BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(6)

Genome-wide integrative analysis revealed a correlation between lengths of copy number segments and corresponding gene expression profile

Ken Miyaguchi[1], Yutaka Fukuoka[2]*, Hiroshi Mizushima[3], Mahmut Yasen[4], Shota Nemoto[1], Toshiaki Ishikawa[5], Hiroyuki Uetake[5], Shinji Tanaka[4], Kenichi Sugihara[6], Shigeki Arii[4], and Hiroshi Tanaka[1]

[1]Department of Bioinformatics, Graduate School; [2]Department of Biosystem Modeling, Graduate School of Biomedical Science, Tokyo Medical and Dental University, Japan; [3]Center for Public Health Informatics, National Institute of Public Health, Japan; [4]Department of Hepato-Biliary-Pancreatic Surgery; [5]Department of Translation Oncology; [6]Department of Surgical Oncology, Graduate School, Tokyo Medical and Dental University, Japan; Yutaka Fukuoka - E-mail: fukuoka.bsm@tmd.ac.jp; Phone: +81-3-5803-4777; Fax: +81-3-5803-4777; *Corresponding author

Received November 05, 2011; Accepted November 08, 2011; Published November 20, 2011

Abstract:
Microarray analysis has been applied to comprehensively reveal the abnormalities of DNA copy number (CN) and gene expression in human cancer research during the last decade. These analyses have individually contributed to identify the genes associated with carcinogenesis, progression, metastasis of tumor cells and poor prognosis of cancer patients. However, it is known that the correlation between profiles of CN and gene expression does not highly correlate. Factors which determine the degree of correlation remain largely unexplained. To investigate one such factor, we performed trend analyses between the lengths of CN segments and corresponding gene expression profiles from microarray data in hepatocellular carcinoma (HCC) and colorectal carcinoma (CRC). Significant correlations were observed in CN gain of HCC and CRC ($p<0.05$). The trend of the CN loss showed a significant correlation in HCC although there was no correlation between the length of CN loss segments and gene expression in CRC. Our findings suggest that the influence of CN on gene expression highly depends on the length of CN region, especially in the case of CN gain. To the best of our knowledge, this is the first study describing the correlation between lengths of CNA segments and expression profiles of corresponding genes.

Keywords: copy number alteration, gene expression, microarray, hepatocellular carcinoma, colorectal carcinoma

Background:
In late years, relations of carcinogenesis and copy number alteration (CNA) attract attention in conjunction with impaired expression of genes. Microarray-based genome-wide copy number analyses have revealed that a tremendous amount of alterations were in malignant genomic DNAs comparing to normal DNAs in various malignancies **[1, 2].** Accumulation of CNAs is thought to be one of triggers for malignant transformation and also for remarkably decreasing prognosis by easily causing recurrence or metastasis **[3, 4, 5].** Although comprehensive analyses for either genomic CNAs or gene expressions have been frequently performed by a number of researchers, the precedence study has rarely reported in genome-wide correlation between CNAs and abnormal gene expressions **[6, 7, 8, 9].** Some reports showed that approximately

15 to 60% of CNAs were positively correlated with gene expression profiles **[10, 11, 12, 13].**

Epigenetic regulation, such as DNA methylation or histone modification including methylation, acetylation and phosphorylation, is thought as one of the factors which may lead a discrepancy between CNAs and gene expression profiles **[14, 15, 16].** However, it may be possible that some of the gene expression changes are more severely influenced by CNAs than epigenetic regulations. Although there should be some biological mechanisms for the correlation between CNAs and gene expressions, the mechanisms are not fully understood.

Generally, each CNA has a difference in the length of DNA sequence which can be short as tens of thousands of base pairs, longer than hundreds of millions of base pairs, or something between them **[1].** Therefore, in addition to CNA profiles such as loss/deletion and gain/amplification, the length of the genomic mutation is possibly another factor to cause expression changes of genes encoded on the genomic region which has an abnormality in DNA copy number. To the best of our knowledge, no comprehensive study has been performed for discovering the association between genomic length of CNAs and gene expression changes. We hypothesized that the expressions of genes encoded on a long CNA segment tend to be more susceptible for copy number changes than genes on a shorter CNA segment because a wide range of chromosomal region covers more valuable sequences for genes to be efficiently translated from genomic DNA. Those gene expressions may show changes positively correlated with patterns of altered copy numbers; down-regulation for loss/deletion, up-regulation for gain/amplification. This is the first report discussing about the association between genomic lengths of genome-wide CNAs and expressions of genes encoded on those CNA segments by analyzing the data from microarrays for human hepatocellular carcinoma (HCC) and colorectal carcinoma (CRC).

Methodology:

Clinical specimens

Fresh-frozen tissue samples were collected from 10 HCC patients (9 male, 1 female) and 7 CRC (3 male, 4 female) patients who had undergone surgical resection at the Tokyo Medical and Dental University (TMDU, Tokyo, Japan) between 2005 and 2008. This study was approved by the institutional review board. Informed consent was obtained from all patients in accordance with the guidelines by the review board. Adjacent nonmalignant tissues corresponding to the malignant tissues were also collected as non-tumor samples. All specimens were collected in cryotubes, frozen just after the resection and stored at −80°C until the time of use. Clinical characteristics of samples are shown in **Table 1 (see supplementary material).**

DNA copy number analyses

Genomic DNAs were extracted from frozen specimens except CRC malignant tissue and isolated using a QIAamp DNA Mini Kit (Qiagen, Valencia, CA). For malignant CRC samples, laser microdissection (LMD) was applied to isolate tumor epithelial cells from sectioned tissues using LMD6000 (Leica Microsystems GmbH, Wetzlar, Germany) and extractions of genomic DNAs were done with Qiagen QIAamp DNA Micro Kit. The experiment was performed by strictly following the assay manual and using GeneChip Human Mapping 250K Sty Arrays (Affymetrix, Santa Clara, CA), which contained 238,229 SNP probe sets.

Gene expression analyses

Total RNAs were extracted from frozen specimens except CRC malignant tissue using RNeasy Mini Kit (Qiagen, Valencia, CA). For malignant CRC samples, LMD was applied to isolate tumor epithelial cells from sectioned tissues using Leica LMD6000 and extractions of total RNAs were done with Qiagen RNeasy Micro Kit. Following the manufacturer's instructions, prepared cocktails were hybridized on GeneChip Human Genome U133 Plus 2.0 Arrays and signal intensities for probe sets were detected by GeneChip Scanner 3000 7G.

Data analysis

The raw intensity data of SNP probe sets were first analyzed by Affymetrix Genotyping Console 4.0 software to produce the genotype data to generate the CNA data by estimation of the copy number for each SNP probe set comparing each intensity data with corresponding normal DNA. Subsequently, we identified continuous CNA regions based on integrated chromosomal positions of SNP probes and annotated these regions with relative gene symbols which were about 23,000 genes taken from the EntrezGene database **[17].** If a gene has multiple probes, the mean of the expression values of all the probes was used for the gene. Then, a log ratio between the gene expressions of the non-tumor and tumor tissues was calculated for each gene. The genes without locus information were excluded and remaining 20,532 genes were used for the subsequent analysis. The CNA data and the log ratios were merged based on locus information. For each continuous CNA region, the numbers of the up- and down-regulated genes in the tumor tissue were counted. Then the data for all CNA regions were merged and classified into 8 classes based on the length of the region: <100Kb, 100-500Kb, 500Kb-1Mb, 1-5Mb, 5-10Mb, 10-50Mb, 50-100Mb and 100-500Mb. A fraction of the up/down regulated genes was calculated in each class and the Cochran-Armitage test was performed to test if the data have a trend.

Results and Discussion:

In terms of an integration analysis between CNAs and corresponding gene expressions, there has been no study describing a higher correlation. Secondary factors may have influence to reflect a copy number change to gene expression. Therefore, we hypothesize that a length of CNA segment may affect transcriptional efficiency and performed a comprehensive correlation study between the lengths of CNA segments and corresponding gene expressions by analyzing microarray data of DNA copy number and gene expression in two different cancers.

First, we investigated the relationship between the genomic and transcriptional profiles by directly comparing array data which were normalized and annotated with gene information (data not shown). However, the direct comparison did not show any significant correlation. In terms of the effect on the transcription, there might be mainly two different types of CNAs: influential CNA and non-influential CNA. Epigenetic regulation such as DNA methylation could be another factor to change the transcriptional switch directly without having

genomic alterations. CNA may have an indirect effect to gene expression through epigenetic regulation.

Secondly, we compared the lengths of CNA segments and the fractions of gene expression within the segments. The correlation between CNA and gene expression was analyzed by the Cochran-Armitage trend test. Results of the Cochran-Armitage trend test revealed a significant association between the length of CNA region of amplification/deletion and the fraction of up/down regulated genes in HCC-gain (p=0.021), HCC-loss ($p=2.7\times10^{8}$) and CRC-gain (p=0.021). There was no significant association in CRC-loss **(Figure 1).** To the best of our knowledge, this is the first time that the correlation between the lengths of CNA segments and the corresponding genes expression profiles were investigated by the genome-wide integration analysis.

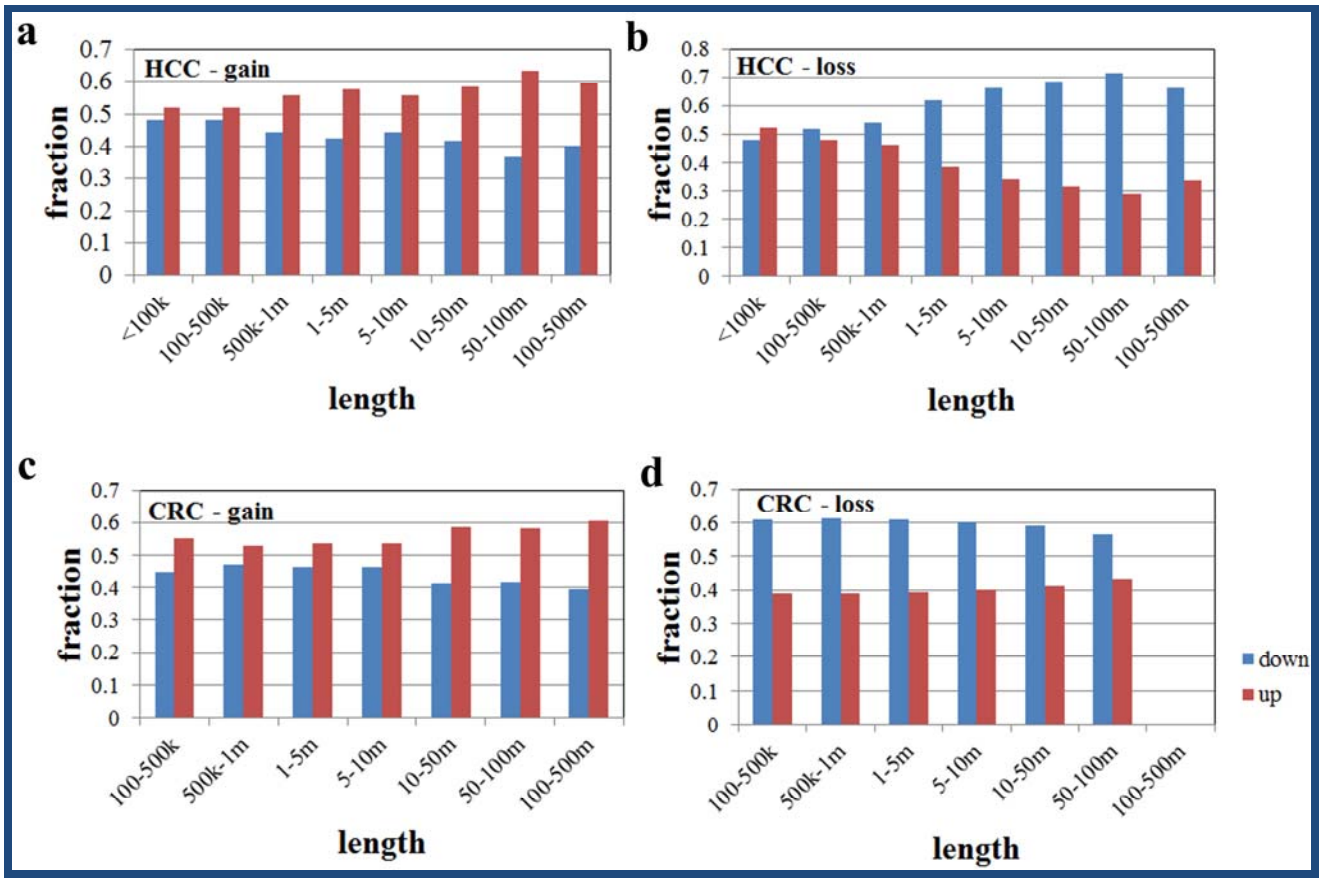

Figure 1: Relationships between the length of CNA region and the number of up/down regulated genes. The horizontal and vertical axes represent the length of chromosomal region of amplification/deletion and the fraction of the up/down regulated genes. Results of the Cochran-Armitage trend test revealed a significant association between the length of CNA region of amplification/deletion and the fraction of up/down regulated genes in HCC-gain (p=0.021), HCC-loss ($p=2.7\times10^{8}$) and CRC-gain (p=0.021). There was no significant association in CRC-loss. These results suggest that more genes are up-regulated as the length of gained region increases. As for the lost region, further analysis is required because of the discrepancy in the results of the two different types of cancers.

For gain of copy number, the lengths of CNA and gene expressions were significantly correlated in HCC and CRC. Genome sequence and secondary and/or higher-order structure might be the causes for a larger influence on a gene expression within a longer CNA region. A wide range of CN gain segments might cover a larger number of gene sequences. In addition, it could increase the copy number of not only a whole gene sequence but also non-coding sequences and a promoter region which is necessary to initiate the transcription of the gene. In contrast, a gene may not be properly transcribed with incomplete sequence of gene or promoter region which is not fully covered by CNA segment. Therefore, a longer gain region tends to cover more complete DNA sequence and up-regulate the gene expression. Since a higher-order structure of genome DNA is formed for a transcriptional event, genome sequence with a wide range of CNA regions may be more likely to form a higher-order structure **[18, 19, 20]**, resulting up-regulation of genes.

For CN loss, the trend analysis in HCC showed a significant correlation but the gene expressions were not significantly correlated with the length of segments in CRC. Whenever a part of the gene sequence or the promoter region is lost, it may become impossible to promote the transcription. As same as the effect of gain, the length of loss might have an influence to structure of genomic DNA, such as an imbalance in a multidimensional formation of chromosomes and unfolded structure of DNA during a gene transcription. It may be

possible that the imbalance affects even gene transcriptions around the genomic region with sequence deletions. A secondary effect of the imbalance could be associated with irregular transcriptions of genes without any alterations in copy number. However, the influence of CN loss is more complicated because the chromosomal location of loss seems also critical for the transcription. We assume that the influence of loss varies depending on which part of sequence missing and the length of it.

Conclusion:

The results of the Cochran-Armitage trend test revealed a significant association between the length of CNA region of amplification/deletion and the fraction of up/down regulated genes in HCC-gain, HCC-loss and CRC-gain. There was no significant association in CRC-loss. These results suggest that more genes are up-regulated as the length of gained region increases. As for the lost region, further analysis is required because of the discrepancy in the results of the two different types of cancers.

Acknowledgement:

This work was supported in part by grant-in-aids from Science and Technology Promotion Adjustment Expenses of Japan (No. 08005234), Integrated Database Project and a scientific research grant (No. 20510184) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan.

References:

[1] Bignell GR *et al. Nature.* 2010 **463**: 7283 [PMID: 20164919]
[2] Garraway LA *et al. Nature.* 2005 **436**: 7047 [PMID:16001072]
[3] Mehta KR *et al. Clin Cancer Res.* 2005 **11**: 5 [PMID:15756001]
[4] Guttman M *et al. PLoS Genet.* 2007 **3**: 8 [PMID:17722985]
[5] Schleiermacher G *et al. J Clin Oncol.* 2010 **28**: 19 [PMID:20516441]
[6] Yusenko MV *et al. BMC Cancer.* 2009 **9** [PMID:19445733]
[7] Newnham GM, *et al. BMC Cancer.* 2011 **11** [PMID:21385341]
[8] Myllykangas S *et al. Int J Cancer.* 2008 **123**: 4 [PMID:18506690]
[9] Haverty PM *et al. Genes Chromosomes Cancer.* 2008 **47**: 6 [PMID:18335499]
[10] Pollack JR *et al. Proc Natl Acad Sci U S A.* 2002 **99**: 20 [PMID: 12297621]
[11] Xu C *et al. Mol Cancer.* 2010 **9** [PMID: 20537188]
[12] Natrajan R *et al. Breast Cancer Res Treat.* 2010 **121**: 3 [PMID: 19688261]
[13] Kim JH *et al. Cancer Res.* 2007 **67**: 17 [PMID: 17804737]
[14] Wilson IM *et al. Cell Cycle.* 2006 **5**: 2 [PMID: 16397413]
[15] Jones PA & Laird PW, *Nat Genet.* 1999 **21**: 2 [PMID: 9988266]
[16] Jones PA & Baylin SB, *Nat Rev Genet.* 2002 **3**: 6 [PMID: 12042769]
[17] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
[18] Georgel PT. *Biochem Cell Biol.* 2002 **80**: 3 [PMID: 12123282]
[19] Bell O *et al. Nat Rev Genet.* 2011 **12**: 8 [PMID:21747402]
[20] Li G & Reinberg D, *Curr Opin Genet Dev.* 2011 **21**: 2 [PMID:21342762]

Edited by P Kangueane
Citation: Miyaguchi *et al.* Bioinformation 7(6): 280-284(2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: Information of clinical tissue samples used for microarray analyses

Characteristic		Cases
Hepatocellular carcinoma		
Gender		
	M	9
	F	1
Age		
	<60	4
	≥60	6
Colorectal carcinoma		
Gender		
	M	3
	F	4
Age		
	<60	2
	≥60	5