Genomics Data 2 (2014) 144–146

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Platform comparison of detecting copy number variants with microarrays and whole-exome sequencing

Joep de Ligt [a,1], Philip M. Boone [b], Rolph Pfundt [a], Lisenka E.L.M. Vissers [a], Nicole de Leeuw [a], Christine Shaw [b,c], Han G. Brunner [a,f], James R. Lupski [b,d,e], Joris A. Veltman [a,f], Jayne Y. Hehir-Kwa [a,*]

[a] *Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences, Institute for Genetic and Metabolic Disease, Radboud University Medical Centre, Nijmegen 6500 HB, The Netherlands*
[b] *Department of Molecular and Human Genetics, Baylor College of Medicine, Houston,TX, United States*
[c] *Roche NimbleGen, Madison, WI, United States*
[d] *Department of Pediatrics, Baylor College of Medicine, Houston, TX, United States*
[e] *Texas Children'sHospital, Houston, TX, United States*
[f] *Department of Clinical Genetics, Maastricht University Medical Centre, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands*

ARTICLE INFO

Article history:
Received 20 May 2014
Received in revised form 6 June 2014
Accepted 7 June 2014
Available online 17 June 2014

Keywords:
Copy number variation
Microarray
Whole exome sequencing

ABSTRACT

Copy number variation (CNV) is a common source of genetic variation that has been implicated in many genomic disorders, Mendelian diseases, and common/complex traits. Genomic microarrays are often employed for CNV detection. More recently, whole-exome sequencing (WES) has enabled detection of clinically relevant point mutations and small insertion—deletion exome wide. We evaluated (de Ligt et al. 2013) [1] the utility of short-read WES (SOLiD 5500xl) to detect clinically relevant CNVs in DNA from 10 patients with intellectual disability and compared these results to data from three independent high-resolution microarray platforms. Calls made by the different platforms and detection software are available at dbVar under nstd84.

Specifications	
Organism/cell line/tissue	*Human blood*
Sex	–
Sequencer or array type	*Affymetrix 250 k, Affymetrix CytoScanHD, N imblegen Custom ExonArray, Solid 5500xl*
Data format	*Analyzed*
Experimental factors	*Normal*
Experimental features	*Positive samples with rare de-novo coding CNV*
Consent	All patients gave their written informed consent before study entry.
Sample source location	*NA*

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd84/ (CNV calls from all platforms/programs)

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46060 (raw 250 k data)

Experimental design, materials and methods

Sample selection

Ten samples were selected that had previously been diagnostically reported as containing at least one clinically relevant, rare de novo CNV associated with intellectual disability (ID), detected by routine microarray based screening within the Department of Human Genetics, Radboud University Medical Centre, Nijmegen [1]. These CNVs were chosen to represent a wide range of clinically relevant CNVs detected by microarray based analysis in our Genome Diagnostics division. The selected CNVs (1) contained at least one coding region, (2) were validated de novo using the same microarray platform on parental DNAs, (3) occurred across a variety of chromosomes, (4) ranged in copy number state from zero to three, and (5) ranged in genomic size from 15 kb to 24 Mb (Table 1).

Eleven of these de novo CNVs were detected using an Affymetrix 250K *NspI* (Affymetrix, Santa Clara, CA) microarray and one, in patient 1, with the Affymetrix 2.7M microarray platform (Table 1).

Whole exome sequencing

Genomic DNA from these 10 samples was isolated from blood using the QIAamp DNA Mini Kit (Qiagen, Venlo, The Netherlands). Whole exome sequencing (WES) was performed at the University Medical

[*] Corresponding author.
E-mail address: Jayne.Hehir-Kwa@radboudumc.nl (J.Y. Hehir-Kwa).
[1] Current position: Hubrecht Institute for Developmental Biology and Stem Cell Research, KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584CT Utrecht, The Netherlands.

http://dx.doi.org/10.1016/j.gdata.2014.06.009

Table 1
Overview of the detection of 12 clinically relevant de novo CNVs.

		Discovery microarray					WES read depth algorithms			
Patient	Chromosome	Estimated start position (kb)	Estimated end position (kb)	CNV size (kb)	Copy number state	Nr. Genes	CONTRA	cn.MOPS	ExonDepth	CoNIFER
1	chr10	89,642.6	89,657.5	14.9	1	1a	–	–	–	–
2	chr19	33,371.1	33,394.2	23.0	0	1	–	–	V	V
3	chr8	77,745.6	77,795.2	49.6	1	1	–	-	V	V
4	chr17	1,203.6	1,516.5	312.9	3	8	-	–	V	V
5	chr16	29,673.2	29,988.3	315.1	1	16	–	–	V	V
6	chr15	43,759.8	44,862.9	1103.2	1	24	–	–	–	V
7	chr2	233,166.3	233,886.7	720.5	3	16	–	–	V	V
8	chrX	6495.3	7951.7	1456.4	0	5	–	–	V	V
9	chr2	239,952.7	241,373.1	1420.5	3	14	–	–	V	V
	chr2	241,442.7	243,001.9	1559.2	1	31	–	–	V	V
	chr15	60,489.7	62,906.5	24,603.6	3	210	–	–	V	V
10	chr20	77,771.0	102,374.6	2416.8	3	91	–	V	V	V

CNVs as detected by the discovery microarray (hg19), genomic location, size, predicted copy number state and the number of genes in the region. a. A single exon deletion. Detection by the different WES approaches; –, CNV is not detected with a minimum overlap of 30%, and V, detected with a minimum overlap of 30%.

Centre Nijmegen (UMCN) according to the manufacturer's guidelines. The samples were enriched using the Agilent SureSelect V2 protocol, and sequencing was performed on a SOLiD 5500xl system (Life Technologies) to a median read depth of 67 across targeted regions. Read correction and mapping were performed with Lifescope v1.3 (Life Technologies), using default settings. After mapping reads with a mapping quality (MAPQ) value below 20 were discarded to select reliably mapped reads. The value of 20 was based on read depth ratio of the X chromosome between female and male samples. This was the lowest quality value that resulted in a 0.5 ratio.

The WES data were analyzed with four different published CNV detection programs; (1) cn.MOPS v1.6.4 [2], (2) CONTRA v2.0.3 [4], (3) CoNIFER v0.2.0 [3], and (4) ExomeDepth v0.8.4 [6], with hg19-based RefSeq gene exon definitions as target regions in the analysis. Overlapping exonic regions were merged resulting in a list with unique genomic regions for further analysis. The tools were selected based on their availability and ability to perform rare CNV detection on .bam files.

CNV segments identified by WES underwent additional merging. CNV calls of the same copy number were merged if they were within 5 Mb distance and fewer than 30 informative data points were between the calls. These values protect against overcalling while allowing for gene deserts, resulting in a more robust and uniform call-set.

Affymetrix 250K NspI & Affymetrix 2.7M microarray

Samples were processed in accordance with the manufacturer's instructions. Hybridization, washing and scanning were performed with appropriate Affymetrix GeneChip products. The 250K microarrays image processing was performed with Affymetrix GeneChip Command Console software. Genotypes were called with Affymetrix Genotyping Console Software v2.1 using the BRLMM algorithm with default-calling threshold of 0.5 and a prior size of 10,000 bases. Samples were required to have a minimum Quality Control SNP call rate of 90%. CNV identification was performed using CNAG v2.0 with default HMM settings [5]. Image processing, CNV calling and merging for the 2.7M microarray was performed via Affymetrix Power Tools v1.14.3 with default settings and calling thresholds. Data viewing and analysis was performed with the Affymetrix Chromosome Analysis Suite (ChAS) software v2.0 and the UCSC genome browser (UCSC Genome Browser on Human Feb. 2009, GRCh37/hg19, NCBI Build 37.3). Samples were required to have minimum quality thresholds of; MAPD ≤ 0.2049, SNP-QC > 1,1 and WavinessSegCount ≤ 10.

Affymetrix CytoScanHD (2.6M) microarray

The Affymetrix CytoScanHD platform with 2.6 million probes was used to serve as a benchmark of the high resolution microarrays currently used in a diagnostic setting at the UMCN genetics department. Experiments were performed in the UMCN according to the manufacturer's specifications. CNVs were called with Affymetrix Power Tools v1.14.3 using default settings and calling thresholds. Data viewing and analysis was performed with ChAS v2.0 and the UCSC genome browser (UCSC Genome Browser on Human Feb. 2009, GRCh37/hg19, NCBI Build 37.3). Samples were required to have minimum quality thresholds of; MAPD < 0.25, SNPQC > 15 and Waviness-SD < 0.12.

NimbleGen custom design ExonArray (4.2 M) microarray

A custom, comparative genomic hybridization (CGH) array with approximately 4.2 million oligonucleotide probes was manufactured for BCM by Roche NimbleGen (RNG); this array, referred to as the “ExonArray”, served as an orthologonal independent experimental approach and a high resolution benchmark for CNV detection. The custom array design included 2.15 million backbone probes and 1.85 million supplemental exonic probes targeted to the RNG “Big Exome” (exome definition includes all exons from RNG Exome v2.0, Agilent SureSelect 50Mb, RefSeq, CCDS, and the BCM Human Genome Sequencing Center HGSC content (both VCRome and HGSCv1 designs)). The aim of the design was to cover each exon (and flanking sequence, if necessary) with at least 8 probes. A series of test arrays was manufactured with 10X oversampling of exon-targeted probes (i.e. 80 per exon) and runs with control DNA to empirically identify the 8 probes per exon with best linear signal response to a range of DNA concentrations, which were included in the final ExonArray design. In the ExonArray, the ideal coverage of 8 or more probes was achieved for >135,000 (~86%) of the targeted exons; 249 (0.16%) of the exons could not be targeted at all.

The 10 patient samples were analyzed with the ExonArray at BCM according to the manufacturer's specifications and using gender matched control DNA (HapMap individuals NA10851 and NA15510). Segmentation was performed with RNG DEVA software, using default settings (requiring a minimum of five probes per segment), to account for the high resolution of the platform the maximum number of segments allowed per chromosome was increased to 500. CNVs were derived from the segments using a $\log_2$ deviation value of ≤ -0.415 (the theoretical $\log_2$ of a 50% mosaic heterozygous loss) for deletions or ≥ 0.322 (the theoretical $\log_2$ ratio of a 50% mosaic heterozygous gain) for duplications and higher order gains.

Segments were merged when they were; 1) within 1 Mb of each other, 2) fewer than 100 probes were located between the events and 3) the average $\log_2$ values had a maximum difference of 0.5.

Evaluating the CNV detection power of WES

The false negative (FN) detection rate of WES was calculated by measuring the number of CNV events detected using the high-resolution microarray platforms that were missed by WES. To prevent overestimation due to platform design (exon targeted vs. whole genome), we accounted for both the exome enrichment targets and the detection power of WES. We selected CNVs that were identified by at least two independent microarray platforms (minimum overlap of 30% of the CNV region, to allow for breakpoint inaccuracies due to the large differences in probe densities) and the CNV had to encompass at least three exons.

For each CNV, the largest region, detected by the CytoScanHD or the ExonArray, was used for further analysis. After applying these selection criteria to the total set of 6074 CNVs identified by the different microarray experiments, the resulting consensus dataset contained 97 CNVs. Of these 97 consensus CNVs, 25 did not occur in the common CNV dataset and were considered rare CNVs. Consensus CNVs were only considered as positively detected by WES if a CNV was called in the same region and overlapped the consensus CNV region for at least 30%.

Discussion

We present a high quality data set of CNVs in 10 individuals. The use of different techniques and algorithms allowed a systematic assessment of detection power and accuracy. The high-resolution array datasets could be studied in more detail for their discrepancies or the mechanisms of genomic instability leading to small (coding) events. The WES callset is useful for developers of calling software to identify the caveats and advantages of the different models evaluated in our study.

Conflict of interest

The authors declare no conflict of interest.

References

[1] J. de Ligt, P.M. Boone, R. Pfundt, L.E. Vissers, T. Richmond, J. Geoghegan, K. O'Moore, N. de Leeuw, C. Shaw, H.G. Brunner, J.R. Lupski, J.A. Veltman, J.Y. Hehir-Kwa, Detection of clinically relevant copy number variants with whole-exome sequencing. Hum. Mutat. 34 (2013) 1439–1448.

[2] G. Klambauer, K. Schwarzbauer, A. Mayr, D.A. Clevert, A. Mitterecker, U. Bodenhofer, S. Hochreiter, cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 40 (2012) e69.

[3] N. Krumm, P.H. Sudmant, A. Ko, B.J. O'Roak, M. Malig, B.P. Coe, A.R. Quinlan, D.A. Nickerson, E.E. Eichler, Copy number variation detection and genotyping from exome sequence data. Genome Res. 22 (2012) 1525–1532.

[4] J. Li, R. Lupat, K.C. Amarasinghe, E.R. Thompson, M.A. Doyle, G.L. Ryland, R.W. Tothill, S.K. Halgamuge, I.G. Campbell, K.L. Gorringe, CONTRA: copy number analysis for targeted resequencing. Bioinformatics 28 (2012) 1307–1313.

[5] Y. Nannya, M. Sanada, K. Nakazaki, N. Hosoya, L. Wang, A. Hangaish, M. Kurokawa, S. Chiba, D.K. Bailey, G.C. Kennedy, S. Ogawa, A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res. 65 (2005) 6071–6079.

[6] V. Plagnol, J. Curtis, M. Epstein, K.Y. Mok, E. Stebbings, S. Grigoriadou, N.W. Wood, S. Hambleton, S.O. Burns, A.J. Thrasher, D. Kumararatne, R. Doffinger, et al., A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 28 (2012) 2747–2754.