methods for CNA identification have been developed (Beroukhim *et al*., 2007; Diskin *et al*., 2006; Guttman *et al*., 2007; Hsu *et al*., 2005; Hupe *et al*., 2004; Jong *et al*., 2004; Lai *et al*., 2005a; Lai *et al*., 2005b; Lipson *et al*., 2006; Nilsson *et al*., 2009; Olshen *et al*., 2004; Picard *et al*., 2005; Rouveirol *et al*., 2006; Rueda *et al*., 2009; Shah *et al*., 2006; Shah *et al*., 2007). These methods can be roughly categorized into two classes: CNA identification for individual samples and RCNA identification across multiple samples. In contrast to the CNA analysis of individual samples, RCNA analysis involving multiple samples is more difficult and its application still remains limited due to the diversity of CNA patterns among individuals and the computational burden resulting from the high density of signals and large sample size. Published approaches for RCNA analysis include significance testing for aberrant copy number (STAC) (Diskin *et al*., 2006), geometic family algorithm (GFA) (Lipson *et al*., 2006), minimal alteration region (MAR) method (Rouveirol *et al*., 2006), Multiple Chain Hidden Markov Model (HMM) (Shah *et al*., 2007), genomic identification of significant targets in cancer (GISTIC) method (Beroukhim *et al*., 2007), multiple sample analysis (MSA) (Guttman *et al*., 2007), probabilistic recurrent copy number region method A (pREC-A) (Rueda *et al*., 2009), etc. Despite the use of different algorithms, most of these approaches adopt a two-step strategy that requires signal discretization (smoothing, binarization, segmentation, HMM modeling, cutoff definition, etc.) for individual samples prior to cross-sample analysis. Although discretization provides a useful CNA pattern for individuals, it may result in loss of the raw distribution information during the conversion of continuous signals into discretized data. As a result, the overall statistical power of RCNA detection may be diminished. Furthermore, individual

*To whom correspondence should be addressed.

sample analysis, along with cross-sample analysis, may produce a heavy computational burden that could impede the application of these methods in genome-wide studies with high resolution and large sample size. Especially in the near future, the next generation sequencing data from large populations will present more challenges to these methods.

In this article, we propose a computationally efficient and statistically powerful approach, correlation matrix diagonal segmentation (CMDS), which directly analyzes the intensity ratio data to identify RCNAs across a large set of samples, with no prior need of single-sample analysis. First, we introduce the rationale behind CMDS and the procedure for its implementation. Next, we investigate the statistical power of CMDS under a variety of configurations via simulation, and compare CMDS with single-sample analysis-based methods in terms of power and computational time. Finally, we demonstrate the applicability and efficiency of CMDS by analyzing two real datasets of lung cancer and brain cancer from Affymetrix and Illumina high-resolution array platforms.

2 METHODS

2.1 Proposed method

When a RCNA exists in the same chromosomal region across individuals (Fig. 1A), it produces CN co-variation between neighboring chromosomal sites within the region. As a result, a correlation block can be observed in the CN correlation matrix of chromosomal sites (Fig. 1B). Therefore, RCNAs can be identified by detecting the corresponding correlation blocks along the diagonal of the correlation matrix. Based upon this rationale, we propose the CMDS approach, described below and as illustrated in Figure 1.

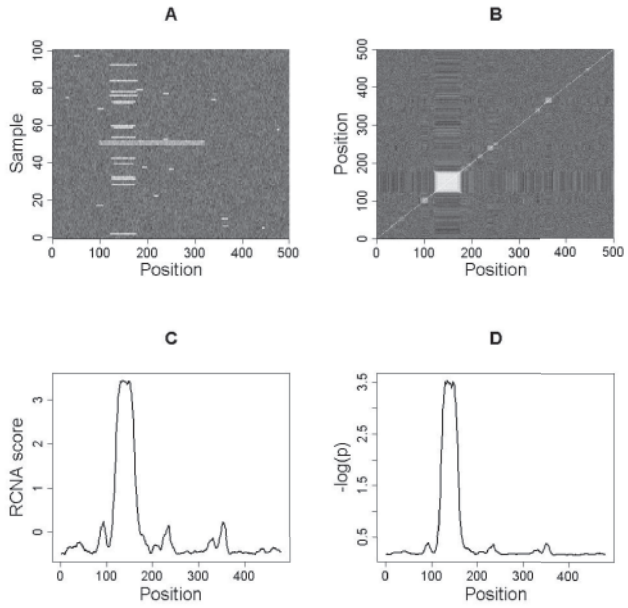

Fig. 1. Illustration of the CMDS approach. (**A**) Visualization of CN values of 500 chromosomal sites and 100 samples. White indicates amplification (CN > 2). (**B**) Visualization of the CN correlation matrix of 500 sites. The white block corresponds to a RCNA region. (**C**) The RCNA scores obtained by the DT. (**D**) Negative $\log_{10}(p)$-values from significance test of RCNA scores.

2.1.1 Input data As input CMDS takes the CN data of m physically ordered chromosomal sites of n individual samples, denoted by a $n \times m$ matrix ($\boldsymbol{X}$), in which the element x_{ij} is the CN value of individual i at chromosomal site j. Usually, x_{ij} is measured by the intensity ratio (or $\log_2$ ratio) between a target sample and reference sample(s).

2.1.2 CN correlation matrix The first step of CMDS is to calculate Pearson's correlation coefficients between chromosomal sites based on the CN matrix $\boldsymbol{X}$ via the formula:

$$r_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{ij}-\bar{x}_j}{s_j}\right)\left(\frac{x_{ik}-\bar{x}_k}{s_k}\right) \quad (1)$$

where r_{jk} is the correlation coefficient between sites j and k; $\bar{x}_j$, $\bar{x}_k$, s_j and s_k are CN means and SDs of sites j and k across n individuals. For the convenience of probability calculation, r_{jk} is normalized by Fisher's transformation

$$z_{jk} = \frac{\sqrt{n-3}}{2}\log\frac{1+r_{jk}}{1-r_{jk}} \quad (2)$$

The z_{jk} values of all possible pairs of m sites compose a diagonally symmetric $m \times m$ correlation matrix ($\boldsymbol{Z}$), which is the basis of the CMDS analysis. The statistical scenario underlying CMDS is, if no CNA exists across samples (i.e. the variation in $\boldsymbol{X}$ is resulted only from random experimental noises), the expected r_{jk} should be 0 (i.e. no correlation between sites) and $\boldsymbol{Z}$ will randomly follow the standard normal distribution (SND). In contrast, if CNA recurs at a chromosomal region across samples, it will cause positive correlation (i.e. $r_{jk} > 0$) between the sites within the region. As a result, the expected z_{jk} values corresponding to the region will be higher than others and result in a square block along the diagonal of $\boldsymbol{Z}$ (Fig. 1B). Thus, RCNA regions can be detected by searching square blocks along the diagonal of $\boldsymbol{Z}$.

2.1.3 Diagonal transformation Usually, identification of square blocks in a matrix can be treated as an image recognition problem. However, the size of $\boldsymbol{Z}$ in genome-wide studies could be too large for direct image processing, and image recognition does not provide probability measures for significance testing of such correlation blocks. Since we are only interested in the blocks along the diagonal, it is unnecessary to search the entire matrix. We therefore propose a computationally efficient and easily implemented approach that performs a diagonal transformation (DT) to convert the 2D problem into a 1D problem. Along the diagonal of the correlation matrix $\boldsymbol{Z}$, DT calculates a RCNA score (R_h) for each chromosomal site h based on the correlation values of $b+1$ sites surrounding (and including) the site h:

$$R_h = \frac{2}{b(b+1)}\sum_{j=(h-b/2)}^{h+b/2-1}\sum_{k=j+1}^{h+b/2} z_{jk} \quad (3)$$

In the DT formula above, the average of the z_{jk} values of $b+1$ sites around the site h (b/2 left sites, b/2 right sites, plus the site h *per se*) are used to measure the degree of RCNA of the site h, which is equivalent to sliding a $(b+1)\times(b+1)$ square block along the diagonal of $\boldsymbol{Z}$ and averaging the z_{jk} values within each block (excluding self correlation values, i.e. $j \neq k$). Since DT only needs to calculate the z_{jk} values within small blocks (not all the elements in $\boldsymbol{Z}$), it is very low cost in terms of both computer time and memory.

2.1.4 Significance test of the RCNA score Under the null hypothesis that there is no CNA (i.e. no correlation between chromosomal sites), the RCNA score R_h will randomly follow a normal distribution with a mean of $\mu_R = 0$ and a variance of

$$\sigma_R^2 = \frac{2}{b(b+1)} \quad (4)$$

In real data analysis, μ_R and σ_R^2 may deviate from their theoretical values due to artifactual correlations between samples and/or adjacent chromosomal

positions, which usually are the results of experimental noise, batch effect, genomic wave effect (Diskin *et al*., 2008; Marioni *et al*., 2007), non-recurrent CNA and/or CNA of a whole chromosome or chromosomal arm in some individuals. Therefore, we propose to use the observed mean and variance of R_h, $\hat{\mu}_R$ and $\hat{\sigma}^2_R$, to replace μ_R and σ^2_R, and then test if the R_h value is different from $\hat{\mu}_R$. The final statistic for the RCNA test of site h is:

$$t_h = (R_h - \hat{\mu}_R)/\hat{\sigma}^2_R \quad (5)$$

here t_h is the standardized R_h value and follows Student's t distribution. Since the df of t_h is large in genome-wide analysis and the t distribution is very close to the SND, P-values can be obtained from SND. We choose right-tailed probability for this test, because theoretically there is no true negative correlation in $\mathbf{Z}$ and the expected t_h value can only be ≥ 0.

We have investigated the distribution of R_h values and found that the standardization step is important for real data analysis. Direct testing based on unstandardized R_h values may lead to a substancial increase of false positives (Supplementary Fig. 1).

2.2 Datasets

Both simulated and real datasets were used in this study.

2.2.1 Simulated data Multiple datasets were simulated for statistical power analysis. Given a set of chromosomal sites and a CN value c, we simulated $\log_2$ intensity ratios based on a normal distribution $N(\mu_c, \sigma^2)$, where $\mu_c = \log_2(c/2)$ and $\sigma = 0.509$ (this is an empirical estimate from lung cancer data described later). The length of the CNA region for individual samples was simulated based on a Poisson distribution with mean $= 3t$, where t is the length of RCNA region (i.e. the overlapping region across individuals). To make the simulation more realistic, we randomly selected 5% of samples from each simulated set and added one non-recurrent CNA region to each sample. The lengths and positions of the non-recurrent regions were generated using the Poisson distribution and uniform distribution, respectively.

2.2.2 Lung cancer dataset DNA CN data (tumor/normal intensity ratios) for 371 lung cancer (adenocarcinoma) patients were used. This dataset was generated from the Affymetrix Human Mapping 250K STY SNP Array platform in the Tumor Sequencing Project (TSP) and is publicly available at http://www.broad.mit.edu/cancer/pub/tsp/. More details can be found elsewhere (Weir *et al*., 2007).

2.2.3 Brain cancer dataset DNA CN data (tumor/normal intensity ratios) for 213 brain cancer (glioblastoma) patients was also used. This dataset was generated from the Illumina Infinium 550K BeadChip platform in The Cancer Genome Atlas (TCGA) project and is publicly available at http://cancergenome.nih.gov. More details can be found elsewhere (The TCGA Research Network, 2008).

2.3 Methods for comparison

In order to evaluate the performance of CMDS, we compared it with a typical single-sample calling based RCNA testing approach, STAC (Diskin *et al*., 2006) and a HMM-based RCNA analysis method, pREC-A (Rueda *et al*., 2009). Since STAC requires input data in the form of a binary matrix that indicates whether the CN of each chromosomal site for each sample is normal or aberrant (i.e. CNA status), we first segmented the data for individual samples using two different methods, adaptive weights smoothing (AWS) (Hupe *et al*., 2004) and circular binary segmentation (CBS) (Olshen *et al*., 2004); then we defined different matrices of binary CNA status using three cutoffs (CN $>$ 2.25, 2.50 and 2.75) and performed cross-sample RCNA tests using STAC. According to the two different methods used in the first step for single-sample analysis, we refer to the STAC analyzes as two different methods, AWS-STAC and CBS-STAC. Statistical power and computational time were compared for CMDS, AWS-STAC, CBS-STAC and pREC-A using simulated data.

3 RESULTS

3.1 Statistical power of CMDS

The statistical power of CMDS depends on multiple factors, including the block size (b) chosen for DT, sample size (n), frequency of RCNA in population (f), total number of chromosomal sites (m) to be tested, length (t) and CN (c) of the RCNA region, etc. To understand how these factors affect the power, we investigated simulations under a variety of configurations.

Given a set of n, f, c, m, t and b, we simulated $\log_2$ intensity ratios and performed the CMDS analysis on the simulated data. For each configuration, false positive rate (FPR) and true positive rate (TPR) were counted at different P-value cutoffs and then averaged over 500 replications of simulation. The FPR versus TPR plots, also called receiver operating characteristic (ROC) curves, are presented in Figure 2.

Among the factors determining the power of CMDS, b is the only parameter that needs to be specified for the CMDS analysis. Our simulations show that the optimum of b for CMDS is around 20 (Fig. 2A). Suboptimal values for b will deteriorate the power, because smaller values will produce more noise and larger values will lose sensitivity in the estimation of RCNA score. It is clear from Figure 2B–F that the power of CMDS increases with factors n, f, m, t and the CN amplitude of RCNA region (i.e. the absolute value of c-2). For instance, for detection of a RCNA region with $c = 3$, $f = 0.1$, $t = 30$ and $m = 1000$, the power of CMDS (with $b = 20$ and α level$=0.01$) will increase from $\sim$50% to $>$90%, if the sample size n is quadrupled from 25 to 100.

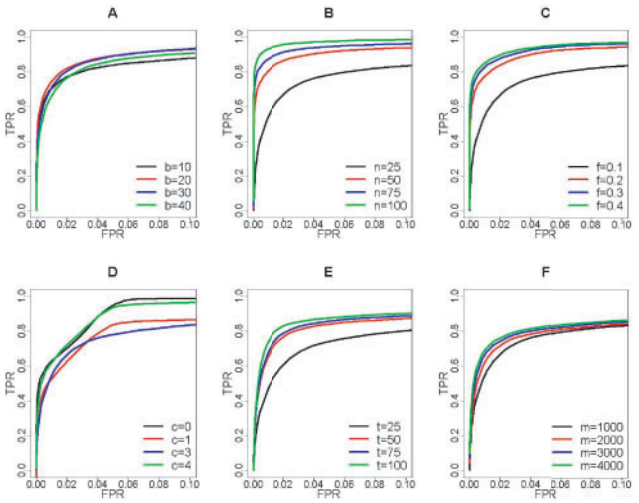

Fig. 2. ROC curves of CMDS under different configurations. Results are based on 500 replications of simulation. Since only the power at small type I error levels is of interest, FPR is presented only from 0 to 0.1. Simulation and analysis parameters are as follows: (**A**) $b = 5$–50, $n = 50$, $f = 0.1$, $c = 3$, $m = 1000$, $t = 10$–50 (randomly chosen from uniform distribution). (**B**) $b = 20$, $n = 25$–100, $f = 0.1$, $c = 3$, $m = 1000$, $t = 30$. (**C**) $b = 20$, $n = 50$, $f = 0.1$–0.5, $c = 3$, $m = 1000$, $t = 30$. (**D**) $b = 20$, $n = 50$, $f = 0.1$, $c = 0$–5, $m = 1000$, $t = 30$. (**E**) $b = 20$, $n = 50$, $f = 0.1$, $c = 3$, $m = 1000$, $t = 25$–100. (**F**) $b = 20$, $n = 50$, $f = 0.1$, $c = 3$, $m = 1000$–5000, $t = 30$.

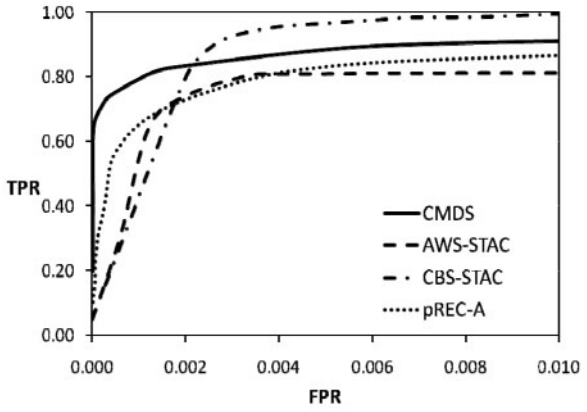

Fig. 3. Power comparison of CMDS, STAC and pREC-A. The result is based on 500 replications of simulation under a configuration of $m=1000$, $t=30$, $n=50$, $f=0.2$ and $c=3$ or 4 (3 and 4 are randomly chosen). $b=20$ is used for the CMDS analysis. A CN cutoff of 2.5 is used to define amplification status for individual sample data before STAC analysis. TPR and FPR are averaged over replications, and FPR is presented only from 0 to 0.01.

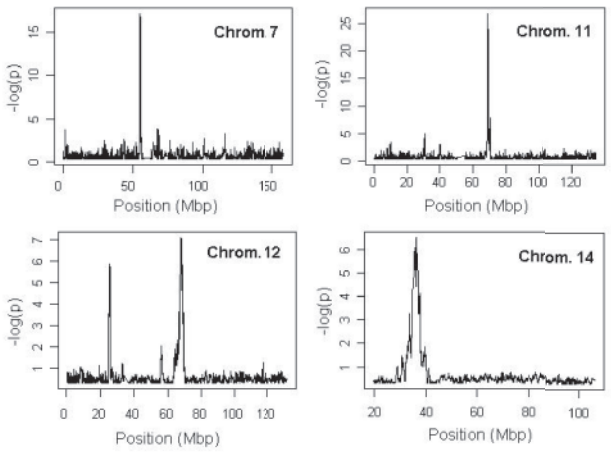

Fig. 4. Negative $\log_{10}(p)$-values from the CMDS analysis of the TSP lung cancer data, demonstrating five significant RCNA regions, 7p11.2, 11q13.3, 12p12.1, 12q15 and 14q13.3, where five known cancer genes, *EGFR*, *CCND1*, *KRAS*, *MDM2* and *TITF1*, are located.

3.2 CMDS versus other approaches

3.2.1 Power Through 500 replications of simulation under a configuration of $m=1000$, $t=30$, $n=50$, $f=0.1$ and $c=3$ or 4 (3 and 4 are randomly chosen for individual samples with CNAs), we compared the statistical power of CMDS against two single-sample calling based two-step approaches, AWS-STAC and CBS-STAC, and a HMM-based approach, pREC-A. Since different cutoffs (CN >2.25, 2.50 and 2.75) for both AWS and CBS produce minor differences of power in the final analysis of STAC, we only present here the STAC results of applying the cutoff CN > 2.5. The ROC curves (Fig. 3) indicate that CMDS is more powerful than AWS-STAC and pREC-A. Although CBS-STAC shows the highest power when FPR is relatively high, CMDS outperforms it when FPR is less than about 0.002. In the analysis of real high-resolution data, since a FPR level <0.002 is usually required due to the problem of multiple testing and the need to control for false discovery rate (FDR), CMDS will be more powerful in practice.

3.2.2 Execution time To compare the execution time of CMDS against AWS-STAC, CBS-STAC and pREC-A, we used three existing R packages, RJaCGH 2.0.0, GLAD 1.18 and DNAcopy 1.18 (downloaded from http://www.bioconductor.org) for the pREC-A, AWS and CBS analyzes, respectively, and a Java program STAC1.2 from http://cbil.upenn.edu/STAC for the STAC concurrence test. To analyze a dataset consisting of 10 000 chromosomal sites and 100 samples, our R version of CMDS takes only 13 s, which is about 0.02, 0.3 and 0.1% of the time used by pREC-A, AWS-STAC and ABS-STAC, respectively. For more details about this comparison of execution time, please refer to Supplementary Table 1.

It should be noted that the processing time for STAC is a rough estimation, because it is sensitive to the cutoff used in the AWS or CBS step (here we choose CN > 2.5. STAC would require more time with the use of a smaller cutoff). In addition, STAC is a permutation-based testing approach. We limited the number of permutations to 10 000 in this comparison. In practical analysis of genome-wide data, since *P*-values need to be estimated with higher precision for the purpose of multiple testing adjustment (e.g. Bonferroni correction or FDR calculation), the required permutation number could be larger and the total time of AWS-STAC or CBS-STAC analysis would increase substantially.

3.3 Applications

To evaluate the applicability of CMDS, we applied it to the TSP lung cancer data and the TCGA brain cancer data. With an option of b=30 and using a cutoff of FDR < 0.05, CMDS identifies 1406 significant sites (39 chromosomal regions) in lung cancer (Supplementary Table 2) and 3792 significant sites (37 chromosomal regions) in brain cancer (Supplementary Table 3). The most significant regions include 7p11.2, 11q13.3, 12p12.1, 12q15, 14q13.3 for lung adenocarcinoma (Fig. 4) and 1p33, 1p36, 4q12, 7p11.2, 12q14.1, 12q15 for glioblastoma (Fig. 5). These regions are important RCNA regions that have been previously reported and/or validated and encompass (or closely adjacent to) tumor-related genes, such as *EGFR*, *CCND1*, *KRAS*, *MDM2*, *TITF1*, *CDK4*, *PDGFRA*, *CDKN2C* (Weir *et al.*, 2007; The TCGA Research Network, 2008). These results indicate that our approach is applicable to real data.

Based on the CMDS analysis, we performed direct comparison of the significant CNAs in lung adenocarcinoma and glioblastoma. Our analysis revealed some common focal changes, such as the amplification of *EGFR* (7p12), *CDK4* (12q14) and *MDM2* (12q15), as well as the deletion of CDKN2A in both cancer types. This suggests that those molecules are key players in important pathways during the carcinogenesis of both tumor types. Furthermore, treatments targeting these molecules may be potentially useful for treating multiple cancer types. On the other hand, several focal events are found exclusively in either lung adenocarcinoma or glioblastoma. For example, *KRAS* (12p12) and *CCND1* (11q13) amplification occurs strictly in lung adenocarcinoma, indicating its specific relevance to smoking. In addition, *TITF1* (14q13) amplification is also specifically identified in lung adenocarcinoma and this is consistent with the high level expression of *TITF1* in lung and its developmental role in lung epithelial differentiation. However, *PDGFRA* (4q11–q13), amplified in glioblastoma, shows

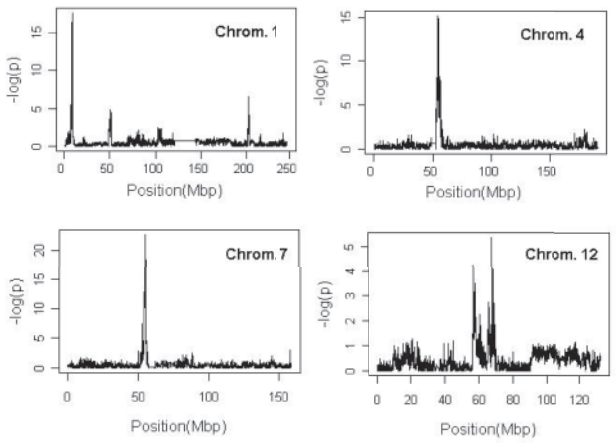

Fig. 5. Negative $\log_{10}(p)$-values from the CMDS analysis of the TCGA brain cancer data, demonstrating seven significant RCNA regions, 1p33, 1p36, 1q32.1, 4q12, 7p11.2, 12q14.1, 12q15, where six known cancer genes, *CDKN2C*, *MDM4*, *PDGFRA*, *EGFR*, *CDK4*, *MDM2*, are located.

no sign of CNA in lung adenocarcinoma. With the amount of genomic data for different cancer types increasing rapidly, CMDS will provide a valuable and very efficient tool for this type of cross-comparison.

The analyzes above were performed chromosome by chromosome using our R version of CMDS on a single PC (Dell Optiplex 755 PC, with 3 GHz CPU and 4 GB RAM, under Windows XP and R 2.9.1), which took about 17 and 36 min for analyzing TSP data and TCGA data, respectively, including data file reading, calculations and output of result. Our C version of CMDS on a single LINUX machine took only ∼3 and 5 min, respectively. It is also possible to run CMDS on a per-chromosome basis in a parallelized fashion; doing so reduced execution time to a few minutes for the R version and less than a minute for the C version.

4 DISCUSSION

Developing CNA analysis methods that are both statistically powerful and computationally efficient is a necessity due to the constant increase in experiment resolution and sample size. We have developed the CMDS approach for identifying RCNAs using genome-wide and population-based data, and demonstrated its power, efficiency and applicability in processing both simulated and real data. In contrast to most existing methods, CMDS directly uses raw CN data and does not require pre-analysis (smoothing and segmentation etc.) of individual samples. To avoid calculating a huge matrix, CMDS adopts a quick DT method to construct a RCNA score, the significance of which can be quickly tested based on a normal distribution. These features make CMDS a more efficient and powerful approach. It provides a fast and easily implementable tool for RCNA analysis and is especially well-suited for analysis of emerging genome-wide studies with larger sample size and higher data resolution.

CMDS reports significant RCNA regions at a population level; however, there is sometimes a need to obtain CNA calls from individual samples for downstream analyzes such as a correlation test between CNA and clinical features. In such a case, single-sample calls (SSC) can be obtained for the significant RCNA regions identified by CMDS, using a threshold or statistical test (e.g. *t*-test or permutation test on CN means of candidate regions). Thus, the information from multiple samples can be used for SSC, and computational time can be substantially reduced by excluding regions with no significant RCNA. Of course, similar to Guttman's conserved genomic aberration detection through the MSA method (Guttman *et al*., 2007), this multiple sample-based strategy for SSC may lose power for identifying non-recurrent or broad CNA regions.

An important feature of CMDS is that correlations between chromosomal positions are used to measure the CNA concordance in a population. In a recent study (McCarroll *et al*., 2008), a similar correlation-based approach was used to investigate common copy number variants (CNVs) in the HapMap human populations. This suggests that although CMDS was originally developed for CNA analysis of cancer populations, it may be applicable to data from normal populations. Our preliminary studies indicate that CMDS can be applied not only to array data but also to whole-genome sequencing data with extremely high resolution, to quickly identify common CNV regions in normal populations (Supplementary Fig. 2). Of course, since CNVs in normal populations differ from CNAs in cancer populations in terms of frequency, amplitude, length, etc., further investigation on the statistical property of CMDS for CNV analysis is warranted.

ACKNOWLEDGEMENTS

We wish to thank the TSP and TCGA research networks for providing the real datasets, and thank John Osborne for downloading and organizing the data. We also thank Dr Feng Gao and Ms Ling Lin for careful testing of the program and constructive feedback.

Funding: National Human Genome Research Institute grants (to R.K.W.).

Conflict of Interest: None declared.

REFERENCES

Beroukhim,R. *et al*. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA*, **104**, 20007–20012.

Diskin,S.J. *et al*. (2006) STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res*., **16**, 1149–1158.

Diskin,S.J. *et al*. (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.*, **36**, e126.

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

Hsu,L. *et al*. (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.

Hupe,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Jong,K. *et al*. (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20**, 3636–3637.

Lai,W.R. *et al*. (2005a) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Lai,Y. *et al*. (2005b) A statistical method to detect chromosomal regions with DNA copy number alterations using SNP-array-based CGH data. *Comput. Biol. Chem.*, **29**, 47–54.

Lipson,D. *et al*. (2006) Efficient calculation of interval scores for DNA copy number data analysis. *J. Comput. Biol.*, **13**, 215–228.

Marioni,J.C. *et al*. (2007) Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol*., **8**, R228.

McCarroll,S.A. *et al*. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.

Nilsson,B. *et al*. (2009) Ultrasome: efficient aberration caller for copy number studies of ultra-high resolution. *Bioinformatics*, **25**, 1078–1079.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Picard,F. *et al*. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Rouveirol,C. *et al*. (2006) Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 849–856.

Rueda,O.M. *et al*. (2009) Detection of recurrent copy number alterations in the genome: taking among-subject heterogeneity seriously. *BMC Bioinformatics*, **10**, 308.

Shah,S.P. *et al*. (2006) Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, **22**, e431–e439.

Shah,S.P. *et al*. (2007) Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics*, **23**, i450–i458.

The TCGA Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Weir,B.A. *et al*. (2007) Characterizing the cancer genome in lung adenocarcinoma, *Nature*, **450**, 893–898.