CNAmet: an R package for integrating copy number, methylation and expression data

Riku Louhimo and Sampsa Hautaniemi*
Computational Systems Biology Laboratory, Institute of Biomedicine and Genome-Scale Biology Research Program, University of Helsinki, Finland
Associate Editor: Martin Bishop

ABSTRACT
Summary: Gene copy number and DNA methylation alterations are key regulators of gene expression in cancer. Accordingly, genes that show simultaneous methylation, copy number and expression alterations are likely to have a key role in tumor progression. We have implemented a novel software package (CNAmet) for integrative analysis of high-throughput copy number, DNA methylation and gene expression data. To demonstrate the utility of CNAmet, we use copy number, DNA methylation and gene expression data from 50 glioblastoma multiforme and 188 ovarian cancer primary tumor samples. Our results reveal a synergistic effect of DNA methylation and copy number alterations on gene expression for several known oncogenes as well as novel candidate oncogenes.

Availability: CNAmet R-package and user guide are freely available under GNU General Public License at http://csbi.tkk.helsinki.fi/CNAmet.

1 INTRODUCTION
Genomic instability is a hallmark of cancer and characterization of copy number changes in tumors has lead to identification of a number of genes that contribute to tumor progression and drug response (Kan et al., 2010). In addition to copy number alterations, DNA methylation of CpG islands regulates gene expression patterns in cancers (Esteller, 2008; Kim et al., 2010). Copy number aberrations, methylation patterns and gene expression profiles can be measured genome-scale with microarrays, which enables integration of these data and further identification of genes that are crucial to cancer progression.

When integrating copy number and expression data, the major goal is to identify genes that are both amplified and upregulated or deleted and downregulated (Pinkel and Albertson, 2005). Gene upregulation can also be due to hypomethylation (decrease in methylation of cytosine and adenosine residues in DNA) and downregulation due to hypermethylation. As both copy number and methylation changes may affect gene regulation, integration of these data should result in improved characterization of genes essential in cancer progression (Sadikovic et al., 2008; Stransky et al., 2006).

We introduce an R package (CNAmet) that integrates high-throughput copy number, methylation and expression data. Our primary goal is to identify genes that are amplified, hypermethylated and upregulated, or deleted, hypermethylated and downregulated, though all combinations between copy number, methylation and expression levels are calculated. To our knowledge CNAmet is the first software package for copy number, methylation and expression integration. To demonstrate the utility of CNAmet we analyze copy number, methylation and expression data from 50 patients with glioblastoma multiforme (GBM), which is the most aggressive type of brain cancer, as well as 188 ovarian cancer (OV) patients (The Cancer Genome Atlas Research Network, 2008).

2 METHODS
The CNAmet algorithm consists of three major steps. In the weight calculation step, the signal-to-noise ratio statistic (Hautaniemi et al., 2004) is used to link expression values to copy number and methylation aberrations. In the score calculation step, the weight values are combined to a score that indicates genes whose expression alterations are due to changes in DNA methylation and copy number levels. In the significance evaluation step, corrected P-values of the scores are calculated with a permutation test.

Let \( n \) denote the number of genes and \( n \) the number of samples. In general, \( n \) can vary between the datasets but is subsequently assumed the same for notational convenience. Inputs to CNAmet are labeling matrices for copy number (\( cn \)) and methylation (\( me \)) data \( M_{n \times n}, M_{n \times n} \in \{0, 1\}^{n \times n} \). For example, when searching for genes whose upregulation is likely due to hypomethylation and high copy number status, ‘1’ denotes amplification and ‘0’ lack of amplification in \( M_{cn} \). Similarly, ‘1’ denotes hypomethylation and ‘0’ lack of hypomethylation in \( M_{me} \).

In order to calculate weights for the \( i \)-th gene we first take the \( i \)-th row in \( M_{cn} \). Let \( m_{cn,i}^{0} \) and \( \sigma_{cn,0} \) be the mean and SD of the expression values of samples labeled with ‘1’ for the \( i \)-th gene in \( M_{cn} \), and \( m_{cn,i}^{1} \) and \( \sigma_{cn,1} \) are calculated with samples labeled with ‘0’. The values \( m_{me,i}^{0} \) and \( \sigma_{me,0} \) are calculated similarly from \( M_{me} \) for methylation data. Now, for the \( i \)-th gene we calculate the weight for methylation and expression data as

\[
W_{me,i} = \frac{m_{me,i}^{1} - m_{me,i}^{0}}{\sigma_{me,1} + \sigma_{me,0}} \quad \text{if } m_{me,i}^{1} > 0, \sigma_{me,0} > 0.
\]

Equation (1) is used similarly to calculate the weight \( W_{cn,i} \) for copy number data. By default, the weights are calculated for genes that have ‘1’ in at least two samples in both copy number and methylation data. Weights with a negative sign are denoted as NA. Events where all samples are labeled with ‘1’ in either methylation or copy number data are listed separately.

In order to combine the weight values we define \( T \) to be the total number of samples and \( U \) the number of samples in the intersection of samples with ‘1’ in \( M_{cn} \) and \( M_{me} \) for the \( i \)-th gene. The optional correction term \( s_{i} \) of the CNAmet score for the \( i \)-th gene is calculated as \( s_{i} = \frac{T}{U} \). The correction term forces the CNAmet score to favor genes that have abundant methylation...
Our analysis of the GBM data using CNAmet resulted in four genes better (Supplementary File 2).

Fig. 1. Expression differences in patients with different EGF R methylation and copy number statuses. Black bars denote medians and filled rectangles contain values between 25th and 75th percentile. Patients with increased hypomethylation (met = 1) and amplification (cna = 1) display higher EGF R expression levels than patients with only an amplification (P < 3.8 × 10⁻⁸).

3 RESULTS

Our analysis of the GBM data using CNAmet resulted in four lists of genes (Supplementary File 1). In the hypomethylation and amplification analysis the top scoring six genes (P < 0.05) included MDM2, EGF R and PDGFRA that are well-known oncogenes. We also compared CNAmet to ANOVA and our results indicate that CNAmet identifies more oncogenes and prioritizes the resulting genes better (Supplementary File 2).

Based on these results, we hypothesized that the effect of methylation and copy number expression for these genes is synergistic. To test this hypothesis we grouped the samples based on their methylation and amplification status gene by gene (Supplementary File 2) of which EGF R is shown in Figure 1. Samples with hypomethylated and non-amplified EGF R show almost normal expression, while samples with EGF R amplification result in upregulated EGF R expression when compared to solely amplified samples (t-test, P < 3.8 × 10⁻⁸). This demonstrates a synergistic function of methylation and copy number changes in EGF R expression in GBM. The results are similar for MDM2 and PDGFRA (Supplementary File 2). The ability of CNAmet to detect such synergetic effects is also demonstrated in an analysis of OV data (Supplementary Files 2 and 3).

Amplification and overexpression of EGF R are controversial prognostic factors in GBM (Phillips et al., 2006). We compared the age-independent survival of GBM patients with hypomethylated and amplified EGF R to patients with only hypomethylated EGF R. Patients with hypomethylated EGF R had marginally better prognosis than the patients with hypomethylated and amplified EGF R (logrank test P < 0.06; Supplementary File 2). This survival effect would have been undetected when using copy number data only (Supplementary File 2), which illustrates the benefits of integrating methylation and copy number data.

4 CONCLUSION

We have designed and implemented a novel and versatile method, CNAmet, to facilitate the integration of copy number, methylation and expression data. We applied CNAmet to GBM and ovarian cancer data and our results demonstrate the added value of integrating these three data sources.

ACKNOWLEDGEMENTS

We are grateful to Tiia Pelkonen for proof-reading.

Funding: Academy of Finland (project 125826); Sigrid Juselius Foundation, Finnish Cancer Associations and Biocentrum Helsinki.

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. 57, 289–300.

Esteller,M. (2008) Epigenetics in Cancer. N Engl. J. Med. 358, 1148–1159.

Hautaniemi,S. et al. (2004) A strategy for identifying putative causes of gene expression variation in human cancers. J. Franklin Inst. 341, 77–88.

Kim,Z. et al. (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466, 695–733.

Kim,M. et al. (2010) DNA methylation markers in colorectal cancer. Cancer Metastasis Rev. 29, 181–206.

Ovaska,K. et al. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. Genome Med. 2, 56.

Phillips,H.S. et al. (2006) Molecular subclones of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 9, 157–173.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. Nat. Genet. 37, S11–S17.

Sadikovic,B. et al. (2008) In vitro analysis of integrated global high-resolution DNA methylation profiling with genomic imbalance and gene expression in osteosarcoma. PLoS ONE 3, e2834.

Stranský,N. et al. (2006) Regional copy number-independent deregulation of transcription in cancer. Nat. Genet. 38, 1386–1396.

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