European Journal of Human Genetics (2014) 22, 652–659

www.nature.com/ejhg

ARTICLE

Mate pair sequencing for the detection of chromosomal aberrations in patients with intellectual disability and congenital malformations

Sarah Vergult[1,5], Ellen Van Binsbergen[2,5], Tom Sante[1,5], Silke Nowak[1], Olivier Vanakker[1], Kathleen Claes[1], Bruce Poppe[1], Nathalie Van der Aa[3], Markus J van Roosmalen[2], Karen Duran[2], Masoumeh Tavakoli-Yaraki[2], Marielle Swinkels[2], Marie-José van den Boogaard[2], Mieke van Haelst[2], Filip Roelens[4], Frank Speleman[1], Edwin Cuppen[2], Geert Mortier[1,3], Wigard P Kloosterman*[,2,5] and Björn Menten*[,1,5]

Recently, microarrays have replaced karyotyping as a first tier test in patients with idiopathic intellectual disability and/or multiple congenital abnormalities (ID/MCA) in many laboratories. Although in about 14–18% of such patients, DNA copy-number variants (CNVs) with clinical significance can be detected, microarrays have the disadvantage of missing balanced rearrangements, as well as providing no information about the genomic architecture of structural variants (SVs) like duplications and complex rearrangements. Such information could possibly lead to a better interpretation of the clinical significance of the SV. In this study, the clinical use of mate pair next-generation sequencing was evaluated for the detection and further characterization of structural variants within the genomes of 50 ID/MCA patients. Thirty of these patients carried a chromosomal aberration that was previously detected by array CGH or karyotyping and suspected to be pathogenic. In the remaining 20 patients no causal SVs were found and only benign aberrations were detected by conventional techniques. Combined cluster and coverage analysis of the mate pair data allowed precise breakpoint detection and further refinement of previously identified balanced and (complex) unbalanced aberrations, pinpointing the causal gene for some patients. We conclude that mate pair sequencing is a powerful technology that can provide rapid and unequivocal characterization of unbalanced and balanced SVs in patient genomes and can be essential for the clinical interpretation of some SVs.
European Journal of Human Genetics (2014) **22,** 652–659; doi:10.1038/ejhg.2013.220; published online 9 October 2013

Keywords: intellectual disability; mate pair sequencing; array CGH; structural variation

INTRODUCTION

Structural variations (SVs) have been recognized as an important cause of intellectual disability and multiple congenital abnormalities (ID/MCA) for many years.[1,2] An SV is defined as a difference in the DNA copy-number, orientation or location of relatively large genomic segments (typically >1 kb)[3] and may include deletions, duplications, insertions, inversions and translocations.[4] Genomic microarrays have been instrumental for the identification of one type of SV being submicroscopic copy-number variants (CNVs) in patients with idiopathic ID and congenital anomalies (reviewed in Vissers *et al*[5]). The diagnostic yield in studies using genomic microarrays for these patients is around 14–18%,[6–10] which is a major improvement compared with conventional karyotyping. The genetic cause, however, remains elusive in a large proportion of patients with ID/MCA, and it is generally assumed that these patients' genomes harbor hitherto undetected genomic alterations.

In the last few years, next-generation sequencing (NGS) has emerged as a very powerful technology and has led to the identification of the causal gene for many rare Mendelian disorders.[11–13] In this study, paired-end mapping or mate pair sequencing was used. In comparison with conventional paired-end sequencing, the whole genome can be interrogated for structural variations with less sequence reads, while reaching the same physical coverage. Further, mate pair sequencing facilitates mapping across small repetitive regions because of its longer insert sizes.[14] For this technique, genomic DNA is fragmented into preset fragment lengths (= insert size, eg, 3 kb) of which the ends (= mates) are sequenced by paired-end sequencing. This technology allows the discrimination between concordant mates (reads that map 3 kb from each other on the reference genome with correct orientation) and discordant mates (= reads that map closer or further than 3 kb and/or with incorrect orientation). In this way, structural variations, both balanced as well as unbalanced, can be detected. Korbel *et al*[15] were the first to use NGS to map structural variations in the human genome, and several other groups have used this technique to finemap the breakpoints of specific structural aberrations (mostly apparently

[1]Center for Medical Genetics, Ghent University, Ghent, Belgium; [2]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands; [3]Department for Medical Genetics, University Hospital of Antwerp, Antwerp, Belgium; [4]Heilig Hart Ziekenhuis, Roeselare, Belgium
[5]These authors contributed equally to this work.
*Correspondence: Professor WP Kloosterman, Department of Medical Genetics, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands. Tel: +31 88 756 8082; Fax: +31 88 756 8479; E-mail: W.Kloosterman@umcutrecht.nl
or Dr Professor B Menten, Center for Medical Genetics, Ghent University, De Pintelaan 185, 9000 Ghent, Belgium. Tel: +32 9 332 52 84; Fax: +32 9 332 65 49; E-mail: Bjorn.Menten@ugent.be
Received 29 March 2013; revised 13 August 2013; accepted 29 August 2013; published online 9 October 2013

balanced chromosomal aberrations) in patients with ID/MCA.[16–23] Therefore, the technique has proven its usability in characterizing individual SVs. Here we describe the first systematic comparison between mate pair sequencing, genomic microarrays and karyotyping in a large cohort of ID/MCA patients referred to our diagnostic departments.

Our aim was threefold: first, we determined whether mate pair sequencing enables the identification of all previously detected balanced, unbalanced and complex chromosomal aberrations. Second, we explored the additional clinical value of mate pair sequencing in determining the precise structure of pathogenic SVs and the effects on underlying genes. Third, we evaluated whether the high resolution of mate pair sequencing could lead to an improved diagnostic yield.

MATERIALS AND METHODS

Collection of patients

DNA samples from patients with ID and/or congenital anomalies were collected from the genetic centers of Ghent (Belgium), Antwerp (Belgium) and Utrecht (The Netherlands). From five patients, the parent samples were also collected and investigated with mate pair sequencing. Informed consent was obtained from parents of all the patients.

Chromosome analysis

Analysis of G-banded metaphase chromosomes was performed on short-term lymphocyte cultures using standard procedures. For fluorescent *in situ* hybridization (FISH), probes were labeled with SpectrumGreen or SpectrumOrange with the nick translation kit (Abbott Molecular, Ottignies, Belgium) according to the manufacturer's instructions. FISH was performed as previously described.[24]

Array CGH

Copy-number profiling was performed using 105 K (amadid#019015) or 180 K (amadid#023363) Human Genome CGH Microarray slides from Agilent Technologies (Santa Clara, CA, USA) following the manufacturer's protocols.

Data analysis and visualization was done with our in-house developed webtool Vivar (http://www.medgen.ugent.be/vivar/; Sante *et al*, in preparation). DNA CNVs were identified by circular binary segmentation.[25] Interpretation of CNV data was performed as described in Buysse *et al*.[7]

Mate pair sequencing

Illumina library preparation. Illumina libraries were made with the Illumina 2–5 kb mate pair protocol v2 with minor modifications (Supplementary Methods).

Samples were pooled into groups of 4. Each pool was sequenced (2×50 bp) on a single lane of the HiSeq2000 (Illumina, San Diego, CA, USA).

SOLiD library preparation. SOLiD mate pair libraries were generated according to the SOLiDv4 and SOLiD5500 long mate pair library preparation manuals (Life Technologies, Carlsbad, CA, USA). Libraries were sequenced on one quadrant of a SOLiDv4 sequencer or one lane of a SOLiD5500xl sequencer.

Analysis of mate pair data. Mapping of the data was done using Stampy[26] for the Illumina data and BWA[27] for the SOLiD data. Only uniquely mapped reads were further analyzed and all duplicate reads were removed (Supplementary Table S1).

Cluster analysis. Cluster analysis of discordant mate pairs was performed using an in-house developed script (Vivar, Sante *et al* in preparation).

DOC analysis. Coverage analysis was performed using CNV-Seq.[28] As a reference pool, experiments were grouped according to GC-bias for normalization.

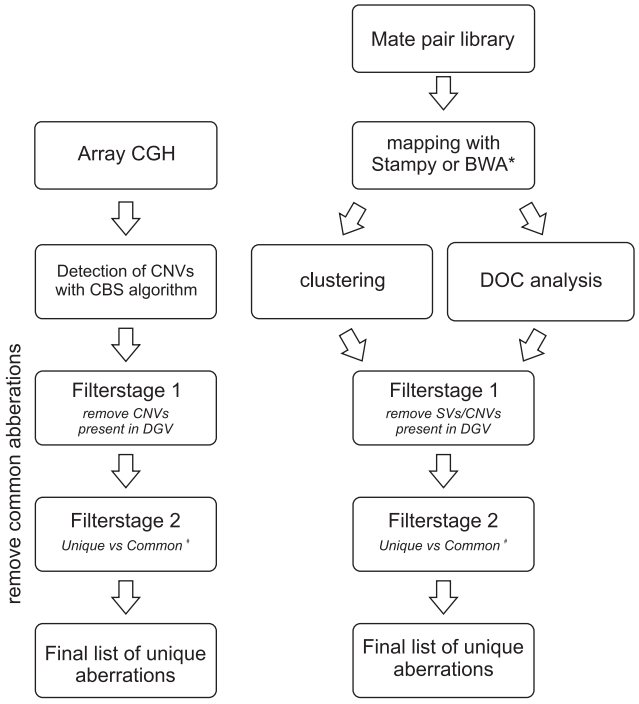

Figure 1 Flowchart of the analysis of the array CGH, cluster and coverage analysis data. Two main filtering steps were used: comparison with DGV followed by a comparison of the remaining aberrations to the aberrations detected in the patient pool (#). Stampy and BWA were, respectively, used for Illumina and SOLiD data (*).

Filtering strategies. To differentiate between possible pathogenic aberrations and benign SVs, several filtering steps were introduced. An overview of the different filtering steps used in the cluster, coverage and array CGH analysis is given in Figure 1.

A more detailed description of the cluster analysis, depth of coverage (DOC) analysis and filtering strategies is given in the Supplementary Methods. Sequencing files were deposited to the European Nucleotides Archive (http://www.ebi.ac.uk/ena) under accession number PRJEB4453. Sequencing data from patients 10, 11 and 13 were already deposited under accession number ERP001438.

Sanger sequencing of the breakpoints and qPCR

For the remaining aberrations after cluster analysis and filter steps 1 and 2 in the patients with a normal array profile, validation was performed by PCR amplification and capillary sequencing. Other selected breakpoints were also confirmed in this way (Supplementary Table S3). In the patients with a normal array profile, the remaining aberrations after DOC analysis and filter steps 1 and 2 affecting coding regions were validated by quantitative PCR (qPCR). More information can be found in the Supplementary Methods.

RESULTS

A systematic comparison was made between mate pair sequencing *versus* array CGH and karyotyping in a cohort of 50 ID/MCA patients who were referred to our diagnostic departments. This cohort was selected to represent the variety of SVs found in ID/MCA patients and contains 21 patients with deletions or duplications (recurrent or nonrecurrent), 6 patients carrying an apparently balanced aberration, 3 patients with a complex chromosomal rearrangement and 20 patients without a causal aberration (Tables 1 and 2, and Supplementary Table S2). For five patients, the parents were included

Table 1 Overview of the detected aberrations in 30 patients with ID/MCA

		Genomic position (GRCh37)			Detected by							
					Inheritance	Array		Coverage		Trio		Genes in break
Patients	*Aberration*	*Karyotyping*	*Array*	*Mate pair*	*pattern*	*CGH*	*Clustering*	*analysis*	*LCRs*	*analysis*	*Genes in region*	*point region*
Nonrecurrent aberrations												
Patient 1	del	46,XY	chr16:21806318–22448172	chr16:21795323–22563293	Unknown	+		+	+		11 genes	*LOC653786*
Patient 2	Tandem dup	46,XX	chr15:57656004–57763906	chr15:57644360–57778338	Maternal	+	+	+	–		*CGNL1*	*CGNL1*
	Tandem dup		chr18:77326777–77529494	chr18:77308690–77531555	Paternal	+	+	+	–		*CTDP1*	–
Patient 3	Tandem dup	46,XY	chr12:24771140–25398348	chr12:24769808–25401031	Maternal	+	+	+	–		5 genes	*KRAS*
Patient 4	del	46,XY	chr12:73241458–73357773	chr12:73227890–73369478	Unknown	+	+	+	–		-	–
	del		chr12:121297228–121618126	chr12:121289330–121625939	Maternal	+	+	+	–		4 genes	*SPPL3, P2RX7*
Patient 5	dup	46,XY	chrX:130608951–130950183	chrX:130580626–130958687	Maternal	+		+	+		*OR13H1*	*IGSF1, LOC286467*
Patient 6	del	46,XX	chr7:110520319–110722966	chr7:110511762–110707768	Paternal	+	+	+	–		*IMMP2L*	*IMMP2L*
Patient 7	Tandem dup	46,XX	chr11:107779704–108157951	chr11:107777875–108164503	*De novo*	+	+	+	–		4 genes	*ATM*
Patient 8	del	46,XX	chr8:141294618–141403896	chr8:141293153–141399016	Unknown	+		+	–		*TRAPPC9*	*TRAPPC9*
Patient 9	Interspersed dup	46,XX	chr4:172013456–172341313	chr4:172013259–172366770	*De novo*	+	+	+	–		–	–
			chr4:174790209–175298661	chr4:174806588–175303104					–		*FBXO8, CEP44*	–
Patient 10	Tandem dup	46,XY	chr9:131395721–132073928	chr9:131395733–132070993	*De novo*	+	+	+	–	+	20 genes	*SPTAN1*
Patient 11	Tandem dup	46,XY	chr19:54187022–54602946	chr19:54184569–54603393	*De novo*	+	+	+	–	+	C19MC cluster, 9 coding genes	*NDUFA3*
Patient 12	del	46,XY	chr1:92834493–93617701	chr1:92826568–93627209	*De novo*	+	+	+	–	+	7 genes	*RPAP2, TMED5*
	Tandem dup		chr1:93794057–93857706	chr1:93789807–93856418	*De novo*	+	+	+	–		*DR1*	*LOC100131564*
Patient 13	Tandem dup	46,XY	chr4:103533732–10400665	chr4:103522991–104009288	*De novo*	+	+	+	–	+	5 genes	*NFKB1, BDH2*
	Tandem dup		chr8:53505877–55140480	chr8:53503328–55161235	*De novo*	+	+	+	–		8 genes	–
	Tandem dup		chr8:80536950–81934189	chr8:80538184–81944843	*De novo*	+	+	+	–		6 genes	*STMN2, PAG1*
	dup		chr14:71023626–72238032	chr14:71027050–72258758	*De novo*	+		+	–		7 genes	–
Microdeletion/duplication syndromes												
Patient 14	del	46,XX	chr7:72685734–74172854	chr7:72714024–74155970	*De novo*	+		+	+		WBS region	
Patient 15	dup	47,XY, + del(15)(q13.1)	chr15:20190547–28305525	chr15:20186026–28290176	Unknown	+		+	+		PWS/AS region	
Patient 16	del	46,XY	chr17:16663504–18853898	chr17:16695511–19139094	*De novo*	+		+	+		SMS region	
Patient 17	del	46,XY	chr17:34823124–36248858	chr17:34723963–36357595	*De novo*	+		+	+		region associated with renal cysts and diabetes	
Patient 18	del	46,XX	chr17:44683–2444275	chr17:10802–2473314	*De novo*	+		+	–		Miller–Dieker region	
Patient 19	del	46,XY	chr22:18894834–21505357	chr22:18918731–21469569	Maternal	+		+	+		VCFS region	
Patient 20	del	46,XY	chr22:18953012–20311704	chr22:18891194–23003094	*De novo*	+		+	+		VCFS region	
Patient 21	dup	46,XY	chr22:18706001–21505358	chr22:18918992–21448778	*De novo*	+		+	+		VCFS region	
Apparently balanced translocations/inversions												
Patient 22	Translocation	46,X,t(X;1)(q26;q25)	–	chr1:175823563–175825129	*De novo*		+		–			–
				chrX:130560260–130562008								*IGSF1*
Patient 23	Translocation	46,XX,t(3;5)(p24;q14)	–	chr3:29299498–29300608	*De novo*		+		–			–
				chr5:88597544–88598474								*MEF2C*
Patient 24	Translocation	46,Y,t(X;3)(p11.2;q13.1)	–	chrX: 47104033–47104034	*De novo*		+		–			*USP11*
				chr3:107477586–107477586								*BBX*
Patient 25	Translocation	46,X,t(X;7)(q12;q22)	–	-	*De novo*				–			–
Patient 26	Inversion	46,XY,inv(4)(q21;q31.3)	–	chr4:82872172–155552374	*De novo*		+		–			*DQ266889*
Patient 27	Complex translocation	46,XY,t(2;14)(p22;q24.3)	–	chr2:38344860–38345669	*De novo*		+		–			
				chr14:79133247–79136029	*De novo*							*NRXN3*
				chr14:71207318–82091783	*De novo*							*MAP3K9*
Complex rearrangements												
Patient 28	dup	46,XX	chrX:129569453–130234890	chrX:129553399–130251499	*De novo*	+	+	+	–		*BC043223, FAM45A, ENOX2, ARHGAP36*	–
	dup		chrX:153297312–153676716	chrX:153298107–153679735		+		+	–		13 genes	*MECP2, FAM50A*
	dup		chrX:153828822–154351545	chrX:153875203–154368527		+		+	–		10 genes	-
Patient 29	Chromothripsis/ complex trisomy 21	46,XX,der(21)t(21;21)	See Figure 4	See Figure 4 + Table 2	*De novo*	+	+	+	–			
Patient 30	Chromothripsis/ complex trisomy 18	46,XX,der(18)	See Figure 4	See Figure 4 + Table 2	*De novo*	+	+	+	–			

For each patient, the inheritance pattern, the karyotyping result and the genomic position obtained with array CGH, cluster analysis and/or coverage analysis of the mate pair data are given. Columns 7–9 display whether an aberration was detected by array CGH, cluster analysis or coverage analysis. When low copy repeats (LCRs) are present at the breakpoint regions, this is shown in column 10. For five patients, trio analysis was performed, which is highlighted in the last column.

Table 2 Genomic coordinates (in GRCh37) of identified clusters on chromosome 21 (patient 29) and chromosome 18 (patient 30)

	Aberrant cluster positions (bp)					
Patient	*chr*	*Startmin*	*Startmax*	*chr*	*Stopmin*	*Stopmax*
Patient 29	21	23582821	23585530	21	39078296	39079749
	21	26809065	26811519	21	43192025	43195016
	21	28322124	28325564	21	40843095	40846302
	21	14606928	14609297	21	40846863	40848238
	21	20625794	20626993	21	38512674	38514710
Patient 30	18	36825392	36829386	18	49026176	49028702
	18	12178182	12179663	18	46679884	46682308
	18	37221170	37223903	18	62969186	62971647
	18	47226591	47229543	18	62101244	62103679
	18	48703124	48706218	18	62928684	62931584

in the analysis allowing the immediate identification of inherited and *de novo* aberrations.

Mate pair sequencing detects aberrations previously detected by array CGH and karyotyping

Our first aim was to investigate whether mate pair sequencing was able to detect the aberrations previously detected by array CGH and karyotyping. The mate pair data of the 50 ID/MCA patients and 10 parents were analyzed using DOC analysis and by clustering of the discordant mate pairs (see Materials and Methods and Supplementary Materials and Methods).

In a first analysis, we evaluated the ability of mate pair sequencing to detect copy-number changes based on DOC measurements. To assess the resolution at which copy-number changes can be detected using our mate pair data, we calculated the distribution of window sizes across the genome (in nucleotides) using 250 reads per window (Supplementary Figure S2A), yielding an average window size between 5.7 and 20.5 kb, depending on the amount of sequencing reads generated per sample (Supplementary Figure S2B). On the basis of these estimates, the resolution of DOC analysis is comparable with or higher than the 180 K CGH arrays, with an average probe spacing of $\sim$13 kb, used in this study.

Our patient cohort included 31 possibly relevant DNA copy-number changes (12 deletions and 19 duplications ranging in size from 66 kb to 8.1 Mb) divided over 22 patients (patients 1–21, patient 28). Using DOC analysis of the mate pair data of these 22 patients, we identified between 9 and 217 copy-number changes per sample (Supplementary Table S2). Next, we applied two filtering strategies to identify private possible pathogenic rearrangements among the predicted SVs in each of the patient genomes based on the DOC analysis. In the first step, we removed all variants that were overlapping with variants in the Database of Genomic Variants (DGV; Figure 1). This resulted in a reduction of the possible pathogenic aberrations by $\sim$75%. The second cross-sample filtering step involved comparison with other patients in our cohort (as explained in the Supplementary Materials and Methods), which further reduced the data sets to an average of 15 'private' variants per patient (Supplementary Table S2). Based on this analysis scheme, we could readily identify all 31 copy-number changes in the 22 patients but no additional causal events were detected.

We tested the use of healthy parents as a control in family-trio-based detection of copy-number changes for patients 10 and 13, which carry *de novo* duplications. For each of the patients, we determined the overlapping copy-number changes that were found with respect to both parents. We found that only the expected duplications could be detected, showing that parents are optimal controls for the detection of *de novo* copy-number changes (Supplementary Figure S3). Overall, we conclude that the DOC signature of mate pair data is a robust alternative for array CGH for the clinical detection of copy-number changes.

As a second approach, we used clustering of anomalous read pairs to identify the known aberrations in our patients. We performed the same filtering procedures as described above and detected an average of four private variants per patient. In 18 patients, we were able detect the previously identified aberrations (both balanced as well as unbalanced) using cluster analysis (patients 2–4, 6–7, 9–13, 22–24 and 26–30). In 12 patients, cluster analysis failed to detect the aberration (ie, 8 recurrent and 4 nonrecurrent aberrations; Table 1) because of flanking segmental duplications (patients 1, 5 and 14–21) or other repetitive regions such as centromeric satellite repeats near the breakpoints (patient 25). In these cases, the flanking repeats impede the unambiguous mapping of short sequencing reads and hence the proper localization of the exact breakpoints.

Altogether, all the SVs that were previously detected by either karyotyping or array CGH in the patients from our cohort were readily detected in the mate pair data as based on clustering of discordant pairs and/or DOC analysis (Table 1) with the exception of a balanced whole-arm translocation (patient 25).

Mate pair sequencing defines the precise structure of SVs previously found by array CGH or karyotyping and reveals underlying gene defects

Although array CGH can efficiently detect regions with DNA copy-number changes, it does not provide the precise breakpoint junctions that underlie such changes. This information could be important in determining whether an SV is pathogenic or not. For all rearrangements that were identified based on clustering of discordant mate pairs, we delineated the molecular structure of the rearrangements. This included the genomic architecture of 15 duplications in 9 patients (patients 2–3, 7, 9–13 and 28). In 10 cases, the duplicated segment resulted from non-inverted tandem duplications. This was confirmed in two patients by capillary sequencing (Supplementary Table S3). In patient 9, two seemingly independent duplications on chromosome band 4q34.1 were detected by array CGH. Cluster analysis revealed the complex nature of this rearrangement, pinpointing one single event (Figure 2a). In patient 28, array CGH revealed the presence of three duplications on the long arm of the X chromosome. Mate pair sequencing allowed the direct reconstruction of the genomic architecture of this complex chromosomal rearrangement (Figure 2b), which was confirmed by FISH analysis (Supplementary Figure S4).

Mate pair sequencing further enabled the refinement of five out of six apparently balanced aberrations (patients 22–24 and 26–27). None of these aberrations showed evidence for gains or losses near the breakpoints after array CGH and DOC analysis. Breakpoints were sequenced by capillary sequencing, revealing microhomology (2–3 bp) and insertions (2–13 bp) at the breakpoints suggestive for nonhomologous end repair as the primary mechanism (Supplementary Table S3). Mate pair sequencing revealed a complex chromosomal rearrangement in patient 27 instead of an apparently balanced translocation, as was apparent based on conventional chromosome analysis. This complex aberration consists of four breakpoints, involving an inversion on chromosome 14 and a translocation between chromosome 2 and 14 (Figure 3).

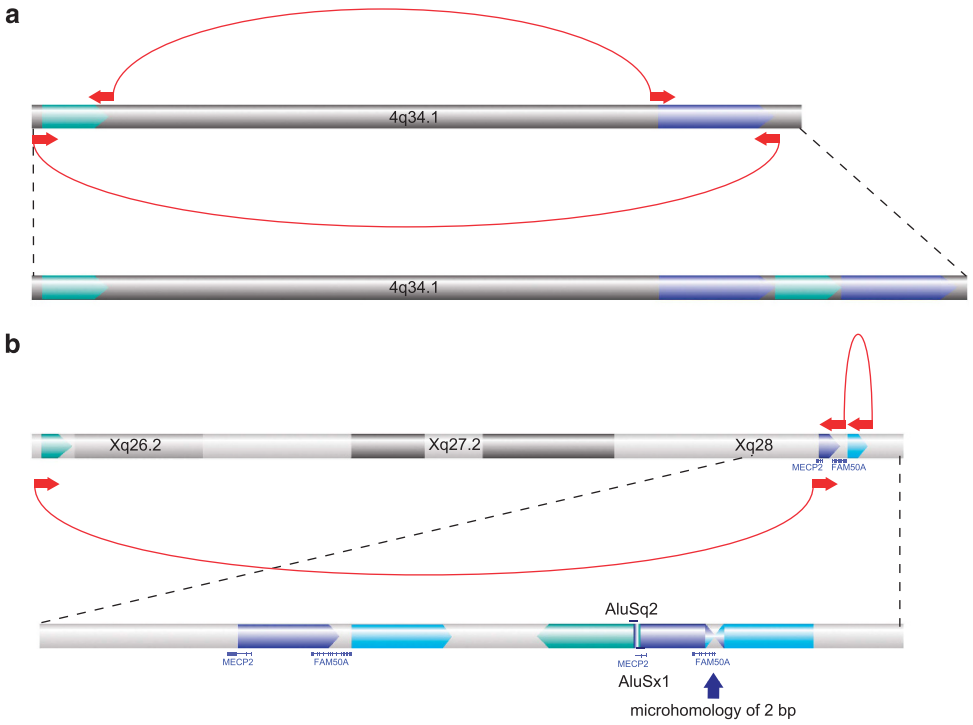

Figure 2 Genomic nature of the duplications observed in patients 9 and 28. (**a**) Cluster analysis revealed the position of the duplications on chromosome band 4q34.1 in patient 9. (**b**) The duplications on chromosome X in patient 28 lead to a complex rearrangement, which was resolved by cluster analysis. A microhomology of 2 bp was observed at the first boundary and at the second Alu elements with a similarity of 93% were noted. Clusters of aberrant mates are indicated at the bottom as blue arrows linked together by the dashed lines. The direction of the arrows indicates the orientation of the reads (both experiments were done on the HiSeq).

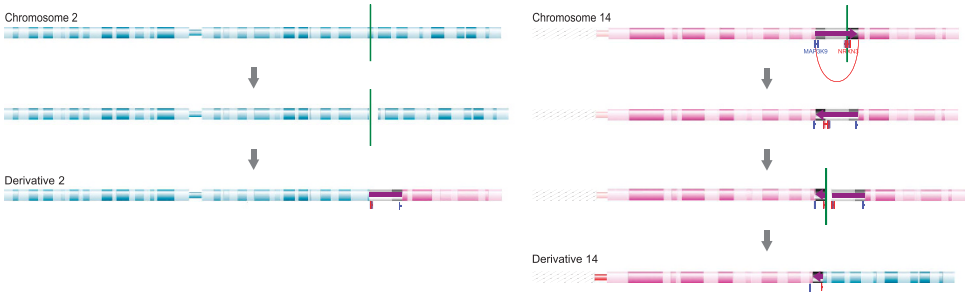

Figure 3 Complex rearrangement in patient 27. Cluster analysis reveals a complex chromosomal rearrangement disrupting the *NRXN3* and *MAPK9* gene.

In patient 29 and 30, a chromothripsis-like event was observed with array CGH involving, respectively, chromosome 21 and chromosome 18.[29] Figure 4 gives an overview of the results for all the detection techniques (ie, conventional karyotyping (panels a and b), array CGH (panels c and d) and mate pair sequencing (panels e and f)) nicely showing that cluster analysis revealed many breakpoint junctions between remote genomic regions on chromosome 21 (patient 29) and chromosome 18 (patient 30; Table 2).

In 14 patients, cluster analysis enabled the identification of a gene that was disrupted by a balanced rearrangement or a duplication event (Table 1).

No additional pathogenic SVs detected by mate pair sequencing in the ID/MCA patients in our cohort

The third aim of this study was to evaluate the possibility to identify clinically relevant genomic rearrangements beyond the resolution of routine array CGH analysis. To this purpose, we included 20 patients with a 'chromosomal phenotype' (indicating abnormalities in multiple organs and severe mental retardation) without causal aberration and a normal karyotype. Using a combination of cluster analysis and DOC measurements SVs could be detected. In this context, it is important to define the boundaries (resolution) with which SVs can be detected using the mate pair technology that we use here.

The detection limits of mate pair cluster and DOC analysis are mainly influenced by two factors. First, the mate pair library insert size determines the resolution for detection of SVs based on cluster analysis. In our experiments, we used an insert size of 3 kb, which provides a good balance between resolution and the ability to detect breakpoints across repetitive regions.[14,30] We estimated the resolution that can be reached based on clustering analysis by calculating the distribution of the sizes of deletions predicted by our analysis (Supplementary Figure S5). This shows that the majority of deletions is smaller than 10 kb and the lower detection limit is ~1 kb.

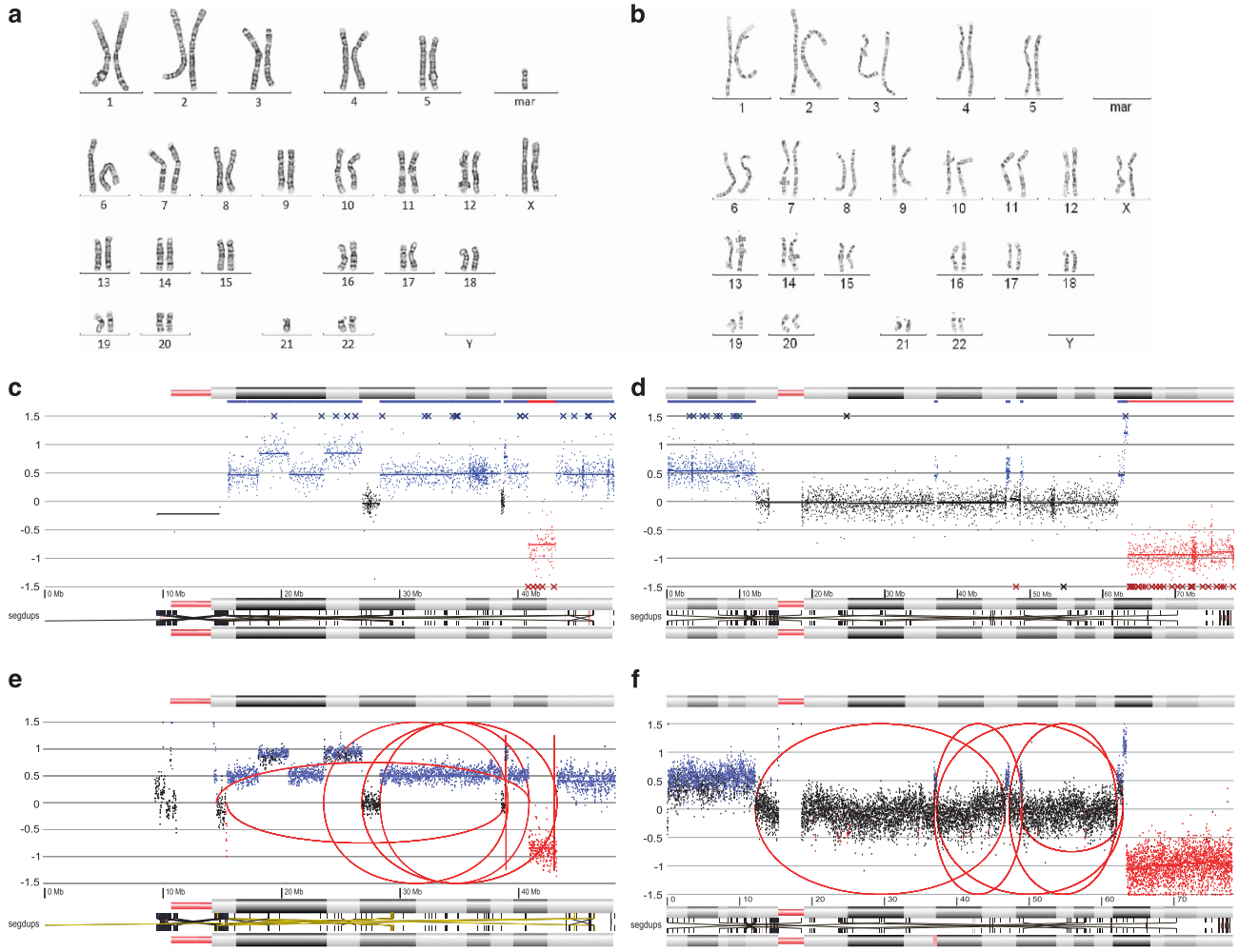

Figure 4 Chromothripsis-like events in patients 29 and 30. (**a**, **b**) Karyogram of patient 29 (**a**) and patient 30 (**b**). (**c**, **d**) array CGH profile of chromosome 21 (patient 29, **c**) and chromosome 18 (patient 30, **d**). (**e**, **f**) Cluster and coverage profiles of the rearranged chromosomes. Aberrant clusters are depicted as red arches. Segmental duplications on these chromosomes are shown at the bottom of each profile (ie, segdups).

The human genome contains many more small deletions (<1 kb) than large deletions.[31] Thus, a major fraction of small deletions remains undetected using read-pair analysis of libraries with 3 kb insert size as used here.

A second factor that primarily determines the resolution of detection for DOC analysis is the overall amount of sequencing reads generated for a sample.[31] As described above, we estimate that the resolution for detection of copy-number changes in our data set varies between 5.7 and 20.5 kb depending on overall amount of sequencing reads generated for a sample.

Finally, we should note that a substantial fraction of the genome is refractory to sequencing using short reads, such as (sub)telomeric and centromeric regions. We calculated the distribution of physical genomic coverage across 7.9 Mb of non-repeat masked genomic sequences and we found that >95–100% of the sequences are covered at more than 1x and 80–100% are covered at more than 20x for each of the samples examined (Supplementary Figure S6).

With these resolution and coverage parameters for mate pair sequencing in mind, we examined the SV calls in the 20 patients without array CGH diagnosis. We found that >75% of these aberrations were also present in the DGV or in at least four other patients from our cohort. For the remainder of the rearrangements detected with cluster or DOC analysis, respectively, PCR across the breakpoint junctions in the patients and their respective parents or qPCR was performed to find out whether any of these rearrangements had occurred *de novo*. Although some of the rearrangements could not be confirmed by conventional PCR or qPCR (possibly false positive), the other rearrangements were inherited from one of the parents. We therefore conclude that mate pair sequencing did not reveal any *de novo* SVs in the samples beyond the resolution of array CGH and thus led to the same diagnostic results as the array CGH profiling.

DISCUSSION

SVs are an important cause of ID and congenital malformations. However, the current diagnostic tools to detect SVs have limited resolution (karyotyping) or cannot detect copy neutral rearrangements (array CGH) implying that improved technologies may have benefits. Previous work has shown that paired-end mapping or mate pair sequencing has unprecedented resolution for the detection of SV breakpoints of balanced genomic rearrangements in patients with ID/MCA and may thus be a valuable tool for diagnostic implementation.[16–23] Here we made a systematic comparison between mate pair sequencing *versus* array CGH and karyotyping.

As a first important conclusion from this work, we demonstrate that all types of pathogenic SVs previously found with array CGH or karyotyping could also be found using a combined strength of the cluster and DOC signatures of mate pair sequencing. Even when breakpoints reside in repetitive regions (centromere or segmental duplications), DOC analysis could still reveal the breakpoint.

A major strength of mate pair sequencing is in the elucidation of the precise genomic architecture of the SVs (Table 1). For example, cluster analysis showed that most duplications are in tandem with a head-to-tail orientation. In the two patients where the duplications were not tandem, the precise genomic locations of the duplicated segments could be defined (Figure 2). In 14 patients, cluster analysis enabled the identification of a gene that was disrupted by a balanced rearrangement or a duplication event. In five of these cases, the disrupted gene could explain the phenotype of the patient (Table 1). Especially in patient 27, mate pair sequencing was of intrinsic value. The t(2;14) translocation in this patient disrupts the *NRXN3* gene, but this does not explain the cardiac defects in patient 27. The second aberrant cluster on chromosome 14 due to an inversion disrupts the *MAP3K9* gene, which is an interesting candidate gene for the cardiac phenotype observed in the patient.[32] Without mate pair sequencing, this crucial information would not have been revealed. The enhancement in resolution of mate pair sequencing and hence the ability of directly pinpointing disease genes has great diagnostic value.

Both array CGH analysis and mate pair sequencing may be hampered in accurately assessing genomic regions with highly repetitive regions such as centromeric and telomeric regions. For example, recurrent aberrations (surrounded by low copy repeats) were only detected by coverage analysis and not by cluster analysis. Balanced whole chromosome arm translocations cannot be detected because of the repetitive nature of the centromeric breakpoints; these aberrations, however, have no immediate causal relationship to the phenotype of the patient.

The resolution of mate pair sequencing is merely dependent on the insert size and the amount of sequence reads. In our analysis, we noted big differences between different samples and between different runs, resulting in practical resolutions of 20.5 kb down to 5.7 kb for DOC analysis and a lower detection limit of $\sim$1 kb for cluster analysis. In 2011, Cooper *et al*[33] compared CNVs in 15 767 children with ID and various congenital defects to 8329 adult controls, and concluded that in 14.2% of these children the disorder is caused by CNVs $>$ 400 kb. On the basis of this, Vermeesch *et al*[34] suggested that arrays should aim to detect at least any imbalance larger than 500 kb, which is definitely the case for the arrays used in this study ($\sim$13 kb). Still a large fraction of ID/MCA cases remains without a conclusive genetic diagnosis. We show that mate pair sequencing at the resolution used here did not enhance the diagnostic yield for these undiagnosed patients. This finding should fuel further efforts in order to search for smaller *de novo* coding and noncoding mutations as the underlying cause of the disease in a subset of these patients by means of exome or whole-genome sequencing.[11,35,36] Other approaches, such as improvements of mate pair protocols, higher sequence coverage, combinations of different library insert sizes or screening of larger patient cohorts could possibly also lead to higher detection rates.

Altogether, we made a systematic comparison between mate pair sequencing and array CGH/karyotyping for the genetic diagnosis of patients with ID/MCA. We demonstrate that mate pair sequencing enables the rapid identification and delineation of structural variants and has added value for the identification of disease genes in these patients. Further improvements in sequencing throughput will allow the identification of the whole spectrum of genomic mutations from single nucleotide changes up to large SVs by means of (paired-end) whole-genome sequencing, making sequencing a holistic detection platform.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We are indebted to all patients, their families and the clinicians involved for their cooperation. We thank Lies Vantomme and Shalina Baute for expert technical assistance. Sarah Vergult was supported by a PhD fellowship from the Research Fund Flanders (FWO) and is now supported by a postdoctoral grant from the Special Research Fund (BOF) from Ghent University. This work was supported by Grant SBO60848 from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) and a Methusalem grant of the Flemish Government. Bruce Poppe is Senior Clinical Investigator at the Research Foundation – Flanders (FWO) and Geert Mortier was Senior Clinical Investigator at the Research Foundation – Flanders (FWO) until 2010. This article presents research results of the Belgian program of Interuniversity Poles of attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming (IUAP).

1 Girirajan S, Campbell CD, Eichler EE: Human copy number variation and complex genetic disease. *Annu Rev Genet* 2011; **45**: 203–226.
2 Stankiewicz P, Lupski JR: Structural variation in the human genome and its role in disease. *Annu Rev Med* 2010; **61**: 437–455.
3 Feuk L, Carson AR, Scherer SW: Structural variation in the human genome. *Nat Rev Genet* 2006; **7**: 85–97.
4 Conrad DF, Pinto D, Redon R *et al*: Origins and functional impact of copy number variation in the human genome. *Nature* 2010; **464**: 704–712.
5 Vissers LE, de Vries BB, Veltman JA: Genomic microarrays in mental retardation: from copy number variation to gene, from research to diagnosis. *J Med Genet* 2010; **47**: 289–297.
6 Miller DT, Adam MP, Aradhya S *et al*: Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* 2010; **86**: 749–764.
7 Buysse K, Delle Chiaie B, Van Coster R *et al*: Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *Eur J Med Genet* 2009; **52**: 398–403.
8 Koolen DA, Pfundt R, de Leeuw N *et al*: Genomic microarrays in mental retardation: a practical workflow for diagnostic applications. *Hum Mutat* 2009; **30**: 283–292.
9 de Leeuw N, Hehir-Kwa JY, Simons A *et al*: SNP array analysis in constitutional and cancer genome diagnostics–copy number variants, genotyping and quality control. *Cytogenet Genome Res* 2011; **135**: 212–221.
10 Hochstenbach R, Buizer-Voskamp JE, Vorstman JA, Ophoff RA: Genome arrays for the detection of copy number variations in idiopathic mental retardation, idiopathic generalized epilepsy and neuropsychiatric disorders: lessons for diagnostic workflow and research. *Cytogenet Genome Res* 2011; **135**: 174–202.
11 Vissers LE, de Ligt J, Gilissen C *et al*: A de novo paradigm for mental retardation. *Nat Genet* 2010; **42**: 1109–1112.
12 O'Roak BJ, Deriziotis P, Lee C *et al*: Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* 2011; **43**: 585–589.
13 Veltman JA, Brunner HG: De novo mutations in human genetic disease. *Nat Rev Genet* 2012; **13**: 565–575.
14 Medvedev P, Stanciu M, Brudno M: Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 2009; **6**: S13–S20.
15 Korbel JO, Urban AE, Affourtit JP *et al*: Paired-end mapping reveals extensive structural variation in the human genome. *Science* 2007; **318**: 420–426.
16 Kloosterman WP, Tavakoli-Yaraki M, van Roosmalen MJ *et al*: Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep* 2012; **1**: 648–655.
17 Kloosterman WP, Guryev V, van Roosmalen M *et al*: Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum Mol Genet* 2011; **20**: 1916–1924.
18 Chen W, Ullmann R, Langnick C *et al*: Breakpoint analysis of balanced chromosome rearrangements by next-generation paired-end sequencing. *Eur J Hum Genet* 2010; **18**: 539–543.
19 Chen W, Kalscheuer V, Tzschach A *et al*: Mapping translocation breakpoints by next-generation sequencing. *Genome Res* 2008; **18**: 1143–1149.
20 Talkowski ME, Rosenfeld JA, Blumenthal I *et al*: Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* 2012; **149**: 525–537.

21 Talkowski ME, Ernst C, Heilbut A *et al*: Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am J Hum Genet* 2011; **88**: 469–481.
22 Chiang C, Jacobsen JC, Ernst C *et al*: Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat Genet* 2012; **44**: S391.
23 Schluth-Bolard C, Labalme A, Cordier MP *et al*: Breakpoint mapping by next generation sequencing reveals causative gene disruption in patients carrying apparently balanced chromosome rearrangements with intellectual deficiency and/or congenital malformations. *J Med Genet* 2013; **50**: 144–150.
24 De Weer A, Poppe B, Vergult S *et al*: Identification of two critically deleted regions within chromosome segment 7q35-q36 in EVI1 deregulated myeloid leukemia cell lines. *PLoS One* 2010; **5**: e8676.
25 Olshen AB, Venkatraman ES, Lucito R, Wigler M: Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 2004; **5**: 557–572.
26 Lunter G, Goodson M: Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 2011; **21**: 936–939.
27 Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–1760.
28 Xie C, Tammi MT: CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 2009; **10**: 80.
29 Liu P, Erez A, Nagamani SC *et al*: Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* 2011; **146**: 889–903.
30 Hillmer AM, Yao F, Inaki K *et al*: Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. *Genome Res* 2011; **21**: 665–675.
31 Mills RE, Walter K, Stewart C *et al*: Mapping copy number variation by population-scale genome sequencing. *Nature* 2011; **470**: 59–65.
32 Wright EM, Kerr B: RAS-MAPK pathway disorders: important causes of congenital heart disease, feeding difficulties, developmental delay and short stature. *Arch Dis Child* 2010; **95**: 724–730.
33 Cooper GM, Coe BP, Girirajan S *et al*: A copy number variation morbidity map of developmental delay. *Nat Genet* 2011; **43**: 838–846.
34 Vermeesch JR, Brady PD, Sanlaville D, Kok K, Hastings RJ: Genome-wide arrays: quality criteria and platforms to be used in routine diagnostics. *Hum Mutat* 2012; **33**: 906–915.
35 Rauch A, Wieczorek D, Graf E *et al*: Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 2012; **380**: 1674–1682.
36 de Ligt J, Willemsen MH, van Bon BW *et al*: Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 2012; **367**: 1921–1929.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)