BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 8 2011, pages 1052–1060
doi:10.1093/bioinformatics/btr106

Genome analysis

Performance assessment of copy number microarray platforms using a spike-in experiment

Eitan Halper-Stromberg[1,2], Laurence Frelin[3], Ingo Ruczinski[1], Robert Scharpf[1], Chunfa Jie[4], Benilton Carvalho[5], Haiping Hao[4], Kurt Hetrick[6], Anne Jedlicka[7], Amanda Dziedzic[7], Kim Doheny[6], Alan F. Scott[6,10], Steve Baylin[8], Jonathan Pevsner[3,9,*], Forrest Spencer[10,*] and Rafael A. Irizarry[1,*]

[1]Department of Biostatistics, Bloomberg School of Public Health, [2]Program in Human Genetics and Molecular Biology, Johns Hopkins University School of Medicine, [3]Department of Neurology, Hugo W. Moser Research Institute at Kennedy Krieger, [4]JHMI Microarray Core, High Throughput Biology Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA, [5]Department of Oncology, University of Cambridge, CRUK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK, [6]Center for Inherited Disease Research, Johns Hopkins University, [7]Department of Molecular Microbiology and Immunology, Johns Hopkins Malaria Research Institute, Johns Hopkins Bloomberg School of Public Health, [8]Department of Oncology, Johns Hopkins University School of Medicine, [9]Department of Neuroscience, Johns Hopkins School of Medicine and [10]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Changes in the copy number of chromosomal DNA segments [copy number variants (CNVs)] have been implicated in human variation, heritable diseases and cancers. Microarray-based platforms are the current established technology of choice for studies reporting these discoveries and constitute the benchmark against which emergent sequence-based approaches will be evaluated. Research that depends on CNV analysis is rapidly increasing, and systematic platform assessments that distinguish strengths and weaknesses are needed to guide informed choice.

Results: We evaluated the sensitivity and specificity of six platforms, provided by four leading vendors, using a spike-in experiment. NimbleGen and Agilent platforms outperformed Illumina and Affymetrix in accuracy and precision of copy number dosage estimates. However, Illumina and Affymetrix algorithms that leverage single nucleotide polymorphism (SNP) information make up for this disadvantage and perform well at variant detection. Overall, the NimbleGen 2.1M platform outperformed others, but only with the use of an alternative data analysis pipeline to the one offered by the manufacturer.

Availability: The data is available from http://rafalab.jhsph.edu/cnvcomp/.

Contact: pevsner@jhmi.edu; fspencer@jhmi.edu; rafa@jhu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 29, 2010; revised on January 20, 2011; accepted on February 17, 2011

1 INTRODUCTION

Copy number variant (CNV) loci are a major source of variation observed among human genomes (Conrad *et al*., 2009; Hurles *et al*., 2008). While some CNVs have no known functional consequence, some are associated with inherited microdeletion or microduplication syndromes (Carter, 2007; Lee and Lupski, 2006; McCarroll and Altshuler, 2007; Scherer *et al*., 2007). High-density microarrays permit the measurement of DNA copy number variation across the genome with broad coverage of all chromosomes and high resolution. Approaches based on high-throughput sequencing technology have been tried (Gilad *et al*., 2009) and are very promising. However, at this time genome sequence-based detection of CNVs is not yet a competitive strategy in terms of costs, availability of established analysis tools for general use and wealth of published performance appraisal from experience. The several microarray-based platforms for CNV detection that have been developed are currently subject to active price and performance competition. Research and clinical laboratories are seeking to determine the technology platform(s) that perform best for detection of CNVs: accurately and precisely reporting their length, their copy number and other features such as homozygosity. Here we provide an assessment of the major vendors in the current marketplace based on direct comparison of the performance of samples containing intentional copy number alterations. These are evaluated using company-recommended software tools, as well as independent methods.

CNVs spanning one megabase (Mb) are detectable by cytogenetic techniques such as spectral karyotyping and fluorescence *in situ* hybridization. However, the level of resolution attained by these methods does not permit detection of copy number change in smaller segments. Microarray comparative genomic hybridization (array-CGH) was the first technique developed to achieve a higher resolution (Lucito *et al*., 2003; Pinkel *et al*., 1998; Pollack *et al*., 1999). In this technology, DNA from a test sample and a reference sample are labeled using different fluorophores (Cy3 and Cy5), and hybridized to probes printed on a glass slide. The log-ratio of the fluorescence intensity of the test to that of the reference DNA is calculated and used as a measure of

*To whom correspondence should be addressed.

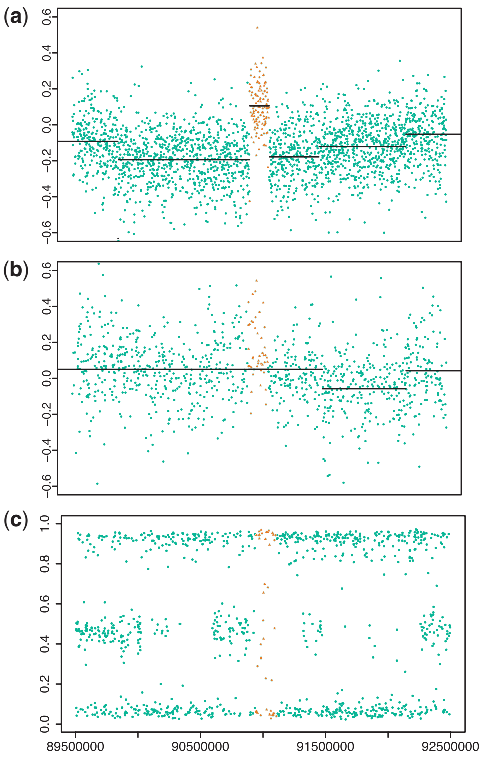

Fig. 1. Raw data plots of a genomic region altered by spike-in to an expected dosage of 2.5. (**a**) Microarray raw log-ratios (M-values) plotted against chromosome locations for the NimbleGen 2.1M platform. The red points represent the region altered by spike-in, green points are surrounding genomic DNA. The horizontal lines show the result of a smoothing technique. (**b**) For Illumina 1M Duo, the BAC spike-in is not detected by the sum of SNP intensities. (**c**) However, the BAF for Illumina 1M Duo shows a pattern of values distributed away from heterozygosity, reflecting the presence of the spike-in.

relative dosage. However, the level of random variation in these log-ratios does not permit copy number calls at the individual feature level. Therefore, analysts use smoothing, a statistical procedure that averages neighboring measurements to improve precision (Fig. 1a). With smoothing techniques, precision, and therefore specificity, can be improved by increasing the size of the averaging window, but at a cost of averaging over, and therefore missing, smaller CNVs. Competing statistical techniques provide different solutions to this sensitivity–specificity tradeoff. Examples are hidden Markov models (Colella *et al.*, 2007; Greenman *et al.*, 2010; Korn *et al.*, 2008; Scharpf *et al.*, 2008; Wang *et al.*, 2007) and segmentation algorithms such as circular binary segmentation (CBS) (Olshen *et al.*, 2004). Note that specificity can be improved without detrimentally affecting sensitivity by increasing the resolution of the array, because the number of averaged points increases when keeping the neighborhood sizes the same (averaging window sizes). Recently, Agilent and NimbleGen have developed high-resolution microarrays, with roughly 1 million or 2 million features respectively, in which the oligonucleotide features are chosen to represent a tiling path through the human genome.

In parallel development, genotyping arrays from Illumina and Affymetrix originally designed for genome-wide genotyping of SNPs (Di *et al.*, 2005; Kennedy *et al.*, 2003) were adapted to provide CNV calls providing both polymorphic and non-polymorphic elements (Redon *et al.*, 2006). The most recent platforms contain features for >1 million SNPs, and translate to a resolution of one feature per ∼3000 bps. For each SNP, these platforms provide an intensity-based measurement for each allele, denoted generally as A and B. While the ratio of these intensities (A:B) is used for genotyping, the sum (A+B) provides a quantitative measure of copy number dosage. If a reference sample is available, then we can form a ratio as in array-CGH and use the same smoothing techniques (Hupe *et al.*, 2004; Lai *et al.*, 2005). The Affymetrix 6.0 platform includes 906 600 SNP probes as well as 945 806 non-polymorphic probes used strictly for CNV detection to roughly double the resolution. In their latest product, the Affymetrix 2.7M array includes many more CNV features (2 394 920 compared to 400 103 SNP features). Algorithms developed by Illumina and Affymetrix make clever use of the bi-allelic information from SNP probes to detect copy number change. For example, Illumina defines the B-allele frequency (BAF) as the estimated number of B alleles divided by the sum of both alleles at a given SNP location. This estimate is based upon interpolation from plotted canonical clusters of the AA, AB and BB genotypes at a given SNP, with the x-axis representing B to A allelic intensity ratios and the y-axis representing B + A intensity sums (Staaf *et al.*, 2008; Steemers and Gunderson, 2007). The expectation, when copy number dosage is 2, is to see three clouds of points when plotting BAF across the genome: one cloud for each genotype AA, AB and BB. CNVs can be detected by looking for regions with an unexpected number of clusters in the BAF measure even in cases where the dosage measure does not show strong signal (Fig. 1b shows intensity (A+B); Fig. 1c shows BAF).

In this work, we compare performance on the six platforms listed in Table 1. This set consists of one SNP genotyping, three array-CGH, and two hybrid array platforms, i.e. platforms in which a substantial number of probes are non-polymorphic. To guide the comparison we developed a spike-in experiment in which known genomic fragments were added in varying quantities to diploid lymphoblastoid genomic DNA samples from two individuals with known large deletions. Real-time quantitative PCR (qPCR) was used to estimate the concentration of bacterial artificial clones (BACs) containing between 141 and 217 kb of human genomic DNA. Four spike-in cocktails were prepared according to a modified Latin square design. Actual relative representation of BAC insert in each cocktail was determined as read number per base in single-end sequence output from a Solexa Genome Analyzer. The four spike-in reagents were added to lymphoblastoid genomic DNA, and each DNA sample was hybridized in replicate to each of the six platforms (Table 2, Methods in Supplementary Material). Sample labeling, hybridization, and scanning were performed at publicly accessible laboratory service sites with recognized expertise, and all protocols conformed to company-specified recommendations. Data were analyzed in two modes: one based on company-recommended software for detection of CNV calls and one independent of company-recommended software to evaluate the specificity and sensitivity of CNV and dosage measurements. We also evaluated the fidelity of the start and end estimates of CNV regions detected in each platform.

Table 1. General properties of six examined platforms

Official name (abbreviation)[a]	Array type	Lab Information[b]	Software[c]	Number of probes	Number of regions[d]	Distance[e]
Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix 6.0)	Hybrid	Center for Inherited Disease Research	Genotyping Console 3.0	1 852 406	981	1516
Affymetrix Cytogenetics-Whole-Genome 2.7M Array (Affymetrix 2.7M)	Hybrid	Gene Array Core of the Johns Hopkins Malaria Research Institute	Chromosome Analysis Suite	2 795 023	515	1021
Agilent SurePrint G3 Human CGH 1x1M (Early Access) (Agilent 1M)	Array CGH	JHMI Microarray Core of Johns Hopkins School of Medicine	DNA Analytics Software	967 029	1496	2825
Illumina Human1M-Duo BeadChip (Illumina 1M Duo)	SNP genotyping	Center for Inherited Disease Research	Bead Studio CNV partition or PennCV	1 153 974	1351	2455
NimbleGen CGH 2.1M (NimbleGen 2.1M)	Array-CGH	NimbleGen	NimbleScan v2.4	2 161 679	481	1314
NimbleGen CGH 3x720K (NimbleGen 3x720K)	Array-CGH	NimbleGen	NimbleScan v2.4	719 690	3957	3780

[a] Official name from company website with abbreviated name in parentheses.

[b] Location of hybridization, labeling and scanning.

[c] Company-recommended software.

[d] Probes were divided into regions such that the largest gap within these was <25 kb.

[e] Mean distance between probes within regions.

2 METHODS

To facilitate raw data comparisons across platforms, we created a log-ratio statistic, in $\log_2$ scale, proportional to the copy number dosage. For consistency with other microarray comparison publications (Irizarry *et al*., 2005) we denoted the statistic with M. A value of $M=0$ was associated with a copy number 2 and for every dosage doubling/halving; M was expected to increase/decrease by one. We obtained an M-value for each sample on each array feature. We also ran the default CNV detection algorithms provided by the array vendors. To support in-depth evaluation of platform performance, we distinguished between the feature-level measurements provided by our M statistic and the lists of genomic segments with associated dosage estimates provided by the default algorithms.

These studies were performed with approval of the Johns Hopkins Institutional Review Board and with informed consent of the families from whom DNA was obtained.

2.1 Experimental design

The overall experimental design is summarized in Table 2 with details in Methods in Supplementary Material. Briefly, two human genomic DNAs received spike-in mixes containing BAC clones in a modified Latin Square configuration. The experimental design also included technical replicates, i.e. independent labeling and array hybridizations of the same preparation of genomic DNA containing a spike-in panel.

Table 2. A modified Latin square design was used to generate four human genomic samples (rows, Tubes 1–4) with known change at specific loci

Spike-in mixture tube	Chromosome	Start	End	1928 female 22q		1133 male 21q	
				1	2	3	4
205J11	chr1	114314371	114488274	3	4	8	2
241C3	chr4	113064719	113225602	2.5	3	4	8
760L22	chr4	150206489	150389519	3	4	8	2
174B4	chr9	76502429	76697246	4	8	2	2.5
371D21	chr10	51551730	51706743	8	2	2.5	3
702N13	chr11	90945439	91112354	4	8	2	2.5
148N19	chr12	90464294	90631281	8	2	2.5	3
281G19	chr13	31804434	32001249	2	2.5	3	4
1149F2	chr15	45396506	45537976	2	2.5	3	4
43A6	chr21	17336776	17489124	2	3	2	2.5
886F4	chr21	27423603	27571602	3	2	2.5	2
NA	chr21	41229212	46909417	2	2	1	1
NA	chr22	43690910	45137025	1	1	2	2
293B4	chr22	45137026	45331620	1	1.5	2	2.5
NA	chr22	45331621	45807686	1	1	2	2
766H19	chr22	45807687	46018513	2	1	2.5	2
NA	chr22	46018514	47394477	1	1	2	2
957L9	chr22	47394478	47611023	1.5	2	3	2
NA	chr22	47611024	49691432	1	1	2	2

In this study, 11 loci modified by addition of BAC clones (indicated by Roswell Park accession number, top row) were analyzed in regards to TP versus coverage detection as well as actual versus estimated start and end of spiked-in regions. These are the first 11 BAC clones listed in this table, from top. All BAC clones listed in this table were analyzed to determine dosage estimates. Values indicate nominal final copy number at each position after mixture with the cell line DNAs indicated. 'NA' in the accession number row indicates deletion loci detected using fluorescence in situ hybridization (FISH). BAC clones spiked-in over the deletion loci, which are the last three BAC clones listed in this table, were analyzed in regards to TP versus coverage detection of deletion regions.

2.2 Preparation of DNA samples

Lymphoblastoid cell lines obtained from anonymized individuals were chosen for the presence of large copy number aberrations characterized by methods other than microarray hybridization (Pevsner,J., unpublished). Cell line 1133 was from a male with a hemizygous deletion on Chromosome 21. Cell line 1928 was from a female with a hemizygous deletion on Chromosome 22 as well as an amplification on Chromosome 6p. Bacterial stocks containing clones from the human male BAC library RPCI-11 were purchased from the Roswell Park Cancer Institute. DNA was isolated by standard methods (Qiagen Inc., Chatsworth, CA), and purity was assured by the presence of BamHI digest fragments at equimolar representation, and by unambiguous sequence reads from the BAC ends using T7 and SP6 primers. DNA concentrations were determined by spectrophotometer at A_{260}, and by real-time qPCR using a universal primer pair that amplified a vector segment. The qPCR measurements were used to adjust each BAC concentration to achieve the same number of molecules per microliter.

Four mixtures of BAC DNAs were assembled for addition to genomic DNA in Tubes 1–4 (Methods in Supplementary Material). Within each BAC mix, the relative representation of four different BAC DNAs was determined by qPCR based on primer pairs that recognize sites in human genomic DNA and that have similar reaction efficiencies. Then, the BAC mixes were added to genomic DNA and qPCR was again used to check the relative representation of four BAC locations within each genomic DNA sample.

2.3 Second-generation BAC sequencing

For each of the four BAC mixtures, the DNA sequence was obtained using Solexa/Illumina 1G (Illumina Inc., San Diego, CA) at the Johns Hopkins Genetics Core Resources Facility. For this, a library was made for each spike-in mix using the Illumina genomic DNA sample preparation kit according to instructions.

2.4 Values used in accuracy assessment

In the Section 3, we plot observed versus expected dosage estimate. For the observed values we calculated the average M-value across all points within the spiked-in region. For the expected values we used second generation BAC sequencing. For two of the BAC mixtures, DNA sequence was obtained using Solexa/Illumina 1G (Illumina Inc., San Diego, CA) at the Johns Hopkins Genetics Core Resources Facility. For this, a single-end library was made for the spike-in mixes using the Illumina genomic DNA sample prep kit according to manufacturer's instructions. For two BAC mixtures, we obtained single lane yields of 145 658 and 110 542 kb (with read lengths of 36 nucleotides) and average fold depth of coverage of 39.5× and 29.6×. The reads from each sample were mapped to the spiked-in regions on the hg18 build using Bowtie (Langmead *et al*., 2009) with default parameters. Note that under default parameters, if a read had multiple matches, one was chosen at random. We then computed the reads per kilobase per million (RPKM) for each region. We added an offset to RPKM values, for plotting purposes, to account for lack of genomic DNA in the sequenced BAC mixtures (explained further in Methods in Supplementary Material). For the results presented in Section 3.1 we included only autosomal data.

2.5 Microarray data acquisition

We submitted DNA samples to core facilities that are accessible to the biomedical research community. In all cases, the laboratories running the arrays were not provided access to the information in Table 2, and performed the experiments in a blind fashion. A reference DNA was recommended for two-color analyses on NimbleGen and Agilent arrays, and male DNA from Promega (catalog #G147A) was used. For each platform, a company-recommended analysis was performed to produce lists of detected CNV regions. For Affymetrix, Agilent and Illumina platforms, array data, analysis parameters and results were shared with the company for assent that data quality was acceptable and analysis parameters were within recommendation. NimbleGen performed analysis on site as part of their microarray service. An independent analysis was performed in parallel using the following steps.

2.6 Preprocessing

For the array-CGH arrays, NimbleGen and Agilent, we used loess normalization (Yang *et al*., 2002) via LIMMA: Linear Models for Microarray Data (Smyth, 2005). The SNP and hybrid arrays were preprocessed using CRLMM: the Corrected Robust Linear Model with Maximum Likelihood Distance (Carvalho *et al*., 2007; Ritchie *et al*., 2009) to obtain allele A and B intensity summaries. For the SNP features we considered the sum (I = A + B) as a dosage estimate. To obtain M-values for the Affymetrix 6.0 and Illumina 1M Duo platforms we used a library of reference arrays, obtained from the microarray core, to create a pseudo-reference array. This library included 75 individuals for Affymetrix 6.0 and 39 individuals for Illumina 1M Duo, taken from HapMap samples. The Illumina 1M Duo library samples were run the same month as the spike-in samples and the Affymetrix 6.0 library samples were run in a range of 1–4 months prior to the spike-in samples. For each feature we obtained the median, across arrays, of the I = A + B values. We then took the $\log_2$ ratio of the test sample versus the pseudo-reference value. For the Affymetrix 2.7M platform, the manufacturer processed the data in a similar way, using their own reference samples and removed 653 155 probes known to be problematic.

2.7 Wave effect removal

For removal of the wave effect, described in detail in Section 3.1, features on each platform were divided into groups so that the largest gap between any two probes was <25 kb. For each of these groups we then fitted a curve to the M-values plotted across the genome using loess (Cleveland, 1979) with a neighborhood span of 1 Mb. The estimated curves were then subtracted from M to create wave-corrected M-values. For CNV detection, we created lists of regions based on our own pre-processed data by applying CBS, with default parameters, to wave-corrected M-values. The mean M-value in each of the detected regions was used as an estimate of percent dosage increase (in $\log_2$ scale).

2.8 CNV detection sensitivity and specificity

As some of the company-recommended algorithms did not include procedures to detect CNVs in the X and Y chromosomes, we removed these spiked-in BACs from this analysis. Furthermore, we focused on the regions known to have amplifications because every algorithm easily found all, or nearly all, deletions. We combined the results from all eight samples, which resulted in a total of 64 true positive (TP) regions. All platforms contained probes in all spiked-in regions and probe density was sufficient on all platforms to detect spike-ins for all regions on at least some of the eight samples.

3 RESULTS

3.1 Waves

The presence of large-scale technical artifacts, referred to as waves, has been reported for array-CGH data (Marioni *et al*., 2007). These authors report that removing this artifact is critical for the development of a novel model-based CNV calling algorithm. Others report the same observation for high-resolution arrays (van de Wiel *et al*., 2009) and SNP genotyping arrays (Diskin *et al*., 2008). However, except the software for Affymetrix 2.7, we found that company-recommended default algorithms did not include data analytic steps to remove waves (Fig. 2a). We estimated and removed the wave effects using a loess smoother (Marioni *et al*., 2007) as described in detail in the Section 2. The estimated waves curve for Agilent 1M had the largest amplitude followed by the NimbleGen 2.1M, NimbleGen 3x720K and Affymetrix 6.0 arrays (Fig. 2b). The wave effect was substantially smaller for the Illumina 1M Duo and Affymetrix 2.7M platforms; apart from Illumina 1M Duo and Affymetrix 2.7M, we also found high correlation between the estimated wave effects in all platforms and genomic GC-content (Fig. 2b and c, Table 3). An interesting finding was that, in general, Illumina data did not show wave effects except for one sample (Tube 2, replicate B) for which the correlation with GC-content was 0.79. For the feature-level analyses presented in this section, we removed the estimated wave effects for all platforms. Note that for the Agilent arrays the effect was higher in one set of replicates (Fig. 2b).

3.2 Accuracy

The spike-in experiment permitted an assessment of accuracy; an assessment not currently found in the literature. We used second generation sequence data to support a precise measurement of BAC elements within the spike-in reagents as described in the Section 2. Therefore, each spiked-in region had an expected relative dosage within each sample. We plotted the observed versus expected log (base 2) dosages and noticed the expected linear pattern for all platforms (Fig. 3). Note that because doubling of expected dosage should result in doubling of the observed dosage, the slope of these lines should be one. We refer to the observed slopes as amplification dosage sensitivity. This dosage sensitivity greatly varied between platforms, although the best performing platforms achieved less than three-quarters the desired slope of one. Note that this discrepancy between true and measured dosage is common in microarrays and is explained by the Langmuir adsorption model as applied to probe saturation (Hekstra *et al*., 2003). Measuring a

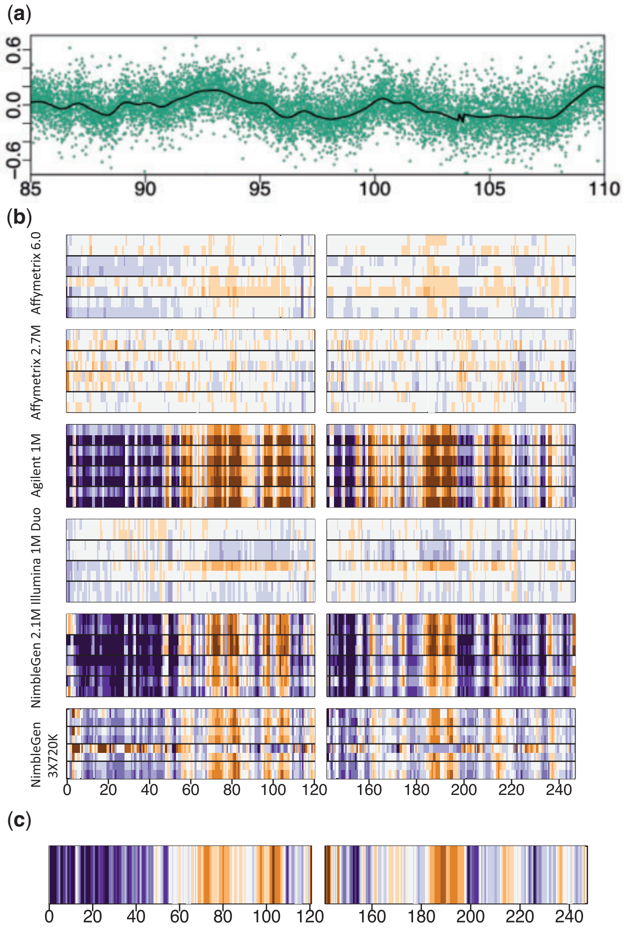

Fig. 2. Appearance of the wave artifact in many array platforms. (**a**) Microarray raw log-ratios (M-values) were plotted against chromosome locations for a 25 Mb stretch on Chromosome 1 for NimbleGen 2.1M. The black curve represents the fitted wave curve. (**b**). For the entire Chromosome 1, the estimated wave curve for each array on each platform is shown in color (blue represents peaks, red represents valleys). (**c**) For the entire Chromosome 1, genomic GC content is shown (blue represents peaks, red represents valleys). We computed the percent of G or C nucleotides in moving windows of size 1 Mb from human build hg18.

95% confidence interval around our regressed lines, the predicted ranges of NimbleGen 2.1M, NimbleGen 3x720K and Agilent 1M all overlapped one another at most observation values (Fig. 3 and Table 3). These three substantially outperformed the other platforms with Illumina 1M Duo performing worst.

We also assessed sensitivity with respect to deletions. For this we simply compared the pre-wave-corrected M-values for regions expected to have copy number 2 to regions with known deletions and therefore copy number 1 (Fig. 4). We refer to the difference in the median of these two sets of values as the indel sensitivity. Here the Illumina platform also underperformed. All other platforms showed similar results with NimbleGen slightly outperforming.

3.3 Probe density

We measured per region dosage sensitivity across the eight samples on a given platform to see whether probe density affected platform accuracy (Supplementary Fig. S1). Although both probe density and dosage sensitivity in spiked-in regions ranged greatly across and within platforms, we did not see a correlation between them.

Table 3. Summarized results

Platform	Wave-GC correlation	Dosage sensitivity (± range for 95% confidence interval of slope)	Indel sensitivity	Replicate variability	Null region variability (wave removed)	Fidelity median
Affymetrix 6.0	0.59	0.46 (± 0.03)	0.56	0.29	0.24 (0.04)	1748
Affymetrix 2.7M	0.01	0.33 (± 0.05)	0.53	0.27	0.18 (0.08)	2771
Agilent 1M	0.83	0.72 (± 0.06)	0.54	0.23	0.22 (0.05)	2467
Illumina 1M Duo	0.00	0.20 (± 0.03)	0.39	0.16	0.15 (0.04)	2332;3.11; 2979
NimbleGen 2.1M	0.74	0.68 (± 0.05)	0.67	0.25	0.22 (0.04)	700
NimbleGen 3x720K	0.67	0.65 (± 0.07)	0.59	0.24	0.18 (0.05)	3825

The first fidelity mean listed for Illumina 1M Duo corresponds with results from Bead Studio CNV partition and the second with results from PennCV.

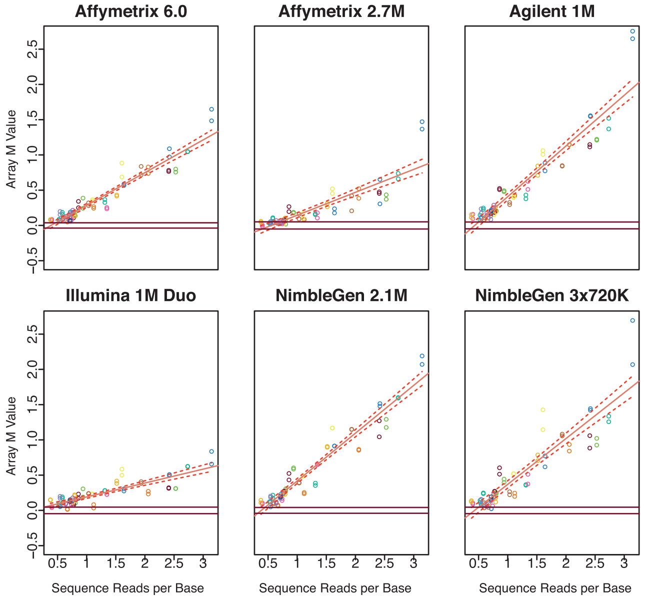

Fig. 3. Dosage sensitivity plot comparing microarray intensity to nominal copy number determined by sequence. For each platform we computed the wave-corrected, average log-ratio (M-values) within each spiked-in region. For these same regions we obtained reads per base from sequence data reflecting the relative representation of each spike-in component. Here we plot the microarray dosage estimates against the log (base 2), genomic DNA corrected, sequence dosage estimates. Data from all eight arrays are shown. Each color represents a region present at varying concentration on different arrays. Solid lines are the fitted regression and dashed lines are the 95% confidence intervals. The horizontal lines represent the typical range, one median absolute deviation (a statistic similar to SD but robust to outliers) in each direction from zero, of non-spiked-in regions.

3.4 Precision

We obtained our first measure of precision comparing the prewave-corrected M-values between technical replicates. As each sample was hybridized in replicate on each platform, we had two vectors of

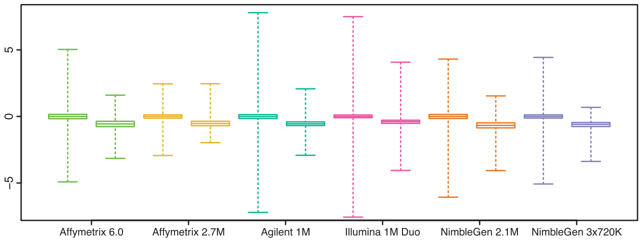

Fig. 4. Boxplots of pre-wave-removed M-values for regions expected to have copy number 2 (left) and copy number 1 (right) for each platform (represented by colors).

M-values for each sample that, in principle, should be the same. The SD of these differences serves as a measure of specificity that can be interpreted as the typical copy number percentage change in two arrays when none should be seen. We refer to this summary as the replicate variability. All platforms were comparable in this measure (Supplementary Fig. S2 and Table 3) although Illumina 1M Duo slightly outperformed the others. We obtained the second measure of precision by considering the variability among regions with known copy number 2. Although the vast majority of the non-spiked-in regions without previously reported deletions were expected to have copy number 2, we needed to allow for the possibility of some unknown CNVs being present. We therefore used an estimate of SD of the measurements in these regions that was robust to outliers: the median absolute deviation (MAD). We refer to this summary as the null region variability. This measure was interpreted as the observed percent change in dosage when none is expected. All platforms were comparable in this measure as well, although only after removing the wave effects (Supplementary Fig. 3a compared to Fig. 3b before removing wave effects and Table 3).

3.5 CNV detection sensitivity and specificity

The results presented above relate to general characteristics of the feature-level data. We also studied the downstream results based on the CNV detection algorithms provided by the vendors. For each platform, we used company-recommended software with recommended parameters to generate a list of CNV calls. Commercially available software support generally provides flexibility to users for the purpose of generating alternative results lists. We used reasonable recommended parameters for CNV detection under conditions where both software operators and companies were blind to spike-in positions, event sizes and dosages. No attempt was made to optimize detection of our BAC spike-in locations. Each vendor reviewed the software parameters that we used, and a list of regions predicted to be CNV along with the reported dosage estimates. With the exception of Affymetrix 6.0, each vendor's software attached a level of statistical significance to each reported region. For each level of statistical significance we selected the regions above that level and counted the number of spiked-in regions that overlapped them. This number was denoted as TPs. Any amount of overlap sufficed to qualify as a TP but 75% of regions had >50% agreement. In Section 3.5 we evaluate the resolution of these overlaps. We also recorded the sum of lengths of all non-spiked-in regions detected at each confidence level and denoted this as coverage. To avoid penalizing technologies finding real CNVs not associated with our spike-in experiment, we discarded regions found in the Database of Genomic Variants (Iafrate *et al.*, 2004). Regions cataloged in this database are likely TPs and should therefore not be counted as false positives. We also performed an alternative analysis in which we left out regions that do not have probes in at least two array platforms. Undocumented real CNVs in these regions could be especially unfairly detrimental to the one platform having probes within such regions. That one platform would be singly penalized for reporting a real CNV since it would be counted as a false positive and no other platform would have the opportunity to encounter it.

We plotted TP versus coverage for each platform preferring algorithms that achieved high TP counts without spending coverage. We analyzed amplifications (Supplementary Fig. S4) and deletions (Supplementary Fig. S5) separately. These results, which used

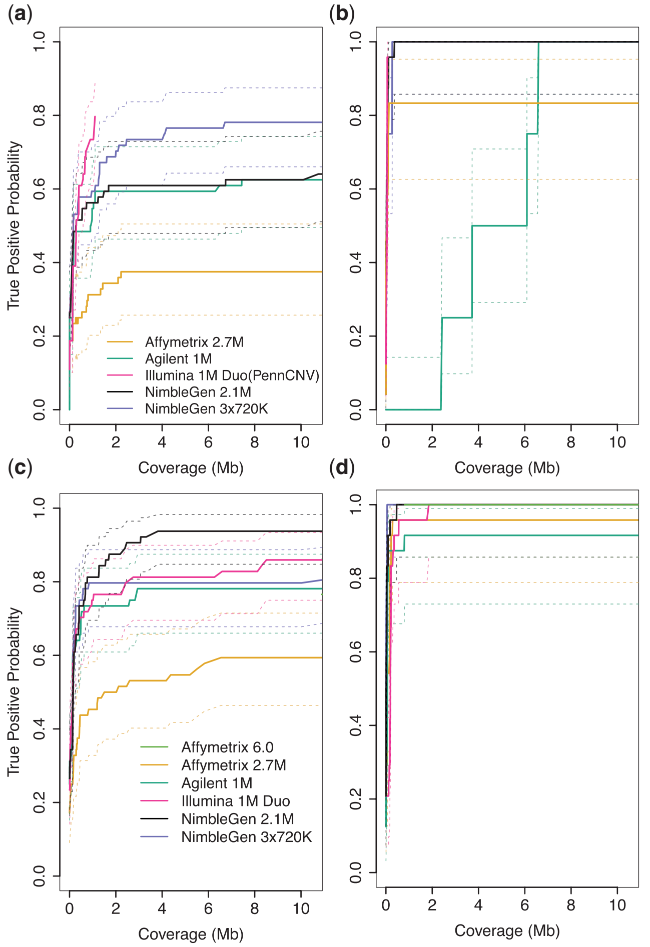

Fig. 5. TP versus coverage plots. Here we plot TP versus coverage for the data obtained from the company default algorithms as well as our alternative approach: removing waves and applying CBS in the case of amplifications, and in the case of deletions applying CBS without wave correction. We did this analysis for amplifications and deletions separately and then obtained a single curve (solid line) and 95% confidence intervals (dashed lines) by applying a binomial test upon detected TPs (explained further in Methods in Supplementary Material). The plots are from (**a**) company algorithms for amplifications, (**b**) company algorithms for deletions, (**c**) alternative analysis for amplifications and (**d**) alternative analysis for deletion.

Table 4. Binomial probabilities, with 95% confidence intervals in parentheses, for TPs detected using company-recommended algorithms, corresponding to three coverage values

Platform	Coverage					
	Company algorithms for amplifications			Company algorithms for deletions		
	0.1 Mb (lower, upper)	0.5 Mb (lower, upper)	1 Mb (lower, upper)	0.1 Mb (lower, upper)	0.5 Mb (lower, upper)	1 Mb (lower, upper)
Affymetrix 2.7M	0.19 (0.1, 0.3)	0.25 (0.15, 0.37)	0.31 (0.2, 0.44)	0.71 (0.49, 0.87)	0.83 (0.63, 0.95)	0.83 (0.63, 0.95)
Agilent 1M	0.25 (0.15, 0.37)	0.48 (0.36, 0.61)	0.55 (0.42, 0.67)	0 (0, 0.14)	0 (0, 0.14)	0 (0, 0.14)
Illumina 1M Duo	0.19 (0.1, 0.3)	0.61 (0.48, 0.73)	0.73 (0.61, 0.84)	0.96 (0.79, 1)	0.96 (0.79, 1)	0.96 (0.79, 1)
NimbleGen 2.1M	0.44 (0.31, 0.57)	0.52 (0.39, 0.64)	0.56 (0.43, 0.69)	0.88 (0.68, 0.97)	1 (0.86, 1)	1 (0.86, 1)
NimbleGen 3x720K	0.31 (0.2, 0.44)	0.58 (0.45, 0.7)	0.59 (0.46, 0.71)	0.75 (0.53, 0.9)	1 (0.86, 1)	1 (0.86, 1)

Data graphed in Figure 5a and b. For Affymetrix 6.0 there is only one TP count with corresponding coverage. The binomial probability and lower/upper confidence bounds are 0.61 and 0.48/0.73 for amplifications, corresponding to a coverage value of 4.6 Mb, and for deletions 1.0 and 0.86/1.0 corresponding to a coverage value of 365 kb.

Table 5. Binomial probabilities, with 95% confidence intervals in parentheses, for TPs detected using alternative analysis, corresponding to three different coverage values

Platform	Coverage					
	Alternative analysis for amplifications			Alternative analysis for deletions		
	0.1 Mb (lower, upper)	0.5 Mb (lower, upper)	1 Mb (lower, upper)	0.1 Mb (lower, upper)	0.5 Mb (lower, upper)	1 Mb (lower, upper)
Affymetrix 6.0	0.36 (0.24, 0.49)	0.69 (0.56, 0.8)	0.77 (0.64, 0.86)	0.04 (0, 0.21)	1 (0.86, 1)	1 (0.86, 1)
Affymetrix 2.7M	0.25 (0.15, 0.37)	0.44 (0.31, 0.57)	0.45 (0.33, 0.58)	0.54 (0.33, 0.74)	0.96 (0.79, 1)	0.96 (0.79, 1)
Agilent 1M	0.38 (0.26, 0.5)	0.7 (0.58, 0.81)	0.73 (0.61, 0.84)	0.83 (0.63, 0.95)	0.88 (0.68, 0.97)	0.92 (0.73, 0.99)
Illumina 1M Duo	0.44 (0.31, 0.57)	0.7 (0.58, 0.81)	0.75 (0.63, 0.85)	0.21 (0.07, 0.42)	0.92 (0.73, 0.99)	0.96 (0.79, 1)
NimbleGen 2.1M	0.34 (0.23, 0.47)	0.73 (0.61, 0.84)	0.81 (0.7, 0.9)	0.92 (0.73, 0.99)	1 (0.86, 1)	1 (0.86, 1)
NimbleGen 3x720K	0.34 (0.23, 0.47)	0.75 (0.63, 0.85)	0.8 (0.68, 0.89)	1 (0.86, 1)	1 (0.86, 1)	1 (0.86, 1)

Data graphed in Figure 5c and d.

company-recommended analyses, demonstrate that Illumina 1M Duo is the clear leader.

In comparing platforms using company-recommended guidelines, reasonable parameters used in analysis on the various platforms were quite different from one another. Thus, software attributes and array performance were seriously confounded, and a direct comparison of array performance was still needed. Moreover, we wanted to know how platforms compared using data where measurement artifacts from waves had been removed. For this reason, and for the purpose of controlling other pre-processing steps with demonstrated room for improvement (Dunning *et al*., 2008), we created an alternative data analysis pipeline. For amplification detection, our pipeline applied CBS to the wave-removed M-value data and for deletion detection our pipeline applied CBS to pre-wave-corrected M-data as the deletion regions were larger than the amplification regions and wave removal was not necessary for detection. As before, we plotted TP versus coverage for each platform for amplifications (Supplementary Fig. S6) and deletions (Supplementary Fig. S7) separately. To examine the range of performance, we applied a binomial test to combined curves from all arrays and provided an estimated curve along with point-wise confidence intervals, again separating out amplifications (Fig. 5a and b, and Table 4) and deletions (Fig. 5c and d, and Table 5). The results of this pipeline outperformed the results of our second alternative analysis, in which we discarded regions not having probes in at least two platforms (Supplementary Figs S8 and S9), and which appeared similar to the company-recommended analysis. This may be attributed to the fact that many more hundreds of megabases of genomic space were discarded when accounting for documented CNVs than when accounting for regions not covered by at least two platforms. Overall, performance was greatly improved with use of our alternative data analysis pipeline. Specifically, the use of this analytical approach substantially improved the performance of Agilent 1M, NimbleGen 2.1M, NimbleGen 3x720K, Affymetrix 6.0 and Affymetrix 2.7M. Under these conditions, performance was comparable across platforms with NimbleGen 2.1M slightly outperforming the rest and Affymetrix 2.7M performing as well for deletions (Fig. 5).

3.6 CNV region resolution

The molecular details of probe and sample constitution diverge substantially across the platforms we compare. A common assumption is that ‘per probe’ performance is similar on different

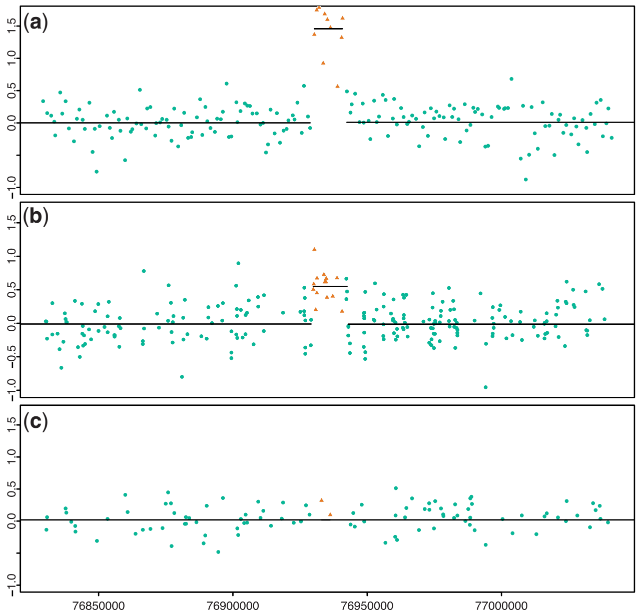

Fig. 6. Small CNV detection. (**a**) Microarray raw log-ratios (M-values) plotted against chromosome locations for the NimbleGen 2.1M Platform. The red points represent a region called as a small CNV by two technologies. (**b**) The same region for Affymetrix 6.0. (**c**) The same region for Illumina 1M Duo.

platforms, and that a larger probe number should provide CNV definition of increased resolution. We tested this assumption across platforms. For the spiked-in regions that were detected, we compared the true start and end of each BAC to the CNV start and end estimates provided by the vendor-recommended algorithms. We calculated the distance between true and estimated positions, and summarized these in a boxplot (Supplementary Fig. S10). The higher resolution arrays, NimbleGen 2.1M and Affymetrix 2.7M, clearly outperformed Illumina 1M Duo, NimbleGen 3x720K, Agilent 1M and Affymetrix 6.0 in this measure. Moreover, we examined non-spiked-in regions that were detected by more than one platform, and we noticed dozens of regions that were detected by NimbleGen 2.1M and just one other array platform, with Affymetrix 6.0 being the second platform most often, but closely followed by Agilent 1M and NimbleGen 3x720K. Close inspection revealed that the lower resolution of the other platforms allowed fewer, and sometimes zero, points in these regions (Fig. 6). As they were detected by multiple platforms and across replicates we believe that these are real CNVs. These two observations strongly support the common assumption of greater resolution from denser coverage, even across platforms.

4 DISCUSSION

We have performed a thorough comparison of six CNV detection platforms: Agilent 1M, NimbleGen 2.1M, NimbleGen 3x720K, Illumina 1M Duo, Affymetrix 6.0, and Affymetrix 2.7M. We used a carefully designed spike-in experiment to facilitate the data analysis. We found that removal of strong wave artifacts greatly improved analysis, in particular for Agilent 1M, NimbleGen 2.1M, and Affymetrix 2.7M arrays. Precision from replicates was similar across all platforms. Greater probe density does result in higher breakpoint resolution among platforms tested here, even though diverse technologies are used for sample labeling and signal detection.

Our analysis demonstrated that, when using company-recommended software, the Illumina 1M Duo and Affymetrix 6.0 platforms performed well at detecting CNVs while not performing well in dosage sensitivity. This presented an apparent paradox. It is explained by the use of the BAF measure, which is quite sensitive to copy number changes in cases where the dosage-related measures perform poorly (e.g. Fig. 1). We found various examples where dosage measures do not detect CNVs but BAF measures do (Supplementary Fig. S11 includes four examples). However, the use of SNP allele frequency to detect copy number alterations introduces two obstacles to CNV analysis. First, the accuracy of dosage estimates is below that of the CGH arrays from NimbleGen and Agilent. Secondly, the density of naturally occurring SNPs limits genotyping array density. This, in turn, hinders the detection ability of Illumina 1M Duo and Affymetrix 6.0 for small CNVs (less than ~50–100 kb) which arrays of high density, and NimbleGen 2.1M in particular, are able to detect.

After consideration of dosage sensitivity, precision, specificity, sensitivity and CNV border definition, the NimbleGen 2.1M platform demonstrated the best overall performance. However, careful removal of the wave artifact is a necessity. For researchers who depend on current company-recommended analysis algorithms, the Illumina 1M Duo platform performed best as it is able to detect CNVs with high sensitivity and specificity if they are >100 kb. However, in addition to its insensitivity to small CNVs, this platform underperforms at estimation of dosage and CNV start/end location.

We expect new high-throughput CNV detection platforms to emerge and new technologies, such as analysis of second generation sequence data, to become practical in the near future. Our study design and statistical analysis strategies are not specific to the platforms addressed here, and this approach can be readily applied to incorporate comparisons of newer technologies as they enter the research and clinical marketplace. Technical assessments comparing extant and emergent methods can support informed platform choice for a given research or clinical purpose, provide context for cross-platform results comparison where needed, and sharpen focus on the relative technical limitations and strengths that drive improvements in competing technologies.

ACKNOWLEDGEMENTS

We thank the Johns Hopkins Fund for Medical Discovery for seed funding, and in-kind donations of materials or labor from Agilent, Affymetrix, NimbleGen, JHMI Microarray Facility, Johns Hopkins Malaria Institute and Center for Inherited Research, Johns Hopkins Genetics Core Resources Facility.

Funding: F.S. was supported by R01GM077456, R.I. by R01GM083084 and R01RR021967, E.H.-S. by T32GM974906, J.P. by HD24061 and I.R. by UL1 RR 025005. All grants were received from the National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Carter,N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39**, S16–S21.

Carvalho,B. *et al*. (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, **8**, 485–499.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.

Colella,S. *et al*. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*, **35**, 2013–2025.

Conrad,D.F. *et al*. (2009) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

Di,X. *et al*. (2005) Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics*, **21**, 1958–1963.

Diskin,S.J. *et al*. (2009) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.*, **36**, e126.

Dunning,M.J. *et al*. (2008) Statistical issues in the analysis of Illumina data. *BMC Bioinformatics*, **9**, 85.

Gilad,Y. *et al*. (2009) Characterizing natural variation using next-generation sequencing technologies. *Trends Genet.*, **25**, 463–471.

Greenman,C.D. *et al*. (2010) PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, **11**, 164–175.

Hekstra,D. *et al*. (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res.*, **31**, 1962–1968.

Hupe,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Hurles,M.E. *et al*. (2008) The functional impact of structural variation in humans. *Trends Genet.*, **24**, 238–245.

Iafrate,A.J. *et al*. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.,* **36**, 949–951.

Irizarry,R.A. *et al*. (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, **2**, 345–350.

Kennedy,G.C. *et al*. (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol.*, **21**, 1233–1237.

Korn,J.M. *et al*. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, **40**, 1253–1260.

Lai,W.R. *et al*. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lee,J.A. and Lupski,J.R. (2006) Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron*, **52**, 103–121.

Lucito,R. *et al*. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.*, **13**, 2291–2305.

Marioni,J.C. *et al*. (2007) Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol.*, **8**, R228.

McCarroll,S.A. and Altshuler,D.M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, S37–S42.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pinkel,D. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Pollack,J.R. *et al*. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, **23**, 41–46.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Ritchie,M.E. *et al*. (2009) R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. *Bioinformatics*, **25**, 2621–2623.

Scharpf,R.B. *et al*. (2008) Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *Ann. Appl. Stat.*, **2**, 687–713.

Scherer,S.W. *et al*. (2007) Challenges and standards in integrating surveys of structural variation. *Nat. Genet.*, **39**, S7–S15.

Smyth,G.K. and Speed,T.P. (2003) Normalization of cDNA microarray data. *Methods*, **31**, 265–273.

Staaf,J. *et al*. (2008) Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics*, **9**, 409.

Steemers,F.J. and Gunderson,K.L. (2007) Whole genome genotyping technologies on the BeadArray platform. *Biotechnol. J.*, **2**, 41–49.

van de Wiel,M.A. *et al*. (2009) Smoothing waves in array CGH tumor profiles. *Bioinformatics*, **25**, 1099–1104.

Wang,K. *et al*. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.

Yang,Y.H. *et al*. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.