ORIGINAL ARTICLE

Molecular karyotyping in patients with mental retardation using 100K single-nucleotide polymorphism arrays

Juliane Hoyer, Alexander Dreweke, Christian Becker, Ina Göhring, Christian T Thiel, Maarit M Peippo, Ralf Rauch, Michael Hofbeck, Udo Trautmann, Christiane Zweier, Martin Zenker, Ulrike Hüffmeier, Cornelia Kraus, Arif B Ekici, Franz Rüschenhoff, Peter Nürnberg, André Reis, Anita Rauch

J Med Genet 2007;44:629–636. doi: 10.1136/jmg.2007.050914

See end of article for authors' affiliations

Correspondence to:
Dr med Anita Rauch, Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander-University Erlangen-Nuremberg, Schwabachanlage 10, 91054 Erlangen, Germany; arauch@humgenet.uni-erlangen.de

Received 4 April 2007
Revised 19 June 2007
Accepted 20 June 2007
Published Online First
29 June 2007

Background: Using array techniques, it was recently shown that about 10% of patients with mental retardation of unknown origin harbour cryptic chromosomal aneusomies. However, data analysis is currently not standardised and little is known about its sensitivity and specificity.

Methods: We have developed an electronic data analysis tool for gene-mapping SNP arrays, a software tool that we call Copy Number Variation Finder (CNVF). Using CNVF, we analysed 104 unselected patients with mental retardation of unknown origin with a genechip mapping 100K SNP array and established an optimised set of analysis parameters.

Results: We detected deletions as small as 20 kb when covered by at least three single-nucleotide polymorphisms (SNPs) and duplications as small as 150 kb when covered by at least six SNPs, with only one false-positive signal in six patients. In 9.1% of patients, we detected apparently disease-causing or de novo aberrations ranging in size from 0.4 to 14 Mb. Morphological anomalies in patients with de novo aberrations were equal to that of unselected patients when measured with de Vries score.

Conclusion: Our standardised CNVF data analysis tool is easy to use and has high sensitivity and specificity. As some genomic regions are covered more densely than others, the genome-wide resolution of the 100K array is about 400–500 kb for deletions and 900–1000 kb for duplications. The detection rate of about 10% of de novo aberrations is independent of selection of patients for particular features. The incidental finding in two patients of heterozygosity for the 250 kb recurrent deletion at the *NPH1* locus, associated with autosomal recessive juvenile nephronophthisis, which was inherited from a healthy parent, highlights the fact that inherited aberrations might be disease-related even though not causal for mental retardation.

Mental retardation is defined as a significant impairment of cognitive and adaptive functions with onset at <18 years of age and has a prevalence of about 2–3%.¹ However, establishing an aetiological diagnosis is usually a challenge, and the underlying cause remains unknown in 65–80% of patients.^{2–4} As chromosomal aberrations are the most common known causes of mental retardation, genome-wide aneusomy screening methods have been developed using comparative genomic hybridisation (CGH) on 2.2–3.5K bacterial artificial chromosome (BAC) arrays with an average of about one clone per 1 Mb,^{5–9} so-called “tiling-path” 32K BAC arrays,¹⁰ and 10K and 100K single-nucleotide polymorphism (SNP) arrays initially designed for genotyping purposes.^{4, 11–13} However, little is known about the sensitivity and specificity of the various methods. As data generated with SNP arrays are suitable not only for analysis of copy-number variation, but also for the detection of uniparental disomies and for association studies based on SNP genotypes, we considered them the most powerful tools for research into mental retardation. However, the available analysis methods for detecting copy-number variation are time-consuming and non-standardised, usually relying on visual detection of aberrations in data plots. We therefore developed an easy to use data-analysis tool that we call Copy Number Variation Finder (CNVF), and analysed 104 patients with mental retardation of unknown origin for copy-number variations on a 100K SNP array and explored the sensitivity and specificity of different parameters for CNV definition.

METHODS

Patients

In total, 104 patients referred for developmental delay /mental retardation with or without multiple congenital anomalies were

recruited without further selection, if clinical evaluation with respective targeted molecular tests revealed no aetiological diagnosis. The tests were: karyotyping after GTG banding with a resolution of about 500 bands according to ISCN 2005;¹⁴ and subtelomeric screening by metaphase fluorescence in situ hybridisation (FISH), using the optimised BAC/P1-derived artificial chromosome (PAC) clone set described by Knight *et al.*¹⁵ Patients were categorised using the de Vries score,¹⁶ which measures the presence of malformations, dysmorphism, growth anomalies and familial occurrence of mental retardation as frequent signs in chromosomal aberrations. Fifteen patients also had normal results by molecular karyotyping with the 10K array (Affymetrix GeneChip 10K Mapping Array, V.1; Affymetrix, Santa Clara, USA).⁴

Methods

Hybridisation to mapping 100K arrays

From each patient, a sample of 500 ng of DNA extracted from fresh blood samples was processed according to the manufacturer's instructions (Affymetrix GeneChip Human Mapping 100K Assay Manual). In brief, 250 ng of high-quality genomic DNA per 50K array was digested with *Xba*I or *Hind*III and ligated to *Xba*I or *Hind*III adaptors, respectively. After PCR

Abbreviations: BAC, bacterial artificial chromosome; CGH, comparative genomic hybridisation; CNAG, Copy Number Analyser for GeneChip; CNAT, Copy Number Analysis Tool; CNVF, Copy Number Variation Finder; DFA, deterministic finite automata; FISH, fluorescence in situ hybridisation; GSA, genomic smoothing algorithm; OMIM, Online Mendelian Inheritance in Man; PAC, P1-derived artificial chromosome; qPCR, quantitative PCR; SNP, single-nucleotide polymorphism; SPA, single-point analysis

Figure 1 Screen shot of the Copy Number Variation Finder (CNVF) output using the optimised parameter set in patient 12. Note indication of the 17q and 22q confirmed deletions depicted in fig 2, and a duplication on chromosome 2 that was already present in the genomic-variation database. The lines flanking an aberration cluster give the flanking SNPs (separated by a semicolon for easy access to genome information on the respective aberration) in the University of California Santa Cruz genome browser. Columns contain the Copy Number Analysis Tool (CNAT) output information. Left to right: probe set, chromosome, physical position, call (ie genotype), single-point analysis copy number, single-point analysis p value, genomic smoothing algorithm copy number, genomic smoothing algorithm p value, loss of heterozygosity value. Note that the chromosomal position of SNPs given in the array data may differ from the actual genome browser information. The feature "homozygote only" means that SNP intensity values indicating a deletion are only labelled as deletion if the respective genotype is homozygous.

amplification, random fragmentation and labelling, samples were hybridised to 50 K arrays (GeneChip Human Mapping 50K *Xba*240 and *Hind*240 arrays; Affymetrix) in a hybridisation oven (Affymetrix Hybridisation Oven 640; Affymetrix). Washes and staining of the arrays were performed using a fluidics station (Affymetrix Fluidics Station 450; Affymetrix), and images were obtained using a gene-chip scanner (Affymetrix GeneChip Scanner 3000; Affymetrix). Call rates had to exceed 98%.

Data analysis

The raw data was analysed using the CNAT V.2.0.0.9 (Affymetrix) using 0.5 Mb sliding windows for the genomic smoothing algorithm (GSA). CNAT V.2 relies on the Affymetrix GDAS program but uses the same algorithm as CNAT V.3, which is implemented in the Affymetrix GTYPE program. The CNAT output was exported to a table, which was then filtered for clusters of SNPs fulfilling particular criteria using a self-programmed, modular-designed Java program, CNVF. The data processing is split into three steps. In the first step, a parser goes

over the input data and converts the data line by line into a format that can be processed by the other modules. During the transformation process, the parser classifies all intensity or probability values for each SNP into the categories (1) deletion, (2) duplication, (3) normal and (4) invalid. An SNP is classified as a deletion if its copy number or p value is below a particular configurable threshold. All SNPs that have a copy number above another threshold are marked as duplicated. Therefore, all SNPs neither below the first threshold nor above the second threshold are classified as normal SNPs. SNPs that could not be genotyped by the GDAS program ("no call") are marked as invalid SNPs and will be ignored by the following processing modules. After the parser has converted the input data and all SNPs have been classified, the next module, the matcher, processes the data. The matcher searches either for blocks of continuous SNPs or for patterns of at least three of five SNPs that fulfil particular conditions. The matchers are based on deterministic finite automata (DFA). These DFAs are automatically built at run time out of given regular expressions similar to those that are used in Perl for matching strings. The main difference, however, is that the regular expressions and

the resulting DFAs are able to work on objects instead of strings. All objects that are matched by the DFAs are stored for further processing. After the matcher has found a number of SNPs that fit the given pattern, a validator module checks for particular constraints that each cluster must fulfil. All data that do not pass the various conditions are filtered out.

After visual analysis of data values from the CNAT program in a patient with known 1.4 Mb PMP22 duplication, a patient with known 1.5–1.9 Mb 7q11.23 deletion, and four patients with known 3 Mb 22q11.2 deletion, two different parameter settings, one with a higher stringency to limit the number of possible aberrations (A1) and one with a lower stringency to allow the detection of smaller aberrations (A2) were established for this study. While the less stringent parameter set A2 would have allowed the detection of the known aberration in all six patients, the more stringent parameter set A1 would have detected the 1.4 Mb duplication, but would have missed the 1.5–1.9 Mb deletion and the 3 Mb deletion in three of four patients. For A1, GSA p values, which are $\pm \log_{10}$ values, were filtered for clusters of at least six adjacent SNPs with a value < -2.5 containing at least three SNPs with a value < -5 for deletions, or at least six adjacent SNPs with a value $> +2.5$ and containing at least three SNPs with a value $> +5$ for duplications. For A2, GSA p values were filtered for clusters of at least three adjacent SNPs with a value of at least ± 2.5 containing at least three SNPs with a single-point analysis (SPA) copy number value of ≤ 1.5 for deletions and ≥ 2.5 for duplications.

Validation of aberrations

Deletions of an expected maximum size of at least 100 kb were analysed by two-colour FISH with locus-specific BAC or PAC probes from the National Center for Biotechnology Information human genome map and subtelomeric control probes on metaphase spreads prepared from blood samples. Probes for the regions of interest were directly labelled with Cy3-dCTP (Amersham, Buckinghamshire, UK), and subtelomeric control

probes were directly labelled with SpectrumGreen-dUTP (Vysis, Des Plaines, Illinois, USA) using a nick-translation kit (Roche Diagnostics, Mannheim, Germany), according to the suppliers' instructions. Hybridisation was performed as described previously.¹⁷ For duplications and selected deletions apparently < 100 kb, quantitative PCR was performed using Taqman assays with minor-groove binder probes. The assays were performed with fourfold measurements, each multiplexed with an albumin probe as an internal reference, and calculation of copy numbers by the $\Delta\Delta C_t$ method as described previously.¹⁸

RESULTS

When applying the more stringent CNVF parameter set (A1), molecular karyotyping with the 100K SNP array revealed 10 unique deletions, 23 unique duplications and 9 different recurrent duplications (figs 1, 2 and 3A), whereas parameter set A2 revealed 42 unique deletions, 151 unique duplications, 53 different recurrent deletions and 314 recurrent duplications (fig 3B). Set A1 indicated a total of 92 aberrations in 53 of 104 patients (average of 0.9 aberrations per patient in the total study group; fig 3C), whereas set A2 indicated at least 1 of 1792 aberrations in every patient, with an average of 17.2 aberrations per patient (fig 3D). Distribution of maximum sizes of the respective aberrations is shown in figure 3 (E, F). Depending on the given SNP coverage of a particular region and the size of an aberration, the minimum aberration size as defined by the SNPs showing a particular copy-number variation may differ significantly from the maximum aberration size as defined by flanking SNPs with normal values.

Using FISH with locus-specific BAC/PAC probes, we verified 46 different potential unique and rare deletions with an estimated maximum size of ≥ 100 kb in 52 patients. All 10 deletions (maximum size 0.8–14 Mb) indicated by parameter set A1 were confirmed by FISH, whereas only 9 of 36 further investigated possible deletions indicated by set A2 could be confirmed by FISH or quantitative PCR (qPCR). However, 24 of

Figure 2 Examples of plots of $\pm \log_{10}$ GSA p values calculated with Affymetrix Copy Number Analysis Tool V.2 for copy numbers of SNPs of the Affymetrix GeneChip Human Mapping 100K SNP array in patient 12 using a GnuPlot program. Genotypes of individual SNPs are indicated by coloured bars at the bottom (magenta, homozygous; blue, heterozygous; brown, no call). Upper: plot of $\pm \log_{10}$ GSA p values for SNPs covering chromosome 17 in patient 12, with a 0.4 Mb de novo deletion in 17q21.31 indicated by Copy Number Variation Finder (CNVF) parameter set A1 and confirmed by fluorescence in situ hybridisation (arrow). Note magenta stretch of homozygosity corresponding to the deleted region. Lower: plot of $\pm \log_{10}$ GSA p values for SNPs covering chromosome 22 in the same patient showing a 250 kb inherited deletion in 22q11.23 indicated by CNVF parameter set A2 and confirmed by FISH (arrow). Note that the resolution of the plot does not allow determination of the number of SNPs passing the thresholds of ± 2.5 . This is achieved by the CNVF, which indicated in this specific case that 16 SNPs passed the threshold in the 17q21.31 deletion and 7 in the 22q11.23 deletion (fig 1). *Duplication identified in 11.5% of patients, which overlaps with a deletion at 28.1 Mb described as a normal variant.

Figure 3 Overview of results of molecular karyotyping with the 100K array using the Copy Number Variation Finder (CNVF) with parameter sets (A, C, E) A1 and (B, D, F) A2. (A, B) White bars, deletions; black bars, duplications. (A) Set A1 revealed 10 unique deletions, 23 unique duplications and 9 different recurrent duplications; (B) set A2 revealed 42 unique deletions, 151 unique duplications, 53 different recurrent deletions and 314 recurrent duplications. (C, D) Number of patients with a particular number of aberrations. (C) Using the stringent parameter set A1, 51 patients showed no aberration, whereas 53 of the 104 patients showed at least 1 of 92 aberrations (average of 0.9 aberrations per patient in the total study group); (D) A2 indicated at least 1 of 1792 aberrations in every patient, with an average of 17.2 aberrations per patient. (E, F) Size distribution of aberrations indicated by sets (E) A1 and (F) A2.

these possible deletions (18 with a minimum size of 769 bp to 80 kb and 6 with a minimum size of 140–370 kb) could not be confirmed by FISH. They were not further investigated by qPCR because of their small size and lack of gene content or overlap with known polymorphisms. Of 20 unique duplications indicated by parameter set A1 (size 0.2–4 Mb), 6 were confirmed by qPCR, whereas the remaining 14 possible duplications (minimum size 0.8 bp to 224 Mb; 3–10 SNPs) could not be confirmed. Three duplications indicated by A1 were not further investigated because they overlapped with normal gains described elsewhere. Rare duplications indicated only by the less stringent parameter set A2 were not further investigated, because the rate of false-positive signals was already 70% for the more stringent parameter set A1. As nine of the rare deletions confirmed by FISH or qPCR were indicated only by A2 (table 1), an optimised setting would thus use A2 for detection of deletions and A1 for duplications. This approach would result in 3 of 22 (14%) indicated rare deletions and 14 of 20 (70%) indicated rare duplications being false positive, with an average of 1 false-positive signal per 6 patients. However, if only deletions of at least 100 kb and duplications of at least 900 kb were followed, specificity would be 100%, with the loss of all four inherited duplications and only one inherited deletion.

One 14 Mb deletion containing 59 genes was considered to be disease-causing because of its size and gene content, but the

parents of the proband withdrew consent to have themselves test. Eight of the further 18 (44%) confirmed deletions and 2 of the 6 (33%) confirmed duplications occurred de novo, giving a total de novo rate of 42% (tables 1 and 2). Inherited aberrations had maximum sizes between 0.2 and 2 Mb (minimum sizes 20 kb–2 Mb), and de novo aberrations had maximum sizes between 0.85 and 12.9 Mb (minimum sizes 0.4–12.8 Mb) and were thus statistically significantly larger ($p = 0.001$, Wilcoxon test). Note that the actual size of a confirmed aberration was defined according to the copy number of SNPs within and flanking a cluster with high GSA p values and not according to the GSA p values of the SNPs themselves, which were only used to find aberrations. We have previously shown that p values may overestimate or underestimate aberration sizes, whereas copy-number values correspond very well to FISH results.¹¹

Array data from patients with confirmed aberrations were also investigated with the program Copy Number Analyser for GeneChip (CNAG),¹⁹ which analyses data from the two 50 kb arrays (*Hind* and *Xba*), separately. One de novo and 6 inherited aberrations were missed by both arrays, and 4 de novo and 6 inherited aberrations were indicated by one array only (table 1). Apart from the confirmed rare aberrations, the CNAG program indicated 105 small aberrations in these 23 patients, with only 1 overlap with the CNVF data (mean 5 aberrations/patient; range 1–13). We used the 100K array on the 15 patients who had

Table 1 Overview of detected aberrations verified by FISH or quantitative PCR

Patient	CNV type	Location	Origin	5' flanking SNP (location in bp)	3' flanking SNP (location in bp)	Minimum size (Mb)	CNVF			CNAG		Reported
							SNPs	A 1	A 2	Xba	Hind	
104	Deletion	1p32.2–31.1	*	rs10493211 (56728933)	rs10493473 (70560236)	14.00	814	+	+	+	+	
77	Deletion	5q14.3–15	De novo	rs10514301 (87975410)	rs9314105 (93668872)	5.60	234	+	+	+	+	
96	Deletion	6q16.1–21	De novo	rs726488 (99645914)	rs2010315 (112529340)	12.80	577	+	+	+	+	
45	Deletion	7p22.3–22.2	De novo	rs10487884 (2678041)	rs314593 (4257670)	1.50	76	+	+	+	+	Friedman ¹² (3.6 Mb)
42	Deletion	16p12.1	De novo	rs999499 (24061164)	rs10521131 (24919925)	0.70	22	0	+	0	0	
72	Duplication	17p11.2	De novo	rs8074464 (16475319)	rs10512425 (20564580)	3.70	31	+	+	0	+	Potocki ²⁰
12	Deletion	17q21.31	De novo	rs10491142 (40523389)	rs3912061 (41520916)	0.40	31	+	+	0	+	Multiple ^{7 21–23}
5	Deletion	18q21.2	De novo	rs4800947 (50552638)	rs784395 (51742365)	1.20	62	+	+	+	0	
85	Deletion	19p13.11–p12	De novo	rs876982 (18643377)	rs10518280 (23846921)	4.50	65	+	+	+	+	
85	Duplication	19q13.11– q13.12	De novo	rs7249990 (39870683)	rs7255938 (42606994)	2.70	31	+	+	+	0	
100	Deletion	21q22.11– q22.12	De novo	rs8129832 (32329260)	rs10483024 (35362570)	3.00	116	+	+	+	+	
63	Duplication	1p36.13	Maternal	rs794230 (18034558)	rs10489573 (18430508)	0.44	10	+	+	0	0	Loss, Wong ²⁴
9	Duplication	2p22.3	Paternal	rs458628 (32427049)	rs1902030 (33232242)	0.63	31	+	+	+	0	Gain, Simon- Sanchez ²⁵
55, 85	Deletion	2q13	Paternal	rs951954 (109816880)	rs1002016 (111333143)	0.13	4	0	+	0	0	Gain, Redon ²⁶
51	Deletion	2q32.1	Paternal	rs9288111 (185889502)	rs850889 (186033754)	0.13	5	0	+	0	0	
24	Deletion	3q24	Maternal	rs6770324 (146954460)	rs3804663 (147293301)	0.23	20	0	+	0	0	Loss, Redon ²⁶
61	Duplication	5q33.2	Maternal	rs153431 (153629119)	rs1875473 (153840497)	0.15	7	+	+	0	0	
57	Deletion	5q34	Maternal	rs10515859 (162777146)	rs6864913 (164840549)	2.00	120	+	+	+	+	Menten ²⁷
29	Deletion	7q31.1	Maternal	rs1014799 (111153998)	rs2249151 (111580754)	0.35	23	0	+	0	+	Gain, Simon- Sanchez ²⁵
92	Deletion	9p23	Maternal	rs983282 (11834120)	rs1416572 (12177104)	0.20	21	0	+	+	0	Loss, Redon ²⁶
56	Deletion	9p21.1	Paternal	rs16915373 (29850957)	rs6476179 (29975815)	0.02	7	0	+	+	0	Redon ²⁶
31	Duplication	11p11.2	Paternal	rs3913231 (49516614)	rs1580785 (51150171)	0.85	23	+	+	0	+	Loss, Redon ²⁶
41	Deletion	13q21.33	Maternal	rs10492467 (70639393)	rs1841359 (71488074)	0.80	59	+	+	+	+	Redon ²⁶
102	Deletion	19q13.2	Paternal	rs9304608 (47889421)	rs10518262 (48326010)	0.12	3	0	+	0	0	Loss, Redon, ²⁶ Simon- Sanchez ²⁵
12	Deletion	22q11.23	Maternal	rs10483117 (23956820)	rs3730109 (24399560)	0.26	9	0	+	0	+	Redon ²⁶

*Parents were not available for investigation.

+, Aberration indicated by the respective analysis; 0, aberration not indicated by the respective analysis.

Location is given using coordinates from National Center for Biotechnology information assembly 35 of the human genome sequence.

normal results with the 10K SNP array in our pilot study;⁴ one (patient 12), was found to carry a de novo 0.4 Mb microdeletion in 17q21.31 and a maternally inherited 260 kb deletion in 22q11.23, and another patient (patient 51) had a 130 kb paternally inherited deletion in 2q32.1.

Morphological anomalies measured by the de Vries score (range 0–10) in the 104 patients of our entire study group showed a similar distribution and a similar average score of 3.1 (range 0–8) to those observed in patients with de novo or apparently disease-causing aberrations, who had an average score of 3.0 (range 1–5) (fig 4).

DISCUSSION

Our results show that the CNVF is an easy tool to use for standardised detection of copy-number variations, and has high

sensitivity and specificity using intensity data from GeneChip Mapping SNP arrays as calculated by the Affymetrix copy-number tool (V2 or V3). In contrast to other approaches such as the Affymetrix visualisation tool for the copy-number tool data or the CNAG program,¹⁹ our CNVF tool circumvents time-consuming and non-standardised visual data analysis. We established an optimised set of analysis parameters for the 100K array, resulting in detection of deletions when covered by at least three SNPs on the array and of duplications when covered by at least six SNPs, with a low false-positive rate of only one false aberration per six patients. Specificity of indicated aberrations can even be increased to 100% without loss of de novo or apparently disease-causing aberrations if deletions with a minimum size of 100 kb and duplications with a minimum size of 900 kb only are followed. The low

Table 2 Overview of clinical findings in patients with de novo aberrations

Patient	Aberration	Size (Mb)	Short stature	Major morphological anomalies	Seizures	De Vries score
104	Del 1p32.2-31.1*	14.00	+	TOF, HC, DK, JA, HCC	—	5
77	Del 5q14.3-15	5.60	—	HCC	+	1
96	Del 6q16.1-21	12.80	—	DORV	—	2
45	Del 7p22.3-22.2	1.50	—	CP, scoliosis, TOF	—	4
42	Del 16p12.1	0.70	—	—	—	1
72	Dup 17p11.2	3.70	+	—	—	3
12	Del 17q21.31	0.40	+	VSDs	+	4
5	Del 18q21.2	1.20	+	HSCR, HCC, MC ³¹	+	3
85	Del 19p13.11-p12	4.50	—	TOF, macrocephalus	—	4
	Dup 19q13.11-q13.12	2.70				
100	Del 21q22.11-q22.12	3.00	+	—	—	3

CP, cleft palate; DK, double kidney; DORV, double outlet right ventricle; HC, hydrocephalus; HCC, hypoplasia of corpus callosum; HSCR, Hirschsprung disease; JA, jejunal atresia; MC, microcephaly; TOF, tetralogy of Fallot; VSD, ventricle septum defect.

*Data from the parents of this patient were not available, but because of the large deletion size, this aberration was considered to be disease-causing.

false-positive rate is especially important, as confirmation of aberrations with independent methods is laborious and time-consuming. Using visual analysis of our confirmed aberrations with the CNAG program, we were unable to detect one de novo and six inherited aberrations identified in our study (table 1). This might be explained by the separate analysis of the two 50K datasets by this program, which increases the gaps in genome-wide SNP coverage. The CNAG program is also based on a different algorithm to calculate copy-number values, and requires at least four adjacent SNPs with a similar copy-number change to reach a false-positive rate of ≤ 4 per patient.¹² This shows that our CNVF tool has excellent specificity with the described settings, although its sensitivity with respect to CNAG could not be established. However, as the SNPs on the 100K array are not evenly distributed and thus some areas of the genome are covered more densely than others, the genome-wide resolution of this array of our approach is >100 kb. Only 92% of the genome is within 100 kb of an SNP and 83% is within 50 kb of a SNP, thus the genome-wide resolution is more likely to be about 400–500 kb for the detection of deletions requiring three adjacent SNPs, and about 900–1000 kb for duplications requiring six adjacent SNPs.

We were able to detect apparently disease-causing or de novo cryptic chromosomal aberrations in 10 of 104 unselected patients with mental retardation of unknown aetiology after normal karyotyping and subtelomeric screening. Minimum

sizes of aberrations varied from 0.4 to 14 Mb. In total, 15 of these patients had not had aberrations found when analysed with the 10K SNP array.⁴ One of these patients was found with our program to carry a de novo 0.4 Mb microdeletion in 17q21.31. When correcting for the patient with a 5 Mb de novo 6q22 deletion who had already been identified in that pilot study, we obtained a detection rate of de novo non-subtelomeric aberrations of 10.5% (11 of 105 patients), which is in the same range as the 11% previously reported in 100 patients with mental retardation studied with the GeneChip Mapping 100K array¹² and the 10% reported in 100 patients studied with a tiling 32K BAC array.¹⁰ Despite the approximately 10 times lower resolution, a similar detection rate of 9.1% non-subtelomeric disease-related aberrations was also achieved with 2.6–3.5K BAC arrays in a total of 374 patients.^{5–8, 27–29} This might be explained by our data and previously published data¹⁰ showing that the majority of de novo and thus probably disease-related aberrations are >1 Mb and as a group are significantly larger than inherited variants. It is also noteworthy that we observed the same rate of de novo aberrations in unselected patients as found in previous studies that had selected patients for a particular number of major and minor anomalies. In contrast, patients diagnosed with monogenic disorders, recurrent microdeletion syndromes and subtelomeric aberrations showed higher average de Vries scores (3.4, 3.6 and 3.8, respectively).⁴ Thus, although a de Vries score of ≥ 3 might be useful for selecting patients with a higher probability of having a subtelomeric aberration,¹⁶ this does not seem to be a good predictor for the presence of an interstitial aberration.

The majority of de novo aberrations we detected are unique, but two of our patients showed microaberrations that have recently been reported. One of the latter overlaps with the up to 0.6 Mb 17q21.31 microdeletion reported to occur with a frequency of 0.3% in patients with mental retardation, and results in a novel distinct phenotype (OMIM 610443).^{21–23} The second corresponds to the 3.7 Mb duplication of the Smith–Magenis syndrome region at 17p11.2, which is associated with a less distinct phenotype also referred to as Potocki–Lupski syndrome.^{20, 30} The 1.5 Mb de novo deletion in 7p22.3–22.2 in patient 45 overlaps with a 3.6 Mb de novo deletion described by Friedman *et al* in a patient with mental retardation.¹² One of our patients with a 1.2 Mb deletion in 18q21.2 led to the identification of the underlying gene defect of the Pitt–Hopkins syndrome, which was a poorly established entity with no OMIM number.³¹

Our de novo aberrations represented 42% of the confirmed aberrations. This proportion is much lower than the 65–89% de

Figure 4 Distribution of de Vries scores in all 104 patients (black areas) and those with apparently disease-causing or de novo aberration (white).

novo aberrations reported by others using 3.5K BAC arrays,^{9, 27} which is probably explained by the higher resolution of the 100K SNP array and the fact that de novo aberrations are usually larger than polymorphisms. Apart from the 11 apparently disease-causing or de novo aberrations observed in 10 patients, we obtained an average of 1.5 copy-number variations per patient with a median of 1 using our optimised data-analysis parameter set. Interestingly, 36 patients (35%) showed no copy-number variation at all. These numbers are relatively low compared with the average of 104 copy-number variations detected with tiling-path BAC arrays in normal people,²⁴ which is probably explained by the stringency of our algorithm to avoid false positives but possibly also to the fact that the 100K SNP array does not cover highly polymorphic repeat regions. Molecular karyotyping has recently been criticised for the need to follow up many rare aberrations, with a high number found to be inherited from healthy parents.³² This problem, if one at all, is only temporary, as is evident from the fact that at the beginning of our study none of the inherited rare aberrations observed was known, but since then, most have been described and deposited in the database of genomic variants (table 1). However, one of the inherited deletions (2q13), which is also described as a normal genomic variant in the database, corresponds to the 250 kb recurrent deletion of the *NPH1* locus, which is the major mutation in autosomal recessive juvenile nephronophthisis (OMIM 256100).³³ This *NPH1* deletion occurred in the heterozygous state in two patients and one of each respective parent, none of whom had signs of nephronophthisis, giving a 2% heterozygous carrier frequency in our study group of European descent. This finding emphasises that inherited aberrations can be recessive or oligogenic mutations, as also observed in the recessive Peters Plus syndrome (OMIM 261540)³⁴ or the thrombocytopenia and absent radius (TAR) syndrome with complex inheritance.³⁵ Thus, inherited copy-number variations might not be associated with the disorder this investigation was intended to study and might represent at-risk alleles for late-onset disorders such as breast cancer,³⁴ an issue that should be taken into account when obtaining informed consent from patients and their families.

Our data show that the CNVF, which is available on request, is a powerful tool to analyse data from the GeneChip Mapping 100K array for molecular karyotyping. We detected verifiable rare deletions as small as 20 kb and duplications as small as 150 kb with a high specificity of only one false-positive aberration in six patients. Using our optimised parameter set and given the average SNP spacing of 23.6 kb, the theoretical resolution is 48 kb for deletions and 120 kb for duplications. However, as some genomic regions are covered more densely than others and some regions including the subtelomeric regions being poorly covered in this array, the actual genome-wide resolution is estimated to about 400–500 kb for deletions and 900–1000 kb for duplications.

Electronic database information

- National Center for Biotechnology Information (NCBI): <http://www.ncbi.nlm.nih.gov/>
- Genome Browser of the University of California Santa Cruz (UCSC): <http://genome.ucsc.edu/cgi-bin/hgTracks>
- Online Mendelian Inheritance in Man (OMIM): <http://www.ncbi.nlm.nih.gov/Omim>.
- Database of Genomic Variants: <http://projects.tcag.ca/variation/>

ACKNOWLEDGEMENTS

We thank the family members for their kind participation, and Michaela Kirsch and Sabrina Schulz for excellent technical assistance. This work was supported by grant RA 833/7-1 (AnR, PN) funded by the Deutsche Forschungsgemeinschaft (DFG).

Authors' affiliations

Juliane Hoyer, Ina Göhring, Christian T Thiel, Udo Trautmann, Christiane Zweier, Martin Zenker, Ulrike Hüffmeier, Cornelia Kraus, Arif B Ekici, André Reis, Anita Rauch, Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany

Alexander Dreweke, Computer Science Department 2, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany

Christian Becker, Peter Nürnberg, Cologne Center for Genomics and Center for Molecular Medicine Cologne, University Cologne, Germany

Maarit M Peippo, Department of Medical Genetics, Family Federation of Finland, Helsinki, Finland

Ralf Rauch, Michael Hofbeck, Pediatric University Hospital Tübingen, Tübingen, Germany

Franz Rüschemdorf, Gene Mapping Center, Max-Delbrück-Center for Molecular Medicine Berlin-Buch, Germany

Competing interests: None declared.

REFERENCES

- 1 Ropers HH, Hamel BC. X-linked mental retardation. *Nat Rev Genet* 2005;6:46–57.
- 2 Flint J, Knight S. The use of telomere probes to investigate submicroscopic rearrangements associated with mental retardation. *Curr Opin Genet Dev* 2003;13:310–16.
- 3 Hunter AG. Outcome of the routine assessment of patients with mental retardation in a genetics clinic. *Am J Med Genet* 2000;90:60–8.
- 4 Rauch A, Hoyer J, Guth G, Zweier C, Kraus C, Becker C, Zenker M, Hüffmeier U, Thiel C, Rüschemdorf F, Nürnberg P, Reis A, Trautmann U. Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. *Am J Med Genet* 2006;140A:2063–74.
- 5 Miyake N, Shimokawa O, Harada N, Sosonkina N, Okubo A, Kawara H, Okamoto N, Kurosawa K, Kawame H, Iwakoshi M, Kosho T, Fukushima Y, Makita Y, Yokoyama Y, Yamagata T, Kato M, Hiraki Y, Nomura M, Yoshiura K, Kishino T, Ohta T, Mizuguchi T, Niikawa N, Matsumoto N. BAC array CGH reveals genomic aberrations in idiopathic mental retardation. *Am J Med Genet A* 2006;140:205–11.
- 6 Schoumans J, Ruivenkamp C, Holmberg E, Kyllerman M, Anderlid BM, Nordenskjöld M. Detection of chromosomal imbalances in children with idiopathic mental retardation by array based comparative genomic hybridisation (array-CGH). *J Med Genet* 2005;42:699–705.
- 7 Shaw-Smith C, Redon R, Rickman L, Rio M, Willatt L, Fiegler H, Firth H, Sanlaville D, Winter R, Colleaux L, Bobrow M, Carter NP. Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J Med Genet* 2004;41:241–8.
- 8 Vissers LE, de Vries BB, Osoegawa K, Janssen IM, Feuth T, Choy CO, Straatman H, van der Vliet W, Huys EH, van Rijk A, Smeets D, van Ravenswaaij-Arts CM, Knoers NV, van der Burg I, de Jong PJ, Brunner HG, van Kessel AG, Schoenmakers EF, Veltman JA. Array-based comparative genomic hybridization for the genome-wide detection of submicroscopic chromosomal abnormalities. *Am J Hum Genet* 2003;73:1261–70.
- 9 de Ravel TJ, Balikova I, Thienpont B, Hannes F, Maas N, Frys JP, Devriendt K, Vermeesch JR. Molecular karyotyping of patients with MCA/MR: the blurred boundary between normal and pathogenic variation. *Cytogenet Genome Res* 2006;115:225–30.
- 10 de Vries BB, Pfundt R, Leisink M, Koolen DA, Vissers LE, Janssen IM, Reijmersdal S, Nillesen WM, Huys EH, Leeuw N, Smeets D, Sistermans EA, Feuth T, van Ravenswaaij-Arts CM, van Kessel AG, Schoenmakers EF, Brunner HG, Veltman JA. Diagnostic genome profiling in mental retardation. *Am J Hum Genet* 2005;77:606–16.
- 11 Rauch A, Rüschemdorf F, Huang J, Trautmann U, Becker C, Thiel C, Jones KW, Reis A, Nürnberg P. Molecular karyotyping using an SNP array for genome-wide genotyping. *J Med Genet* 2004;41:916–22.
- 12 Friedman JM, Baross A, Delaney AD, Ally A, Arbour L, Armstrong L, Asano J, Bailey DK, Barber S, Birch P, Brown-John M, Cao M, Chan S, Charest DL, Farnoud N, Fernandes N, Flibotte S, Go A, Gibson WT, Holt RA, Jones SJ, Kennedy GC, Krzywinski M, Langlois S, Li HI, McGillivray BC, Nayar T, Pugh TJ, Rajcan-Separovic E, Schein JE, Schnerch A, Siddiqui A, Van Allen MI, Wilson G, Yong SL, Zahir F, Eyedoux P, Marra MA. Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation. *Am J Hum Genet* 2006;79:500–13.
- 13 Ming JE, Geiger E, James AC, Cipraro KL, Nimmakayalu M, Zhang Y, Huang A, Vaddi M, Rappaport E, Zackai EH, Shaikh TH. Rapid detection of submicroscopic chromosomal rearrangements in children with multiple congenital anomalies using high density oligonucleotide arrays. *Hum Mutat* 2006;27:467–73.

- 14 **Shaffer L**, Tommerup N. *ISCN: An international system for human cytogenetic nomenclature*. Basel: Karger, 2005.
- 15 **Knight SJ**, Lese CM, Precht KS, Kuc J, Ning Y, Lucas S, Regan R, Brenan M, Nicod A, Lawrie NM, Cardy DL, Nguyen H, Hudson TJ, Riethman HC, Ledbetter DH, Flint J. An optimized set of human telomere clones for studying telomere integrity and architecture. *Am J Hum Genet* 2000;**67**:320–32.
- 16 **de Vries BB**, White SM, Knight SJ, Regan R, Homfray T, Young ID, Super M, McKeown C, Splitt M, Quarrell OW, Trainer AH, Niermeijer MF, Malcolm S, Flint J, Hurst JA, Winter RM. Clinical studies on submicroscopic subtelomeric rearrangements: a checklist. *J Med Genet* 2001;**38**:145–50.
- 17 **Rauch A**. Epithelial cells from buccal smears and urine. In: Rautenstrauss B, Liehr T, eds. *FISH technologies. Springer lab manual*. Heidelberg: Springer, 2002;97–108.
- 18 **Thiel CT**, Kraus C, Rauch A, Ekici AB, Rautenstrauss B, Reis A. A new quantitative PCR multiplex assay for rapid analysis of chromosome 17p11.2–12 duplications and deletions leading to HMSN/HNPP. *Eur J Hum Genet* 2003;**11**:170–8.
- 19 **Nannay Y**, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC, Ogawa S. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res* 2005;**65**:6071–9.
- 20 **Potocki L**, Bi W, Treadwell-Deering D, Carvalho CM, Eifert A, Friedman EM, Glaze D, Krull K, Lee JA, Lewis RA, Mendoza-Londono R, Robbins-Furman P, Shaw C, Shi X, Weissberger G, Withers M, Yatsenko SA, Zackai EH, Stankiewicz P, Lupski JR. Characterization of Potocki-Lupski syndrome (dup(17)(p11.2p11.2)) and delineation of a dosage-sensitive critical interval that can convey an autism phenotype. *Am J Hum Genet* 2007;**80**:633–49.
- 21 **Koolen DA**, Vissers LE, Pfundt R, de Leeuw N, Knight SJ, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M, Schinzel A, Baumer A, Anderlid BM, Schoumans J, Knoers NV, van Kessel AG, Sistermans EA, Veltman JA, Brunner HG, de Vries BB. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet* 2006;**38**:999–1001.
- 22 **Sharp AJ**, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, Eichler EE. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* 2006;**38**:1038–42.
- 23 **Shaw-Smith C**, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S, Curley R, Cumming S, Dunn C, Kalaitzopoulos D, Porter K, Prigmore E, Krepischi-Santos AC, Varela MC, Koiffmann CP, Lees AJ, Rosenberg C, Firth HV, de Silva R, Carter NP. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat Genet* 2006;**38**:1032–7.
- 24 **Wong KK**, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL. A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet* 2007;**80**:91–104.
- 25 **Simon-Sanchez J**, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, de Vries BB, Peckham E, Gwinn-Hardy K, Crawley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A. Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum Mol Genet* 2007;**16**:1–14.
- 26 **Redon R**, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MU, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy number in the human genome. *Nature* 2006;**444**:444–54.
- 27 **Menten B**, Maas N, Thienpont B, Buysse K, Vandesompele J, Melotte C, de Ravel T, Van Vooren S, Balikova I, Backx L, Janssens S, De Paepe A, De Moor B, Moreau Y, Marynen P, Fryns JP, Mortier G, Devriendt K, Speleman F, Vermeesch JR. Emerging patterns of cryptic chromosomal imbalance in patients with idiopathic mental retardation and multiple congenital anomalies: a new series of 140 patients and review of published reports. *J Med Genet* 2006;**43**:625–33.
- 28 **Rosenberg C**, Knijnenburg J, Bakker E, Vianna-Morgante AM, Sloos W, Otto PA, Kriek M, Hansson K, Krepischi-Santos AC, Fiegler H, Carter NP, Bijlsma EK, van Haeringen A, Szuhai K, Tanke HJ. Array-CGH detection of micro rearrangements in mentally retarded individuals: clinical significance of imbalances present both in affected children and normal parents. *J Med Genet* 2006;**43**:180–6.
- 29 **Tyson C**, Harvard C, Locker R, Friedman JM, Langlois S, Lewis ME, Van Allen M, Somerville M, Arbour L, Clarke L, McGilvray B, Yong SL, Siegel-Bartel J, Rajcan-Separovic E. Submicroscopic deletions and duplications in individuals with intellectual disability detected by array-CGH. *Am J Med Genet A* 2005;**139**:173–85.
- 30 **Potocki L**, Chen KS, Park SS, Osterholm DE, Withers MA, Kimonis V, Summers AM, Meschino WS, Anyane-Yeboah K, Kashork CD, Shaffer LG, Lupski JR. Molecular mechanism for duplication 17p11.2, the homologous recombination reciprocal of the Smith-Magenis microdeletion. *Nat Genet* 2000;**24**:84–7.
- 31 **Zweier C**, Peippo MM, Hoyer J, Sousa S, Bottani A, Clayton-Smith J, Reardon W, Saraiva J, Cabral A, Göhring I, Devriendt K, de Ravel T, Bijlsma EK, Hennekam RCM, Orrico A, Cohen M, Dreweke A, Reis A, Nürnberg P, Rauch A. Haploinsufficiency of TCF4 causes syndromal mental retardation with intermittent hyperventilation (Pitt-Hopkins syndrome). *Am J Hum Genet* 2007;**80**:994–1001.
- 32 **Subramonia-Iyer S**, Sanderson S, Sagoo G, Higgins J, Burton H, Zimmern R, Kroese M, Brice P, Shaw-Smith C. Array-based comparative genomic hybridization for investigating chromosomal abnormalities in patients with learning disability: Systematic review meta-analysis of diagnostic and false-positive yields. *Genet Med* 2007;**9**:74–9.
- 33 **Konrad M**, Saunier S, Heidet L, Silbermann F, Benessy F, Calado J, Le Paslier D, Broyer M, Gubler MC, Antignac C. Large homozygous deletions of the 2q13 region are a major cause of juvenile nephronophthisis. *Hum Mol Genet* 1996;**5**:367–71.
- 34 **Lesnik-Oberstein SA**, Kriek M, White SJ, Kalf ME, Szuhai K, den Dunnen JT, Breuning MH, Hennekam RC. Peters Plus syndrome is caused by mutations in B3GALT1, a putative glycosyltransferase. *Am J Hum Genet* 2006;**79**:562–6.
- 35 **Klopocki E**, Schulze H, Strauss G, Ott CE, Hall J, Trotier F, Fleischhauer S, Greenhalgh L, Newbury-Ecob RA, Neumann LM, Habenicht R, König R, Seemanova E, Megarbane A, Ropers HH, Ullmann R, Horn D, Mundlos S. Complex inheritance pattern resembling autosomal recessive inheritance involving a microdeletion in thrombocytopenia-absent radius syndrome. *Am J Hum Genet* 2007;**80**:232–40.