

Genome-Wide Detection of Human Copy Number Variations Using High-Density DNA Oligonucleotide Arrays

Daisuke Komura¹ **Shumpei Ishikawa**¹ **Fan Shen**²
komura@hal.rcast.u-tokyo.ac.jp shumpei@genome.rcast.u-tokyo.ac.jp Fan_Shen@affymetrix.com

Kunihiro Nishimura¹ **Sigeo Ihara**¹ **Hiroshi Nakamura**¹
kuni@genome.rcast.u-tokyo.ac.jp ihara@genome.rcast.u-tokyo.ac.jp nakamura@hal.rcast.u-tokyo.ac.jp

Keith W. Jones² **Jing Huang**² **Hiroyuki Aburatani**¹
kjones@affymetrix.com Jing_Huang@affymetrix.com haburata-tky@umin.ac.jp

¹ Research Center for Advanced Science and Technology, The University of Tokyo.
4-6-1 Komaba, Meguro, Tokyo 153-8904, Japan

² Affymetrix, Inc. 2420 Central Express way, Santa Clara CA 95051, USA

Keywords: copy number variations, SNP genotyping microarray

1 Introduction

Recent studies report that copy number variations (CNVs) may be very common events in the human genome and relevant to various diseases[1]. Among a number of molecular techniques for genome-wide CNV detection, high density SNP genotyping microarrays, which have been used for detection of genome-wide DNA copy number changes in tumor cells as well as SNP genotyping, have been paid attention because of its potential of detecting small CNVs that cannot be detected by BAC-based array CGH and providing information about two kinds of genomic variations (SNP and CNV) simultaneously in a single experiment.

However, the algorithms for the detection of copy number changes in tumor cells cannot be directly applied to CNV analysis since the use of a single reference sample is limited by the inability to explicitly determine whether a signal intensity ratio change between a test and reference sample is due to a gain in the test sample versus a loss in the reference sample or is due to a loss in the test sample versus a gain in the reference sample. Moreover, direct comparison of signals between two individuals with different SNP genotypes causes significant skews, which reduce the detection power, because of differences in the allele-specific probe affinities.

2 Method

To overcome these problems, we have developed an algorithm for CNV detection using probe intensity information from the GeneChip Human Mapping 500K Early Access arrays[2]. The algorithm 1) reduces the skews derived from the probe affinity differences by estimating and correcting them using signal intensity information from multiple samples, and 2) estimates absolute copy numbers by incorporating probe intensity and SNP information from multiple reference samples. The algorithm consists of three major parts, intensity pre-processing, CNV detection and copy number inference. Intensity pre-processing step includes probe selection, noise reduction, normalization. The affinity differences are estimated here by Gaussian mixture clustering. In CNV detection step, pair-wise comparisons of probe intensities for all possible pairs of samples are carried out and then merged to extract

candidate CNV regions for each sample. The copy number inference step utilizes signal ratios and SNP information to more precisely define CNV boundaries and the copy number within each region. The final step uses a maximum clique algorithm to define the diploid samples for any given region based on the results from the large reference data. Through a comparison of the test sample to the diploid sub-set, precise boundaries and accurate copy number inferences can be achieved. The details of our algorithm will be shown in our poster.

3 Results and Discussion

In order to estimate the false positive rate of our algorithm, we used quantitative PCR (qPCR) and Mass Spectrometry for experimental validation and compared replicates of the same DNA sample. Experimental validation of CNVs called in three replicate experiments for two DNA samples (each compared to the HapMap reference set from 270 individuals) indicated that the average false positive rate was 2.5%. Similarly, self-self comparisons of ten HapMap samples, each done in triplicate, identified an average of 0.73 CNVs per experiment. The above data indicate that the false positive signal due to intensity variability is less than five percent, and that the reproducibility is consistently high. The CNV calling algorithm, with an optimized density threshold, was applied to 270 HapMap samples and 6,473 sample-level CNVs that led to 1203 unique CNV events (Figure 1) in total were identified. The algorithm described here offers an optimal solution by providing both SNP genotype information as well as CNV profiling in a single experiment.

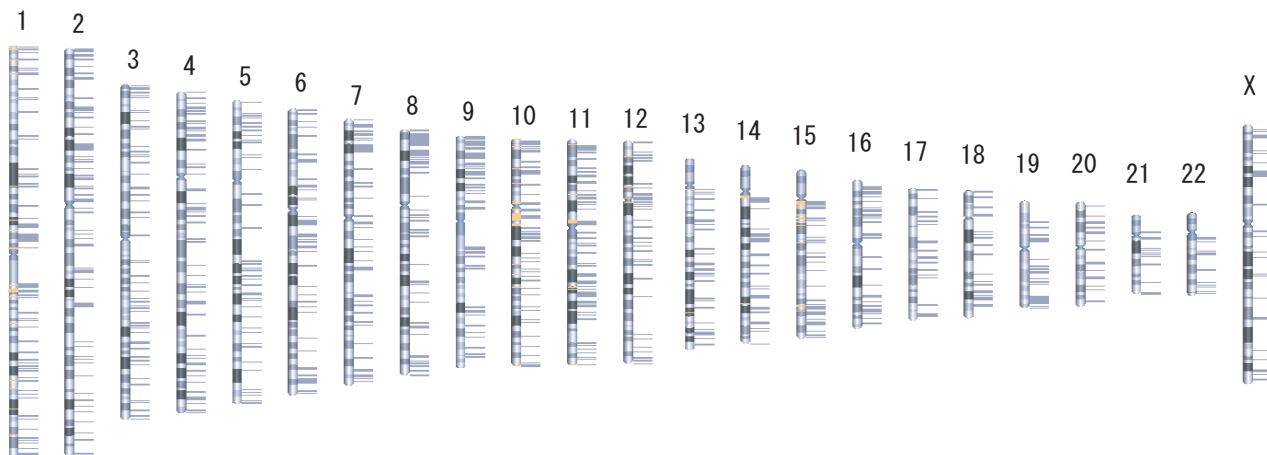


Figure 1: CNVs in Hapmap 270 samples detected by our algorithm. Horizontal bars indicate the presence of CNVs.

References

- [1] Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.W., Dallaire, S., Freeman, J., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J., Marshall, C., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Tchinda, J., Valsesia A., Yang, F., Zhang, J.J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W., and Hurles, M.E., Global variation in copy number in the human genome. *Nature*, *in press*.
- [2] Komura, D., Shen, F., Ishikawa, S., Fitch, K.R., Chen, W., Zhang, J., Liu, G., Ihara, S., Nakamura, H., Hurles, M.E., Lee, C., Scherer, S.W., Jones, K.W., Shapero, M.H., Huang, J., and Aburatani, H., Genome-wide detection of human copy number variations using high density DNA oligonucleotide arrays, *Genome Research*, *in press*.