Genomics & Informatics Vol. 8(3) 138-141, September 2010

Comparison of the Affymetrix SNP Array 5.0 and Oligoarray Platforms for Defining CNV

Ji-Hong Kim[1,2], Seung-Hyun Jung[1,2], Hae-Jin Hu[1,2], Seon-Hee Yim[1] and Yeun-Jun Chung[1,2]*

[1]Integrated Research Center for Genome Polymorphism, [2]Department of Microbiology, The Catholic University of Korea School of Medicine, Seoul 137-701, Korea

Abstract

Together with single nucleotide polymorphism (SNP), copy number variations (CNV) are recognized to be the major component of human genetic diversity and used as a genetic marker in many disease association studies. Affymetrix Genome-wide SNP 5.0 is one of the commonly used SNP array platforms for SNP-GWAS as well as CNV analysis. However, there has been no report that validated the accuracy and reproducibility of CNVs identified by Affymetrix SNP array 5.0. In this study, we compared the characteristics of CNVs from the same set of genomic DNAs detected by three different array platforms; Affymetrix SNP array 5.0, Agilent 2X244K CNV array and NimbleGen 2.1M CNV array. In our analysis, Affymetrix SNP array 5.0 seems to detect CNVs in a reliable manner, which can be applied for association studies. However, for the purpose of defining CNVs in detail, Affymetrix Genome-wide SNP 5.0 might be relatively less ideal than NimbleGen 2.1M CNV array and Agilent 2X244K CNV array, which outperform Affymetrix array for defining the small-sized single copy variants. This result will help researchers to select a suitable array platform for CNV analysis.

Keywords: copy number variation (CNV), single nucleotide polymorphism (SNP)

Introduction

Recent progress of human genome mapping has facilitated the understanding of inter-individual differences in various phenotypes such as disease susceptibility and responsiveness to drugs (Estivill and Armengol, 2007). Together with single nucleotide polymorphism (SNP), the large-scale genomic variations, which are named copy number variation (CNV), are recognized to be the major component of human genetic diversity (Freeman *et al*., 2006) and used as a genetic marker in many disease association studies.

Since the first discovery of CNV using BAC array platform, the microarray platforms for defining the global CNVs have been enormously improved. Iafrate *et al* identified 255 CNVs from 39 individuals using 3K BAC array and Sebat *et al.* defined 76 CNVs in 20 individuals using oligonucleotide array based ROMA approach (representational oligonucleotide microarray analysis) (Iafrate *et al*., 2004; Sebat *et al*., 2004). Due to the improvement of technology and array resolution, the size of CNVs has been getting smaller and the number of CNVs detected per individual genome increasing. However, defining CNV is sensitive to choice of array platforms and detection algorithms. Indeed, the number and size of CNVs were diverse between the studies using different platform (Scherer *et al*., 2007).

Among many whole-genome CNV analysis platforms, SNP arrays have been commonly used for CNV discovery due to its ubiquitous genome coverage and advantageous resolution. Also in Korea, SNP arrays have been frequently applied for CNV-disease association studies. The Korea Association Resource (KARE) consortium used the Affymetrix SNP array 5.0 genotyping data of 8,848 individuals provided by Korea National Health Institute (KNIH) for studying the associations between genetic variations and phenotypes. Using the same dataset, we previously reported the evolutionary and functional implications of the CNVs in Korean population (Yim *et al*., 2010).

Despite the popularity of SNP arrays in studying CNVs, there are concerns regarding possibility of false discovery that can be caused by one-dye SNP arrays. There have been several reports comparing the array-CGH platforms for detecting CNVs (Baumbusch, 2008; Curtis and Lynch, 2010; Hester *et al*., 2009), but there has been no report that validated the accuracy and reproducibility of CNVs identified by Affymetrix SNP array 5.0. In this study, we compared the characteristics of CNVs from the same set of genomic DNAs detected by three different array platforms; Affymetrix SNP array 5.0, Agilent 2X244K CNV array and NimbleGen 2.1M CNV array.

*Corresponding author: E-mail yejun@catholic.ac.kr
Tel +82-2-2258-7343, Fax +82-2-596-8969
Accepted 10 September 2010

Methods

Study subject

Genomic DNA was isolated from two HapMap cell lines, GM10851 and GM15510 (Coriell, Camden, NJ, USA) with DNeasy® Blood & Tissue Kit (Qiagen, Hilden, Germany).

Array hybridization and data processing

For CNV defining using Affymetrix Genome-wide SNP 5.0 platform, array hybridization and all the downstream data preprocessing procedures including allele correction, summarization and background correction were performed as described previously (Yim *et al*., 2010).

For Agilent 2X244K CNV array, array hybridization and all the downstream data preprocessing were performed according to the manufacturer's instructions. In brief, 2ug of genomic DNA from NA10851 and NA15510 were digested with heat fragmentation method and the NA10851 cell line was labeled with Cy3-dUTP and NA15510 was labeled with Cy5-dUTP using Agilent Genomic DNA Enzymatic Labeling Kit (Agilent Technologies, Santa Clara, CA). Labeled DNAs were purified by Amicon Ultra-0.5 purification column (Millipore, Billerica, MA) and then pre-annealed with a blocking reagent (Agilent technologies) containing Cot-1 DNA (ConnectaGen, Seoul, Korea). Array hybridization was performed for 40 hours at 65°C and 20 rpm. After hybridization, the arrays were washed and scanned with Agilent GA2565C scanner (Agilent technologies). Array images were analyzed with Feature Extraction software (Agilent Technologies) with the CGH-v4.95 protocol for normalization.

For NimbleGen 2.1M CNV array, array hybridization and all the downstream data preprocessing were performed according to the manufacturer's instructions. In brief, 2.5 ug of genomic DNA from the NA10851 cell line was labeled with Cy3-labeled NimbleGen validated Random 9mer (TriLink BioThchknologies, San Diego, CA) and NA15510 was labeled with Cy5-labeled NimbleGen validated Random 9mer (TriLink BioTechnologies) using Klenow Fragment (NEB, Ipswich, MA). Labeled reference and test DNA were combined, denatured and pre-annealed with 2X Hybridization Buffer (Roche NombleGen, Madison, WI). The mixed DNA was hybridized onto the arrays for 72 hours in a MAUI hybridization machine (BioMicro, Utah) at 42°C. After hybridization, the arrays were washed and scanned with GenePix 4000B scanner (Molecular Devices, Sunnyvale, CA). Array images were analyzed with NimbleScan v2.6 (Roche NombleGen).

Defining CNVs

All the data sets from the three different platforms were analyzed using the rank segmentation method of NEXUS software (BioDiscovery) under the same threshold. NEXUS is the algorithm that recursively divides chromosomes into segments of common intensity distribution functions like the circular binary segmentation (CBS).

Result and Discussion

Based on the rank segmentation algorithm, we defined the CNVs between NA10851 and NA15510 using the three different array platforms. Details of the identified CNVs are summarized in Table 1. Although the same DNA was analyzed by the same detection algorithm, the number of CNVs varied from 102 to 608 depending on the types of the array used. The smallest number of CNVs were identified by Affymetrix SNP 5.0, while Agilent 2X244K and NimbleGen 2.1M arrays defined 3 to 6 times more CNVs. Gain/loss ratio of the CNVs identified by Affymetrix SNP 5.0, Agilent 2X244K and NimbleGen 2.1M arrays were 0.89 (48/54), 0.99 (188/189), and 3.57 (475/133), respectively. In spite of the large difference in the number of identified CNVs among the platforms, the total CNV coverage of the genome was all similar in three the platforms, ranging from 6.7% to 9.9% of the genome. The median size of the CNVs was largest for Affymetrix SNP 5.0 and smallest for Agilent 2X244K array.

Of the 102 CNVs identified by Affymetrix 5.0, 76 CNVs (75%) were consistently detected by the other two platforms. Among them, 71 (93%) CNVs overlap the CNV in DGV. Fig. 1A illustrates the example of a CNV identified by all three platforms. This 35 kb sized copy number gain on 13q is located within the previously reported CNV in DGV and its boundaries were almost same boundaries regardless of the platforms.

Table 1. General characteristics of the CNVs identified by the three platforms*

	Affymetrix 5.0	Agilent 244Kx2	NimbleGen 2.1M
Total CNVs	102	377	608
Total coverage	193,342,543	299,372,362	235,211,553
Average size	1,895,515	794,091	386,861
Median size	139,099	50,240	69,969
CNV not in DGV	5	4	95

*DGV database used hg18,v9, Mar.2010 version.

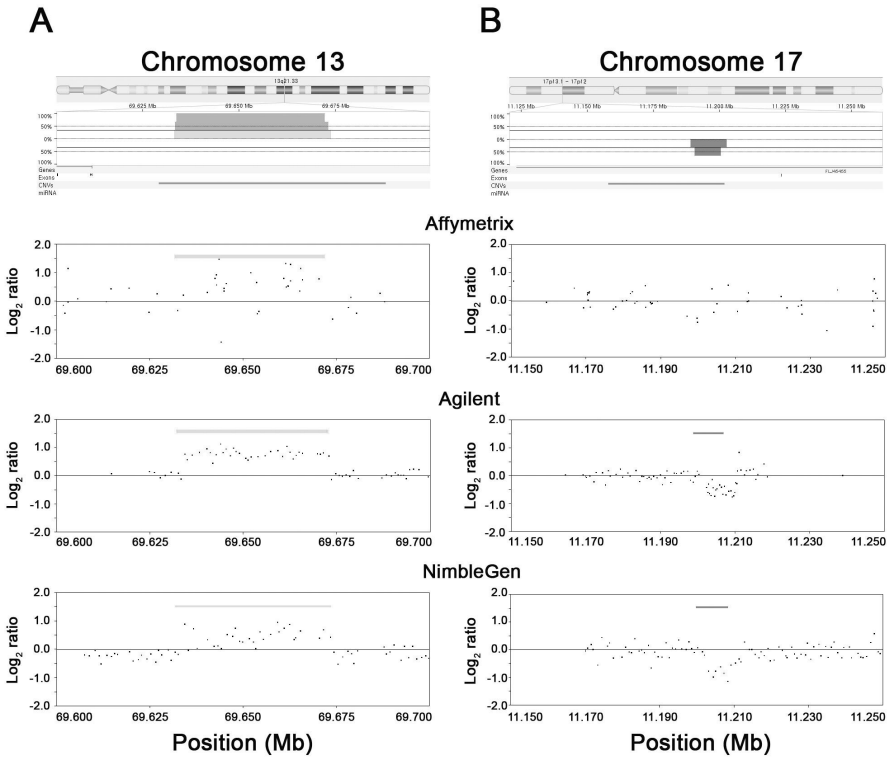

Fig. 1. (A) A CNV on Chromosome 13q is consistently detected across the three platforms. (B) A CNV on Chromosome 17p is detected by Agilent 2X244K and NimbleGen 2.1M arrays but not by Affymetrix SNP 5.0 array. CNV was defined using NEXUS software.

Forty four per cent (166 out of 377) of CNVs identified by Agilent 2X244K array and 56.4% (343 out of the 608) CNVs identified by NimbleGen 2.1M array were not detected by Affymetrix 5.0. Fig. 1B illustrates the example of a CNV identified by Agilent 2X244K and NimbleGen 2.1M arrays but not by Affymetrix 5.0 array. This 9 kb sized copy number loss on 17p also overlaps the CNV in DGV. Its boundaries were defined almost identically by Agilent 2X244K and NimbleGen 2.1M arrays. This data reflects relatively low resolution of Affymetrix 5.0 array, which makes the platform less efficient for detecting small sized CNVs.

Due to the widespread availability, SNP array has been frequently used for genome-wide CNV discovery and CNV-disease association studies. Affymetrix Genome-wide SNP 5.0 is one of the commonly used SNP array platforms for SNP-GWAS as well as CNV analysis in Korea. In our analysis, Affymetrix SNP array 5.0 seems to detect CNVs in a reliable manner, which can be applied for association studies. However, for the purpose of defining CNVs in detail, Affymetrix Genome-wide SNP 5.0 might be relatively less ideal than NimbleGen 2.1M CNV array and Agilent 2X244K CNV array, which outperform Affymetrix array for defining the small-sized single copy variants.

Acknowledgement

National Institute of Health, Korea Centers for Disease Control and Prevention gratefully provided the KARE genotype and epidemiological data and also supported this work through the KARE Analysis Consortium. This study was supported by grants from Korea Research Foundation Grant funded by the Korean Government (KRF-2008-220-E00025).

References

Baumbusch, L.O. (2008). Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics* 9, 379.

Curtis, C., and Lynch, A. (2009). The pitfalls of platform comparison: DNA copy number array technologies assessed, *BMC Genomics* 10, 588.

Estivill, X., and Armengol, L. (2007). Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* 3, 1787-1799.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., Carter, N.P., Scherer, S.W., and Lee, C. (2006). Copy number variation: new insights in genome diversity. *Genome Res.* 16, 949-961.

Hester, S.D., Reid, L., and Nowak, N. (2009). Comparison of comparative genomic hybridization technologies across

microarray platforms. *J. Biomol. Tech.* 20, 135-151.
Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat. Genet.* 36, 949-951.
Scherer, S.W., Lee, C., Birney, E., Altshuler, D.M., Eichler, E.E., Carter, N.P., Hurles, M.E., and Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nat. Genet.* 39(7 Suppl):S7-15.
Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T.C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science* 305, 525-528.
Yim, S.H., Kim, T.M., Hu, H.J., Kim, J.H., Kim, B.J., Lee, J.Y., Han, B.G., Shin, S.H., Jung, S.H., and Chung, Y.J. (2010). Copy number variations in East-Asian population and their evolutionary and functional implications. *Hum. Mol. Genet.* 19, 1001-1008.