Benign copy number changes in clinical cytogenetic diagnostics by array CGH

H. Whitby^a A. Tsalenko^b E. Aston^a P. Tsang^b S. Mitchell^c
P. Bayrak-Toydemir^{c,d} C. Hopkins^b G. Peters^b D.K. Bailey^b L. Bruhn^b
A.R. Brothman^{a, c-e}

^aUniversity of Utah CGH Microarray Laboratory, Department of Pediatrics, Salt Lake City, UT

^bAgilent Technologies, Santa Clara, CA; ^cARUP Laboratories, ^dUniversity of Utah Department of Pathology

^eUniversity of Utah Department of Human Genetics, Salt Lake City, UT (USA)

Accepted in revised form for publication by H. Kehrer-Sawatzki and D.N. Cooper, 15 July 2008.

Abstract. A database of apparently benign copy number variants (bCNVs) detected by a Spectral Genomics Inc./PerkinElmer BAC array platform has been maintained through the University of Utah Comparative Genomic Hybridization laboratory since 2005. The target population for this database represents 1,275 patients with abnormal phenotypes, primarily children referred for developmental delay and mental retardation. These bCNVs are independent of any identified copy number abnormality detected. The most common 35 bCNVs observed and their frequencies are reported here, and a subset of ten of the patients studied was evaluated on a new oligonucleotide CNV array set designed by Agilent Technologies. There was a 76% concordance of

calls for these 35 bCNVs detected by both array platforms in the same patients. The higher resolution of the Agilent oligonucleotide array compared to the BAC array allowed determination of the precise breakpoints of the observed CNVs, in addition to documentation of additional CNVs of smaller sizes. As expected, observed CNVs and their frequencies were generally consistent with those of other previously published and available databases, including the Database of Genomic Variants (<http://projects.tcag.ca/variation/>). The availability of these data should assist other clinical laboratories in the evaluation of CNVs of unknown clinical significance.

Copyright © 2009 S. Karger AG, Basel

It has been estimated that as much as one fifth of the normal human genome contains copy number variants (CNVs), as assessed by array comparative genomic hybridization (aCGH) (Iafrate et al., 2004; Carter, 2007). The detection and interpretation of these CNVs in a clinical laboratory setting has presented challenges, especially in determining whether or not they are pathogenic when analyzing patients with abnormal phenotypes (Lee et al., 2007). The discovery

of new pathogenic CNVs has also led to the identification of new genetic syndromes (Koolen et al., 2006; Krepischi-Santos et al., 2006; Sharp et al., 2006; Shaw-Smith et al., 2006). It has been estimated that a large fraction (~40%) of CNVs detected in children with developmental delay or mental retardation are actually benign CNVs and inherited from a parent (de Ravel et al., 2006). For consistency in this special issue of the journal, 'bCNVs' will be defined as benign copy number variants; but we otherwise generally refer to these as benign copy number *changes* (bCNCs) to distinguish them from pathogenic ones (pCNCs) or those of unknown clinical significance (uCNC) (Aston et al., 2008).

Benign CNVs have been defined differently in different scientific disciplines. Typically, CNVs would represent sequences of at least 1 kb in size, and if they are true population polymorphisms would be present in at least 1% of a particular population (Scherer et al., 2007). Traditional cytogenetic analysis, on the other extreme, identifies polymorphic regions which are visible at the microscopic level

This work was supported by the University of Utah CGH Microarray Laboratory and the ARUP Institute for Clinical and Experimental Pathology, ARUP Laboratories, Salt Lake City, UT (USA).

H.W. and A.T. contributed equally to this manuscript.

Request reprints from Arthur R. Brothman, Ph.D., FACMG
Cytogenetics and Molecular Cytogenetics, ARUP Laboratories
500 Chipeta Way, Salt Lake City, UT 84108-1221 (USA)
telephone: +1 801 584 5128; fax: +1 801 584 5036
e-mail: art.brothman@genetics.utah.edu

(Brothman et al., 2006) which would generally be greater than 5 Mb. The Database of Genomic Variants (DGV) (<http://projects.tcag.ca/variation/>) has become a resource for information regarding the frequency of particular bCNVs, yet recent studies indicate that the majority of bCNV loci are smaller in size than those reported by the DGV (de Smith et al., 2007; Korbelt et al., 2007a; Kidd et al., 2008; Perry et al., 2008). The availability of detailed information regarding bCNVs which are not associated with any pathogenic phenotype is extremely important for interpretation of aCGH findings on clinical specimens.

There have been numerous reports of CNVs in various populations (e.g. Redon et al., 2006; Qiao et al., 2007; Jakobsson et al., 2008), but these are almost exclusively from cohorts of phenotypically *normal* individuals and often cell lines rather than DNA derived directly from patient specimens. The Cytogenetics Laboratory at the University of Utah has been using aCGH to evaluate DNA derived from blood of patients with developmental delay, mental retardation and 'syndromic' or dysmorphic phenotypes since 2005 (Aston et al., 2008). Initial work was done using the Spectral Genomics Inc./PerkinElmer BAC-based aCGH, and more recently has encompassed oligonucleotide-based aCGH (Agilent Technologies). A database of changes detected in all patients studied has been maintained, and the data were analyzed in conjunction with publicly available databases (such as the DGV) to categorize genomic regions which likely contain bCNVs.

We present here a summary of findings describing 35 of the most frequently observed bCNVs detected in a cohort of 1,275 clinically affected individuals, initially studied using a BAC array of about 1 Mb resolution. These data are compared with publicly available databases, and a subset was compared with that obtained when the same patient DNA was hybridized to a high resolution oligonucleotide array set designed to specifically assess CNV regions throughout the human genome.

Methods

The University of Utah Cytogenetics/Microarray Laboratory data include results from two bacterial artificial chromosome (BAC) array chips purchased from Spectral Genomics Inc./PerkinElmer (SGI/PE) (Turku, Finland): the Constitutional Chip™ (CC) and the SpectralChip 2600™ (1Mb). Patients were initially studied using either the 1Mb, CC, or both. The CC is a targeted array designed to screen for 41 recognized microdeletion or microduplication syndromes plus 41 unique subtelomeric regions. Clones are concentrated in the areas of interest with a sparse backbone of clones in other areas of the genome. This design results in large gaps between clones in some regions. The version of the chip most recently used (Constitutional Chip™ 3.0) was printed with 604 BAC clones. The 1Mb is a whole genome array. The most recent chip used (SC2600-B33) spaced 2,610 BAC clones along the genome at roughly 1-Mb intervals (Aston et al., 2008). The majority of the CNVs identified by the University of Utah were identified on the 1Mb chip. Both chips were based on NCBI Build 35.1.

Clones detecting copy number changes that map to regions of known CNVs as well as single clones of unknown significance were added to this database. An entry was made each time a patient exhibited a copy number change of a BAC in any of these regions. Abnor-

malities with known clinical significance were excluded. From this database, we were able to determine which clones represented sequences which were frequently deleted or duplicated in this patient population. We identified the most common CNV regions and then selected DNAs from ten patients who exhibited copy number changes of those common regions for further study with the Agilent CNV microarray set.

All protocols followed the manufacturer's recommendations with minor modifications. DNA was isolated from whole blood or fixed cell pellets with the Puregene Genomic DNA Purification Kit (Gentra Systems, Minneapolis, MN). DNA concentration was measured using a NanoDrop Spectrophotometer (NanoDrop Technologies, Inc, Wilmington, DE).

The genomic BAC microarrays were performed using Spectral Genomics chips (SGI/PE 1Mb or CC) and reagents. Normal male or female DNA (representing pooled individuals) from Promega (Madison, WI) was used as a sex-mismatched control. The BAC microarrays were run in duplicate using a dye-reversal strategy. A copy number variant was called only if confirmed on both hybridizations. Chips were scanned using either the Gene Pix 4000B or the Agilent Microarray scanner and the Gene Pix Pro 6.0 software package (Axon Instruments, Union City, CA). The data were analyzed using the SpectralWare software program (SGI/PE).

The Agilent CNV microarray set G4423B (AMADIDs 018897 and 018898) contains 392,824 oligonucleotide probes targeting 19,400 disjoint intervals spanning 28% of the genome and printed on two slides containing 244K probes each. These probes were selected from Agilent's database of 24.3 million oligonucleotide probes suitable for aCGH at an average spacing of 100 bp within the non-repeat masked regions of the genome. The 19,400 intervals were derived from 4,902 CNV regions and 16,110 in-dels from the DGV as well as 8,061 regions from the UCSC segmental duplication track. Median probe spacing within targeted regions varied from approximately 200 bp for smaller intervals to as high as 2,400 bp for larger intervals (average 982 bp). An additional 69,785 probes were selected from the remaining regions of the genome to provide whole genome coverage with a median spacing of 27 kb. The array probes were selected using an empirical model that utilizes scores for homology, thermodynamics, secondary structure and sequence complexity. To ensure adequate coverage of regions duplicated in the genome, homology criteria were relaxed to allow probes with multiple perfect matches in the reference human genome assembly. Probes for chromosomes 1–8 were assigned to array 1, and probes for chromosomes 9–22, X and Y were assigned to array 2. About 10,000 probes were repeated on both arrays for normalization.

A set of ten samples which showed common bCNVs based on the BAC array data set was chosen for analysis on the Agilent Human Genome CNV microarray. Normal male or female DNA (representing pooled individuals) from Promega was used as a sex-mismatched control. Sample and control DNAs were labeled with Cy5 or Cy3, respectively, using the Agilent DNA labeling kit. Following manufacturer's recommended hybridization and washes, the arrays were scanned at 5 µm resolution using the Agilent microarray scanner and analyzed using Feature Extraction v9.5.3.1, DNA Analytics v4.0 (Agilent Technologies), and prototype software tools using similar statistical methods. In addition to the ten clinical specimens, two self-self experiments for two patient samples were analyzed to estimate a false positive rate and to fine tune the parameters used for data analysis. A replicate hybridization was performed on one patient to determine reproducibility.

Data from the two arrays in each CNV array set were merged and treated as a single dataset for each experiment. Log2 ratios of fluorescent signals for each measured probe sequence were assigned to all perfect genomic matches of this sequence in the NCBI 36.1 build of the reference genome resulting in probes being assigned to a total of 500,974 genomic locations. This expanded set of log2 ratios was analyzed with a modified version of the ADM2 statistical algorithm (Lipson et al., 2005, 2006). In brief, the ADM2 algorithm uses an iterative procedure to identify all genomic regions for which the weighted average of the measured log2 ratios from probes in the region deviates from its expected value of 0 by more than a given threshold. In the version of the ADM2 algorithm used for this study, the statistical score of each

Table 1. Thirty-five of the most commonly observed CNVs in the Utah data set of 1,275 patients using the BAC array described, listed in descending frequency. The number of oligonucleotides represented on the Agilent CNV array set for the corresponding genomic location, and frequencies of CNVs observed at the University of Utah are listed. For comparison, listing of these regions noted in the Database of Genomic Variants (DGV, April 2008) is also shown.

Cytogenetic location	Start	End	Size (bp)	BAC clone	No. of probes on Agilent CNV array set	No. of patients with gain or loss	CNV frequency	
							Utah	DGV
19p13.2	8,714,331	8,864,039	149,708	RP11-79F15	126	342	0.268	0.535
8p23.1	7,295,547	7,305,838	10,291	RP11-122N11	9	182	0.143	0.156
8q21.2	86,898,278	86,954,755	56,477	RP11-96G1	46	142	0.111	0.485
5p15.1	17,465,419	17,636,603	171,184	RP11-88L18	180	128	0.100	0.182
14q12	28,581,577	28,768,137	186,560	RP11-125A5	40	115	0.090 ^a	0.145
5q23.1	69,777,034	69,816,329	39,295	RP11-551B22	17	82	0.064	0.426
1p36.13	16,880,178	16,998,245	118,067	RP1-163M9	125	71	0.056	0.168
9q32	114,865,611	115,042,807	177,196	RP11-9H12	110	70	0.055 ^a	0.084
16p11.2	34,289,161	34,490,212	201,051	RP11-488I20	70	63	0.049	0.164
16p11.1	34,476,094	34,627,143	151,049	RP11-80F22	47	59	0.046	0.166
10q26.3	135,078,853	135,240,498	161,645	RP11-108K14	155	52	0.041	0.054
10qter	135,070,014	135,168,517	98,503	GS-261-B16	97	45	0.035	0.102
2q37.3	242,552,130	242,710,332	158,202	RP11-341N2	144	44	0.035	0.068
1p21.1	103,945,236	104,113,271	168,035	RP11-259N12	131	43	0.034 ^a	0.377
13q21.2	56,600,596	56,729,961	129,365	RP11-100C24	136	38	0.030 ^a	0.118
10q11.22	46,487,806	46,642,258	154,452	RP11-314P12	110	36	0.028	0.302
17q21	42,087,497	42,244,517	157,020	RP11-79O18	28	34	0.027	0.228
4p13	34,453,716	34,603,120	149,404	RP11-81N11	108	32	0.025	0.162
7p21.3	14,338,566	14,452,359	113,793	CTD-2004K7	26	29	0.023	0.100
15q11.2	19,316,573	19,534,174	217,601	RP11-810K23	96	27	0.021 ^a	0.564
8q24.2	137,773,471	137,919,464	145,993	RP11-17M8	68	27	0.021	0.020
2q21	131,997,314	132,171,474	174,160	RP11-89B17	98	26	0.020	0.122
4q35.2	190,938,587	191,019,567	80,980	RP5-963K6	96	19	0.015	0.013
12q24.33	132,109,084	132,272,519	163,435	GS-221-K18	129	18	0.014	0.007
11q22.3	103,803,542	103,965,030	161,488	RP11-179B7	40	16	0.013	0.003
16p11.2	32,562,031	32,772,923	210,892	RP11-499D5	117	15	0.012	0.245
1q31.3	194,871,057	195,039,112	168,055	RP11-91G12	101	15	0.012	0.087
1p36.3	5,189,626	5,308,620	118,994	RP4-703E10	58	15	0.012	0.004
15q12	20,361,317	20,536,186	174,869	RP11-80H14	132	12	0.009	0.376
6p25.3	89,701	213,735	124,034	RP3-416J7	80	12	0.009	0.174
19p13.3	3,064,304	3,266,968	202,664	RP11-554A7	96	11	0.009	0.194
15q25	80,867,007	81,024,206	157,199	RP11-152F13	148	11	0.009	0.127
15q14	32,346,899	32,537,408	190,509	RP11-602M11	114	10	0.008	0.329
7q35	143,246,098	143,375,452	129,354	RP4-669B10	72	10	0.008	0.182
1p36.33	1,609,515	1,720,123	110,608	RP1-283E3	74	10	0.008	0.081

^a These BACs were removed from recent versions of the arrays, and thus the corresponding University of Utah frequencies may be skewed. For comparison with other published population studies, see Supplemental Table 1.

genomic interval uses the measured log₂ ratio of each probe sequence only once, even if a given genomic sequence has multiple perfect matches in this interval. The modified ADM2 algorithm was applied with a threshold of 5, minimum absolute average log₂ ratio in called intervals of 0.2 and minimum of 2 probes.

Results

The initial set of BAC aCGH data from the SGI/PE platform was evaluated as part of a clinical assay for 1,275 patients referred for suspected genetic abnormalities (primar-

ily developmental delay or mental retardation) (Aston et al., 2008). CNVs thought to be bCNVs were entered into the local University of Utah database for categorizing location and frequency. Examples of the most frequently observed (presumably) benign CNVs detected by the SGI/PE platform in ten of the 1,275 patients are shown in Table 1, along with information on frequencies with which these regions have been reported in the DGV. Additional information on how these regions compare with other databases and reported frequencies of gains and losses is included in Supplemental Table 1 (for supplemental material, see www.karger.com/doi/10.1159/000184696).

A subset of ten patient DNA samples from the University of Utah sample set was selected for further characterization using the Agilent CNV array set. The number of oligonucleotide probes in each of the 35 most frequently observed bCNVs from the BAC aCGH data set is shown in Table 1. Using the modified ADM2 algorithm described in the Methods section, we detected on average 546 CNV regions on the autosomes in each of the ten patients (Fig. 1 and Supplemental Table 2). As none of the 35 most frequent bCNVs found from the BAC aCGH data set were observed on either sex chromosome, only the autosomes were included in this analysis. In the self-self experiments, we detected nine and 64 calls, respectively. The patient sample with 64 calls in the self-self experiment also had the largest number of calls in the hybridization vs. the reference sample (834 gains and losses). This sample had been stored at 4°C for two years prior to running on the Agilent array while other samples were stored for less than one year, a factor that could impact the performance of the assay. This sample was excluded from subsequent analysis.

CNV regions detected using the Agilent CNV array set in the ten patients overlapped 28 of the 35 most frequent bCNV regions (80%) from the Utah database, with no copy number variations observed for the remaining seven regions (clones) in this set of patients (Supplemental Table 1). The discordant regions may be due to differences in the reference DNAs used in the different studies, as the source of the normal individuals used in the pooled DNA lots may have varied. Other explanations include possible mis-mapping of BACs on early genome reference builds, or false positive/negative results on either array. There are also several regions where significant differences were seen between the CNV frequencies from the University of Utah data and the frequencies reported in the DGV (Table 1). While some of these differences may be due to population sampling, an explanation for differences in five of the regions noted in Table 1 is that these regions were not represented on newer versions of the BAC array, and hence would be under-represented in the later patients studied.

The median size of CNV calls in the Agilent CNV array data set is 4.9 kb (Fig. 2), similar to size distributions previously reported in high resolution studies (de Smith et al., 2007; Korbel et al., 2007b; Kidd et al., 2008; Perry et al., 2008). 746 CNVs detected in ten samples were larger than 50 kb (14%), and 716 CNVs were smaller than 1 kb (13%). Known abnormalities were detected in two patients with both platforms used: one on chromosome 7 at 72.3 Mb (the patient had Williams syndrome) and one showing a complex subtelomere abnormality on chromosome 22 at 44.7 Mb and at 47 Mb (data not shown).

The total of 5,460 CNVs detected in the ten patients on the autosomes (Fig. 1 and Supplemental Table 2) represent 1,547 copy number variant regions, of which 948 regions were detected in more than one patient (Supplemental Table 3). Six hundred sixty-four (70%) of the CNV regions called in more than one sample had consistent ($\pm$ five probes) breakpoints while some showed different breakpoints within a CNV region (see examples in Figs. 3–5 and Supplemen-

Fig. 1. Total number of CNVs detected in each sample using the Agilent Whole Genome CNV array set. 546 gains and losses were detected in each sample on average on the autosomes, compared to 9 calls in a self vs. self experiment. Sample 7 was hybridized in a replicate experiment to determine variability in total calls.

Fig. 2. Distribution of CNV sizes in 10 patients. Size distribution is shown on the log scale, with 10-fold multiples of 1 and square root of (10). The median size of detected CNVs is 4.9 kb. 746 CNVs larger than 50 kb (14%) and 716 CNVs smaller than 1 kb (13%) were detected.

tal Tables 3 and 4). Of the 28 most frequent bCNV regions from the Utah database that were observed in the Agilent CNV array data set, eight were detected in multiple samples and showed variants with consistent breakpoints (up to five probes). The lengths of the variant regions observed in dif-

Fig. 3. Example of the most commonly observed CNV (at 19p13.2) showing similar breakpoints in six different specimens. Profiles of log₂ ratio data from the Agilent CNV array set on chromosome 19 from 8.704 to 8.764 Mb (59 kb) covering approximately the same region as BAC clone RP11-79F15. Gains (right of zero) or losses (left of zero) called by ADM2 using CGH Analytics are indicated by shaded bars, encompassing a similar set of 98–103 probes in each patient. Repeat sequences are visualized by exact repetitive patterns in the log₂ ratios in many of the aberrant regions shown.

Fig. 4. Example of a commonly observed CNV (at 13q21.2) covering approximately the same region as RP11-100C24 with variable size and breakpoints in eight different specimens. Profiles of log₂ ratio data from the Agilent CNV microarray set on chromosome 13 from 56.612 to 56.687 Mb. Gains (right of zero) or losses (left of zero) called by

ADM2 using CGH Analytics are indicated by shaded bars encompassing distinctly variable size CNVs and variable CNV breakpoints in different individuals. Repeat sequences are visualized by exact repetitive patterns in the log₂ ratios in many of the aberrant regions shown.

ferent individuals in 15 of these 28 regions indicated greater than 50% variability. The differing breakpoints could potentially represent different variants in the same region and/or regions with complex CNV structure (Perry et al., 2008).

Discussion

CNVs have been reported and noted in numerous publications and within multiple publicly available databases; the samples from which they have been ascertained generally represent a plethora of individuals. Previous focused studies of these widespread variant regions have included different

Fig. 5. Examples of four of the more commonly observed bCNVs (other than those in Figs. 3 and 4) showing BAC clone locations and copy number variations detected in individual samples by the Agilent CNV microarray set together with the corresponding oligonucleotide locations. Note that size scales on the X axis are different for each of the four plots. BACs are illustrated by blue lines; gains are shown in red

and losses are in green. The upper two plots (BACs RP11-9H12 and RP11-96G1) show consistent boundaries for the samples in which they were observed, while the lower two plots (BACs RP11-88L18 and RP11-551B22) show variable regions. For a complete listing of all overlapping regions, see Supplemental Tables 3 and 4.

populations of individuals, but all presumably with *normal* phenotypes. Array CGH analysis of patients with congenital abnormalities has become a main (and sometimes a first tier) assay for clinical cytogenetics and genetics laboratories, yet how to distinguish normal variants from those of clinical significance has created a dilemma. A diagnosis of a normal variant (bCNV) should not generally be considered an abnormality, nor should it be (by itself) a major factor in interpretation of a laboratory result. Of note, one should consider the possibility that bCNVs can be observed in homozygous form, and these may in fact not be clinically insignificant (Curry et al., 2008).

New syndromes are now being discovered at one of the most rapid paces in the history of our field (Krepischi-Santos et al., 2006), and these are primarily due to the recognition of new pathogenic copy number changes (pCNVs). Furthermore, there has been the association of multiple, individually rare *de novo* CNVs, with neurodevelopmental disorders such as autism spectrum disorders (Sebat et al., 2004) and schizophrenia (Walsh et al., 2008). What has commonly happened in clinical laboratory settings when initiating aCGH testing, was that newly identified copy number changes were seen in the evaluation of patients with anomalies, but the interpretation of those findings could be troublesome if they involved novel sites in the genome; it was difficult to determine whether these changes were

pathogenic or benign. Examination of parental samples helps in this assessment, but these studies are both costly and time consuming. Distinguishing the difference between these pCNVs and bCNVs is essential in such clinical evaluation (Lee et al., 2007).

The University of Utah Cytogenetics Laboratory has maintained a database of all copy number changes on all patients studied by aCGH since 2005. As the number of patients evaluated increased, it became apparent that many of the CNVs detected were consistent with those reported in public databases such as the DGV, while new ones were also observed. The DGV contains gains and losses reported in various studies in normal individuals in all 35 of the most frequent (presumably) bCNV regions defined by clones in the Utah database. Even CNVs detected in individual studies overlap with a significant proportion of these regions. Namely, 25 overlap CNV regions reported in Perry et al. (2008), and 17 CNV regions reported in de Smith et al. (2007) (Supplemental Table 1).

Clinically significant findings were observed in approximately 17% of cases amongst the 1,275 patients analyzed with the SGI/PE BAC array platform (Aston et al., 2008) yet, the patients which showed these pCNVs (some showing duplications/deletions of >10 Mb chromosomal regions), often showed the same bCNVs observed in normal patients. As we have shown here, use of a high density CNV array set

dramatically increases the number of calls of CNVs detected. The overall resolution of clinical array platforms is increasing, and the number of detected CNVs in clinical studies would be expected to rise. Goals of this study were to compare the initial entries in the Utah database with those of other reported CNVs, and to further characterize the most commonly observed CNVs using a high resolution array set designed to measure CNVs.

As we anticipated, bCNVs in the patient population studied appear to be similar to bCNVs described in the public domain from normal individuals, yet actual frequencies of gains and losses differ in some instances (Table 1 and Supplemental Table 1). Similar but not exact frequencies were observed in the subset of patients analyzed using the Agilent CNV array set. This is not surprising and can likely be attributed to the small sampling studied. While gain vs. loss information is generally helpful in assessing potential pathogenicity of a newly identified CNV (Lee et al., 2007), we are cautious to over-interpret gain/loss values in this current study as the sources of reference DNAs are likely different from those used in other comparative studies. Also, as expected, a much greater number of CNVs was detected with the high density Agilent platform when compared with the SGI/PE BAC array, with a total of 440–721 calls per individual (Fig. 1). In the subset of ten patients evaluated at high density, the modal size of CNVs was in the 1-kb range, with approximately 7% of CNVs 100 kb or larger (Fig. 2). These findings are consistent with those of other reports (de Smith et al., 2007; Korbelt et al., 2007b; Kidd et al., 2008; Perry et al., 2008).

This report represents summary data from 1,275 initial patients, with only a fraction of those (ten patients) studied with the high density CNV array set. Expansion of these data is inevitable as more laboratories perform aCGH for diagnostics. The CNVs listed in Table 1 are certainly not a comprehensive list, but represent those most frequently observed in the University of Utah aCGH database, and also

represent loci which we believe are bCNVs. As noted above, the clinically significant abnormality detection rate by aCGH at Utah has been approximately 17% using the SGI/PE BAC array platform. The ten patients used in this study were a sampling of all patients studied, and were picked as they exhibited the most frequently seen bCNVs. Indeed, two of these patients (20%) showed clinically significant abnormalities *in addition to* the bCNVs.

Characterization of the sizes of the CNVs observed in the high resolution oligonucleotide array data indicates that, similar to data from other studies, some CNVs are extremely stable (without any detectable variability of breakpoints), while others appear to have variable breakpoints (Fig. 5 and Supplemental Table 4). The overall size of the copy number variable regions and the locations of specific breakpoints, e.g. those that impact functional genomic sequences, could influence the phenotypic manifestations of specific CNVs. It should be considered that there may be other factors involved in phenotypic associations of CNVs such as incomplete penetrance and/or instances where CNVs (or mutations, epigenetic effects, etc.) elsewhere in the genome may affect the clinical manifestation of specific CNVs. Whether bCNVs truly are benign remains a key question.

In the medical cytogenetics community, it is important that clinical correlations with detailed copy number change data are assessed and documented whenever possible. Data sharing and communication between laboratories performing aCGH (or copy number assessment by other methods) regarding findings and clinical correlates will improve our understanding of these variants, and how they should be interpreted.

Acknowledgements

We would like to thank Drs. Sarah South and Jennifer Warren for their thoughtful discussions and critical reading of this manuscript.

References

- Aston E, Whitby H, Maxwell T, Glaus N, Cowley B, et al: Comparison of targeted and whole genome analysis of postnatal specimens using a commercially available array based comparative genomic hybridisation (aCGH) microarray platform. *J Med Genet* 45:268–274 (2008).
- Brothman AR, Schneider NR, Saikovich I, Cooley LD, Butler MG, et al: Cytogenetic heteromorphisms: Survey results and reporting practices of Giemsa-band regions that we have pondered for years. *Arch Pathol Lab Med* 130:947–949 (2006).
- Carter NP: Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 39:S16–21 (2007).
- Curry C, Mao R, Aston E, Mongia S, Treisman T, et al: Homozygous deletions of a copy number change detected by array CGH: A new cause for mental retardation? *Am J Hum Genet Part A* 146A:1903–1910 (2008).
- de Ravel TJ, Balikova I, Thienpont B, Hannes F, Maas N, et al: Molecular karyotyping of patients with MCA/MR: The blurred boundary between normal and pathogenic variation. *Cytogenet Genome Res* 115:225–230 (2006).
- de Smith AJ, Tsalenko A, Sampas N, Scheffer A, Yamada NA, et al: 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 (2007).
- Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al: Detection of large-scale variation in the human genome. *Nat Genet* 36:949–951 (2004).
- Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al: Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998–1003 (2008).
- Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al: Mapping and sequencing of structural variation from eight human genomes. *Nature* 453:56–64 (2008).
- Koolen DA, Vissers LE, Pfundt R, de Leeuw N, Knight SJ, et al: A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet* 38:999–1001 (2006).
- Korbelt JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al: Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318:420–426 (2007a).
- Korbelt JO, Urban AE, Grubert F, Du J, Royce TE, et al: Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proc Natl Acad Sci USA* 104:10110–10115 (2007b).
- Krepischi-Santos AC, Vianna-Morgante AM, Jehee FS, Passos-Bueno MR, Knijnenburg J, et al: Whole-genome array-CGH screening in undiagnosed syndromic patients: Old syndromes revisited and new alterations. *Cytogenet Genome Res* 115:254–261 (2006).

- Lee C, Iafrate AJ, Brothman AR: Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* 39:S48–54 (2007).
- Lipson D, Tsalenko A, Yakhini Z, Ben-Dor A: Interval scores for quality annotated CGH data. Workshop on Genomic Signal Processing and Statistics (GENSIPS), Newport, Rhode Island (2005).
- Lipson D, Aumann Y, Ben-Dor A, Linial N, Yakhini Z: Efficient calculation of interval scores for DNA copy number data analysis. *J Comput Biol* 13:215–228 (2006).
- Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revena L, et al: The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* 82:685–695 (2008).
- Qiao Y, Liu X, Harvard C, Nolin SL, Brown WT, et al: Large-scale copy number variants (CNVs): Distribution in normal subjects and FISH/real-time qPCR analysis. *BMC Genomics* 8:167 (2007).
- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al: Global variation in copy number in the human genome. *Nature* 444:444–454 (2006).
- Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, et al: Challenges and standards in integrating surveys of structural variation. *Nat Genet* 39:S7–15 (2007).
- Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al: Large-scale copy number polymorphism in the human genome. *Science* 305:525–528 (2004).
- Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, et al: Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* 38:1038–1042 (2006).
- Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, et al: Microdeletion encompassing *MAPT* at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat Genet* 38:1032–1037 (2006).
- Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al: Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320:539–543 (2008).