Genomics 102 (2013) 174–181

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Review

Diagnosis of copy number variation by Illumina next generation sequencing is comparable in performance to oligonucleotide array comparative genomic hybridisation

J.L. Hayes [a,1], A. Tzika [a,1], H. Thygesen [b], S. Berri [c], H.M. Wood [c], S. Hewitt [a], M. Pendlebury [a], A. Coates [a], L. Willoughby [a], C.M. Watson [c], P. Rabbitts [c], P. Roberts [a], G.R. Taylor [c,*]

[a] Department of Clinical Cytogenetics, St James's University Hospital, Leeds Teaching Hospitals NHS Trust, Leeds, LS9 7TF, UK
[b] Leeds Cancer Research UK Centre, St James's University Hospital, Leeds, LS9 7TF, UK
[c] Leeds Institute of Molecular Medicine, University of Leeds, St James's University Hospital, Leeds, LS9 7TF, UK

ARTICLE INFO

Article history:
Received 3 December 2012
Accepted 9 April 2013
Available online 15 April 2013

Keywords:
Copy number
Illumina
Next generation sequencing
CNV

ABSTRACT

Array comparative genomic hybridisation (aCGH) profiling is currently the gold standard for genetic diagnosis of copy number. Next generation sequencing technologies provide an alternative and adaptable method of detecting copy number by comparing the number of sequence reads in non-overlapping windows between patient and control samples.

Detection of copy number using the BlueGnome 8 × 60k oligonucleotide aCGH platform was compared with low resolution next generation sequencing using the Illumina GAIIx on 39 patients with developmental delay and/or learning difficulties who were referred to the Leeds Clinical Cytogenetics Laboratory. Sensitivity and workflow of the two platforms were compared.

Customised copy number algorithms assessed sequence counts and detected changes in copy number. Imbalances detected on both platforms were compared.

Of the thirty-nine patients analysed, all eleven imbalances detected by array CGH and confirmed by FISH or Q-PCR were also detected by CNV-seq. In addition, CNV-seq reported one purported pathogenic copy number variant that was not detected by array CGH.

Non-pathogenic, unconfirmed copy number calls were detected by both platforms; however few were concordant between the two.

CNV-seq offers an alternative to array CGH for copy number analysis with resolution and future costs comparable to conventional array CGH platforms and with less stringent sample requirements.

Contents

Abbreviations: CNV, copy number variation; FISH, fluorescent in situ hybridisation; GaII x, Illumina Genome Analyzer IIx; MLPA, multiplex ligation-dependent probe amplification; OMIM, Online Mendelian Inheritance in Man; NGS, next generation sequencing; Q-PCR, Quantitative PCR; E–M, Expectation–maximization.

* Corresponding author at: Department of Pathology, University of Melbourne, Victoria 3010, Australia.
E-mail address: graham.taylor@unimelb.edu.au (G.R. Taylor).

[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.ygeno.2013.04.006

1. Introduction

Constitutional chromosomal abnormalities are a frequent cause of congenital structural malformation and developmental delay disorders in children [1]. Standard investigation involves karyotyping G-banded metaphase cells with confirmation of detected rearrangements using FISH, MLPA or other molecular based techniques [2]. Targeted molecular testing may be an option if the phenotype is well described, but limited coverage of genome-wide targets with low-throughput capability means that this cytogenetic approach has largely been replaced in the postnatal setting by high-resolution whole genome DNA array CGH (aCGH) analysis [3,4]. With selectable resolution according to microarray design, this technology has heralded a new era of genomics and assisted in the identification of multiple novel microdeletion syndromes [4,5]. It has become the gold standard for constitutional genetic diagnosis of copy number variation (CNV). As aCGH has developed from a research to a routine application, issues with the reproducibility and standardisation of microarray experiments have been raised [6].

The ultimate resolution for genomic interrogation is at the level of the base pair. Advances in DNA sequencing technology mean that detection of copy number variation by sequencing (CNV-seq) is now a realistic option for whole genome copy number analysis. Initially described in the analysis of cancer genomes [7,8], high-resolution mapping of copy number variation requires alignment of sequenced reads to a reference genome [9]. Distribution of the aligned reads is then analysed on a segmental or genomic window-by-window basis to determine differences in read-depth between the test and reference genomes. An increase in sample read-depth across a window, when compared to the control sample, represents a gain in genomic material; a reduction in read-depth suggests a loss [8]. Initial work was based on a high number of reads with deep coverage across the genome [8–12], a level not compatible with the high throughput, cost-sensitive requirements of diagnostic service.

We have previously described a method of sample multiplexing to determine dosage changes in tumour DNA [13]. In addition, we have compared the copy number profile generated by next generation sequencing (NGS) with that obtained by aCGH in DNA from cell lines and demonstrated that these appear to be almost identical [13]. In this study we expand this observation to clinical samples by comparing the results of CNV-seq analysis with the widely used UK gold standard 8×60k oligonucleotide microarray (oligoarray) for copy number detection in 39 phenotypically abnormal children. Firstly, we analysed the raw data of both techniques, without removing poor quality oligonucleotide probe data prior to microarray analysis. This was performed to compare the technologies directly, including their ability to detect imbalances above the noise produced from the technique. Secondly, we use a CNV-seq analysis method that closely resembles the method of analysis of the microarray to compare the platforms in a diagnostic setting.

2. Results

Table 1a and b shows a cross-tabulation of CNV calls at the window level. For the generation of this table, each microarray probe was compared to the NGS window to which it maps. NGS calls were based on regions of three adjacent windows, with applied thresholds constructed from posterior probabilities of ± 0.5. These were compared to the microarray probe that mapped the middle of the three windows. Microarray calls were made on the basis of three consecutive probes deviating from the commonly used microarray thresholds of ± 0.3. In some cases NGS windows were counted more than once in these tables, as occasionally a number of consecutive microarray probes map the same window in the targeted regions of the oligoarray. Comparison of the two platforms based on individual windows showed that the oligoarray produce the most calls. The read-depth across the genome was plotted and compared to the microarray probe location. These plots were then compared to the distribution of pathogenic genes (genes defined as morbid in Online Mendelian Inheritance in Man, OMIM). Fig. 1 shows distributions across chromosome 1. It is clear that the NGS reads are

Table 1
a. Losses and gains across the genome in a subset of 11 samples from the study (the remaining samples could not be included as their corresponding NGS and array data were analysed on different genome builds however all abnormalities are described in genome build 19). NGS calls were based on regions of three adjacent windows and compared to the microarray probe that mapped the middle of the three windows. Microarray calls were made on the basis of three deviating consecutive probes. More 'sets' of microarray probes were called as lost or gained than CNV-seq window 'sets', which suggests that the detection of copy number is more optimal using NGS read counts than fluorescence intensities. Commercial microarray software applies quality scores to each probe, and therefore most of these probes are not actually called. b. Direct comparison of one sample with noisy array data against NGS data showed significantly higher number of array calls compared to any other array case examined (this outlier was not included in a).

	NGS loss	NGS normal	NGS gain
a			
Microarray loss	302	1128	0
Microarray normal	194	553,400	85
Microarray gain	0	1250	10
b			
Microarray loss	0	1204	0
Microarray normal	9	48,231	9
Microarray gain	0	1125	1

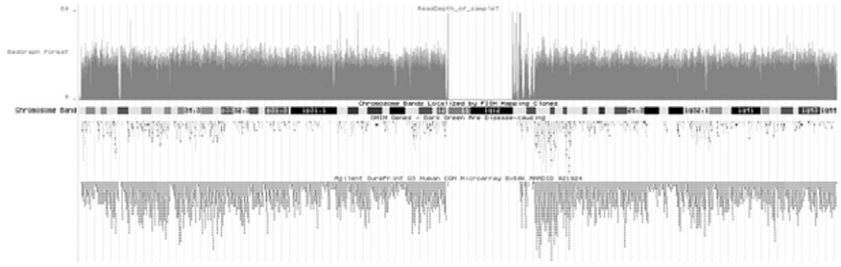

Fig. 1. The distribution of the reads across chromosome 1 (top track) in comparison to the distribution of the microarray probes on the 8 × 60k (bottom track). Pathogenic genes (morbid OMIM genes) are shown on the middle track. It is clear that the microarray probes are located to mirror regions of pathogenicity. CNV-seq reads are more evenly spaced across the genome.

evenly spread across the chromosome, whereas the microarray probes on the bottom track more closely mirror the Online Mendelian Inheritance in Man genes on the middle track.

Twelve of the thirty-nine samples analysed with the CNV-seq method were found to contain clinically significant copy number variants. These ranged in size from 49 kb to 7.8 Mb. Imbalances of different orders were detected (deletions, duplications and triplications) and some cases had more than one imbalance (Table 2). A full table of results found in this study is included in the Supplementary material, including breakpoints and follow-up analysis. All purported pathogenic abnormalities were detected across both platforms, with the exception of a 49 kb deletion at 2q23.1 (sample 10), which was only detected by the CNV-seq method. This region did not contain any microarray probes (Fig. 3). The deleted NGS windows within that region are shown in a background of their neighbouring windows (Fig. 3b). All copy number variants containing morbid OMIM genes were confirmed using FISH or Q-PCR in cases where suitable material was available. The 7q11.23 abnormality was detected as a duplication on the microarray ($\log_2$ 0.71), but was identified as a triplication using NGS ($\log^2$ 1.0402) and FISH (Case 3).

A complete list of the copy number variants called based on the applied thresholds is included in the Supplementary Table. Different copy number calls (assumed non-pathogenic) were made on each platform, which may be due to different controls used on each platform. Each of these calls was carefully interrogated against the Database of Genomic Variants [14] to examine the content of the affected region. Most regions were found to contain previously reported non-pathogenic copy number variation and therefore were not followed up. The regions containing OMIM genes that could link to the patient's overall clinical presentation were followed up using FISH or Q-PCR. Of the abnormalities followed up, 6 false positives were detected on oligoarray and 3 were detected using CNV-seq. Comparison of the calls generated by the two tests showed little concordance.

The number of aligned reads obtained per sample ranged between 2 and 7 million reads (median 3.4 million reads) which produced resolutions between 37 and 130 kb (median 76 kb). Read numbers differed due to improving numbers of reads obtained from the sequencer with successive runs and pipetting error when pooling samples.

3. Discussion

3.1. Comparison of raw data; assessment of capability of the technology

Direct comparison of the raw data generated from the two platforms (analysis method #1) showed that the microarray platform results in more spurious calls (Table 1a, b). Since analysis of the same sample using the proprietary oligoarray software produced significantly fewer calls it was assumed that these were discounted due to low quality (i.e., spots which do not fit the model of data are removed from further analysis). These are likely to be non-biological hybridisation signals that occur during an array experiment due to technical noise. The commercial software uses a quality score system that assesses the quality of each spot in relation to its location on the chip and discards those that fall outside the thresholds of a normal distribution. This preliminary analysis highlights the fact that analogue techniques inherently produce more noise.

In the direct comparisons between the raw NGS and oligoarray data, one sample (Table 1b) produced significantly more calls compared to the average array case. This sample had initially failed on the proprietary oligoarray platform due to the high level of noise. These findings provide evidence that with quality control not applied, particularly in a sample of poor quality, there are more calls when comparing oligoarray to CNV-seq

Table 2
A range of abnormalities of different sizes in different locations in the genome were detected across both platforms. All purported pathogenic abnormalities were detected on both platforms and were validated by FISH, with the exception of a (49 kb) deletion of chromosome 2, which was confirmed by Q-PCR.

Sample	Chromosome	Position		Chromosome band	Genomic length	Gain/loss
		Start	End			
1	16	29591395	30172134	16p11.2	580,739	Loss
2	21	43269253	48062257	21q22.3	4,793,004	Loss
	15	98278672	102370904	15q26.3	4,092,232	Gain
3	7	72703446	74129963	7q11.23	1,426,517	Triplication
4	14	68872493	75759128	14q24	6,886,635	Gain
7	16	1	2569135	16p13.3	2,569,134	Gain
		3318842	4678939	16p13.3	1,360,097	Gain
		4938802	5199796	16p13.3	260,994	Gain
10	2	148792468	148842172	2q23.1	49,704	Loss
16	3	62301648	69944162	3p14.1	7,642,514	Loss
19	3	138133378	145939687	3q23	7,806,309	Loss
21	16	4312849	4693138	16p13.3	380,289	Gain
	16	4946452	5237437	16p13.3	290,985	Gain
25	15	30545167	32419660	15q13.3	1,874,493	Loss
	8	9466830	9709038	8p23.1	242,208	Gain
27	17	7065530	7673270	17p13.1	607,740	Loss
28	1	145118135	145728993	1q21.1	610,858	Loss

directly. Analysis of the same sample using CNV-seq produced a successful result. This analysis also allowed a direct comparison of calls of imbalance on the basis of NGS window and oligoarray probe location.

An attempt was made to analyse the NGS data using thresholds constructed from posterior probabilities of 0.5. Using this approach, a large number of calls were generated, making analysis, follow-up and interpretation particularly difficult; therefore it is not recommended to analyse CNV-seq in this way.

Analysis of the raw NGS data showed that the read distribution fitted well with the gamma model. The approach used for the analysis of the 39 patients in comparison to the oligoarray (analysis method #2) could not be modelled in this way because windows were segmented according to similar $\log^2$ ratio. As a segmentation technique is applied in commercial array CGH analysis, it was deemed necessary to base our NGS approach on this method.

3.2. Quality and volume of DNA

Although 2 μg of DNA was used as starting material, a successful library preparation (with comparable sequencing quality and data output) was achieved from as little as 150 ng in a single case; as picogram quantities are loaded onto the sequencer, it is likely that much lower quantities would be sufficient. We have previously reported copy number variation detection in tumour samples with starting quantities as low as 5 ng [13]. CNV-seq was more tolerant of poor quality DNA, with one abnormality failing on oligoarray and producing a successful result on CNV-seq, as previously mentioned. Our group has also successfully carried out copy number analysis on formalin-fixed paraffin-embedded tissue [13], from which isolated DNA is frequently of poor quality. The low DNA and quality requirements of CNV-seq should reduce the need for repeat samples, and also provide results where a repeat sample is not possible.

3.3. Resolution

The detection of an abnormality using CNV-seq that was not detected using oligoarray demonstrates the power of this technique; however it is highly likely that the opposite would be true of the very small imbalances that fall within the targeted regions of the array. The number of data points across the genome is 60,000 probes using the oligoarray and 75,000 windows using CNV-seq at this read-depth. This suggests that more abnormalities would be detected on average across the genome using CNV-seq compared to oligoarray (only two windows/probes are required to call an imbalance). Higher resolution could be achieved by multiplexing fewer samples per lane, and therefore obtaining more reads per sample, reducing the genomic region covered by a 40 test read window.

We found that 8 samples per flow cell lane of the GAIIx yielded sufficient reads to result in resolution comparable to the oligoarray without a substantial increase in cost to the current diagnostic platform of choice. The average resolution (74 kb) of the CNV-seq was significantly better than the backbone resolution of an oligoarray, but lower than the targeted regions of an optimal oligoarray (48 kb). A fair comparison of the resolution obtained by the two platforms cannot be made since the NGS reads are scattered uniformly over the genome whereas microarray probes are concentrated in predefined clinically important regions (Fig. 1). This can often aid in the interpretation of the results from oligoarray because imbalances are more likely to be detected in well-studied regions, whereas CNV-seq may detect imbalances in regions of poor annotation. However, with the advent of CNV-seq it is likely that information in these less well-studied regions will improve.

The oligonucleotide probes are 60 nucleotides in length and the length of a read in this method of CNV-seq is 74 bp. CNV-seq will therefore allow detection of imbalances in more repetitive regions as reads are less likely to map more than one region. Repetitive loci are not included in the design of an oligoarray in order to prevent noise from hybridisation of a probe to more than one locus and therefore imbalances will not be detected in repetitive regions using this platform. With the higher read count of the HiSeq, mapping could be improved even further at no extra cost. Most centres use at least 100 bp reads on this platform, with some up to 150 bp and usually paired end reads are employed, which would further improve mapping into repetitive regions. For diagnostic use however, a balance between the length of time to perform the sequencing, the resolution and the cost is required.

It must be noted that although the majority of laboratories in the United Kingdom use the 8 × 60k microarray for this type of testing, outside of the UK higher resolution microarray platforms are employed. Evaluation of CNV-seq in light of the cost and resolution of these platforms is of obvious importance.

3.4. Controls and non-pathogenic CNV discordance

Non-pathogenic CNVs (not confirmed) detected by the two platforms were often different. One plausible explanation for the observed discordance could be that the backbone resolution of the NGS platform was higher than that of the oligoarray. A larger number of calls would therefore be expected in regions where the array has limited detection capability. It could also be argued that the digital nature of the NGS data would make a false positive less likely to be generated. In comparison, the quality of the array data can fluctuate more and vary between experiments.

The difference may also be due to the different controls used in the two assays. Oligoarray requires a control to be run in each experiment alongside the test case for practical reasons. CNV-seq on the other hand produces digital data and therefore a previously sequenced control prepared in the same way, can be utilised, eliminating batch-associated CNVs and allowing more streamlined analysis from one run to the next. This also allows the option of multiple controls compared to one sample. For instance, if an imbalance is called at a borderline $\log^2$ ratio with use of a specific control, then that test case could be compared to a different control to confirm the abnormality before any follow-up is performed. There is also the option of population-specific controls, in cases where ethnicity is known.

This study was carried out for autosomal analysis only as a mixed sex control was used. Ideally, two controls of different sexes made from multiple samples of mixed race, prepared within the same laboratory, using the same sequencer, would be used. This would reduce noise and decrease the number of those called non-pathogenic copy number variants from population-specific groups. It was noted prior to this study that a high coverage control produced less noise than a control of the same coverage with the test sample. Our control had approximately 6× higher coverage than the median test sample (20 million versus a median of 3.4 million reads); in practice this would mean that control read variability was less profound making true changes in the test sample more noticeable.

3.5. Application of CNV-seq in the diagnostic laboratory

The availability of a sequencing platform is likely to be the stumbling block for the introduction of CNV-seq in diagnostic laboratories. Access to core sequencing facilities makes sequencing more affordable and accessible. Illumina also has another platform available, the MiSeq, which is more affordable for the initial outlay, but sequencing cost is such that it would not be efficient for this test, and the higher throughput platforms such as the GAIIx or the HiSeq are more cost-effective.

The adaptability of CNV-seq is extremely attractive for a diagnostics laboratory. Read numbers could be dramatically reduced compared to those sequenced here, to produce a very low resolution method which would be more suitable in cases where interpretation of the result may be a problem, such as in prenatal diagnosis, or the number of reads could be dramatically increased if a very small imbalance is to be

detected. Read-lengths are likely to be restricted to those offered by the sequencing service available to the user; however those with the option could offer a quick result with short reads, yet improved mappability with longer reads. Different sequencing tests on the same patient could also be merged, so as to improve coverage in certain regions, a feature which has never before been possible in diagnostic testing. The increase in number of reads will improve breakpoint refinement, which is an essential focus of the detection of CNVs. This could also aid in the interpretation of an abnormality showing differing phenotypic expression across family members.

In the future it is likely that CNV-seq could be coupled with other analyses in a single test as was performed in a recent study by Sarhadi et al. [15]. Paired-end sequencing, although not cost effective for a diagnostic service currently, may be used to detect balanced rearrangements in addition to copy number; methods for this analysis have already been developed [16]. Exome sequencing may also be an option, and tools are available for this type of analysis, enabling SNP calling alongside the copy number analysis [17,18]. As the cost of exome sequencing falls, this may become a practical diagnostic option for measuring gene copy number changes, although unlike CNV-seq, it would not be a genome-wide survey.

3.6. Illustrative abnormal cases

3.6.1. Case 3 — 7q triplication

This case was particularly interesting, as a chromosome 7 duplication was reported on the oligoarray but was later found to be a triplication by interphase FISH using the Vysis ELN FISH probe. CNV-seq also revealed the presence of four copies of this locus in the patient genome. The digital nature of the NGS data allows determination of the exact copy number of each locus in the genome examined. This is certainly more challenging with the use of analogue data, such as the oligoarray data. The $\log^2$ ratio value of each segment called by NGS accurately follows the copy number allowing for a precise characterisation of the imbalance involved, i.e. 2 copies (expected $\log^2$, 0.58) and 3 copies (expected $\log^2$, 1). It must be noted however that triplications are detected by microarray in routine practice; it could be that this was of poorer quality.

3.6.2. Case 10 — 2q23.1 microdeletion

A 10-year old patient with developmental delay, microcephaly and a Smith–Magenis request for investigation was initially referred. Analysis on the oligoarray platform showed no copy number changes associated with the patient's clinical presentation. CNV-seq revealed a microdeletion on chromosome band 2q23.1 of approximately 49 kb. This was the only discrepant result between the two platforms. The region is poorly covered by oligonucleotide probes on the array chip (Fig. 3) whereas the CNV-seq method generates read depth information across the whole genome at a reasonably similar resolution. Fig. 3b shows only 3 consecutive windows in the deleted region to present with lower read counts compared to their neighbouring windows, demonstrating the power of the algorithm to accurately detect such small changes. Interrogation of the CNV call against the Database of Genomic Variants [14] showed a mental retardation autosomal dominant I (*MBD5* gene, OMIM #156200) disease gene to be contained within the region. Loss of this gene has been described in the recently characterised 2q23.1 microdeletion syndrome [19]. The patient's phenotype fitted well with that described in the literature. Interestingly, it has been stated that several 2q23.1 microdeletion cases had initially given the clinical impression of Angelman, Rett or Smith–Magenis syndromes, as was the case for our patient [20]. Follow-up Q-PCR studies confirmed the 2q23.1 deletion and when received, Q-PCR analysis of parental samples will be performed to determine the inheritance of this finding.

3.6.3. Case 7 — child with Rubinstein–Taybi syndrome

3.6.3.1. Case 21 — parental analysis. A child with developmental delay was referred for cytogenetic analysis and was included in the current study. NGS revealed a deletion of one chromosome 16 between DNA positions 3318842 and 4678939. This region contained, among other genes, the *CREB* binding protein — *CREBBP* gene (chr16: 3710941–3866165, 16p13.3). On this basis, a diagnosis of Rubinstein–Taybi syndrome type 1 (RSTS1) was made for this patient.

A parent of this child was also analysed with NGS and oligoarray and revealed a more downstream 3′ breakpoint of that deletion, not including the *CREBBP* gene; hence the phenotype of this patient was apparently normal (deletion on chr16: 4312849–4693138). A number of other aberrations were revealed on 16p13.3 in both tests, showing the power of the techniques to delineate more complex abnormalities. The example presented shows the importance of defining accurate breakpoints of copy number changes in a diagnostic service and demonstrates the ability of NGS to address this requirement.

3.7. Conclusion

Table 3 is included to summarise the main attributes of each technology. Next generation sequencing is unique in its potential to offer convergence of existing genetic technologies, with the possibility of a single platform solution for genetic diagnosis of a range of abnormalities from single gene mutations to aneuploidy. Modification of the Illumina operating protocols allows paired-end mapping, producing accurate positional data and possible detection of balanced rearrangements [16]. The digital nature of the comparison means that once a normal control is sequenced, the dataset can be used for multiple patient samples.

Cost of the CNV-seq will be competitive with oligoarray with the use of the HiSeq and can also be adapted according to the level of resolution required. The challenge, in the diagnostic setting, will be obtaining access to a sequencer with availability at suitable times to ensure an appropriate turnaround time.

The use of the NGS technology for the diagnosis of genetic disease has been previously demonstrated [21]. In this study we have assessed the

Table 3
Comparison of CNV-seq with the 8 × 60k oligoarray; main points included.

CNV-seq	Arrays (8 × 60k)
Even coverage across the genome	More targeted resolution in clinically interesting regions
Expected to detect imbalances in poorly annotated regions	Limited information outside of the target regions
Order of multiplexing can be altered to the resolution required for a specific application	'Fixed' resolution in backbone and targeted regions
Sequencing data from previous experiments can be utilised as a control	'Patient vs patient' or 'patient vs control' run in the same experiment
Sequencing data can be obtained from low quality and quantity of genomic DNA	In general, requires higher amount and good quality of DNA
Ability to use paired-end reads for the detection of balanced rearrangements	Only unbalanced rearrangements can be detected
Eventually, the use of one platform for multiple applications, i.e. detection of copy number changes, balanced rearrangements, mutations, UPD etc	Arrays do not offer such a potential
At the moment, practically more difficult to be used as a diagnostic test in cytogenetics, as access to the facility and a lot of expertise in data processing are required	Extensive expertise in the use of arrays in cytogenetics
Cost of running the service is higher in comparison, but expected to drop in time and with the use of higher capacity sequencers	Relatively low cost of the platform has made it possible for arrays to be used as a front-line test in many laboratories across the UK

performance of the Illumina GAIIx platform in a cytogenetic diagnostic setting and compared it against the UK gold standard of oligoarray CGH. We show that clonal sequencing using the Illumina GAIIx platform can perform high throughput copy number detection at a comparable level to the oligoarray 8 × 60k platform. All purported pathogenic copy number variants detected among a cohort of 39 children with phenotypic abnormalities using the oligoarray method were successfully detected by CNV-seq. In addition, a submicroscopic deletion at 2q23.1, which remained undetected by the oligoarray platform, was revealed by NGS. The digital nature of the NGS data means that the exact copy number at each locus in the genome can be inferred. As the cost of sequencing continues to fall, NGS technologies may offer a superior analysis pipeline for the detection of chromosomal and genetic abnormalities on a single platform.

4. Materials and methods

4.1. Sample selection

Thirty-nine patients with developmental delay and/or learning difficulties referred to the Leeds Teaching Hospitals NHS Trust Cytogenetic Department for array CGH testing were selected. Genomic DNA was extracted from whole blood using a salt-precipitation extraction protocol [22] and stored at −20 °C.

4.2. Illumina Genome Analyzer library preparation

DNA concentration and purity were determined using the Quant-iT PicoGreen ds DNA BR assay (Invitrogen, Paisley, UK) and the Agilent Bioanalyzer Genechip (Agilent Ltd, UK). Two micrograms of genomic DNA was used to prepare the DNA libraries for sequencing according to standard protocols. DNA was sheared to approximately 150–200 base pairs using adaptive focused acoustics (Covaris S2, KBioscience, Hertz, UK). Fragments were purified using MinElute columns (Qiagen, Chatsworth, CA) and end repair was performed using the End-It DNA End-Repair kit (EpiCentre, Madison, WI). Eight different adaptors containing 6-nt barcodes were used to index the samples (LigaFast Kit, Promega), using previously described methods [13]. Fragments were size selected to 200 bp ± 25 bp from 2% TBE agarose gels (QiaQuick Gel Extraction Kit, Qiagen). Following PCR amplification, the libraries were purified on a QiaQuick column and quality-checked using an Agilent Bioanalyzer DNA 1000 LabChip, with quantification performed by a Quant-iT PicoGreen dsDNA assay. Equal amounts of each tagged library were then pooled for cluster generation and sequencing using the standard Illumina single-read 76-cycle operating protocol. 8 samples per lane of a flow cell on the Illumina GaII (GaIIx) were sequenced.

4.3. Array CGH processing

DNA was cleaned up using ethanol precipitation [23]. The optimised BlueGnome protocol was used for processing CytoChip 8 × 60k oligoarrays including enzyme digestion, labelling, clean up, hybridisation and washing [24]. Scanning was carried out according to the manufacturer's protocol. Promega male and female controls (ten pooled DNA samples) were used for this analysis.

4.4. Data analysis

Image analysis and base calling were performed using the Illumina CASAVA pipeline. Subsequent analysis of copy number variation was as previously reported [13]. Python scripts first segregated the samples according to their indexing tags. Reads were aligned to the reference sequence using the BWA alignment algorithm [25]. A custom designed Python script was used to perform pair-wise comparisons of each test and control sample; the genome was split into non-overlapping windows of equal numbers of normal reads and the number of patient reads within each window was counted. Prior to the statistical analysis, the read counts were corrected for GC bias, achieved using the locally weighted regression (LOESS) method [26]. An increased test:control ratio indicates a gain whereas a decreased test:control ratio indicates a loss.

The windows were defined in such a way that the read count of the patient sample in each window would be approximately 40 reads; this window size offered a balance between resolution and noise. For the purposes of the comparison, we used a bioinformatically constructed control made of sequencing data from 20 normal, Caucasian individuals of both sexes downloaded from the 1000 genome project [27]. Data was obtained from samples sequenced on Illumina platforms and trimmed to the same read length and merged. We restricted the analysis to autosomes since the reference sample was based on a mixture of males and females.

Two methods of analysis are described below. The first method was a statistical-based method performed on the file obtained after windows were generated for the sequencing and on the raw excel file obtained from BlueGnome array processing.

The second, the proposed analysis protocol for the NGS data, more closely resembles the method of array analysis as it uses a segmentation algorithm and $\log^2$ ratio thresholds to produce copy number calls.

4.5. Analysis method #1 — direct comparison of NGS and oligoarray technologies

The purpose of this was to compare the technology with the same analysis method, before any low quality calls are removed.

A natural model for the read count in test sample i in window w is:

$$y_{i,w} \sim \text{Poisson}\left(c_{i,w}\lambda_i/2\right)$$

where λ_i is the average read count in sample i for a window with a copy number of 2, and $c_{i,w}$ is (unknown) a copy number of window w in sample i. However, since the processed read counts were not integers and (more importantly) since they were overdispersed (which could, to some extent, be expected due to statistical fluctuations of the window sizes), we decided to model the processed read counts as $y_{i,w} \sim \text{gamma}(c_{i,w}\alpha_i/2, \beta_i)$, where the shape parameter α_i and the rate parameter β_i were estimated by fitting the theoretical (0.25, 0.75) quantiles to the corresponding empirical quantiles, utilising the fact that the vast majority of the windows will have a copy number of 2. The fractions $P_{i,c}$ of the windows with copy number c were estimated using the EM algorithm [28]. This model provided a very good fit to the data (Fig. 2).

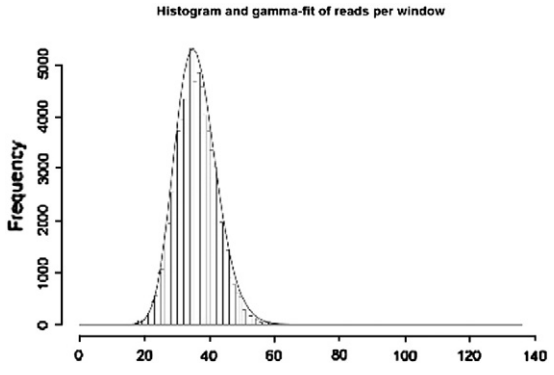

Fig. 2. Number of reads per window fitting into a gamma distribution model.

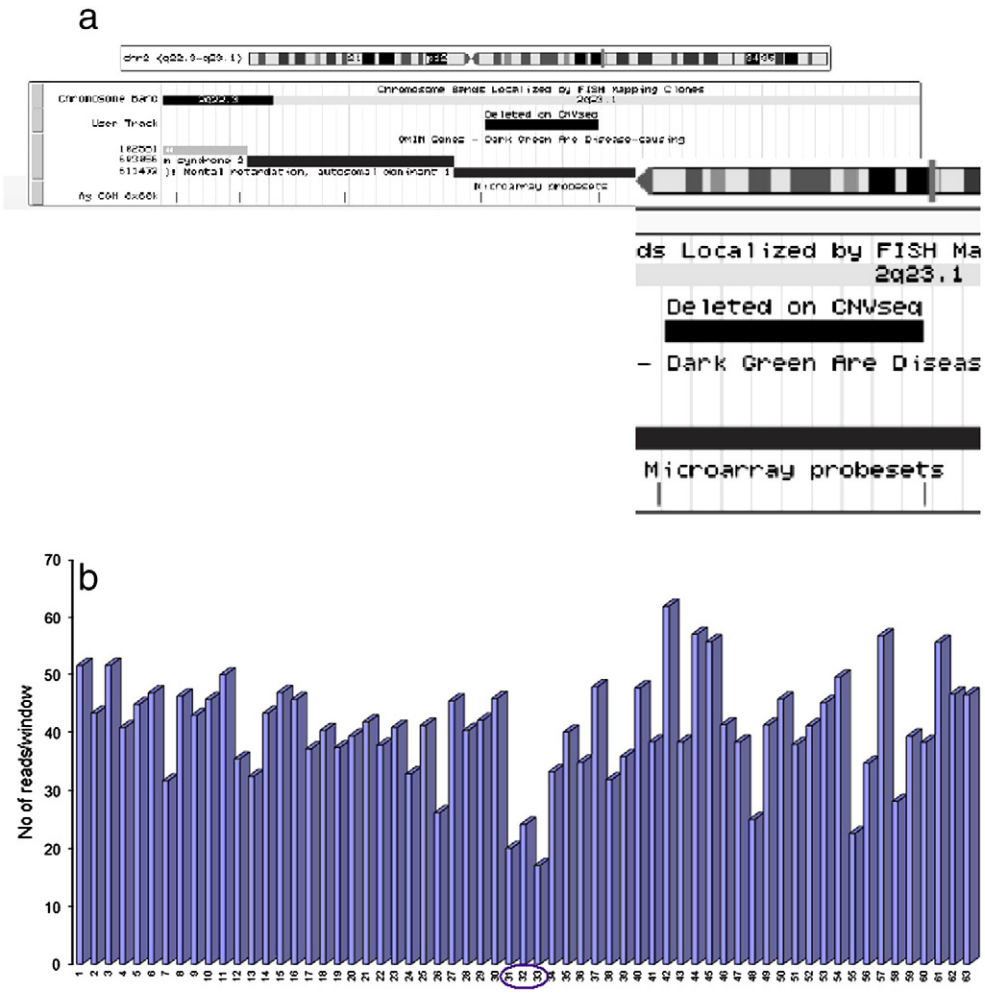

Fig. 3. a. The region of imbalances detected on CNV-seq but not oligoarray. The top track shows the region identified to be deleted on the sequencing. This was detected in three windows. The bottom track shows the microarray probes for the Agilent 8 × 60k platform. There are two probes just outside this region but no probes within the region. This region on chromosome 2 is not a targeted area on the microarray so probe coverage is lower. b. The read numbers in each window across the region of the deleted 2q. Windows 31–33 (circled) have fewer reads in comparison to the other windows. Read numbers across a region can vary due to factors such as GC content and repetitive elements, for this reason comparison to a control is essential.

After the model was fitted we calculated posterior probabilities for copy number aberrations using Bayes' formula:

$$p\left(CN_{i,w} = c\right) = \frac{P_{i,c}d_{i,c}\left(y_{i,w}\right)}{\sum_{k=1:a}P_{i,k}d_{i,k}\left(y_{i,w}\right)}$$

where $d_{i,c}$ is the fitted density function for the read counts corresponding to copy number c in sample i.

From these posterior probabilities, thresholds for CNV detection can be constructed. In order to minimise the number of misclassified genes, we decided to call a CNV when the posterior probability of a CNV exceeds 0.5.

The separation of windows with normal and abnormal copy numbers becomes clearer when the calls are based on larger windows, e.g. by fusing two or three adjacent windows. This is particularly relevant for the gains, which are more difficult to separate from the noise than the deletions. For the purpose of comparison with the microarray data, we generated CNV calls based on regions of three adjacent windows and compared the NGS call to the microarray that mapped the middle of the three windows. This allowed us to compare copy number calls from the two technologies at similar resolution.

4.6. Analysis method #2 — comparison of the proposed method to oligoarray analysed by proprietary software

The Leeds Cytogenetics Laboratory currently uses BlueGnome's proprietary software, BlueFuse, for the analysis of oligoarray-CGH data. In order to compare the data generated from the two platforms, we used a method that more closely resembled the commercial method of analysis of oligoarrays, with segmentation applied. This was then compared to the results of the oligoarray when processed using the BlueFuse software.

Following segmentation into windows across the genome, the proposed NGS method of analysis involved normalisation of the read counts, adjusting for differential read-depths between samples (full code for this analysis is available in the Supplementary material). A $\log^2$ ratio of normalised sample:control read count was then calculated for each window. A mean $\log^2$ ratio was generated for consecutive windows and used to produce graphical representations of copy number variation for each patient. Segments of equal copy number were called using the Bioconductor DNAcopy package [29]. Log^2 ratio thresholds were set based on the midpoint between the baseline (0) and the expected $\log^2$ of a heterozygous deletion (−1) and between the baseline and the expected $\log^2$ of a heterozygous gain (0.58) resulting in thresholds of −0.5 and +0.29. The minimum platform resolution was

defined as being equal to the average genomic length of two read windows as this was the minimum requirement for the generation of a distinct segment with this method.

Follow-up of imbalances detected on either platform was performed only if they contained disease genes and, in some cases, likely to be of clinical significance to the phenotype. It should also be noted that some samples were not followed up due to lack of material.

4.7. Fluorescent in situ hybridisation (FISH) analysis

Metaphase spreads were prepared from peripheral blood lymphocytes using standard methodology. Chromosomes were visualised by counterstaining with DAPI.

4.8. Quantitative PCR analysis

Q-PCR for follow-up of imbalances was carried out using primers designed by and protocols supplied by Primer Design Ltd. Q-PCR follow-up for the discordant abnormality identified in one patient was performed by the Cytogenetics Laboratory at Addenbrookes Hospital as part of a diagnostic test.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2013.04.006.

Acknowledgments

This project was supported by The Leeds Teaching Hospitals Trust Charitable Foundation grant number 9R11/1001.

Dr Kelly Cohen was involved in the early stages of the NGS project. Dr Eamonn Sheridan was responsible for the clinical advice on the 2q23.1 case.

Amanda Clarkson, DipRCPath., (Addenbrookes Hospital Cytogenetics Department) performed Q-PCR follow-up of the 2q23.1 deletion.

Primer Design Ltd was responsible for the design of all primers for Q-PCR follow-up.

The Translational Unit at St James's Hospital helped with the bioinformatic processing of the data.

References

[1] K.K.B. Filkins, Ultrasound and fetal diagnosis, Curr. Opin. Obstet. Gynecol. 17 (2005) 185–195.
[2] J.B. Moeschler, M. Shevell, Clinical genetic evaluation of the child with mental retardation or developmental delays, Pediatrics 117 (2006) 2304–2316.
[3] O.P. Kallioniemi, A. Kallioniemi, D. Sudar, D. Rutovitz, J.W. Gray, F. Waldman, Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors, Semin. Cancer Biol. 4 (1) (1993) 41–46.
[4] L. Vissers, B. de Vries, J. Veltman, Genomic microarrays in mental retardation: from copy number variation to gene, from research to diagnosis, J. Med. Genet. 47 (2010) 289–297.
[5] B.S. Emanuel, S.C. Saitta, From microscopes to microarrays: dissecting recurrent chromosomal rearrangements, Nat. Rev. Genet. 8 (11) (2007) 869–883.
[6] J. Hehir-Kwa, M. Egmont-Petersen, I. Janssen, D. Smeets, A. van Kessel, J. Veltman, Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis, DNA Res. 14 (2007) 1–11.
[7] P.J. Campbell, P.J. Stephens, E.D. Pleasance, S. O'Meara, H. Li, T. Santarius, L.A. Stebbings, C. Leroy, S. Edkins, C. Hardy, et al., Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing, Nat. Genet. 40 (2008) 722–729.
[8] D. Chiang, C. Getz, D. Jaffe, M. O'Kelly, X. Zhao, S. Carter, C. Chad, C. Nusbaum, M. Meyerson, E.S. Lander, High-resolution mapping of copy number alterations with massively parallel sequencing, Nat. Methods 6 (1) (2008) 99–103.
[9] C. Xie, M. Tammi, CNV-seq, a new method to detect copy number variation using high-throughput sequencing, BMC Bioinforma. 10 (2009) 80–89.
[10] A. Magi, L. Tattini, T. Pippucci, F. Torricelli, M. Benelli, Read count approach for DNA copy number variants detection, Bioinformatics 28 (4) (2012) 470–478.
[11] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S.F. Grant, H. Hakonarson, M. Bucan, PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data, Genome Res. 17 (11) (2007) 1665–1674.
[12] R. Xi, A.G. Hadjipanayis, L.J. Luquette, T.M. Kim, E. Lee, J. Zhang, M.D. Johnson, D.M. Muzny, D.A. Wheeler, R.A. Gibbs, R. Kucherlapati, P.J. Park, Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion, Proc. Natl. Acad. Sci. 108 (46) (2011) 1128–1136.
[13] H. Wood, O. Belvedere, C. Conway, C. Daly, R. Chalkley, Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens, Nucleic Acids Res. 38 (14) (2010) 1–11.
[14] A.J. Iafrate, L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, C. Lee, Detection of large-scale variation in the human genome, Nat. Genet. 36 (9) (2004) 949–951, (data. Bioinformatics 23, 657-63).
[15] V.K. Sarhadi, L. Lahti, I. Scheinin, A. Tyybäkinoja, S. Savola, A. Usvasalo, R. Räty, E. Elonen, P. Ellonen, U.M. Saarinen-Pihkala, S. Knuutila, Targeted resequencing of 9p in acute lymphoblastic leukemia yields concordant results with array CGH and reveals novel genomic alterations, Genomics S0888-7543 (13) (2013) 00002–5.
[16] M. Talkowski, C. Ernst, A. Heilbut, C. Chiang, C. Hanscom, A. Lindgren, A. Kirby, S. Liu, B. Muddukrishna, T. Ohsumi, Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research, Am. J. Hum. Genet. 88 (4) (2011) 469–481.
[17] M. Kirwan, A.J. Walne, V. Plagnol, M. Velangi, A. Ho, U. Hossain, T. Vulliamy, I. Dokal, Exome sequencing identifies autosomal-dominant SRP72 mutations associated with familial aplasia and myelodysplasia, Am. J. Hum. Genet. 90 (5) (2012) 888–892.
[18] M. Fromer, J.L. Moran, K. Chambert, E. Banks, S.E. Bergen, D.M. Ruderfer, R.E. Handsaker, S.A. McCarroll, M.C. O'Donovan, M.J. Owen, et al., Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth, Am. J. Hum. Genet. 91 (4) (2012) 597–607.
[19] M.E. Talkowski, S.V. Mullegama, J.A. Rosenfeld, B.W. van Bon, Y. Shen, E.A. Repnikova, J. Gastier-Foster, D.L. Thrush, S. Kathiresan, D.M. Ruderfer, et al., Assessment of 2q23.1 microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder, Am. J. Hum. Genet. 89 (4) (2011) 551–563.
[20] B.W. Van Bon, D.A. Koolen, L. Brueton, D. McMullan, K.D. Lichtenbelt, L.C. Adès, G. Peters, K. Gibson, S. Moloney, F. Novara, et al., The 2q23.1 microdeletion syndrome: clinical and behavioural phenotype, Eur. J. Hum. Genet. 18 (2) (2010) 163–170.
[21] J.E. Morgan, I.M. Carr, E. Sheridan, C.E. Chu, B. Hayward, N. Camm, H.A. Lindsay, C.J. Mattocks, A.F. Markham, et al., Genetic diagnosis of familial breast cancer using clonal sequencing, Hum. Mutat. 31 (2010) 1–8.
[22] E. Travaglini, Methods for the extraction and purification of deoxyribonucleic acids from eukaryote cells, Methods Cell Biol. 7 (1973) 105–127.
[23] J. Wilcockson, The differential precipitation of nucleic acids and proteins from aqueous solutions by ethanol, Anal. Biochem. 66 (1975) 64–68.
[24] BlueGnome Ltd, Capital Park CPC4. Fulbourn, Cambridge, CB21 5XE, 2012.
[25] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (2009) 1755–1760.
[26] W.S. Cleveland, S.J. Devlin, Locally weighted regression: an approach to regression analysis by local fitting, J. Am. Stat. Assoc. 83 (1988) 596–610.
[27] The 1000 genomes project consortium. A map of human genome variation from population-scale sequencing, Nature 467 (2010) 1061–1073.
[28] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. 39 (1) (1977) 1–38.
[29] E. Venkatraman, A.B. Ohlsen, A faster circular binary segmentation algorithm for the analysis of array-CGH data, Bioinformatics 23 (2007) 657–663.