Letters to the Editor

Am. J. Hum. Genet. 78:520–522, 2006

Identification of Frequent Chromosome Copy-Number Polymorphisms by Use of High-Resolution Single-Nucleotide–Polymorphism Arrays

To the Editor:

In the November issue of the *Journal,* Slater et al. (2005) introduced a high-resolution method for the detection of chromosomal abnormalities using high-density synthetic oligonucleotide Affymetrix arrays containing 116,206 SNPs. The authors identified amplifications and deletions of different sizes (1.3–145.9 Mb) in patients by using SNP arrays in combination with the GeneChip Chromosome Copy Number Analysis Tool (CNAT), version 2.0 (Affymetrix). Comparative genomic hybridization and computational fosmid-end-mapping–based approaches have shown that large-scale chromosome copy-number polymorphisms (CNPs) substantially contribute to the genomic variation between normal human individuals (Iafrate et al. 2004; Sebat et al. 2004; Sharp et al. 2005; Tuzun et al. 2005). It has been proposed that CNPs might be associated with complex diseases, such as cancer, neurological disorders, autism, and obesity (Sebat et al. 2004; Check 2005).

Slater et al. (2005) suggested that it is highly likely that multiple SNPs cover CNP regions and could allow their detection. The algorithm (CNAT, version 2.0) that they used for the detection of chromosomal aberrations was developed using a reference set of 110 healthy individuals who also carry CNPs. Slater et al. proposed that the algorithm needs to be improved to detect CNPs. We suggest that an additional improvement of CNP detection should consider the selection criteria of SNPs for the array. The criteria used by Affymetrix consider Mendelian inheritance, Hardy-Weinberg equilibrium (HWE), genotyping accuracy, and reproducibility (Slater et al. 2005), which may lead to a selection of SNPs that is biased against CNP regions, and thus interferes with the detection of frequent CNPs. This limitation cannot be overcome with improvement of the algorithms. SNPs in CNP regions with frequent losses would lead to an accumulation of apparent Mendelian inheritance errors (e.g., if the genotypes of the parents are AA and B0 and the genotype of the child is A0) or deviations from HWE and thus would be rejected by the criteria. SNPs in frequently amplified genomic regions might produce genotype calls of reduced reproducibility (between homozygous and heterozygous calls, if an individual carries an "AAB" or "ABB" genotype) or might lead to Mendelian inheritance errors. An underrepresentation of SNPs in regions known to contain common CNPs will prevent the identification of these common CNPs, because information from multiple SNPs is required to establish a reliable detection.

To test this hypothesis, we determined the SNP coverage of the most-frequent CNP regions (frequency >0.20) published by Iafrate et al. (2004), Sebat et al. (2004), Tuzun et al. (2005), and Sharp et al. (2005) with SNPs on the Affymetrix GeneChip Mapping 100K Array set. Data of 82 CNP regions were retrieved from the Database of Genomic Variations (representing 12.8% of all CNPs in the database [Iafrate et al. 2004]), and the corresponding SNP data were retrieved from the University of California–Santa Cruz (UCSC) Genome Browser (see Web Resources). The mean intermarker distance (ID) of the Affymetrix 100K SNPs located within the borders of each investigated CNP region was determined. After the exact location of the CNPs was mapped, the mean ID was calculated by dividing the length of each CNP region by the number (plus 1) of 100K array SNPs located within the region. In the cases in which CNPs were not covered by any SNPs, the mean ID size corresponded to the CNP length. Of all analyzed CNP regions, 58.5% contained at least one known gene, and all investigated CNPs except one (chr2-cent-2p11.2) were located outside telomeric or centromeric regions. All investigated CNPs with detailed annotations are listed in an HTML file (online only).

Indeed, 81.7% of the investigated CNP regions had a mean ID larger than the overall mean ID of all SNPs on the array (23.6 kb), and 95.1% of the investigated CNP regions had a mean ID larger than the overall median ID of all SNPs on the array (8.5 kb) (table 1). We divided the CNPs into four groups according to their SNP coverage: 0 SNPs (52.4% of CNPs), 1–4 SNPs with mean ID >23.6 kb (33.0%), >4 SNPs with mean ID >23.6 (6.1%), and >4 SNPs with mean ID ⩽23.6 (8.5%) (table 1). Thus, only 14.6% of all investigated CNP regions were covered with >4 SNPs on the array and might be detectable, although half of them had a mean

Table 1

Coverage of the 82 Investigated Most-Frequent CNPs (Frequency >0.20) with SNPs on the Affymetrix GeneChip Mapping 100K Array Set

		Percentage of CNPs			Length of CNPs (kb)	
SNP Coverage of CNPs	No. of CNPs	Total	With Mean ID >23.6 kb	With Mean ID >8.5 kb	Mean	Median
0 SNPs	43	52.4	43.9	47.6	120	141
1–4 SNPs with mean ID >23.6 kb	27	33.0	32.9	32.9	433	180
>4 SNPs:						
With mean ID >23.6 kb	5	6.1	4.9	6.1	744	285
With mean ID ⩽23.6 kb	7	8.5	.0	8.5	213	157
All	82	100.0	81.7	95.1	268	157

Note.—The overall mean ID of all SNPs on the array was 23.6 kb, and the overall median ID of all SNPs on the array was 8.5 kb. In the cases in which the CNP was not covered by any SNPs (0 SNPs), the mean ID size corresponded to the length of the CNP.

ID >23.6 kb. All other analyzed CNP regions (85.6%) were not covered with SNPs or were too sparsely covered with SNPs to achieve an appropriate detectability. The stratification of CNPs according to the different kinds of copy-number variation (loss, gain, or both) revealed that the majority of CNPs with losses (65.6%) or with both losses and gains (57.6%) were not covered by SNPs at all (fig. 1). Most of the CNPs with gains (64.7%) were covered with only 1–4 SNPs with a mean ID >23.6 kb (fig. 1).

Sharp et al. (2005) recently suggested that segmental duplications may be able to serve as catalysts for CNPs in the human genome. Segmental duplications themselves are enriched significantly more than fourfold within regions of CNP. Indeed, 82.9% of the frequent CNPs investigated in the present study were overlap-

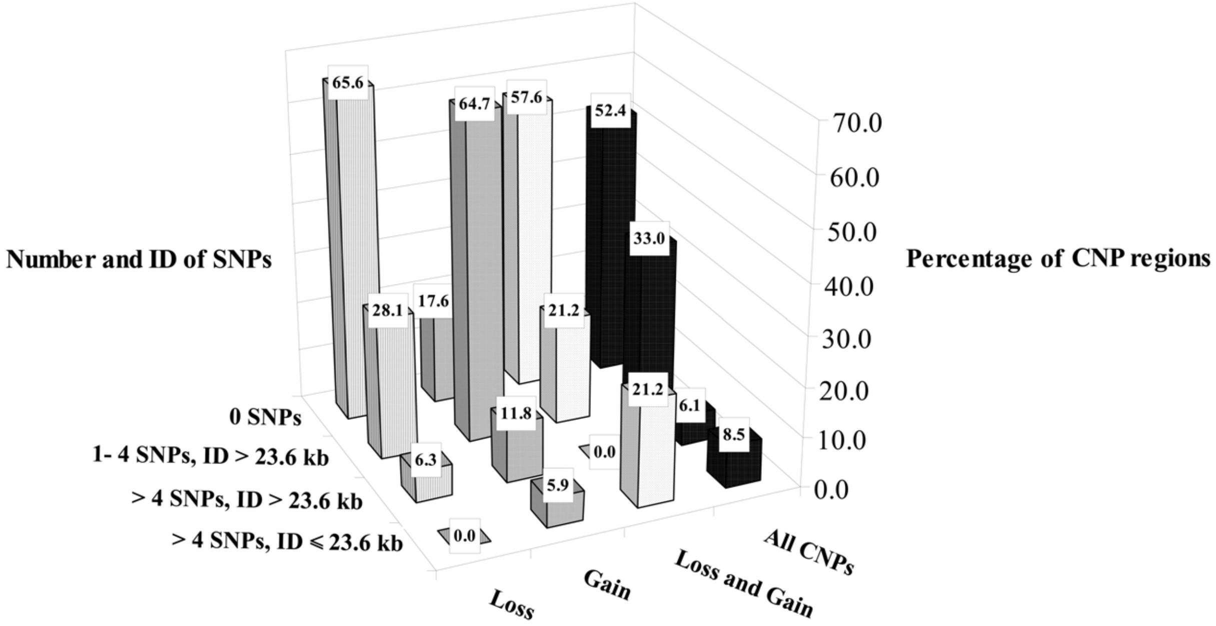

Figure 1 Coverage of the most-frequent CNP regions (frequency >0.20) identified by Iafrate et al. (2004), Sebat et al. (2004), Sharp et al. (2005), and Tuzun et al. (2005) with SNPs on the Affymetrix GeneChip Mapping 100K Array set. The regions were divided into four groups according to SNP coverage: 0 SNPs (not covered), 1–4 SNPs with mean ID >23.6 (covered with 1–4 SNPs with a mean ID larger than the overall mean ID of all SNPs on the array), >4 SNPs with mean ID >23.6 kb, and >4 SNPs with mean ID ⩽23.6 kb. The percentages of CNP regions corresponding to the four groups, stratified according to the kind of copy-number alteration (gain, loss, or both), are displayed. Of all investigated CNP regions, 39.1% had losses, 20.7% had gains, and 40.2% had both gains and losses.

ping segmental duplications. Only five of the most frequent CNP regions investigated in this study (22q11.22, 22q11.21, 19p13.2, 15q14, and 14q32.33) were detected by more than one author group (Iafrate et al. 2004; Sebat et al. 2004; Sharp et al. 2005; Tuzun et al. 2005; data in the HTML file [online only]). This points to the still-unknown significance of the CNPs identified so far (Carter 2004).

Slater et al. (2005) suggested 400 kb as the mean length of CNPs, on the basis of the Database of Genomic Variations. We show here that the most-frequent CNPs (frequency >0.20) investigated in the present study had a mean length of 268 kb and a median length of 157 kb, respectively (table 1). Notably, the CNP regions not covered by SNPs at all were smaller in size (mean length 120 kb; median length 141 kb). However, considering that 91% of the genome is suggested to be within 100 kb of a SNP (Slater et al. 2005), the majority of CNPs should have been covered at least by one SNP on the array.

In conclusion, oligonucleotide-based SNP arrays have been shown to be an excellent tool for analyses of loss of heterozygosity and rare copy-number variation (e.g., Zhao et al. 2004), association studies (e.g., Hu et al. 2005), linkage studies (e.g., Sellick et al. 2005), resequencing applications in humans and other organisms (e.g., Cutler et al. 2001; Maitra et al. 2004; Zwick et al. 2005), and the detection of recombination hotspots (e.g., Wirtenberger et al. 2005). However, the applicability might be somewhat limited with regard to the analysis of frequent CNPs, because of the initial SNP selection. High-density tiling arrays might be an appropriate tool for this kind of analysis. Chip manufacturers may be able to change their SNP selection criteria and provide an updated chip-description file that includes information on the artificially masked SNPs that do not fulfill the selection criteria. But, until they do so, users of high-density SNP arrays in association studies of common diseases should be aware of this limitation.

Michael Wirtenberger,[1] Kari Hemminki,[1,2]
and Barbara Burwinkel[1]

[1]*Division of Molecular Genetic Epidemiology, German Cancer Research Center, Heidelberg, Germany; and* [2]*Department of Biosciences at Novum, Karolinska Institute, Huddinge, Sweden*

Web Resources

The URLs for data presented herein are as follows:

Database of Genomic Variations, http://projects.tcag.ca/variation/
UCSC Genome Browser, http://genome.ucsc.edu/index.html

References

Carter NP (2004) As normal as normal can be? Nat Genet 36:931–932

Check E (2005) Human genome: patchwork people. Nature 437:1084–1086

Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, Chakravarti A (2001) High-throughput variation detection and genotyping using microarrays. Genome Res 11:1913–1925

Hu N, Wang C, Hu Y, Yang HH, Giffen C, Tang ZZ, Han XY, Goldstein AM, Emmert-Buck MR, Buetow KH, Taylor PR, Lee MP (2005) Genome-wide association study in esophageal cancer using GeneChip Mapping 10K Array. Cancer Res 65:2542–2546

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951

Maitra A, Cohen Y, Gillespie SE, Mambo E, Fukushima N, Hoque MO, Shah N, Goggins M, Califano J, Sidransky D, Chakravarti A (2004) The Human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection. Genome Res 14:812–819

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528

Sellick GS, Webb EL, Allinson R, Matutes E, Dyer MJ, Jonsson V, Langerak AW, Mauro FR, Fuller S, Wiley J, Lyttelton M, Callea V, Yuille M, Catovsky D, Houlston RS (2005) A high-density SNP genomewide linkage scan for chronic lymphocytic leukemia-susceptibility loci. Am J Hum Genet 77:420–429

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88

Slater HR, Bailey DK, Ren H, Cao M, Bell K, Nasioulas S, Henke R, Choo KHA, Kennedy GC (2005) High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs. Am J Hum Genet 77:709–726

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727–732

Wirtenberger M, Hemminki K, Chen B, Burwinkel B (2005) SNP microarray analysis for genome-wide detection of crossover regions. Hum Genet 117:389–397

Zhao X, Li C, Paez JG, Chin K, Janne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, Gray JW, Sellers WR, Meyerson M (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 64:3060–3071

Zwick ME, McAfee F, Cutler DJ, Read TD, Ravel J, Bowman GR, Galloway DR, Mateczun A (2005) Microarray-based resequencing of multiple *Bacillus anthracis* isolates. Genome Biol 6:R10

Address for correspondence and reprints: Mr. Michael Wirtenberger, Im Neuenheimer Feld 580, Heidelberg 69120, Germany. E-mail: m.wirtenberger@dkfz.de

0002-9297/2006/7803-0020$15.00