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 \( m \) 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 \( M, \; M_{\text{me}} \subset \{0, 1\}^{m \times n} \). 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 \( M_{\text{cn}} \). Similarly, ’1’ denotes hypomethylation and ’0’ lack of hypomethylation in \( M_{\text{me}} \).

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

\[
W_{\text{me}} = \frac{m_{\text{me}i}^{\text{up}} - m_{\text{me}i}^{\text{down}}}{\sigma_{\text{me}i}^{\text{up}} - \sigma_{\text{me}i}^{\text{down}}} > 0, \; \sigma_{\text{me}i}^{\text{up}} > 0.
\]  

Equation (1) is used similarly to calculate the weight \( W_{\text{cn}} \) 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 \( m_{\text{me}} \times n \)-dimensional score vectors. For copy number data, the correction term forces the CNAmet score to favor genes that have aberrant methylation patterns in combination with copy number alterations.
There were 50 samples with survival data that overlapped between all three and Hochberg, 1995). The expression levels than patients with only an amplification (Benjamini and Hochberg, 1995). The statistical significance for $W_{mc}$, $W_{me}$ and $S$ is calculated by randomly permuting the labeling vectors and recalculating $W_{mc}$, $W_{me}$ and $S$. The $P$-values for $S$ are corrected with the false discovery rate method (Benjamini and Hochberg, 1995). The $H_0$ here is “large score is due to random event”.

CNAmet is available as an R package and as a component in the Anduril framework (Ovaska et al., 2010). The Anduril compliance enables modular analysis via a number of preprocessing methods, such as normalization, copy number segmentation or calling algorithms. This allows the use of different experimental designs and array platforms.

In order to assess the utility of CNAmet we applied it to The Cancer Genome Atlas GBM and OV datasets. GBM dataset consists of 181 patient samples with expression and copy number data, and 64 samples with methylation data (The Cancer Genome Atlas Research Network, 2008). There were 50 samples with survival data that overlapped between all three array types. The OV dataset consists of 188 ovarian cancer samples having data from all three array types.

3 RESULTS

Our analysis of the GBM data using CNAmet resulted in four lists of genes (Supplementary File 1). In the methylation and amplification analysis the top scoring six genes ($P < 0.05$) included MDM2, EGFR 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 on 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 EGFR is shown in Figure 1. Samples with hypomethylated and non-amplified EGFR show almost normal expression, while samples with EGFR amplification result in upregulated EGFR. Strikingly, samples with hypomethylated and amplified EGFR exhibit a substantial upregulation of EGFR expression when compared to solely amplified samples ($t$-test, $P < 3.8 \times 10^{-8}$).

Amplification and overexpression of EGFR are controversial prognostic factors in GBM (Phillips et al., 2006). We compared the age-independent survival of GBM patients with hypomethylated and amplified EGFR to patients with only hypomethylated EGFR. Patients with hypomethylated EGFR had marginally better prognosis than the patients with hypomethylated and amplified EGFR (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


