Genomics & Informatics Vol. 6(4) 231-234, December 2008

Comparison of Normalization Methods for Defining Copy Number Variation Using Whole-genome SNP Genotyping Data

Ji-Hong Kim[1,2], Seon-Hee Yim[2], Yong-Bok Jeong[1,2], Seong-Hyun Jung[1,2], Hai-Dong Xu[1,2], Seung-Hun Shin[1,2] and Yeun-Jun Chung[1,2]*

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

Abstract

Precise and reliable identification of CNV is still important to fully understand the effect of CNV on genetic diversity and background of complex diseases. SNP marker has been used frequently to detect CNVs, but the analysis of SNP chip data for identifying CNV has not been well established. We compared various normalization methods for CNV analysis and suggest optimal normalization procedure for reliable CNV call. Four normal Koreans and NA10851 HapMap male samples were genotyped using Affymetrix Genome-Wide Human SNP array 5.0. We evaluated the effect of median and quantile normalization to find the optimal normalization for CNV detection based on SNP array data. We also explored the effect of Robust Multichip Average (RMA) background correction for each normalization process. In total, the following 4 combinations of normalization were tried: 1) Median normalization without RMA background correction, 2) Quantile normalization without RMA background correction, 3) Median normalization with RMA background correction, and 4) Quantile normalization with RMA background correction. CNV was called using SW-ARRAY algorithm. We applied 4 different combinations of normalization and compared the effect using intensity ratio profile, box plot, and MA plot. When we applied median and quantile normalizations without RMA background correction, both methods showed similar normalization effect and the final CNV calls were also similar in terms of number and size. In both median and quantile normalizations, RMA background correction resulted in widening the range of intensity ratio distribution, which may suggest that RMA background correction may help to detect more CNVs compared to no correction.

Keywords: copy number variation, normalization, SNP, Robust Multiarray Average

Introduction

A wide spectrum of genomic variation is present in the human genome, from single nucleotide polymorphisms (SNP) to microscopically visible, large structural alterations. Recently, the presence of large-scale genomic variation, named copy number variation (CNVs), was uncovered (Iafrate *et al.*, 2004; Sebat *et al.*, 2004). Since the discovery, widespread presence of DNA structural variation in phenotypically normal individuals has been well known (Kim *et al.*, 2008). Accumulating evidence suggests CNV is not just inter-individual variation contributing to diversity of phenotypes in human beings, but also very likely to associate with various human diseases (Estivill & Armengol 2007; McCarroll & Altshuler 2007). Therefore precise and reliable identification of CNV is still important to fully understand the effect of CNV on genome diversity and background of complex diseases.

For defining CNV accurately, resolution is one of the important issues. When CNV was first uncovered, approximately 12 CNVs per genome were identified through both BAC array and oligoarray (Iafrate *et al.*, 2004; Sebat *et al.*, 2004). In 2006, Affymetrix GeneChip Human Mapping 500K early access version was applied to define the CNVs from 269 HapMap individuals (Redon *et al.*, 2006). In that study, ～1500 CNVs were identified and the median size of them was smaller (80 Kb) than those defined by tiling BAC array (230 Kb). In addition to SNP-based CNV analysis, recent higher resolution oligoarray platforms were introduced and revealed that the human genome may contain more CNVs than previously thought and that the average size of CNVs might be smaller than previously reported (de Smith *et al.*, 2007; Perry *et al.*, 2008).

In spite of advance of new technologies, SNP marker has been used frequently to detect CNVs because of several advantages. First, due to large number of known SNP resources, extremely high resolution SNP genotyping chips (＞1 Million) can be designed and currently available. Secondly, accompanying SNP genotype information is useful for disease association study and CNV-SNP combined interpretation can achieve new breakthrough in understanding genetic contribution to

*Corresponding author: E-mail yejun@catholic.ac.kr
Tel +82-2-590-1214, Fax +82-2-596-8969
Accepted 1 December 2008
JHK and SHY contributed equally to this paper.

the complex diseases. But, the analysis of SNP chip data for identifying CNV has not been well established. For example, for CNV detection, signal intensities, instead of genotype, are used, which requires a normalization process to remove systemic errors due to experimental conditions ranging from the array manufacturing process to the quantification of the spot intensities. Since there are various normalization procedures available which can produce different results, it is important to understand the basic characteristics of various methods before applying to the data. In this study, we compared various normalization methods for Affymetrix SNP array 5.0 data for CNV analysis and suggest optimal normalization procedure for reliable CNV call.

Methods

Study subjects

Four normal Koreans (K1-4) and NA10851 HapMap male samples were genotyped using Affymetrix Genome-Wide Human SNP array 5.0. NA10851 DNA was purchased from Coriell Institute for Medical Research (Camden, NJ). Genomic DNA was extracted from blood by using Genomic DNA prep kit (SolGent, Daejeon, Korea). DNA was used for hybridization after quantification and quality check.

Pre-processing SNP array data

Hybridization was performed according to manufacturer's instructions. Affymetrix Genome-Wide Human SNP array 5.0 platforms use single-color detection system in which one sample is hybridized per chip. Therefore, normalization of this platform data is performed between arrays. Among perfect match (PM) and mismatch (MM) probes, only PM probe intensity data were used for CNV analysis. Before applying normalization procedures, we performed allele correction, summarization, and background correction using the software provided by Affymetrix.

Comparing four different normalization methods

There are several approaches to normalize systemic variation of microarray data. We evaluated the effect of median and quantile normalization to find the optimal normalization for CNV detection based on SNP array data. We also explored the effect of Robust Multichip Average (RMA) background correction on each normalization process. It uses the PM distribution to get an overall background level, and transforms values based on the background noise and signal. In total, following 4 combinations of normalization were tried; 1) Median normalization without RMA background correction, 2) Quantile normalization without RMA background correction, 3) Median normalization with RMA background correction, and 4) Quantile normalization with RMA background correction.

SW-ARRAY analysis

CNV was called based on log2 test/reference ratio using SW-ARRAY algorithm (Price *et al*., 2005). We used median value+2.5 MAD (median absolute deviation) and the cutoff of island score was >1 MAD and a threshold of six consecutive probes for calling CNVs.

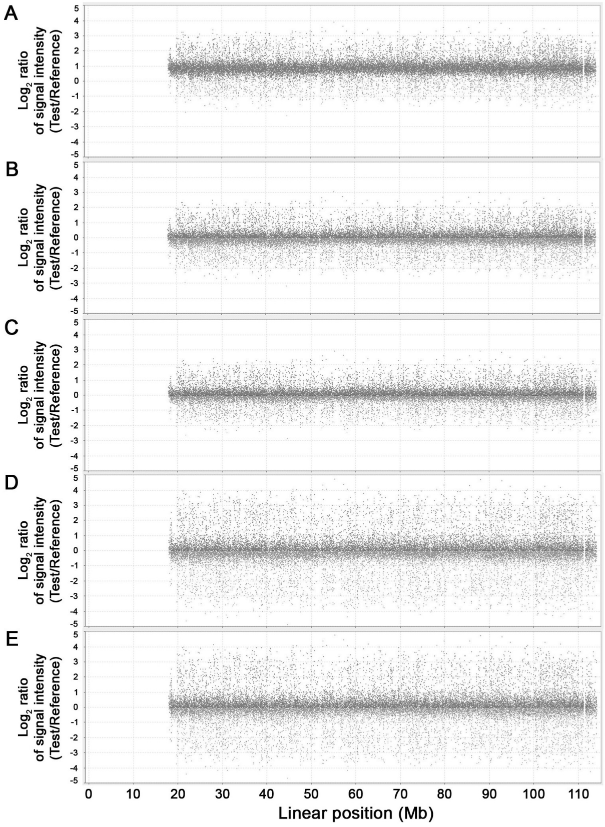

Fig. 1. Comparison of the effect of different normalizations. (A) Rraw intensity ratio values of Chromosome 13 from one of four normal Koreans before normalization. Intensity ratios deviate asymmetrically from zero on a log2 scale. B-E, the intensity ratio profile normalized by (B) median normalization, (C) median normalization with RMA background correction, (D) quantile normalization and (E) quantile normalization with RMA background correction on the same sample.

Results

Comparing the effect of different normalizations

Fig. 1A illustrates the example of intensity ratio plot where intensity ratios deviate asymmetrically from zero on a log2 scale. If CNVs will be defined using this raw data, there must be substantial amount of false CNV calls. We applied 4 different combinations of normalization and compared the effect. As illustrated in Fig. 1B ∼E, the intensity ratios were shifted toward the horizontal zero line. The plots of median (Fig. 1B) and quantile (Fig. 1C) normalized data (without RMA background correction) showed the similar effect. Most of the intensity values are around zero on a log2 scale and range of intensity values was tight, mostly within ±2 in log2 scale. Interestingly, RMA background correction made a visible effect on the data (Fig. 1D and E). Although RMA background corrected data also looked well normalized, the ranges of intensity values (mostly within ±4 on a log2 scale) were wider than those without background correction regardless of the normalization methods.

When we observed the box plot showing intensity ratio distribution, the effects of normalization and RMA background correction were similar as describe above. Fig. 2 illustrates the examples of the effect of median normalization with and without RMA background correction. Raw data before normalization of the 4 Korean samples were highly deviated (Fig. 2A). After median normalization, all 4 intensity ratio distributions looked similar centering around 1 but RMA background corrected data showed wider range of intensity ratio distribution than the data without being RMA corrected (Fig. 2B and C).

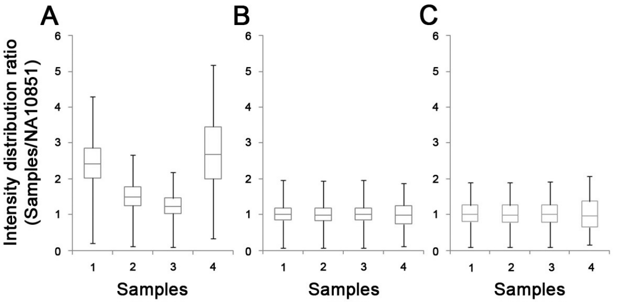

Fig. 2. Box plots showing the distribution of intensity ratios. (A) Raw intensity ratios before normalization of 4 normal Koreans. (B) Intensity ratios after median normalization. (C) Intensity ratios after median normalization with RMA background correction.

We also used the MA plot to visualize the effects of normalization (Fig. 3). Raw intensity data which revealed a deviation from expected horizontal zero line migrated to horizontal zero line after applying median normalization (Fig. 3A, B). When median normalization was performed with RMA background correction, the MA plot showed well-balanced intensity distribution around zero, but the range of intensity ratios was larger than those of median normalization only (Fig. 3C).

Comparing the CNV calls between different normalizations

We then observed the CNV calls using SW-ARRAY algorithm as described in methods from the data which were preprocessed and normalized in 4 different ways (Table 1). Numbers of CNVs identified using both median (n=26) and quantile (n=29) normalized data were similar and most CNVs identified in both normalizations were concordant, 96.2% (25/26) in median; 86.2% (25/29) in quantile. Mean size of CNVs identified through

Table 1. Characters of the CNVs identified by 4 different combinations of normalization

Normalizations	Gain	Loss	Total	Size (Kb)
Median	19	7	26	10,9
Quantile	19	10	29	12,2
RMA- Median	95	64	159	19,0
RMA- Quantile	83	44	127	20,6

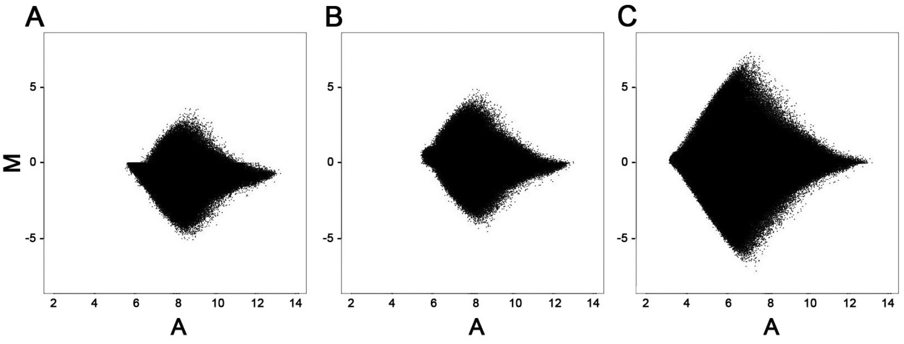

Fig. 3. MA plots showing the distribution of intensity ratios. (A) Raw intensity; (B) after median normalization; (C) after median normalization with RMA background correction.

quantile normalization (12.2 Kb) was slightly bigger than those identified through median normalization (10.9 Kb). Interestingly, more CNVs were identified in RMA background corrected data than in those without background correction. Mean size of CNVs from both median-RMA background correction and quantile-RMA background correction was approximately 20 Kb, which is on average bigger than those identified without background correction. In RMA background correction based normalizations, most CNVs identified in both methods were concordant, 79.9% (127/159) in RMA-median; 100% (127/127) in RMA-quantile. These results imply that the effect of median and quantile normalizations is generally comparable and RMA background correction may help to detect more CNVs compared to without correction.

Discussion

SNP chip based CNV analysis is now one of the most common approaches to detect CNV with higher resolution CNV. However, due to the fundamental difference in design of the probes which were not designed for measuring the signal intensities but for SNP call, data processing including normalization for defining quantitative measurement of structural variation has not been well established. In this study, we compared various normalization methods for Affymetrix SNP array 5.0 data based CNV analysis to select the most reliable normalization methods.

When we applied median and quantile normalizations without RMA background correction, both methods showed similar normalization effect and final CNV calls were also similar in terms of number and size. We applied RMA background correction before the 2 different normalizations. In both median and quantile normalizations, RMA background correction resulted in widening the range of intensity ratio distribution, which may explain why RMA background correction plus median or quantile normalization combination identified approximately 5 times more CNVs than without RMA. Average size of CNV from RMA plus normalization combination (~20 Kb) was slightly bigger than that from normalization without RMA (~12 Kb). However, both sizes are much smaller than those identified through Affymetrix 500K EA (~80 Kb) (Redon *et al.*, 2006). RMA background correction may increase the sensitivity in detecting CNVs, but the actual sensitivity should be estimated through molecular validation of identified CNVs, e.g. MLPA.

There are several limitations in this study. First, since we focused on evaluating the effects of different normalization methods, we did not optimize the parameters for SW-Array algorithm. Second, we did not validate all the CNVs identified by 4 different normalization combinations, which means the sensitivity and specificity of these different combinations were not evaluated. It is hard to make a solid conclusion on whether RMA background correction plus median or quantile normalization is the best optimized normalization method. These limitations will be able to overcome through molecular validation. In conclusion, our results suggest that the effect of different normalization methods is relatively small compared to that of RMA background correction.

Acknowledgement

This work is supported by a grant from the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (01-PJ3-PG6-01GN09-003).

References

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.

Sebat, J., Lakshmi, B., Troge, J., *et al.* (2004). Large-scale copy number polymorphism in the human genome. *Science* 305, 525-528.

Kim, T.M., Yim, S.H., and Chung, Y.J. (2008). Copy number variations in the human genome: potential source for individual diversity and disease association studies. *Genomics & Informatics* 6, 1-7.

McCarroll, S.A., and Altshuler, D.M. (2007). Copy-number variation and association studies of human disease. *Nat. Genet.* 39, S37-S42.

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.* 10, 1787-1799.

de Smith, A.J., Tsalenko, A., Sampas, N., *et al.* (2007). Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.* 16, 2783-2794.

Perry, G.H., Ben-Dor, A., Tsalenko, A., *et al.* (2008). The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.* 82, 685-695.

Price, T.S., Regan, R., Mott, R., *et al.*, (2005) SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res.* 33, 3455-3464.