Open Access

Genome-Wide Analysis of Copy Number Variations in Normal Population Identified by SNP Arrays

Jian Wang[1,2], Tsz-Kwong Man[1,3,4], Kwong Kwok Wong[1,†], Pulivarthi H. Rao[1,3,4],
Hon-Chiu Eastwood Leung[1,3,4,5], Rudy Guerra[6] and Ching C. Lau[*,1,2,3,4]

[1]Texas Children's Cancer Center and Hematology Service, Texas Children's Hospital, [2]Program in Structural and Computational Biology and Molecular Biophysics, [3]Dan L. Duncan Cancer Center, [4]Department of Pediatrics, [5]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas, 77030, USA; [6]Department of Statistics, Rice University, Houston, TX 77004, USA

[†]Present address: Department of Gynecologic Oncology, M.D. Anderson Cancer Center, Houston, Texas 77030, USA

Abstract: Gene copy number change is an essential characteristic of many types of cancer. However, it is important to distinguish copy number variation (CNV) in the human genome of normal individuals from bona fide abnormal copy number changes of genes specific to cancers. Based on Affymetrix 50K single nucleotide polymorphism (SNP) array data, we identified genome-wide copy number variations among 104 normal subjects from three ethnic groups that were used in the HapMap project. Our analysis revealed 155 CNV regions, of which 37% were gains and 63% were losses. About 21% (30) of the CNV regions are concordant with earlier reports. These 155 CNV regions are located on more than 100 cytobands across all 23 chromosomes. The CNVs range from 68bp to 18 Mb in length, with a median length of 86 Kb. Eight CNV regions were selected for validation by quantitative PCR. Analysis of genomic sequences within and adjacent to CNVs suggests that repetitive sequences such as long interspersed nuclear elements (LINEs) and long terminal repeats (LTRs) may play a role in the origin of CNVs by facilitating non-allelic homologous recombination. Thirty-two percent of the CNVs identified in this study are associated with segmental duplications. CNVs were not preferentially enriched in gene-encoding regions. Among the 364 genes that are completely encompassed by these 155 CNVs, genes related to olfactory sensory, chemical stimulus, and other physiological responses are significantly enriched. A statistical analysis of CNVs by ethnic group revealed distinct patterns regarding the CNV location and gain-to-loss ratio. The CNVs reported here will help build a more comprehensive map of genomic variations in the human genome and facilitate the differentiation between copy number variation and somatic changes in cancers. The potential roles of certain repeat elements in CNV formation, as corroborated by other studies, shed light on the origin of CNVs and will improve our understanding of the mechanisms of genomic rearrangements in the human genome.

Keywords: Copy number variation, SNP arrays, genomic variation.

INTRODUCTION

Copy number aberrations have long been known to be associated with cancer. Amplification of oncogenes, such as MYC [1] and ERBB2 [2], as well as homozygous deletion of tumor suppressor genes, such as RB1 [3] and p16 [4] have been shown to contribute to the malignant phenotype. Although gene copy number changes in tumor cells are involved in tumorigenesis, copy number variations also exist in the normal population, as discussed in several recent papers [5-9]. Therefore, it is imperative to distinguish normal copy number variations (CNVs) from aberrant somatic copy number changes that are found in cancers. In addition, identification of CNVs will lead to a better understanding of genetic variation among the normal population and thus provide clues to genome evolution. Correlation of large-scale CNVs with disease phenotype and clinical outcome may also help elucidate the genetic predisposition to cancer and other diseases and could also potentially be used as prognostic markers.

Genome-wide CNVs have been identified using a number of methods. Using the Representational Oligonucleotide Microarray Analysis (ROMA) method that utilizes an oligonucleotide microarray containing 70-mer probes specifically designed for hybridization with Bgl II fragments, Sebat *et al.* [5] identified 76 unique CNVs among 20 individuals with the requirement of at least 3 consecutive probes showing copy number change. Another study using bacterial artificial chromosome (BAC) array comparative genomic hybridization (aCGH) revealed 255 genomic clones with gains/losses among 55 individuals, each of which was regarded as a CNV [6]. While these two studies reported variations of both copy number gains and losses, two other studies have focused on deletion variations in the human genome. Comparison of the sequence of a second genome represented by fosmid end sequences with the reference genome sequence identified 102 deletion sites, where the fosmid pairs span at least 3 standard deviations more than the mean fosmid size [8]. McCarroll *et al.* [7] used SNP genotyping data from the HapMap project to identify segregating deletion variants that harbor physically clustered patterns of null genotypes, apparent Mendelian inconsistencies, and apparent Hardy-Weinberg disequilibrium. In this

*Address correspondence to this author at the Texas Children's Hospital, 6621 Fannin St. MC 3-3320, Houston, TX 77030, USA; Tel: (832) 824-4543; E-mail: cclau@txccc.org

study, at least two SNPs were involved in the aberrant genotyping call in order to define such a segregating deletion variant. These investigators discovered 541 such deletion variants among 269 subjects. Redon *et al.* [9] used a combination of 500K SNP array and BAC aCGH to identify a total of 1447 CNV regions among 270 individuals. For the BAC aCGH data, they defined CNVs as single clones with Log2 hybridization ratios being 6 standard deviations (S.D.) from those of the normal reference and consecutive clones with Log2 hybridization ratios being 4 S.D. from the reference. For the SNP array data, CNVs were only called in regions with at least 4 SNPs and 3 restriction fragments. The CNV regions identified from these two methods were subsequently merged.

In the current study, we utilized the Affymetrix 50K *XbaI* SNP array data generated from 104 healthy individuals comprising of three ethnic groups to carry out CNV analysis in the normal population. Instead of comparing the hybridization signal of the test sample with a reference individual as in the ROMA [5] and BAC aCGH [6] methods, our analysis derived the copy number call for each SNP by comparing hybridization signal of the test sample with that of the mean signal from all test samples. Therefore, our analysis is capable of deriving the copy number more accurately and thus better reflects the variation of gain and loss in the normal population.

MATERIALS AND METHODS

SNP Array Data

The CEL, CHP and DAT files of 50K *XbaI* SNP array data from 104 normal individuals were obtained from Affymetrix (Santa Clara, CA) and they were part of the reference data for the Chromosome Copy Number Tool (CCNT, http://www.affymetrix.com/support/developer/tools/affytools.affx). Reference data .CEL files were accessed through the Affymetrix File SDKs [http://www.affymetrix.com/support/developer/filesdk/GDACFiles/Pages/GDACCELFile.affx]. SNP files were subsequently generated using the Affymetrix GTYPE software. These 104 individuals from the Coriell Cell Repository [10], consisting of 46 males and 58 females, are from three ethnic groups: 20 Asians, 42 African Americans, and 42 Caucasians. Physical positions of the SNPs are based on the May 2004 human genome assembly (UCSC Genome browser, http://genome.ucsc.edu/).

Derivation of SNP Copy Numbers Using dChip

The *dChip* software (Version: 07/31/06) [11] was used to derive the copy number for each SNP, according to the standard procedure for copy number analysis. All samples were used to compute the mean and variation of the signal for 2 copies, using the "invariant set normalization" option to normalize hybridization signals and the PM/MM difference model to calculate model-based expression values. The raw copy number for a SNP in a sample was calculated as 2* Normalized hybridization signal intensity / mean signal intensity of normal sample set of that SNP. Subsequently an array list file was created with single samples separated by "Standardize separators" so that the SNP signal of each sample will be compared with the average of all samples in order to compute the copy number. Finally, a Hidden Markov Model (HMM) with default parameter settings was used to infer the copy number for each SNP.

Identification of CNV Regions

To identify candidate CNV regions, we designed an algorithm that examines the SNP copy numbers along each chromosome in each sample. A candidate CNV region is defined as a chromosomal region containing at least 2 contiguous SNPs with inferred copy numbers changed in the same direction, with no threshold for the minimum required length. A paired t-test was employed to assess the significance of each CNV by comparing the mean raw copy number of SNPs in the CNV region between the sample and the reference set. Insignificant ($P>0.05$) CNVs were filtered out. CNVs were identified from each sample using this algorithm. The boundaries of CNVs from different individuals were then compared and those CNVs with the same start and end chromosomal locations were combined into a single CNV region. Those CNVs from different individuals that overlap with each other were broken into non-overlapping CNV regions. These CNV regions were combined with the other CNVs with no overlaps to form the final list of unique CNV regions.

Comparison between CNVs Identified in this Study with Previous Studies

CNVs from Sebat *et al.* [5], Iafrate *et al.* [6], McCarroll *et al.* [7] and Tuzun *et al.* [8] were downloaded from the Eichler lab webpage (http://humanparalogy.gs.washington.edu/structuralvariation/). The Liftover program at http://genome.ucsc.edu/cgi-bin/hgLiftOver was used to convert the physical positions of CNVs in Sebat *et al.* [5], Iafrate *et al.* [6] and McCarroll *et al.* [7] to May 2004 genome assembly version from the July 2003 Geome Assembly. CNVs from Redon *et al.* [9] were obtained from the Database of Genomic Variants [http://projects.tcag.ca/variation/] [6]. We then compared the CNVs from these previous studies that have at least one SNP on the Affymetrix 50K SNP array with the CNVs identified in this study.

Quantitative PCR

Quantitative PCR (qPCR) was performed using the Taqman assay on an ABI Prism 7000 Sequence Detection System (Applied Biosystems, Foster City, CA, USA) using 96-well optical plates. DNA samples were ordered from the Coriell Cell Repository [10]. Primers and Taqman probes for each region to be validated were designed using the Primer Express Software (v. 3.0, Applied Biosystems). The sequences of primer/probe are shown in Supplemental Table **1**. Each DNA sample was run in triplicates for each of the 3 serial dilutions.

For each of the CNVs chosen for validation, three regions were tested for the copy number by qPCR: a left region spanning 1 Kb downstream from the leftmost SNP in that CNV, a right region spanning 1 Kb upstream from the rightmost SNP in that CNV, and a middle region that is either an exon of a gene within the CNV or spans a 1 Kb sequence encompassing a third SNP in the CNV. Because there are only 2 SNPs in "others CNV 2", only a middle region delimited by the SNPs in that CNV was selected to validate its copy number.

DNA from another sample (NA17261) in the Coriell Cell Repository was utilized to draw standard curves for most of the CNV regions to be validated, since the SNP array data indicated that those regions are diploid in this sample. For the CNV 114 right region and the CNV 119 left region, another sample (NA17059) which was shown by SNP array data as diploid in these two regions was used to generate the standard curves. Similarly, sample NA17254 was used to generate the standard curves for CNV 57.

Analysis of Gene Content in CNVs

RefSeq annotated genes that are fully encompassed in the CNVs were identified. The GOTree Machine [12], a web-based application capable of interpreting interesting gene sets using Gene Ontology was used to identify those gene categories that are over-represented in the list of genes that are fully encompassed in the CNVs.

Analysis of Repeat Elements

To investigate the content of repeat elements in CNVs and their flanking regions, genomic sequences that encompass CNVs and CNV-adjacent regions were identified from the May 2004 assembly of human genome sequence. To identify repeat elements among these genomic sequences, RepeatMasker [13] [version open-3.1.5 in default mode] was run with blastp (version 2.0MP-WashU) against RepBase (Update 20060314). The percentage of repeat elements was calculated as ratio between the number of bases from each category of the repeat elements and the total number of bases in the genomic sequence.

To estimate the content of repeats in the whole genome, the RepeatMasker .out files (May 2004 human genome assembly, downloaded from UCSC Genome Browser at: http://hgdownload.cse.ucsc.edu/goldenPath/hg17/bigZips/) for chromosomes were used. In order to compare the percentage of repeat elements in CNV-flanking regions with that in the rest of the genome, a permutation test was performed by randomly selecting the same number of genomic regions, which are not in the CNV-flanking regions, as the total number of regions flanking the CNVs. We then calculated the percentage of repeat elements in these randomly-selected regions and performed a t-test to compare the percentage of repeat elements between these two groups of genomic sequences. This randomization process was repeated 10 times. The average percentages of repeat

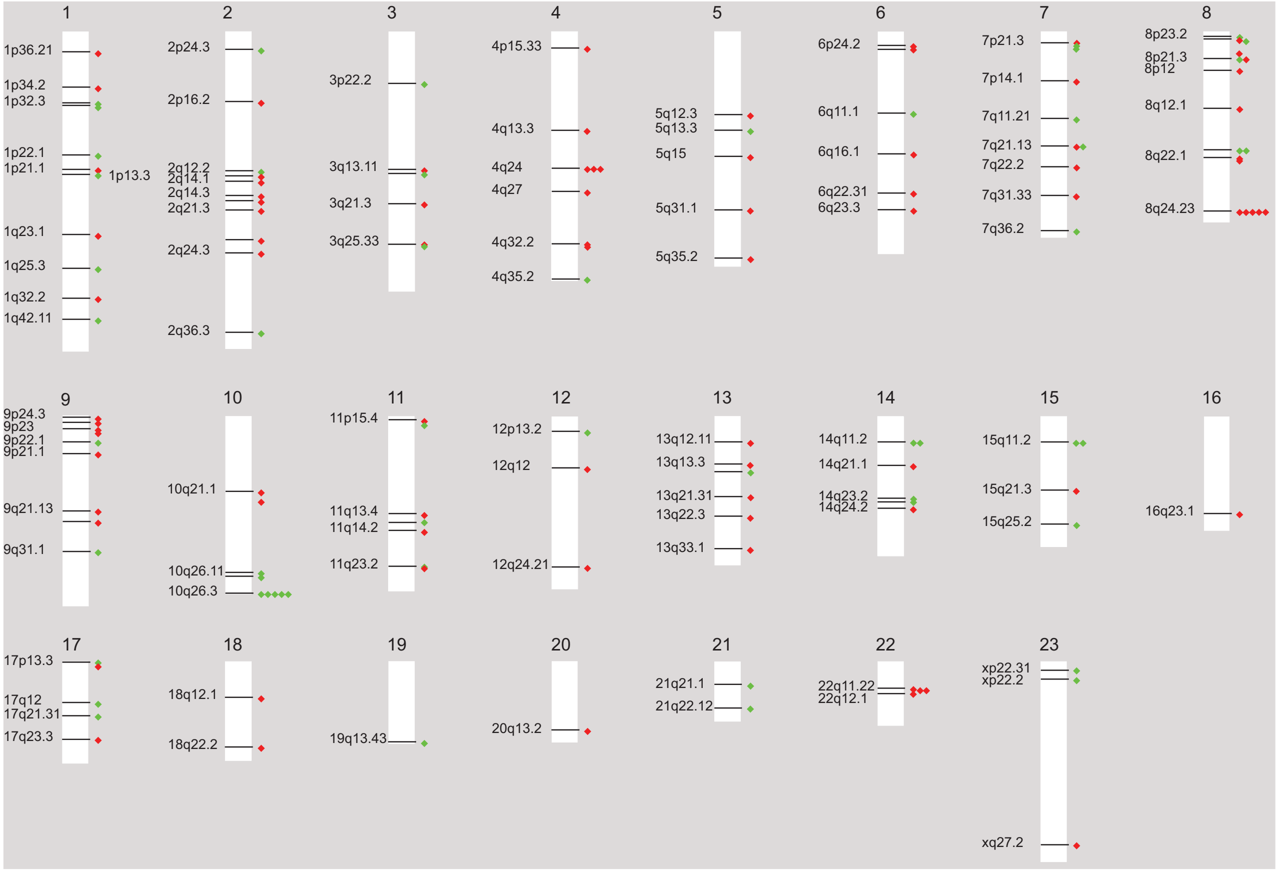


Fig. (1). Distribution of CNVs in the genome. A total of 155 significant CNVs based on the t-test are shown. The start positions of CNVs are depicted as green or red dots, representing gains or losses, respectively, along the chromosome with the cytoband position specified. The number of dots indicates the number of individuals in which the CNV was identified. Results are based on 104 individuals.

elements and P values from the 10 repetitions were subsequently calculated and used to assess the significance level.

RESULTS

Summary of CNVs Identified by our Analysis

The distribution of CNVs among the 23 chromosomes identified by our analysis is shown in Fig. (**1**). A total of 155 CNVs were identified. If any of these 155 CNVs from different individuals overlap, overlapping CNVs would be broken up into non-overlapping genomic regions that we defined as unique CNVs. This procedure facilitated the frequency count of copy number variation events we identified among our sample set. Using this procedure, the original 155 CNVs were reduced to 143 unique CNVs (Supplemental Table **2**), distributed on more than 100 cytobands across all 23 chromosomes. The difference in terms of the number of CNVs (averaged by the number of individuals) among different chromosomes is statistically significant (one-way ANOVA: P = 0.03627) with the number of CNVs in Chr22 being markedly higher than the other chromosomes (Fig. **2**).

About 87% of the unique CNVs were seen only once in the cohort (Table **1**). On average, there are 1.5 (155/104) CNVs per individual. About 37% of the CNVs are gains while 63% are losses. The size of CNV regions based on the 50K Affymetrix array ranges from 68 bp to 18 Mb, with a median length of about 86 Kb and an average length of 421 Kb. The size distribution of CNV regions falls into various ranges, with the largest fraction (32%, combining gain and loss) of CNVs between 100 Kb and 500 Kb (Table **2**). The number of SNPs with CN ≠ 2 present in each CNV region is shown in Fig. (**3**). About 97% of the CNV regions were identified by 3 or more SNPs (CN ≠ 2), and more than half of the CNV regions contain at least 6 SNPs. As we are using inferred SNP copy numbers to identify CNVs, the more non-diploid SNPs present in a CNV, the less likely the CNV is an artifact. To better visualize the locations of SNPs with aberrant copy numbers and the CNV regions in the genome,

Table 1. Frequency of the CNVs detected

	CNV Frequency
CNVs present in 1 individual	125 [87.4%]
CNVs present in 2 individuals	13 [9.1%]
CNVs present in 3 individuals	2 [1.4%]
CNVs present in > 3 individuals	3 [2.1%]

Table 2. Size distribution of CNVs, 37% of which are gains and 63% are losses. The median size of CNVs is ~ 86 Kb

Size Range [Kb]	Number of CNVs [Percentage]	
	Gain	Loss
[0,1]	3 [1.9%]	14 [9.0%]
[1,10]	8 [5.2%]	14 [9.0%]
[10,50]	7 [4.5%]	15 [9.7%]
[50,100]	6 [3.9%]	17 [11.0%]
[100,500]	21 [13.5%]	28 [18.1%]
[500,1000]	7 [4.5%]	4 [2.6%]
>1000	5 [3.2%]	6 [3.9%]
Total	57 [36.8%]	98 [63.2%]

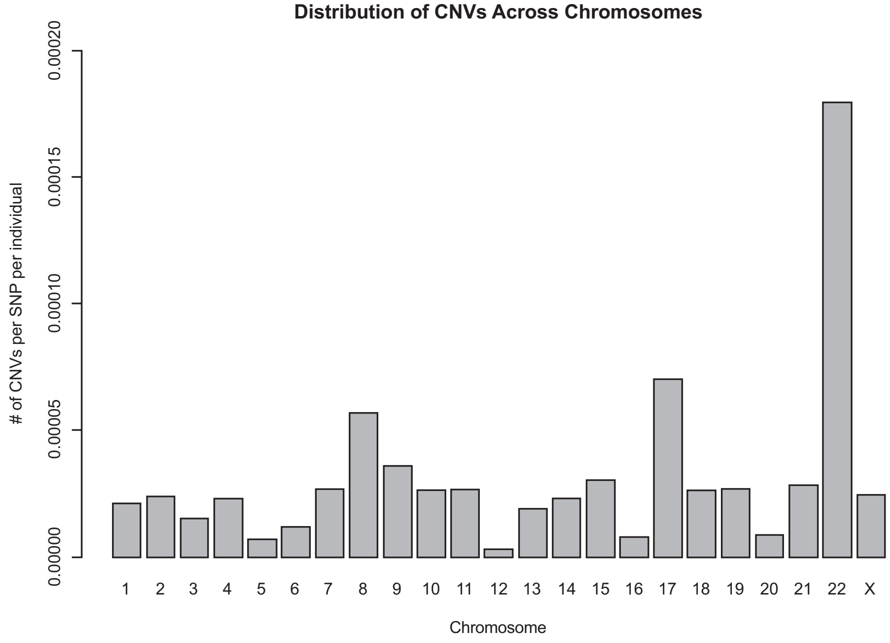


Fig. (2). Number of CNVs per SNPs (averaged by the number of individuals) for each chromosome. Shown here is the average number of CNVs per individual divided by the number of SNPs for each chromosome. A one-way ANOVA test indicated that CNVs are not evenly distributed among chromosomes (P = 0.03627).

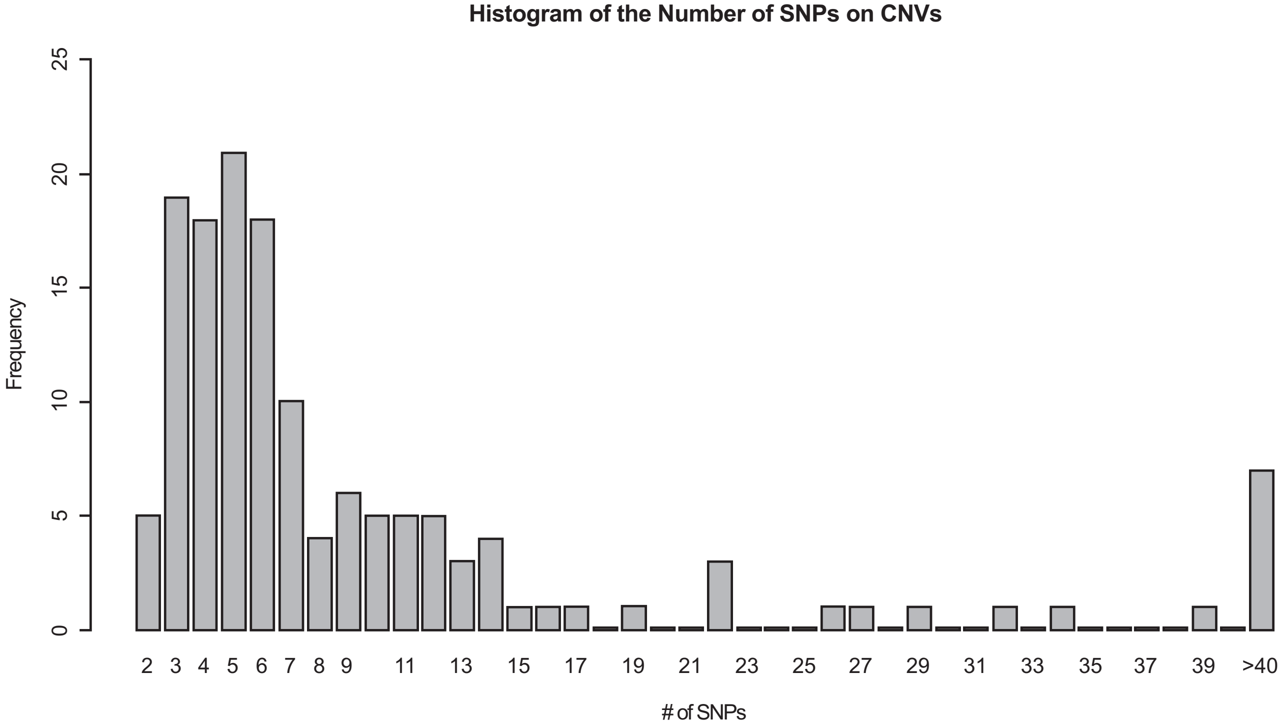


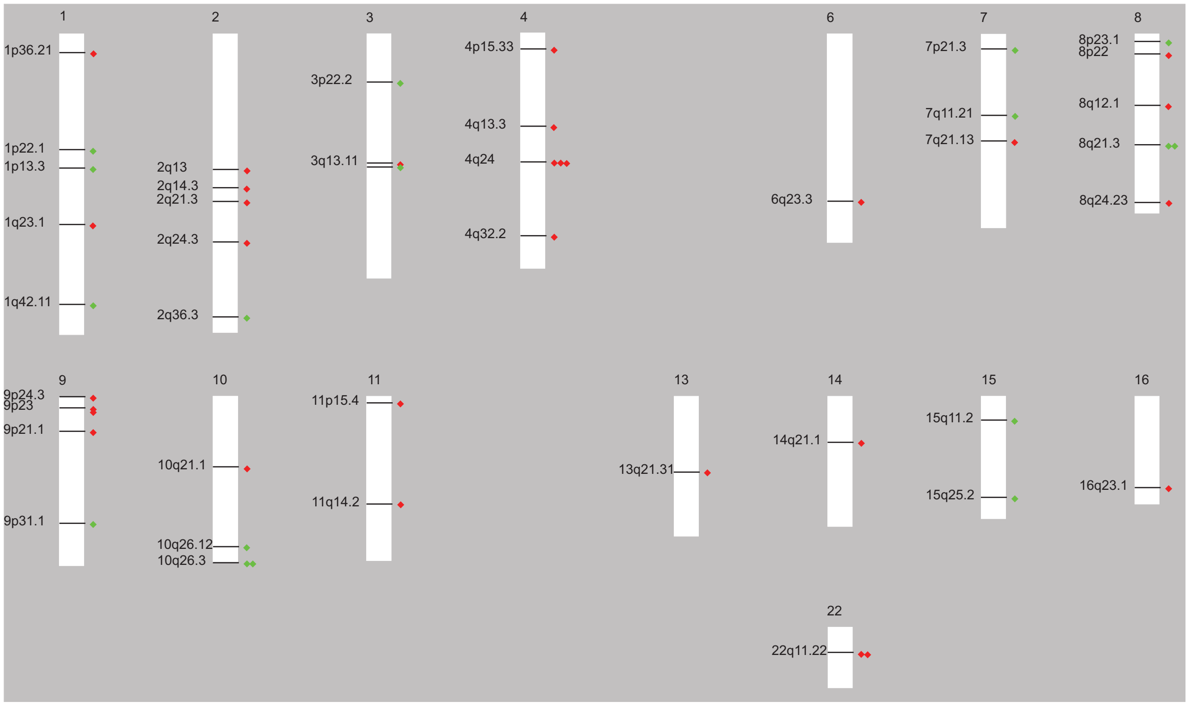
Fig. (3). Histogram of the number of SNPs in each CNV region (143 in total).

we have built a database in Genboree [14] for the CNVs we have identified (Group: jwang_group; Database: CNV using *dChip*).

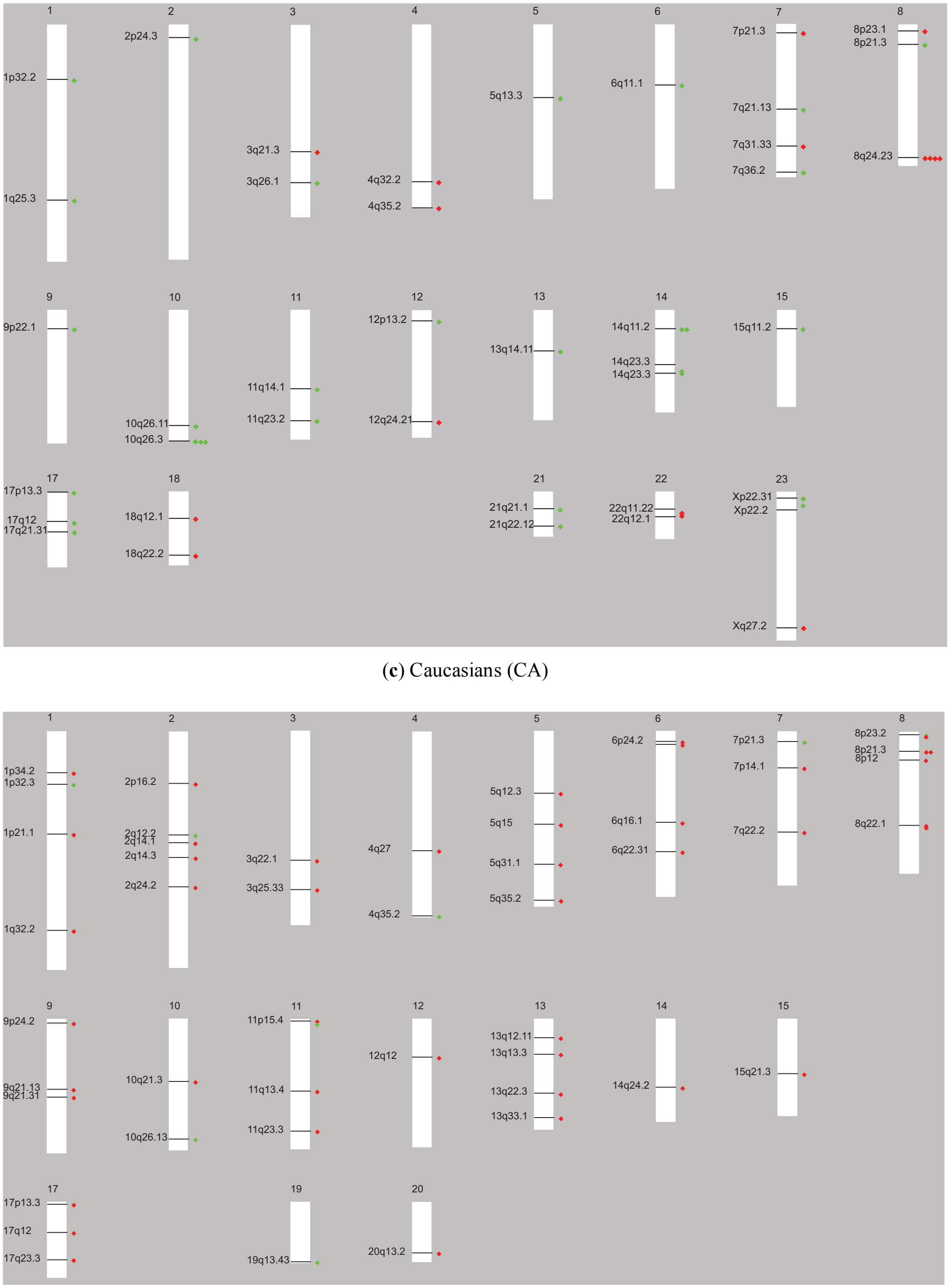
Comparison of CNVs among Different Ethnic Groups

The individuals included in our CNV analysis are from three ethnic groups: Asians (AS), African Americans (AA) and Caucasians (CA). Fig. (**4**) shows that CNVs from these three ethnic groups display different patterns of locations in the genome. Table **3** shows the number of CNVs, divided into gains and losses, from these three ethnic groups. Although the number of assayed individuals from the AS group [24] is only about half of that from the AA group [42] or the CA group [42], the total number of CNVs from these

(**a**) Asians (AS)



(Fig. 4) contd…..



(**b**) African Americans (AF)

(**c**) Caucasians (CA)

Fig. (4). Whole-genome view of CNVs identified from various ethnic groups (**a**) Asians (n=24) (**b**) African Americans (n=42) (**c**) Caucasians (n=42). The start positions of CNVs are depicted as green or red dots, representing gains or losses, respectively, along the chromosome with the cytoband position specified. The number of dots indicates the number of individuals in which the CNV was identified.

three ethnic groups are nearly the same (56 in AS, 50 in AA and 49 in CA). However, the distribution of CNVs by gains and losses varies among these three groups. There are many more losses in the AS (8 gains *vs*. 48 losses) and AA (17 gains *vs*. 33 losses) group, while there are many more gains than losses in the CA (32 gains *vs*. 17 losses) group. Chi-square tests confirmed that the distribution of CNVs with gain/loss in the CA group is significantly different from those in the AS group (p-value = 2.348e-07) and the AF group (p-value = 0.0035). Although both the AS and the AF groups have more losses than gains, the CNV gain/loss profile between these two groups are still significantly different (p-value = 0.03).

Table 3. CNVs of Gains/Losses among Different Ethnic Groups. AS: Asians; AF: African Americans; CA: Caucasians

	AS	AF	CA
Gain [%]	8 [14%]	17 [34%]	32 [65%]
Loss [%]	48 [86%]	33 [66%]	17 [35%]
Total	56	50	49

Comparison of CNVs with other Studies

We compared the CNVs identified in this study with those identified by five previous studies using various methods: ROMA [5], BAC aCGH [6], SNP genotyping [7], Fosmid end reads [8] and a combination of SNP array and BAC aCGH [9].

Based on the consideration of how CNVs from different methods were defined and in reference to the concordance analysis performed by McCarroll *et al.* [7], we set up the following rules to examine concordance between our CNVs and those from the other papers: (1) When we compared the concordance between our CNVs and those from ROMA [5], BAC aCGH [6] or SNP genotyping [7], we consider our CNV to be concordant with other studies if it overlaps with another CNV from these analyses and the overlapped region covers at least 20% of both CNV from this work and CNV from another study; (2) When we compared the concordance between our CNVs and those from the fosmid end reads method, we consider our CNV to be concordant with their CNV only if our CNV completely falls within a CNV from their study; (3) We only compared our CNVs with CNVs from others that have at least one SNP from the 50K *XbaI* SNP array, as this study is unable to identify those CNVs with no SNP representation on our platform. Using these criteria, we found: 8 CNVs identified in our analysis are concordant with 7 [out of 45] CNVs by ROMA; 2 CNVs concordant with 2 (out of 190) CNVs by BAC aCGH; 4 CNVs concordant with 4 (out of 80) CNVs by the SNP genotyping method; 22 CNVs concordant with 21 (out of 979) CNVs by the combined method of SNP array and BAC aCGH (Table **4**). No concordance between our analysis and the fosmid end reads analysis was identified. In total, 30 (21%) CNVs (non-redundant set) from our analysis are concordant with CNVs from these five previous studies.

qPCR Validation of CNVs

Given the large number of CNVs identified in this study, we have chosen to validate by quantitative PCR (qPCR) a subset that includes CNV gains, losses, and some discovered in previous studies, as well as novel CNVs identified in this study. We have validated 6 of the CNVs uncovered in this study. Three of these CNVs overlap with a CNV identified from a previous study [5], while three of them were not concordant with previous studies [5-8]. We have also performed qPCR on two additional Coriell samples in 2 CNV regions that were identified in two other studies [5,6] but were not found in this study. The CNV regions chosen for validation are summarized in Table **5**.

Results of the CNV validation by qPCR are shown in Fig. (**5**). In all of the 6 CNVs uncovered in this study, the copy number fold change from qPCR was consistent with that from the SNP array. For those CNV regions of gain identified by SNP array, qPCR results also indicated significant copy number increase (P (CN > 2) < 0.05). The exceptions were the left region of CNV 119: [P (CN>2) = 0.06], which still showed a strong trend of gain, as well as the middle and right regions of CNV 57, where SNP coverage is sparse compared with the left region of this CNV. For those CNVs of loss, qPCR results always showed

Table 4. Comparison between CNVs identified in this study with those identified from other studies. The numbers of CNVs from this study that are concordant with the studies from Sebat *et al.* 2004 [5], Iafrate *et al.* 2004 [6], Tuzun *et al.* 2005 [8], McCarroll *et al.* 2006 [7] and Redon *et al.* 2006 [9] are: 8, 2, 0, 4 and 22, respectively.
***: A CNV from this work is concordant with a CNV from another study (except for Tuzun *et al.* [8]) if they overlap with each other and the overlapped region is at least 20% of the length of both CNVs. A CNV from this work is concordant with a CNV from Tuzun *et al.* [8] if the former CNV physically falls within the latter one**

Method	Reference	# of Individuals Assayed	CNVs Discovered	CNV Type	CNVs Containing ≥ 1 SNP from the 50K *XbaI* SNP Array	CNVs Concordant with this Study *
ROMA	Sebat *et al.* [5]	20	76	Gains and loss [relative]	45	7
BAC array CGH	Iafrate *et al.* [6]	55	255	Gains and loss [relative]	190	2
Fosmid end reads	Tuzun *et al.* [8]	1	101	Loss	27	0
SNP genotyping	McCarroll *et al.* [7]	269	540	Loss	80	4
SNP array and BAC array CGH	Redon *et al.* [9]	270	1447	Gain and loss	979	21
SNP array	This study	104	143	Gain and loss		

Table 5. CNVs validated by qPCR. CNVs 19, 57, 77, 114, 119 and 138-139 were identified in this study and 3 of them overlap with CNVs identified by the ROMA method [5]. Others CNVs 1-2 were not present among the individuals in this study but were identified in other studies. CNV 114 was identified by ROMA [5] as gain while CNVs 19 and 138-139 were identified as loss by ROMA [5]. DNA from two individuals (NA17260 and NA17269) in our sample set was used to validate others CNVs 1-2 in these individuals as diploid
a: CNVs identified by ROMA [5]
b: CNVs identified by BAC aCGH [6]

CNV from this Study	Cytoband	Copy Number from SNP Array	First_SNP	Last_SNP	Individual	Size [Kb]	# of SNPs	CNV Type in ROMA[a]	Proportion Overlapping with ROMA[a]
CNV 19	2q14.3-2q21.1	1	SNP_A-1648306	SNP_A-1669467	NA17017	1156	22	Loss	100%
CNV 57	7q21.13	3	SNP_A-1664228	SNP_A-1697587	NA17260	1909	75	N/A	0%
CNV 77	8q24.23	1	SNP_A-1658092	SNP_A-1742931	NA17253	72	12	N/A	0%
CNV 114	14q11.2	5	SNP_A-1745948	SNP_A-1710646	NA17254	205	4	Gain	90%
CNV 119	15q11.2	4	SNP_A-1713638	SNP_A-1729906	NA17104	781	5	N/A	0%
CNV 138-139	22q11.22	0	SNP_A-1708192	SNP_A-1724056	NA17160	392	7	Loss	90%
CNV found by other Studies but Not in this Study	**Cytoband**	**Copy Number from SNP Array**	**First_SNP**	**Last_SNP**	**Individual**	**Size [Kb]**	**# of SNPs**	**CNV Type in ROMA**	**CNV Type in BAC aCGH[b]**
Others CNV 1	4q22.3	2	SNP_A-1689701	SNP_A-1702935	NA17260	677	17	Gain	N/A
Others CNV 2	7q35	2	SNP_A-1694209	SNP_A-1736287	NA17269	14	2	Loss	Loss

significant copy number decrease ($P\ (CN < 2) < 0.05$). For the 2 CNVs identified in previous studies but shown in our samples as diploid, qPCR results also indicated no copy number change ($P\ (CN \neq 2) > 0.05$). In conclusion, our qPCR results were consistent with the copy number calls from SNP array and established the validity of these CNVs.

Gene Content in CNVs

The biological significance underlying the presence of CNVs is largely unknown. We have examined the gene content in our 143 unique CNVs and found that 84 of those CNVs encompass or overlap with at least one refSeq annotated gene. We took a similar approach to what was employed by Graubert *et al.* [15] in order to examine whether CNVs are more likely to be located in coding regions than in non-coding regions. We generated 143 pseudo-CNVs, which were randomly picked non-overlapping genomic regions based on the length distribution of those 143 CNVs identified in this study. We generated 1000 sets of 143 pseudo-CNVs from 1000 randomizations, based on which the probability of 84 pseudo-CNVs encompass or overlap with at least one refSeq annotated gene was 0.34. This result indicates that there is no significant enrichment of genes in our CNV regions, compared with the genome.

Next, we asked if there are any categories of genes that are enriched in CNVs. Redon *et al.* [9] dissected the functional categories of genes that overlap with their 1444 CNV regions. They found that cell adhesion, sensory perception of smell and of chemical stimulus, and neurophysiological processes were among those enriched Gene Ontology (GO) categories [9]. There are a total of 432 refSeq annotated genes that are encompassed by or overlap with CNVs identified in our study. GO analysis of those 364 genes that are encompassed by CNVs indicated that 50 GO categories were significantly enriched based on a hypergeometric test (Supplemental Table **3**). Similar to the results of Redon *et al.* [9], our analysis also indicated that genes related to olfactory sensory, chemical stimulus, and other physiological responses were highly enriched in CNVs. Separate analyses on CNVs with gain and loss revealed that these GO categories were also enriched in either type of CNVs (data not shown). The enrichment of these gene categories in CNVs suggests that the natural selection process may promote the fluctuation of copy number of these genes, as one of the potential mechanisms, to cope with various external conditions and stimuli.

Comparison of CNVs in the Human Genome with those in the Mouse Genome

CNVs have been identified not only in humans, but in other mammals, including mice [15-17] and chimpanzees [18]. We have compared the mouse CNVs identified by Graubert *et al.* [15] with our CNVs (using the Liftover program at http://genome.ucsc.edu/cgi-bin/hgLiftOver to convert genomic coordinates) and found only one mouse CNV overlapping with a CNV from this study. Graubert *et al.* [15] also found no significant enrichment of genes in the mouse CNVs. We have also compared the enriched GO

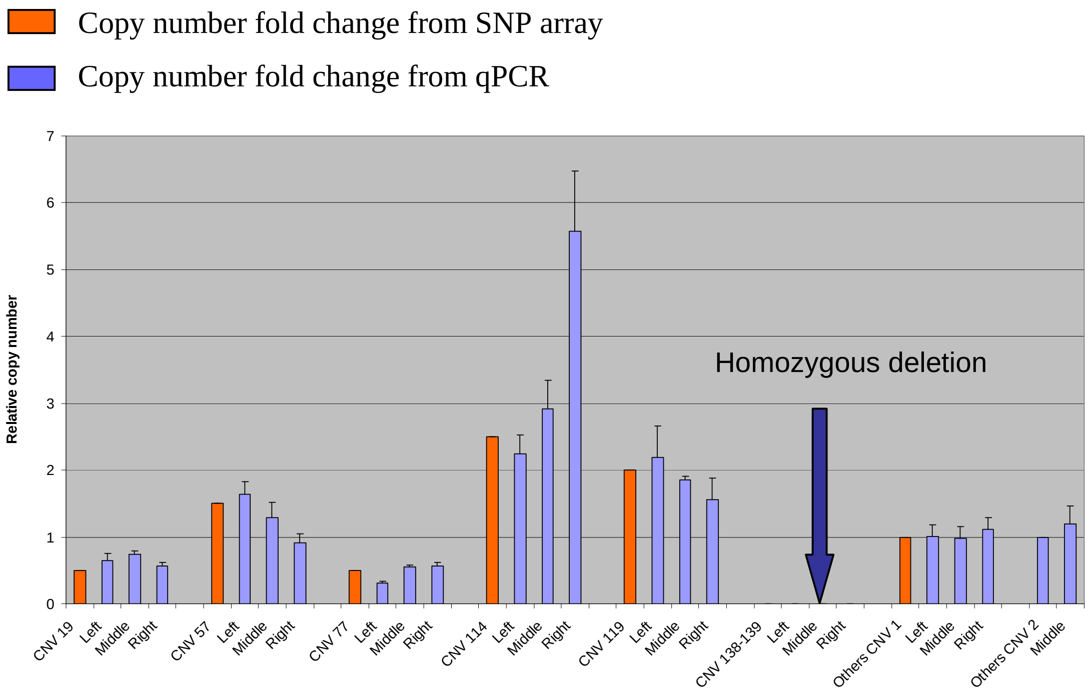


Fig. (5). qPCR validation of CNV copy numbers identified by SNP array. For each CNV, the relative copy number (CN/2) based on SNP array is shown, followed by qPCR result for each region (left, middle, or right) in that CNV. The average and standard error of the relative copy number from 3 serial dilutions for each sample are shown. Standard curves were generated by running qPCR with the primer/probe for each test region in NA17261, except for the CNV 114 right region and the CNV 119 left region, where NA17059 was used, and for CNV 57, for which NA17254 was used. For CNV 138-139, SNP array data indicate homozygous deletion, which was validated by qPCR in all three regions, where the fluorophore signal from the Taqman probe was so weak that the threshold cycle (CT) could not be determined.

categories between mouse CNVs and human CNVs identified in this study. There are several GO categories that are overrepresented in both the mouse CNVs [15] and CNVs in this study including olfactory receptor activity, G-protein-coupled receptor activity and response to stimulus. Conservation of these genes associated with CNVs in mice and humans supports the notion that these genes may be under selection pressure [15] in the course of evolution to vary their dosage in order to effectively respond to various environmental signals.

The Mechanism of CNV Formation

Given the existence of large number of CNVs in the human genome, it is interesting to explore the mechanisms of CNV formation. Segmental duplications [19,20] are enriched in the breakpoints of CNVs detected by 500K SNP array [9] and thus support their role in CNV formation. Examination of our CNVs identified only 1 out of those 143 CNVs that contained a pair of segmental duplications at the ends. This is not surprising since only 593 of the 50K (1%) Xba SNPs are located in segmental duplications, while 17530 (3.5%) of the 500K SNPs, used in the study by Redon *et al.* [9], are covered by segmental duplications. It suggests that the poor coverage of SNPs in segmental duplications has limited our ability to examine whether or not segmental duplications may play a role in CNV formation. It also indicates the clear advantage of higher-density SNP array in exploring the relationship between CNVs and other genomic elements.

Besides segmental duplications, other genomic elements that may promote CNV formation are the array of repetitive sequences that could facilitate non-allelic homologous recombination (NAHR), as suggested by Redon *et al.* [9]. More than 42% of the euchromatic DNA is estimated to be derived from interspersed repeats and other transposable elements in the human genome [21]. The presence of such wide-spread repeat elements makes it a likely candidate to facilitate CNV formation. The presence of repeat elements in the CNV-flanking regions could facilitate non-allelic homologous recombination, which in turn leads to copy number variations in the genome. To test this hypothesis, we examined the content of repeat elements in CNV-adjacent regions at various distances from the CNV boundaries, and compared with those within the CNVs and in the whole genome. As shown in Table **6**, the percentage of all types of repeat elements is higher in all CNV-adjacent regions than those within the CNVs and in the rest of the genome. The maximum percentage of repeat elements is present within the 62.5 Kb regions flanking the CNVs. If broken down into different categories of repeat elements, the same trend holds for the long interspersed nuclear elements (LINEs) and long terminal repeats (LTRs). In contrast, the percentage of short interspersed nuclear elements (SINE) is lower in the CNV-adjacent regions, compared with that in the whole genome. The closer it is to the CNV, the lower percentage of SINE repeats is observed.

In order to test whether these findings are statistically significant, we randomly picked 280 genomic regions of

Table 6. Content of repeat elements in various CNV-adjacent regions, as compared with that in CNVs and in all 22 autosomes. The CNV-adjacent regions are defined by the distance from the CNV boundaries as indicated in the parenthesis. For example, the first category of CNV-adjacent regions includes all regions within 1 Mb from the boundary on either side of each CNV, totaling 280 regions. Repeat elements from all 140 unique CNVs or each category containing the 280 CNV-adjacent regions from 22 autosomes are combined and percentage of repeat elements in each of those categories of regions is calculated using the number of bases in the repeat elements
LINE: Long Interspersed Nuclear Elements; SINE: Short Interspersed Nuclear Elements;
LTR: Long Terminal Repeats; DT: DNA Transposons; Other: Other Repeat Elements

Regions	LINE	SINE	LTR	DT	Other	Total
CNV	20.5%	12.7%	8.1%	2.9%	1.8%	46.0%
CNV_adjacent [1 Mb]	21.3%	12.7%	8.7%	3.1%	1.8%	47.7%
CNV_adjacent [1/2 Mb]	21.5%	12.5%	9.0%	3.1%	1.8%	47.9%
CNV_adjacent [1/4 Mb]	21.5%	12.2%	9.2%	3.1%	1.9%	47.9%
CNV_adjacent [1/8 Mb]	21.9%	11.9%	9.4%	3.1%	1.9%	48.3%
CNV_adjacent [1/16 Mb]	22.1%	11.9%	9.7%	3.1%	1.8%	48.5%
CNV_adjacent [1/32 Mb]	21.3%	11.9%	9.6%	3.3%	1.7%	47.8%
CNV_adjacent [1/64 Mb]	21.3%	11.7%	9.6%	3.2%	1.7%	47.6%
Autosomes	19.1%	13.0%	8.0%	2.8%	1.9%	44.8%

62.5 Kb each in the non-CNV-adjacent regions on autosomes and performed a t-test to compare the percentage of repeat elements, counted in the number of bases, in these randomly-picked regions with that in the 280 62.5 Kb regions flanking the 140 CNVs on autosomes (3 CNV regions on the X-chromosome were excluded in this analysis). This random selection process was repeated 10 times. The percentages of repeat elements and P values from the 10 repetitions were averaged. As shown in Table **7**, the content of LINE and LTR repeats in CNV-adjacent regions is significantly higher than that in the non-CNV-adjacent regions (P-values are 0.0082 and 0.024, respectively), while the content of SINE repeats in CNV-adjacent is lower than that in the non-CNV-adjacent regions (P-value: 0.00085). The overall repeat content in CNV-adjacent regions, however, is not significantly higher than that in non-CNV-adjacent regions (P-value: 0.303).

DISCUSSION

A variety of studies using different platforms [5-9,22] has led to an explosion of CNV discoveries. However, it is important to note the differences of how CNVs were defined in each of these studies [7]. In addition, none of the existing platforms for CNV identification could identify the exact break point of CNVs [7]. The boundaries of CNVs in the previous studies have different relative positions to the underlying true variants [7]. For example, the CNVs identified by ROMA [5] should reside within each variant as the delimiting 70-mer probes indicate the inner boundaries of the variant. Similarly, this limitation also applied to CNVs identified by SNP genotyping [7] and SNP arrays [9]. The use of oligonucleotide probes (e.g. 70-mers for ROMA and 25-mers for SNP arrays) that are far apart from each other to assess a large genomic region for copy number could potentially underestimate or overestimate the size of the

Table 7. Certain repeat elements are significantly enriched in CNV-adjacent intervals. The average percentage of repeat elements counted in the number of bases was calculated in 280 1/16 Mb (62.5 Kb) regions that flank the 140 CNVs on autosomes, as well as 280 randomly picked 1/16 Mb genomic regions that do not flank any CNVs
LINE: Long Interspersed Nuclear Elements; SINE: Short Interspersed Nuclear Elements;
LTR: Long Terminal Repeats; DT: DNA Transposons; Other: other Repeat Elements

	% Repeat Content in CNV-Adjacent 1/16 Mb Region	% Repeat Content in Non-CNV-Adjacent 1/16 Mb Region	P-val
Any repeat	48.5%	47.9%	0.303
LINE	22.1%	19.4%	0.008
SINE	11.9%	14.9%	0.001
LTR	9.7%	8.3%	0.024
DT	3.1%	3.0%	0.325
Other	1.8%	2.3%	0.080

CNV region. As for CNVs identified by BAC aCGH [6], each CNV, as represented by a BAC clone with an average size of 150 Kb, may overlap, reside in, or be encompassed by the BAC clone. On the other hand, the deletion variants identified by fosmid end pairing actually encompass the underlying variants as the boundaries of each deletion variants only indicate the "outer boundaries" for the underlying variant. Such a difference in defining the boundaries of the underlying variant in each of these CNV studies should be taken into consideration when comparing and interpreting these CNV data [7]. In addition, the different statistical metrics used to identify CNVs in these studies should also be considered when comparing these CNV data. It is apparent that the boundaries of CNVs could be improved by higher-resolution array platforms.

It should also be pointed out that each of the previous studies had its own limitations. Both the resequencing method [22] and the SNP genotyping method [7] were inadequate in detecting copy number gains. In addition, since both the ROMA method [5] and BAC aCGH method [6] used a single reference individual to obtain the copy number gain/loss information, the gain/loss identified in their analyses is only a "relative" gain/loss to the reference individual [7]. Some of the regions of copy number loss reported in these two studies could represent regions of copy number gain in the reference sample and normal copy number in the test samples. This is supported by a comparison between the deletion variants from these two studies and those identified using the SNP genotyping data [7]. Only 10 of those more than 300 variants discovered by ROMA [5] and BAC aCGH [6] are concordant with the deletion variants identified using the SNP genotyping data [7]. The study by Redon *et al.* [9] used a high-resolution SNP array and a tiling BAC array to identify CNVs among 270 samples. It was the first report of CNV identification by SNP arrays. In comparing CNV studies based on SNP arrays, it should be pointed out that the various versions of Affymetrix SNP arrays include different set of SNPs that are located on different restriction enzyme digestion fragments in the genome. Thus CNVs identified from the XbaI SNP arrays used in this study, albeit with lower resolution than the Nsp/StyI SNP arrays used by Redon *et al.* [9], could include novel CNVs that were not detectable previously. This is shown by comparing CNVs from the two studies. In addition to 22 of our CNVs that were concordant in both studies, we identified 121 additional CNVs that were not reported by the Redon study, based on our criteria for measuring concordance (Table **4**).

Advantages of Identifying CNVs Using SNP Array

High density SNP array enables a thorough interrogation of genomic DNA copy numbers in an efficient manner, giving much more accurate copy number than the ROMA [5] and BAC aCGH [6] methods, in which copy number gain or loss is only relative to a single reference sample. As arrays containing more SNPs become available (the latest version of Affymetrix Genome-Wide Human SNP Array 6.0 has over 906,600 SNPs and more than 900,000 non-polymorphic probes, according to http://www.affymetrix.com), the genome can be interrogated at higher resolution. The analysis method used in this paper for identifying CNVs can be easily adapted to identify CNVs from SNP arrays containing more SNPs. As a consequence, more CNVs could potentially be identified and the CNVs identified from those SNP arrays will be more accurate and with better defined boundaries.

Ascertainment Bias of SNPs among Different Ethnic Groups

In this study we showed that the distribution of CNVs in terms of gains and losses varies among the three ethnic groups present in the samples. However, it should be noted that the results presented here did not take into consideration the ascertainment bias [23] of SNPs represented on the Affymetrix XbaI 50K SNP array. In case of the complete absence of a particular allele of a SNP in a certain population, the copy number of that SNP in the population would be biased towards deletion. Therefore, correction of such ascertainment bias will be necessary to prove whether our observation of relatively more losses in the AS and AF groups than in the CA groups reflects true inter-population difference.

Possible Mechanisms of CNV Formation

The results of our analysis suggest that specific repetitive elements such as LINE and LTR are enriched in regions flanking the CNVs. Due to the lower resolution of the SNP array used in our study, the CNVs we identified could potentially be broken into smaller ones and subsequently more breakpoints could be identified. Therefore, the analysis of repeat elements in our CNV flanking regions may not necessarily include all the potential breakpoints associated with CNVs. It becomes interesting to assess the content of repeat elements in the flanking regions of CNVs identified from platforms with higher density. To test if this observation is unique with our data set, we examined the content of repeat elements in 62.5 Kb genomic regions flanking the 1390 autosomal CNVs identified by Redon *et al.* [9] using SNP arrays with different SNP contents and in higher resolution. Interestingly, the LINE and LTR elements in those CNV flanking regions are also significantly enriched, while the SINE elements are significantly impoverished, compared with randomly picked genomic regions that do not flank those CNVs (data not shown). This similar result serves as corroboration to our reasoning that certain repeat elements (LINE and LTR) may facilitate the formation of CNVs, presumably through non-allelic homologous recombination. Furthermore, a recent study employed exhaustive paired end mapping and large-scale sequencing to allow for comprehensive detection of structural variants (SVs), including deletions, insertions and inversions in the human genome, of less than 10 Kb in size and enabled detailed dissection of sequence elements within the break points [22]. This study also suggested that the L1 elements, which constitute roughly 7/8 of all the LINE elements in the human genome [24], are significantly enriched in the break points associated with more than 1000 SVs identified in two individuals. The similar conclusion drawn from such a fine mapping and sequencing method provides yet further support of the potential role that certain repeat elements may play in CNV formation. In addition, their study found no significant enrichment of Alu elements, comprising the majority of SINE elements [24], near SVs, which is consistent with our observation that SINE elements are impoverished in the junctions of CNVs identified from our study.

The formation of CNVs and other types of SVs is a complex process and probably involves multiple mechanisms. Besides NAHR, non-homologous end joining (NHEJ), in which no homologous sequences are present in the junctions of break points, has been shown to account for more rearrangements that can cause various genomic disorders [25, 26]. Another replication-based mechanism named replication fork stalling and template switching (FoSTeS) has recently been implicated in facilitating genomic rearrangements [26]. Therefore, repeat elements may only partially explain the mechanism of CNV formation, as suggested by this study and corroborated with data from other reports [9, 22]. Other potential mechanisms, such as NHEJ and FoSTeS, should not be overlooked when exploring the mechanisms of CNV formation.

ACKNOWLEDGEMENTS

This work was supported by generous grants to CCL from the Gillson Longenbaugh Foundation, Carl C. Anderson, Sr. & Marie Jo Anderson Charitable Foundation and Cancer Fighters of Houston, Inc. We would also like to thank John Wilson, James Lupski, Aleksandar Milosavljevic and Kai Yan for helpful discussions.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

ABBREVIATIONS

aCGH	=	Array comparative genomic hybridization
BAC	=	Bacterial artificial chromosome
CNV	=	Copy number variation
LINE	=	Long interspersed nuclear element
LTR	=	Long terminal repeat
NAHR	=	Non-allelic homologous recombination
NHEJ	=	Non-homologous end joining
ROMA	=	Representational oligonucleotide microarray analysis
SINE	=	Short interspersed nuclear element
SNP	=	Single nucleotide polymorphism

REFERENCES

[1] Little CD, Nau MM, Carney DN, Gazdar AF, Minna JD. Amplification and expression of the c-myc oncogene in human lung cancer cell lines. Nature 1983; 306: 194-6.

[2] Di Fiore PP, Pierce JH, Kraus MH, Segatto O, King CR, Aaronson SA. erbB-2 is a potent oncogene when over-expressed in NIH/3T3 cells. Science 1987; 237: 178-82.

[3] Friend SH, Bernards R, Rogelj S, *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. Nature 1986; 323: 643-6.

[4] Kamb A, Gruis NA, Weaver-Feldhaus J, *et al.* A cell cycle regulator potentially involved in genesis of many tumor types. Science 1994; 264: 436-40.

[5] Sebat J, Lakshmi B, Troge J, *et al.* Large-scale copy number variation in the human genome. Science 2004; 305: 525-8.

[6] Iafrate AJ, Feuk L, Rivera MN, *et al.* Detection of large-scale variation in the human genome. Nat Genet 2004; 36: 949-51.

[7] McCarroll SA, Hadnott TN, Perry GH, *et al.* International HapMap Consortium. Common deletion polymorphisms in the human genome. Nat Genet 2006; 38: 86-92.

[8] Tuzun E, Sharp AJ, Bailey JA, *et al.* Fine-scale structural variation of the human genome. Nat Genet 2005; 37: 727-32.

[9] Redon R, Ishikawa S, Fitch KR, *et al.* Global variation in copy number in the human genome. Nature 2006; 444: 444-54.

[10] NIGMS (National Institute of General Medical Sciences). Camden, NJ: Coriell Institute (© 2008). Available from: http://ccr.coriell.org/nigms

[11] Lin M, Wei LJ, Sellers WR, *et al.* dChip-SNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 2004; 20: 1233-40.

[12] Zhang B, Schmoyer D, Kirov S, *et al.* GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 2004; 5: 16.

[13] Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996-2004. Available at: http://www.repeatmasker.org

[14] Genboree. Houston, TX: Bioinformatics Research Laboratory (© 2001-2008). Available from: http://www.genboree.org.

[15] Graubert TA, Cahan P, Edwin D, *et al.* A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet 2007; 3: e3.

[16] Li J, Jiang T, Mao J, *et al.* Genomic segmental polymorphisms in inbred mouse strains. Nat Genet 2004; 36: 952-4.

[17] Snijders A, Nowak N, Huey B, *et al.* Mapping segmental and sequence variations among laboratory mice using BAC array CGH. Genome Res 2005; 15: 302-11.

[18] Perry GH, Tchinda J, McGrath SD, *et al.* Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci USA 2006; 103: 8006-11.

[19] Bailey JA, Gu Z, Clark RA, *et al.* Recent segmental duplications in the human genome. Science 2002; 297: 1003-7.

[20] Cheung J, Estivill X, Khaja R, *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol 2003; 4: R25.

[21] Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 1999; 9: 657-63.

[22] Korbel JO, Urban AE, Affourtit JP, *et al. Paired*-end mapping reveals extensive structural variation in the human genome. Science 2007; 318: 420-6.

[23] Clark AG, Hubisz MJ, Bustamante CD, *et al.* Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 2005; 15:1496-502.

[24] Smit AF. The origin of interspersed repeats in the human genome. Curr Opin Genet Dev 1996; 6: 743-8.

[25] Shaw CJ, Lupski JR. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet 2004; 13: R57-R64.

[26] Lee JA, Carvalho CM, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 2007; 131: 1235-47.

Received: September 01, 2008 Revised: September 26, 2008 Accepted: January 28, 2009

© Wang *et al.*; Licensee *Bentham Open*.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.