CNAmet v.1.2 User Guide

Riku Louhimo

April 16, 2013

1 Introduction

Copy number arrays (array CGH or SNP arrays) and expression microarrays form an essential part of integrative analysis of cancer where the amount of genomic DNA can be linked to the amount of transcription [7]. However, aberrant DNA methylation also affects tumors in ways undetectable by an integrative analysis of copy number aberrations and expression alone [3]. We have designed and implemented the CNAmet algorithm and software package that facilitates the integrative analysis of high-throughput copy number, expression and methylation data. The package is written and distributed in the R statistical language [8].

This user guide contains information on the installation, inputs and outputs of CNAmet as well as a detailed description of the algorithm. We also describe possible ways to analyze the data and how to interpret the results.

2 Getting started

The CNAmet R package can be downloaded from `http://csbi.ltdk.helsinki.fi/CNAmet`.

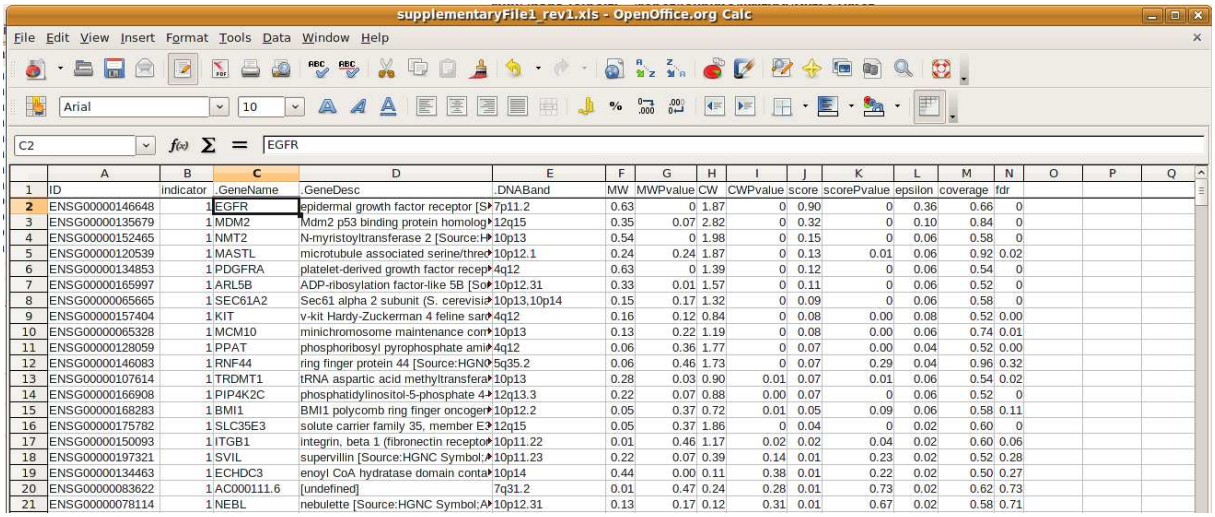
Installation

CNAmet is available in Unix, Windows and Anduril versions.

Unix: The software package is available for 32 and 64bit Unix systems. Download the CNAmet package from the project website. Then, by using the terminal program, go to the directory where the CNAmet package has been downloaded to. Typing `sudo R CMD INSTALL [name of CNAmet package]` will install the CNAmet R package to your environment. Please note that depending on your system settings and configuration you might not need to run the installation as super-user. In this case, remove the 'sudo' keyword from the above installation command.

Windows: Version 1.1 of CNAmet is available for 32bit Windows. If your using the R Windows graphical user interface, select `Rgui > Install pacakges > Install package(s) from local zip-files`. Otherwise, follow the instructions of the R manual Section 6.3.1 'Installing packages - Windows' (`http://cran.r-project.org/doc/manuals/r-release/R-admin.html#Windows-packages`).

Anduril: The CNAmet algorithm has also been implemented as a component in the Anduril component framework [6]. The installation of the Anduril component framework and the CNAmet component is detailed in the Anduril user guide `http://csbi.ltdk.helsinki.fi/anduril`. Anduril is available for Unix and Windows systems. This step is unnecessary if the user is only interested in using the CNAmet standalone R package.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	ID	indicator	.GeneName	.GeneDesc	.DNABand	MW	MWPvalue	CW	CWPvalue	score	scorePvalue	epsilon	coverage	fdr			
2	ENSG00000146648	1	EGFR	epidermal growth factor receptor [S	7p11.2	0.63	0	1.87	0	0.90	0	0.36	0.66	0			
3	ENSG00000135679	1	MDM2	Mdm2 p53 binding protein homolog	12q15	0.35	0.07	2.82	0	0.32	0	0.10	0.84	0			
4	ENSG00000152465	1	NMT2	N-myristoyltransferase 2 [Source:H	10p13	0.54	0	1.98	0	0.15	0	0.06	0.58	0			
5	ENSG00000120539	1	MASTL	microtubule associated serine/thre	10p12.1	0.24	0.24	1.87	0	0.13	0.01	0.06	0.92	0.02			
6	ENSG00000134853	1	PDGFRA	platelet-derived growth factor recep	4q12	0.63	0	1.39	0	0.12	0	0.06	0.54	0			
7	ENSG00000165997	1	ARL5B	ADP-ribosylation factor-like 5B [So	10p12.31	0.33	0.01	1.57	0	0.11	0	0.06	0.52	0			
8	ENSG00000065665	1	SEC61A2	Sec61 alpha 2 subunit (S. cerevisia	10p13,10p14	0.15	0.17	1.32	0	0.09	0	0.06	0.58	0			
9	ENSG00000157404	1	KIT	v-kit Hardy-Zuckerman 4 feline sar	4q12	0.16	0.12	0.84	0	0.08	0.00	0.08	0.52	0.00			
10	ENSG00000065328	1	MCM10	minichromosome maintenance com	10p13	0.13	0.22	1.19	0	0.08	0.00	0.06	0.74	0.01			
11	ENSG00000128059	1	PPAT	phosphoribosyl pyrophosphate ami	4q12	0.06	0.36	1.77	0	0.07	0.00	0.04	0.52	0.00			
12	ENSG00000146083	1	RNF44	ring finger protein 44 [Source:HGN	5q35.2	0.06	0.46	1.73	0	0.07	0.29	0.04	0.96	0.32			
13	ENSG00000107614	1	TRDMT1	tRNA aspartic acid methyltransfera	10p13	0.28	0.03	0.90	0.01	0.07	0.01	0.06	0.54	0.02			
14	ENSG00000166908	1	PIP4K2C	phosphatidylinositol-5-phosphate 4-	12q13.3	0.22	0.07	0.88	0.00	0.07	0	0.06	0.52	0			
15	ENSG00000168283	1	BMI1	BMI1 polycomb ring finger oncoger	10p12.2	0.05	0.37	0.72	0.01	0.05	0.09	0.06	0.58	0.11			
16	ENSG00000175782	1	SLC35E3	solute carrier family 35, member E3	12q15	0.05	0.37	1.86	0	0.04	0	0.02	0.60	0			
17	ENSG00000150093	1	ITGB1	integrin, beta 1 (fibronectin recepto	10p11.22	0.01	0.46	1.17	0.02	0.02	0.04	0.02	0.60	0.06			
18	ENSG00000197321	1	SVIL	supervillin [Source:HGNC Symbol;A	10p11.23	0.22	0.07	0.39	0.14	0.01	0.23	0.02	0.52	0.28			
19	ENSG00000134463	1	ECHDC3	enoyl CoA hydratase domain conta	10p14	0.44	0.00	0.11	0.38	0.01	0.22	0.02	0.50	0.27			
20	ENSG00000083622	1	AC000111.6	[undefined]	7q31.2	0.01	0.47	0.24	0.28	0.01	0.73	0.02	0.62	0.73			
21	ENSG00000078114	1	NEBL	nebulette [Source:HGNC Symbol;A	10p12.31	0.13	0.17	0.12	0.31	0.01	0.67	0.02	0.58	0.71			

Figure 1: Output file of CNAmet with some additional annotations opened in a spreadsheet (Excel-like) program.

Usage

In your R prompt, type `library(CNAmet)` to load the library. The help files for CNAmet are available in the R environment by typing `library(help=CNAmet)` and `?CNAmet`. The R package includes usage examples with example input files. In addition to CNAmet, the software package includes the S2N algorithm for copy number and expression integration [4]. The S2N algorithm can be executed by inputting one of the labeling matrices only. For additional help in installing R packages, please refer to the installation instructions ('Section 6.3 Installing packages') of the R software project manual located at `http://www.r-project.org/`.

Inputs

Inputs to CNAmet are three $m \times n$ matrices, where m is the number of genes and n the number samples. Although the matrix dimensions must be equal, the actual sample sets to be compared need not overlap perfectly. The CNAmet algorithm skips values in the label matrices that are not 0 or 1, and thus defining other values for samples for which, say, methylation status is unknown enables the inclusion of these samples for the copy number weight calculation, and vice versa. However, the number of samples for which expression has been measured should be considered as a logical upper limit for n. Since the three microarray platforms contain non-overlapping probes, the m dimension of the input matrices must match. This is because the problem of mapping measurements (probe to probe mapping) between different array types is not dealt with by CNAmet.

Although CNAmet is meant for the three-way intergration of expression, methylation and copy number data, CNAmet can also be simply used to 1) integrate expression with methylation data or 2) integrate expression with copy number data. In this case, the missing third input matrix is set NULL and CNAmet output only contains the singular integration results.

Outputs

Outputs of the CNAmet R package and algorithm include two weights (one for copy number (column CW), one for methylation (MW)), a score for their combined effect (score), and adjacent p-values to the

weights and score. The p-values are computed by randomly permuting the labels and re-computing the weights and score. Moreover, the score p-value is multiple hypothesis corrected by the false discovery rate estimation. The correction term (epsilon) and the percentage of aberration coverage of MA and CNA samples (coverage) are given as well. An example output file is shown in Figure 1.

Computational time requirements

CNAmet employs a permutation test to assess the statistical significance of results. The time requirements of the procedure increase with the number of rows in data (e.g., genes) and number of tests performed but not with the number of samples. CNAmet has been tested with a data set of 1,700 probes, 50 samples and 1,000 permutations. Using a standard off-the-shelf laptop execution took less than 10 minutes. Another analysis using a different data set with 8,000 probes, 188 samples and 1,000 permutations took 40 minutes.

3 Algorithm overview

Given an expression matrix and two label matrices ($\mathbf{M}_{me}$ for methylation data and $\mathbf{M}_{cn}$ for CNA data), we can define an algorithm.

1. Divide methylation data to normal and hypomethylated, or normal and hypermethylated samples (indicated by 0 and 1 in the label matrix).
2. Divide CNA data to normal and gain, or normal and loss samples similarly to methylation data.
3. For each gene, calculate the signal-to-noise ratio with both methylation (W^i_{me}) and CNA labels (W^i_{cn}).
4. Sum the two statistics together, apply the correction term and output the result score.
5. Randomly permute labels, recalculate all scores, and compute p-values for each score.

The CNAmet algorithm itself consists of three major steps (steps 3-5 above). In the *weight calculation step* the signal-to-noise ratio statistic is used to link expression values to copy number and methylation aberrations [4]. For each gene these analyses result in two weight values denoting the independent association of expression changes to copy number or methylation data. In the *score calculation step* the weight values are combined to a score indicating 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. Outputs of CNAmet include weights for methylation induced expression, copy number induced expression, a score for their combined effect and the adjacent multiple hypothesis corrected p-values.

We next describe the algorithm in more formal terms. A general overview of the different notations used here and their relationships is shown in Image 1. Let m denote the number of genes and n the number of samples. For notational convenience let m (and n) be the same across copy number, methylation and expression data. Inputs to CNAmet are labeling matrices for copy number ('cn' subscript) and methylation ('me' subscript) data $\mathbf{M}_{cn}, \mathbf{M}_{me} \in \{0,1\}^{m \times n}$. 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 $\mathbf{M}_{cn}$, and, similarly, '1' denotes hypomethylation and '0' lack of hypomethylation in $\mathbf{M}_{me}$.

In order to calculate weights for the ith gene we first take the ith row in $\mathbf{M}_{cn}$. Let $m^i_{cn,1}$ and $\sigma^i_{cn,1}$ be the mean and standard deviation of the expression values of samples with '1' for the ith gene in $\mathbf{M}_{cn}$, and $m^i_{cn,0}$ and $\sigma^i_{cn,0}$ be the mean and standard deviation of the expression values of samples having '0'.

The values $m^i_{me,1}$ and $\sigma^i_{me,1}$ are calculated similarly from $\mathbf{M}_{me}$ to methylation data. Now for the ith gene we calculate the weights for methylation and expression data as

$$W^i_{me} = \frac{m^i_{me,1} - m^i_{me,0}}{\sigma^i_{me,1} + \sigma^i_{me,0}}, \quad \sigma^i_{me,1} > 0, \quad \sigma^i_{me,0} > 0. \tag{1}$$

The Eq. 1 is used similarly to calculate the weight W^i_{cn} for copy number data. By default, the weights are calculated for genes that have '1' in at least two samples at both copy number and methylation data. Events where all samples are labeled with '1' in either methylation or copy number data are dealt with separately. In order to combine the weight values we define T to be the total number of samples and U^i the number of samples in the intersection of samples with '1' in $\mathbf{M}_{cn}$ and $\mathbf{M}_{me}$ for the ith gene.

We define the correction term ε_i for Gene i as

$$\varepsilon_i = \frac{U^i}{T}.$$

The correction term ensures that genes with a high number of aberration in both MA and CNA samples (with maximum possible overlap between the two) score high. For instance, consider a sample set of 50 samples. Let gene G_1 have MA and CNA in 20 samples that overlap perfectly. Let gene G_2 have MA and CNA in 20 samples that overlap in 10 samples. Now, G_1 is less affected by the correction term than G_2 ($\varepsilon_{G_1} = 0.4, \varepsilon_{G_2} = 0.2$). This minimizes the impact of genes like G_2 on the result set since they are less informative about the joint effects of MA and CNA in the sample set than genes like G_1.

Now that we have defined the weights and correction term, we calculate the CNAmet score for gene i as

$$S^i = (W^i_{me} + W^i_{cn})\varepsilon_i, \quad W^i_{me} > 0, \quad W^i_{cn} > 0. \tag{2}$$

Outputs of the algorithm include S^i, ε_i, W^i_{me} and W^i_{cn} and the adjacent p-values.

The CNAmet R package includes the signal-to-noise ratio algorithm of Hautaniemi *et al.* which is utilized for the expression to copy number array integration [4]. This is calculated by simply computing the W^i_{cn} term without the error correction. The CNAmet algorithm is also incorporated as a component in the Anduril bioinformatics framework [6].

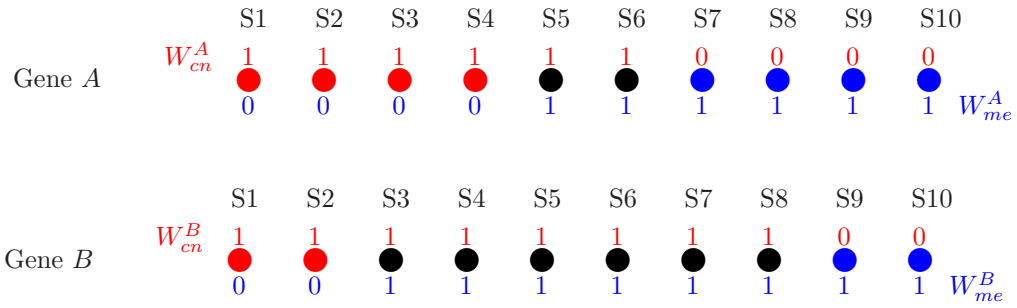


Image 1: Blue circles are samples with MA only. Red circles are samples with CNA only. Black circles have both MA and CNA. W^i_{cn} is calculated by comparing the expression of the samples with a red '1' to the other samples (grouping showed with red 0 and 1). W^i_{me} is calculated by comparing the expression of the samples with a blue '1' to the other samples (grouping showed with blue 0 and 1). U^i is the total number of black samples e.g., here $\varepsilon_A = \frac{2}{10}$ and $\varepsilon_B = \frac{6}{10}$.

4 Possible analysis options

Methylation aberrations (MA) and copy number aberrations (CNA) can be analyzed in multiple ways. The focus in our algorithm is in the integration step and thus we let users define their own method to group the samples.

Input preprocessing options

For CNA analysis, we recommend using a segmentation algorithm such as DNAcopy [5] and calling the CNA either by thresholding (*i.e.,* considering values 2 standard deviations from the median as significant) or using probabilistic or deterministic calling algorithms [1, 10, 2]. Copy number data can originate from a CGH microarray or from a SNP array.

For MA analysis, we recommend using methylation detection algorithms when applicable. Beta values from an Illumina methylation array can also be used to divide the samples by simply splitting them by the mean, median or other value, whichever is pertinent. You can also use more stringent limits, such as quantiles, deciles or similar. Because beta value distributions are skewed towards their extremes (0 and 1), we have also set thresholds based on empirical analysis of the beta value distributions. For example, we often start with a threshold of less than 0.2 for hypomethylation and more than 0.8 for hypermethylation, and then optimize this parameter based on the given beta distribution.

CNAmet analysis

The ability of CNAmet to detect synergetic genes is enhanced by the favorSynergetic parameter. Setting this true, forces CNAmet to compute the correction term epsilon and the W^i_{me} and W^i_{cn} to only use common '0' genes when computing the signal-to-noise ratio. This favors genes whose CNA and MA overlap as much as possible and their normal status samples overlap similarly. However, disabling this parameter might work better for small data sets.

Several different types of combinations of MA and CNA effects on expression can be analyzed with CNAmet. For example

1. Increased hypomethylation with amplification upregulate expression
2. Decreased hypomethylation with amplification upregulate expression
3. Increased hypermethylation with deletion downregulate expression
4. Decreased hypermethylation with deletion downregulate expression

Cases 1 and 3 are straightforward and indicate situations where tumors up- or downregulate a gene either with MA or CNA, or synergistically with both of them. A gene scoring high in an analysis like this would be a central tumor suppressor or oncogene whose deregulation is essential for the cancer. Moreover, the deregulation should happen in a large proportion of the samples, which will also be reflected by a high percentage of samples with either CNA or MA. Considering inputs of CNAmet in case 1, samples with increased hypomethylation (*i.e.,* decreased methylation) are labeled '1' in $\mathbf{M}_{me}$ as are amplified samples in $\mathbf{M}_{cn}$.

Genes scoring high in cases 2 and 4 are more complicated to interpret. For instance, if a gene scores high in case 2 it is most likely unchanged or downregulated in MA samples. However, the disjoint subset, containing in this case the CNA and hypomethylated samples, shows highly concordant expression upregulation. One example of this would be an oncogene that is silenced by methylation in non-aberrant samples, but activated by amplification in CNA samples. Considering inputs of CNAmet in case 2, samples with decreased hypomethylation (*i.e.,* increased methylation) are labeled '1' in $\mathbf{M}_{me}$ as are amplified samples in $\mathbf{M}_{cn}$.

Although amplified and upregulated genes frequently show hypomethylation (and the same applies to tumor suppressor genes), all the four different combinations of MA and CNA have been observed [9]. Interestingly, juxtaposition of hypermethylation and amplification is complicated by the fact that this phenomenon occurs fairly rarely in our experience, which is why our analysis is more concentrated on the relative differences caused by varying amounts of MA and CNA in cancer samples synergistically.

Example: Glioblastoma data analysis

We used CNAmet to analyze cancer data (glioblastoma multiforme) retrieved from the Cancer Genome Atlas. When analyzing hypomethylated and amplified genes, the top scoring genes ($p < 0.05$) included three well-known oncogenes. The genewise expression of the genes when grouped by their methylation and amplification status indicates a synergistic effect (shown in Figure 2). For instance, samples with hypomethylated and amplified *MDM2* show a substantial upregulation when compared to samples with only either aberration or no aberrations. The difference is also statistically significant (t-test $p < 9.62 \times 10^{-}5$). Shown in Figure 3, genes deemed neutral by CNAmet from the same analysis did not reveal a similar pattern.

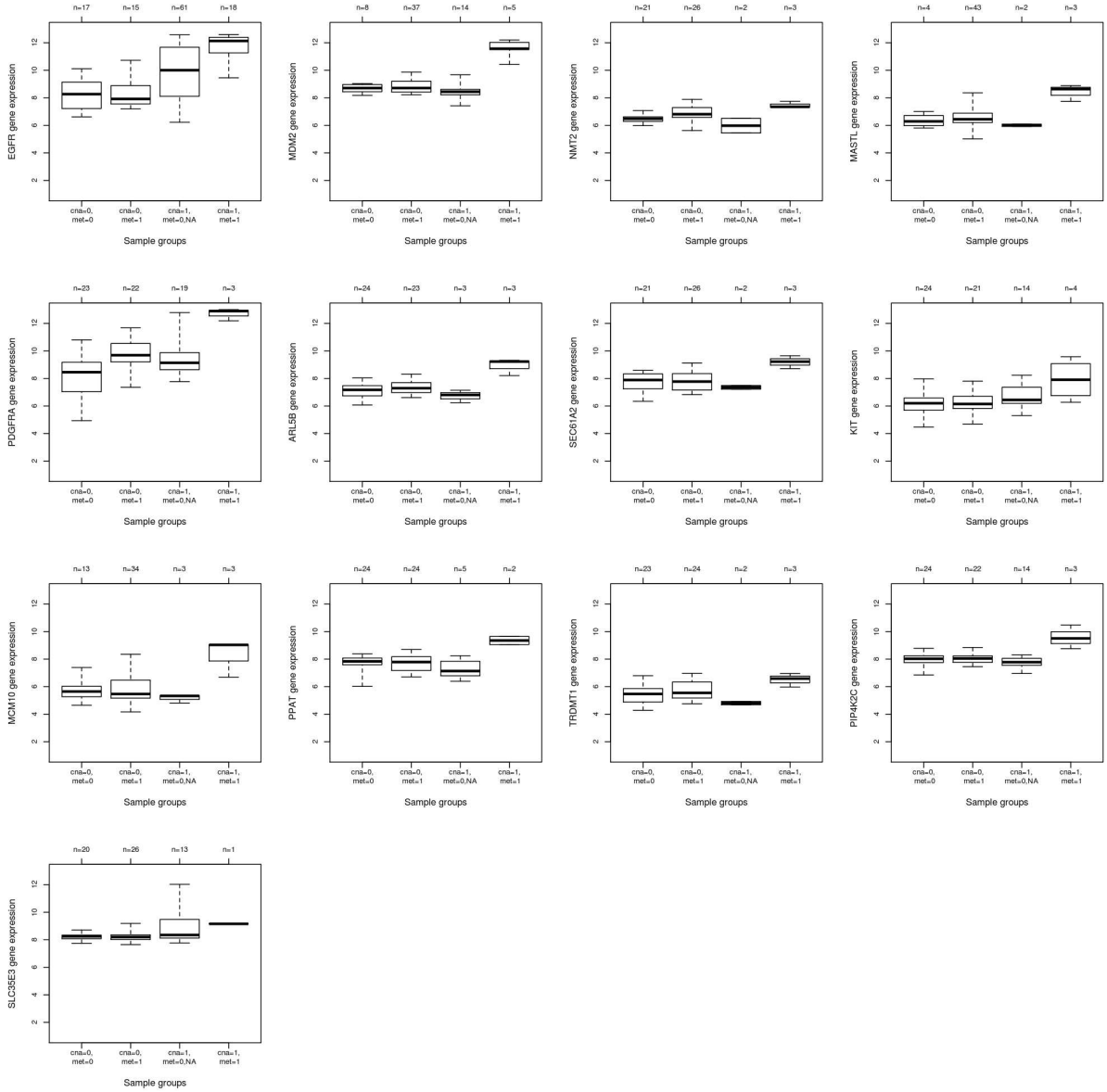


Figure 2: Expression differences in patients with different methylation and copy number statuses. Black bars are group medians. Filled rectangles contain values between 25th and 75th percentile. Patients with increased hypomethylation (met=1) and amplification (cgh=1) display significantly higher expression levels than patients with only an amplification (e.g., for *EGFR* (t-test), $p < 3.8 \times 10^{-8}$).

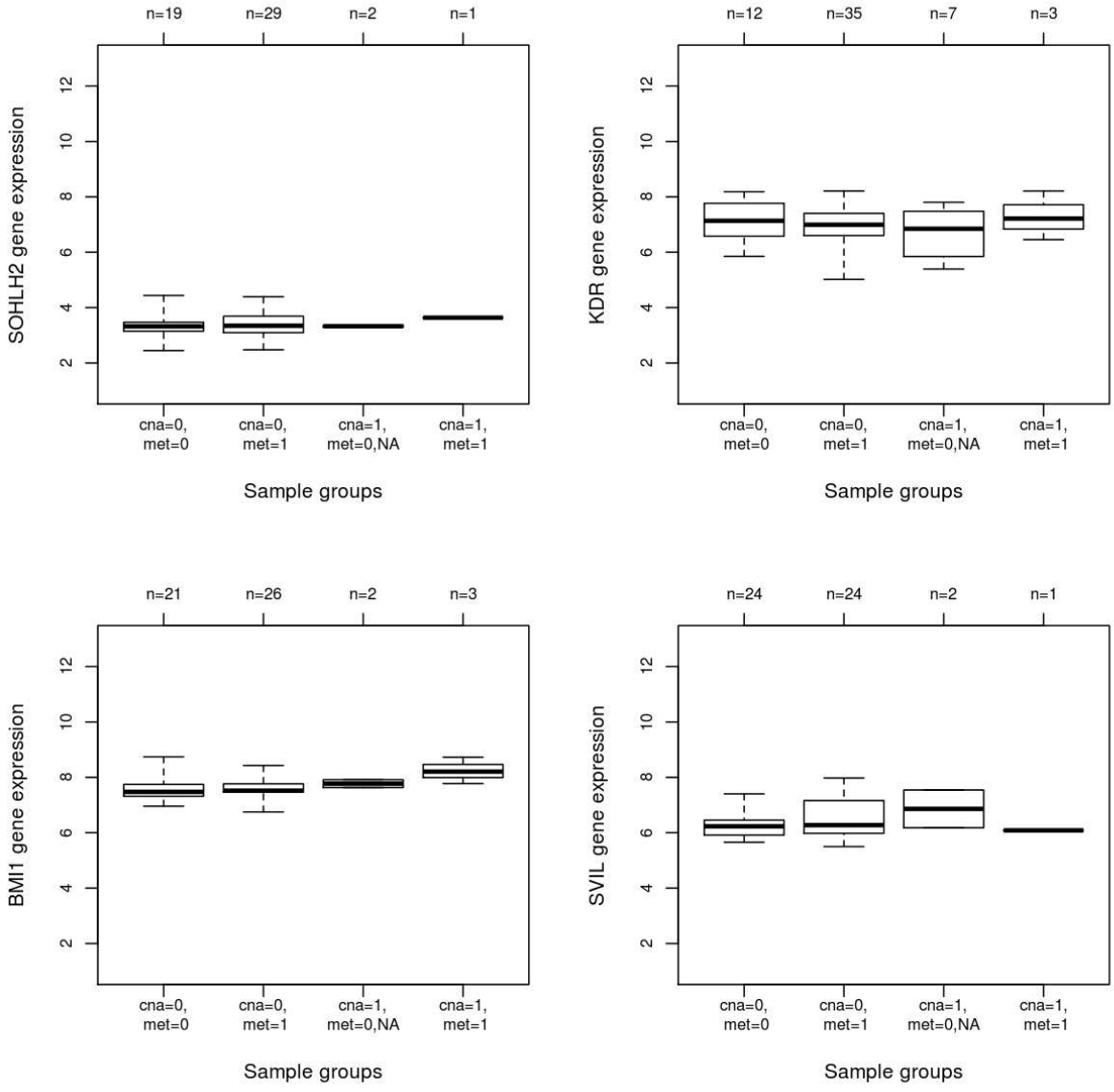


Figure 3: Boxplot of expression patterns for neutral genes.

References

[1] M. Benelli, G. Marseglia, G. Nannetti, R. Paravidino, F. Zara, F. D. Bricarelli, F. Torricelli, and A. Magi. A very fast and accurate method for calling aberrations in array–CGH data. *Biostat*, 11(3):515–518, 2010.

[2] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, J. Du, T. Kau, R. K. Thomas, K. Shah, H. Soto, S. Perner, J. Prensner, R. M. Debiasi, F. Demichelis, C. Hatton, M. A. Rubin, L. A. Garraway, S. F. Nelson, L. Liau, P. S. Mischel, T. F. Cloughesy, M. Meyerson, T. A. Golub, E. S Lander, I. K. Mellinghoff, and W. R. Sellers. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Natl. Acad. Sci. USA*, 104(50):20007–20012, 2007.

[3] Manel Esteller. Epigenetics in Cancer. *N Engl J Med*, 358(11):1148–1159, 2008.

[4] S. Hautaniemi, M. Ringnér, P. Kauraniemi, R. Autio, H. Edgren, O. Yli-Harja, J. Astola, A. Kallioniemi, and O-P Kallioniemi. A strategy for identifying putative causes of gene expression variation in human cancers. *J Franklin Inst*, 341(1–2):77–88, 2004.

[5] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array–based DNA copy number data. *Biostat*, 5(4):557–572, 2004.

[6] K. Ovaska, M. Laakso, S. Haapa-Paananen, R. Louhimo, P. Chen, V. Aittomäki, E. Valo, J. Núñez–Fontarnau, V. Rantanen, S. Karinen, K. Nousiainen, A.-M. Lahesmaa-Korpinen, M. Miettinen, P. Kohonen, J. Wu, J. Westermarck, and S. Hautaniemi. Large–scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med*, 2(9), 2010.

[7] D. Pinkel and D. G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, 37(6):S11–S17, 2005.

[8] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[9] Bekim Sadikovic, Maisa Yoshimoto, Khaldoun Al-Romaih, Georges Maire, Maria Zielenska, and Jeremy A. Squire. In vitro analysis of integrated global high-resolution dna methylation profiling with genomic imbalance and gene expression in osteosarcoma. *PLoS ONE*, 3(7):e2834, 2008.

[10] M. A. van de Wiel, K. I. Kim, S. J. Vosse, W. N. van Wieringen, S. M. Wilting, and B. Ylstra. CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, 23(7):892–894, 2007.