Genomics Data 2 (2014) 55–59

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

A definitive haplotype map of structural variations determined by microarray analysis of duplicated haploid genomes

Tomoko Tahira [a,*], Koji Yahara [b], Yoji Kukita [a,c], Koichiro Higasa [a,d], Kiyoko Kato [e], Norio Wake [e], Kenshi Hayashi [a,*]

[a] Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan
[b] Biostatistics Center, Kurume University, Kurume, Japan
[c] Research Institute, Osaka Medical Center for Cancer and Cardiovascular Diseases, Osaka, Japan
[d] Center for Genomic Medicine, Kyoto University, Kyoto, Japan
[e] Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

ARTICLE INFO

Article history:
Received 11 April 2014
Accepted 11 April 2014
Available online 24 April 2014

Keywords:
Complete hydatidiform moles
Definitive haplotypes
Single nucleotide polymorphism
Copy Number Variation
LD-bin

ABSTRACT

Complete hydatidiform moles (CHMs) are tissues carrying duplicated haploid genomes derived from single sperms, and detecting copy number variations (CNVs) in CHMs is assumed to be sensitive and straightforward methods. We genotyped 108 CHM genomes using *Affymetrix SNP 6.0* (GEO#: GSE18642) and *Illumina 1 M-duo* (GEO#: GSE54948). After quality control, we obtained 84 definitive haplotype consisting of 1.7 million SNPs and 2339 CNV regions. The results are presented in the database of our web site (http://orca.gen.kyushu-u.ac.jp/cgi-bin/gbrowse/humanBuild37D4_1/).

Specifications	
Organism/cell line/tissue	Homo sapiens/complete hydatidiform moles (CHMs)
Sex	Duplicated haploids whose genomes are from single sperms harboring X
Sequencer or array type	*Affymetrix SNP 6.0* and *Illumina 1 M-duo*
Data format	*Affymetrix* Raw data: CEL files, normalized data: SOFT, MINIML and TXT *Illumina* Raw data: GSE54948_ signal_intensities.txt.gz, normalized data: SOFT, MINIML, TXT and GSE54948_ matrix_processed.txt.gz
Experimental factors	Single nucleotide polymorphism (SNP), copy number variation (CNV), LD-bin, CNV segments, CNV regions, definitive haplotypes
Experimental features	Whole genome SNP/CNV haplotyping of 84 duplicated haploid samples
Consent	All patients (donors) gave their written informed consent before study entry.
Sample source location	Japan

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18642
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54948

Experimental design, materials and methods

Samples

Complete hydatidiform mole tissues dissected from patients and the blood sample of one patient served as sources of DNAs for array hybridization experiments as described previously [1]. The informed consent was obtained from each donor. This study was approved by the Institutional Review Board (Ethical Committee of Kyushu University).

SNP genotyping

The raw data files of *Affymetrix SNP 6.0* arrays (CEL files) and sample attribute files of 94 CHM samples and one blood sample that has passed quality control in the previous study [1] were reanalyzed by *Birdseed v2* of *Geotyping Console 4. 1. 1. 834* (*GTC 4.1*), together with CEL files and sample attribute files of 45 *HapMap-JPT* samples (obtained from *Affymetrix*). The locations of markers in genome coordinate of *GRCh37* were according to *GenomeWideSNP_6.na32* that was obtained from

* Corresponding authors at: Division of Genome Analysis, Research Center for Genetic Information, Medical Institute of Bioregulation, Kyushu University, Fukuoka 812–8582, Japan. Tel.: +81 92 642 6171.
E-mail addresses: tomo.tahira@gmail.com (T. Tahira), hayashi.kenshi@gmail.com (K. Hayashi).

http://dx.doi.org/10.1016/j.gdata.2014.04.006

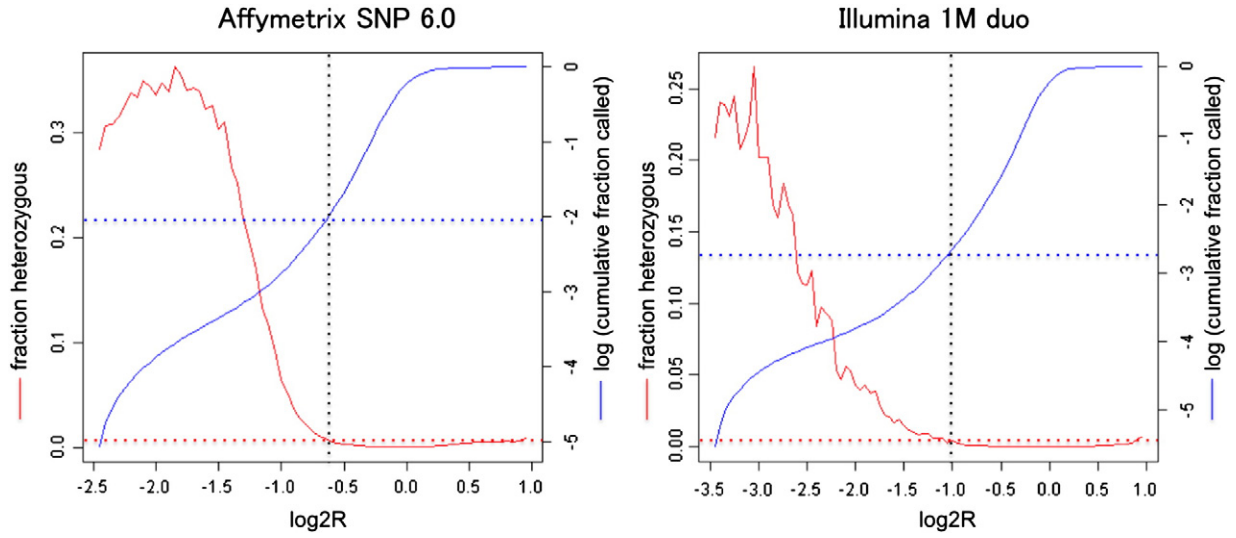

Fig. 1. Increased heterozygosity of calls at a low signal intensity. The genotype calls at the relative signal intensity where heterozygosity was approximately 1% (horizontal red dotted lines) or greater were regarded to contain significant fraction of unreliable calls. Blue horizontal lines indicate the fraction of cumulative calls at the reliability thresholds.

Affymetrix. A total of 905,025 SNP genotypes (excluding chromosome Y and mitochondria) were obtained, at an initial average call rate for the 94 CHMs of 99.2%.

Array hybridization experiments using *Illumina 1 M-duo* was performed for 98 CHM samples that included the 94 samples and one blood samples mentioned above by previously described procedures [1]. The genotypes were called using *GenTrain 2.0* cluster algorithm of *Genome Studio 2011.1*, *Illumina. Human1M-Duov3_H.egt* (based on *GRCh37*) was used as the manifest file and *Human1M-Duov3_H.bpm* as the cluster file. The initial average call rate was 99.5%.

Copy number analysis

The CEL files of *Affymetrix* arrays were subjected to *Copy Number/LOH analysis* module of *GTC 4.1* without regional GC correction. The 94 CHM samples, one blood sample mentioned above and four male samples from *HapMap JPT* (*NA18940*, *NA18943*, *NA18944* and *NA18945*) served as references to obtain "Log2Ratio" (abbreviated as log2R in this paper) data. Then, the data of markers on chromosome Y and mitochondria were excluded and the remaining data were exported as *CNCHP.txt*. The "log R Ratio" (abbreviated as logRR in this paper) data of *Illumina* arrays were calculated by *Genome Studio 2011.1* using the cluster file (*Human1M-Duov3_H.bpm*) as a reference.

Results and discussion

SNP genotyping of haploid samples

CHM genomes are supposed to be genome-widely homozygous. However, the genotypes obtained by the two systems revealed small fractions (0.27% of *Affymetrix* call and 0.01% of *Illumina* call) of heterozygous calls. The dramatic increase of heterozygous calls for the markers at lower relative signal intensities (log2R of *Affymetrix* arrays and logRR of *Illumina* arrays) indicated that the calls were falsely made for the markers at (homozygously) deleted regions where no genotypes should be called, although some of them might be ascribed to the markers in divergent paralogous regions (Fig. 1). These findings provided us an additional quality control measure of SNP genotype calling, that

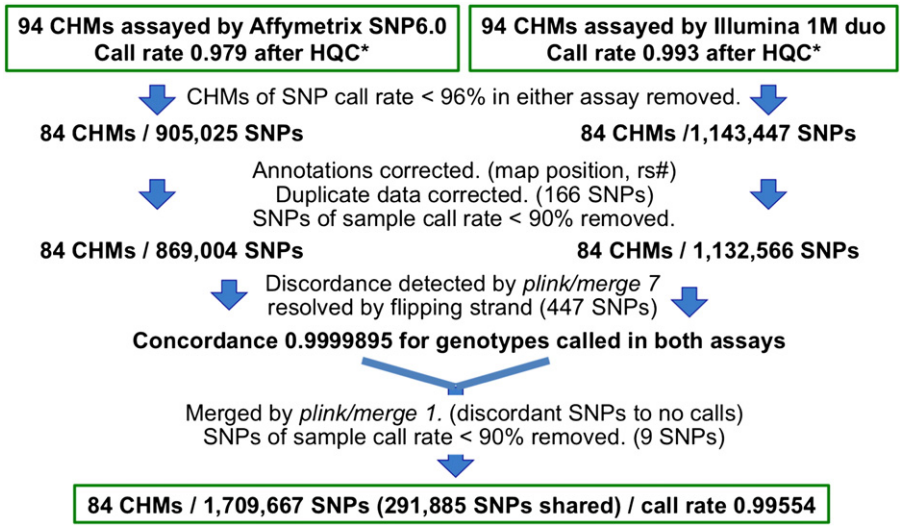

Fig. 2. Overview of SNP genotyping and its quality control. *HQC: haploid quality control, that is, heterozygous calls and weak signal calls were forced to no calls. See text for detail.