Published online 17 June 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 13* ***e80***
doi:10.1093/nar/gkn378

Impact of whole genome amplification on analysis of copy number variants

T. J. Pugh[1], A. D. Delaney[1], N. Farnoud[1], S. Flibotte[1], M. Griffith[1], H. I. Li[1], H. Qian[1], P. Farinha[2], R. D. Gascoyne[2] and M. A. Marra[1],*

[1]Genome Sciences Centre and [2]Department of Pathology, BC Cancer Agency, Vancouver, BC, Canada

Received November 28, 2007; Revised May 27, 2008; Accepted May 28, 2008

ABSTRACT

Large-scale copy number variants (CNVs) have recently been recognized to play a role in human genome variation and disease. Approaches for analysis of CNVs in small samples such as microdissected tissues can be confounded by limited amounts of material. To facilitate analyses of such samples, whole genome amplification (WGA) techniques were developed. In this study, we explored the impact of Phi29 multiple-strand displacement amplification on detection of CNVs using oligonucleotide arrays. We extracted DNA from fresh frozen lymph node samples and used this for amplification and analysis on the Affymetrix Mapping 500k SNP array platform. We demonstrated that the WGA procedure introduces hundreds of potentially confounding CNV artifacts that can obscure detection of bona fide variants. Our analysis indicates that many artifacts are reproducible, and may correlate with proximity to chromosome ends and GC content. Pair-wise comparison of amplified products considerably reduced the number of apparent artifacts and partially restored the ability to detect real CNVs. Our results suggest WGA material may be appropriate for copy number analysis when amplified samples are compared to similarly amplified samples and that only the CNVs with the greatest significance values detected by such comparisons are likely to be representative of the unamplified samples.

INTRODUCTION

Initial analysis of the human genome identified single nucleotide polymorphisms (SNPs) as the primary source of genotypic and phenotypic variation among humans. However, subsequent studies identified large-scale copy number variants (CNV) that apparently impacted millions of nucleotides (1–6). These large-scale variants included polymorphic deletions and duplications that are present in >1% of the population and therefore meet the traditional definition of polymorphism (2). As of November 2007, 4878 CNV loci impacting 808 Mbp of DNA sequence have been identified and these are listed in the Database for Genomic Variants (http://projects.tcag.ca/variation/). CNVs are also features of several human diseases including Alzheimer disease (7), Cri du chat syndrome (8), mental retardation (9) and cancer (10,11). As robust array-based methods for copy number detection continue to mature, increasing numbers of these variants are being identified (2).

Current whole-genome methods to detect CNVs require relatively large input quantities of DNA that are difficult or impossible to obtain from rare cell populations such as biopsies and microdissected tissues. To address this challenge, whole genome amplification (WGA) techniques were developed that increase the amount of DNA for analysis. For example, multiple-strand displacement amplification (MDA) using Phi29 DNA polymerase was used to generate microgram quantities of high molecular weight DNA (>30 kb) from nanograms of high quality input material (12,13). A recent report described a protocol for amplification of picogram quantities of DNA from single cells (14), further expanding the applications for this technique.

The replication fidelity of WGA techniques have been investigated (15–20). Estimates of base-pair incorporation errors resulting from Phi29-mediated amplification have ranged from 2.2×10^{-5} (21) to 9.5×10^{-6} (16) and the concordance of genotypes between unamplified and amplified samples were reported to be >99.8% (16,19). Recurrent WGA-induced copy number biases were observed in previous studies (15–20), and were associated with sequence repeats and proximity to chromosome ends (17–20), increased GC content (17,20), and annotated CNVs (17). Many of these associations were explored descriptively without statistical analysis and there was no consensus on the 92 recurrent regions of bias explicitly defined by three of these studies (16,17,20). A recent study of 532 samples subjected to WGA and subsequent analysis using the Affymetrix 10k Mapping array identified a median of 438 WGA-induced copy number artifacts in

*To whom correspondence should be addressed. Tel: +1 604 877 6082; Fax: +1 604 876 3561; Email: mmarra@bcgsc.ca comparisons between amplified samples and an unamplified reference set (15). While there is a consensus that at least partial compensation of systematic biases can be achieved through the use of an amplified reference (16–20), it is unknown to what degree such comparisons can capture real CNVs detected using more sensitive, higher resolution platforms.

Recently, bias induced by a number of whole genome amplification protocols was examined using a high-throughput, massively parallel whole genome pyrosequencing technique (22). In this comparison, which involved sequencing two bacterial genomes, Phi29 MDA-based approaches generated the most complete genome coverage (50–99%), and introduced the least bias compared to other PCR-based techniques. DNA sequences generated from Phi29-amplified material were 2.9–3.8% lower in GC-content than those from the unamplified material, suggesting a relationship between amplification bias and GC-content. However, over-amplification of certain sequences could not be explained by any of the previously mentioned sources of bias suggesting a need to directly investigate the nature of regions prone to over- or under-amplification. Although the study was of high resolution, direct comparison of the results from this study with those using human samples is difficult due to differences in chromosome organization, size and composition.

In this study, we investigated amplification bias resulting from whole genome amplification on DNA from fresh-frozen human tissues using the Affymetrix 500k Mapping Array Set. We quantified the effects of WGA on microarray signal and background noise, localized and statistically analysed genomic regions of WGA-induced bias, and directly compared the ability to resolve CNVs in comparisons of unamplified and amplified material.

MATERIALS AND METHODS

Tissue material and DNA extraction

Normal lymph nodes from three individuals were fresh frozen in Optimal Cutting Temperature (OCT; Sakura Finetek, Torrance, CA) compound and stored at −80°C by the service pathology laboratory at the BC Cancer Agency. Genomic DNA was extracted from these sources using the Gentra PureGene DNA purification kit (Gentra Systems, Minneapolis, MN). Prior to labelling and microarray hybridization, the genomic DNA was quantified using a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE). Prior to whole genome amplification, the genomic DNA was diluted to ~1.5 ng/µl and quantified using a PicoGreen assay (Invitrogen, Carlsbad, CA). To ensure consistent DNA quality across all samples, the DNA was visualized on an agarose gel to confirm the presence of undegraded, predominantly high molecular weight (>10 kb) DNA.

Whole genome amplification

We used Qiagen's Repli-G Mini whole genome amplification kit and protocol (QIAgen, Valencia, CA) to amplify 7 ng of PicoGreen-quantified DNA from fresh frozen samples to generate >10 µg of high molecular weight DNA. We performed the isothermal amplification reaction in 1.5 ml microcentrifuge tubes incubated in a 30°C water bath for 18 h and inactivated the enzyme by incubating the tubes in a 65°C water bath for 3 min. The amplified products were purified and quantified as described in the previous section and the amplification products were visualized on a 0.8% agarose gel stained with SYBR Green (Invitrogen, Carlsbad, CA).

Labelling and hybridization to the Affymetrix 500k array

500 ng samples of DNA were processed following the instructions in the GeneChip Mapping 500K manual (Affymetrix, Santa Clara, CA). Briefly, 250 ng of DNA was digested using one of two restriction enzymes, Nsp I or Sty I, and ligated to Nsp I or Sty I adaptors. These adaptor-ligated fragments were amplified by PCR and the purified products quantified using a Bio-Tek PowerWave X spectrophotometer and the concentration normalized to 2 µg/µl. The normalized products were then fragmented and labelled as described in the manual. Samples were hybridized to the GeneChip Human Mapping 250K Nsp or Sty array in an Affymetrix Hybridization Oven 640. Washing and staining of the arrays were performed using an Affymetrix Fluidics Station 450. Images of the arrays were obtained using an Affymetrix GeneChip Scanner 3000.

Sample preparation for NimbleGen 385k CGH array

Samples of >2.5 µg of DNA were prepared following the instructions provided by NimbleGen Systems Inc. (NimbleGen Systems Inc, Madison, Wisconsin). Briefly, purified samples were concentrated to 250 ng/µl and analysed for quality on an agarose gel. Samples were then shipped on ice to NimbleGen for subsequent labelling and hybridization to the 385k Human Whole-Genome CGH array.

Genotype and copy number analysis

Genotype calls were derived from microarray images using the GTYPE v4.0 software program (Affymetrix, Santa Clara, CA). We detected CNVs in individual samples using comparisons to a common reference data set and comparisons between pre- and post-amplification sample pairs (Figure 1). These were performed using a software pipeline (Figure 1) that utilizes the Affymetrix Chromosome Copy Number Analysis Tool (CNAT) version 4.0 (Affymetrix, Santa Clara, CA) and an exhaustive *t*-score optimization algorithm.

To analyse sample pairs on the Affymetrix platform, we used CNAT to perform quantile normalization of probe intensities from the samples and calculated $\log_2$ intensity ratios for each probe set on the array. For unpaired analysis of individual samples against a common reference set, we used a set of average probe intensities from the reference set in place of the second sample. The reference set used for this purpose, referred to hereafter as the 'Affy48 reference set', was downloaded from the Affymetrix website (http://www.affymetrix.com/support/technical/sample_data/500k_data.affx) and consisted of

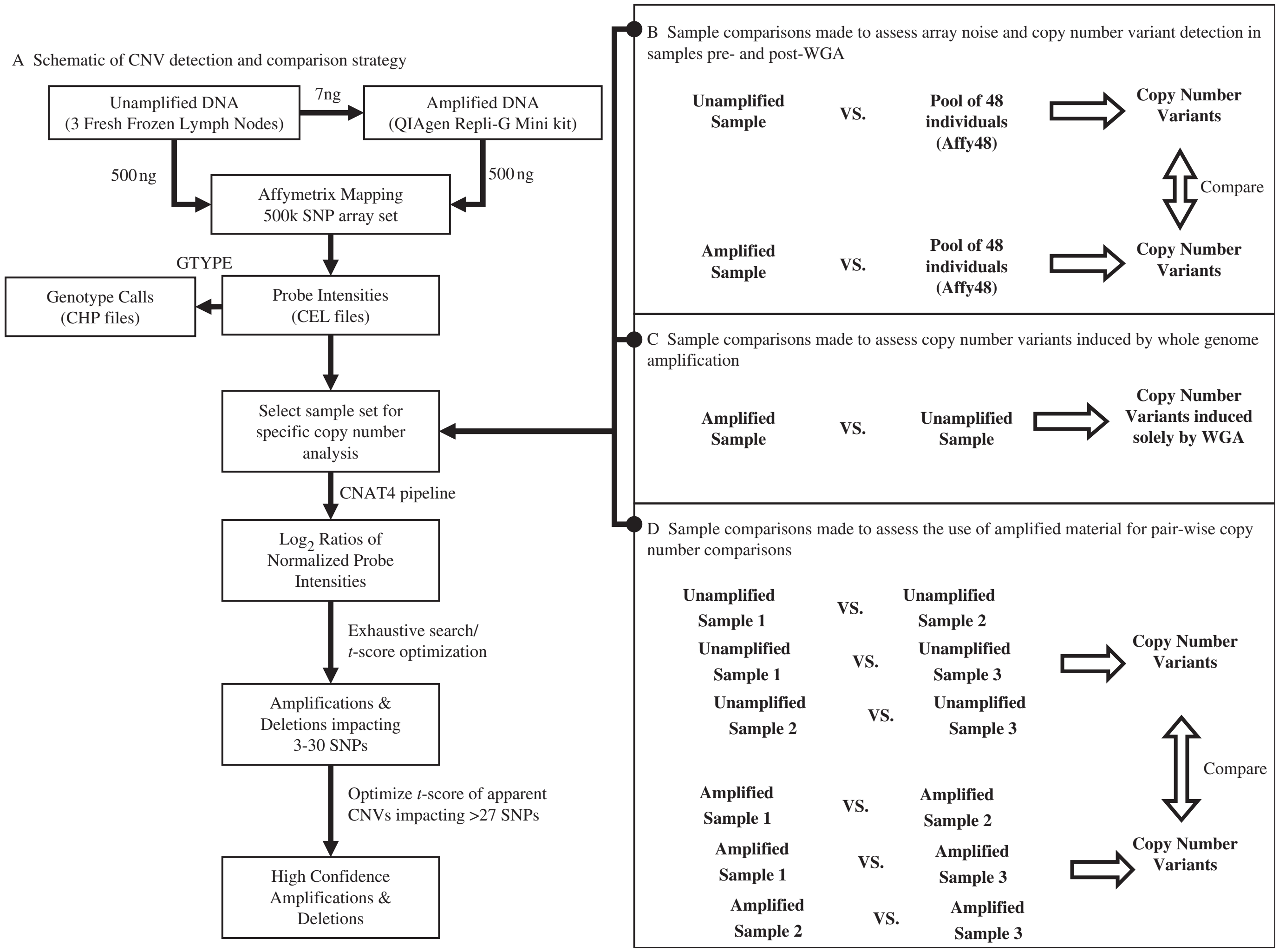
A Schematic of CNV detection and comparison strategy
Unamplified DNA (3 Fresh Frozen Lymph Nodes)
7ng
Amplified DNA (QIAgen Repli-G Mini kit)
500 ng
500 ng
Affymetrix Mapping 500k SNP array set
GTYPE
Genotype Calls (CHP files)
Probe Intensities (CEL files)
Select sample set for specific copy number analysis
CNAT4 pipeline
Log2 Ratios of Normalized Probe Intensities
Exhaustive search/ t-score optimization
Amplifications & Deletions impacting 3-30 SNPs
Optimize t-score of apparent CNVs impacting >27 SNPs
High Confidence Amplifications & Deletions
B Sample comparisons made to assess array noise and copy number variant detection in samples pre- and post-WGA
Unamplified Sample
VS.
Pool of 48 individuals (Affy48)
Copy Number Variants
Compare
Amplified Sample
VS.
Pool of 48 individuals (Affy48)
Copy Number Variants
C Sample comparisons made to assess copy number variants induced by whole genome amplification
Amplified Sample
VS.
Unamplified Sample
Copy Number Variants induced solely by WGA
D Sample comparisons made to assess the use of amplified material for pair-wise copy number comparisons
Unamplified Sample 1
VS.
Unamplified Sample 2
Unamplified Sample 1
VS.
Unamplified Sample 3
Unamplified Sample 2
VS.
Unamplified Sample 3
Copy Number Variants
Compare
Amplified Sample 1
VS.
Amplified Sample 2
Amplified Sample 1
VS.
Amplified Sample 3
Amplified Sample 2
VS.
Amplified Sample 3
Copy Number Variants

Figure 1. Experimental design. (**A**) In this study, we aimed to assess the impact of WGA on the detection of CNVs, to explore copy number biases induced by this technique, and to assess the use of pair-wise analysis to address such biases. To this end, DNA samples from three fresh frozen tissues were subject to WGA and analyzed pre- and post-amplification on the Affymetrix Mapping 500k SNP array set. For each copy number analysis, different sets of microarray data were compared as shown in panels B-D. Log_2 intensity ratios were calculated from the selected data comparisons using a software pipeline based on CNAT v4.0. These ratios were then screened by an 'exhaustive search' algorithm, in which *t*-scores were calculated in 3–30 probe windows and statistically significant aberrations identified above array-specific thresholds defined through permutation. To detect CNVs impacting more than 30 probes, aberrations found to contain more than 27 probes were subject to a *t*-score optimization using larger and larger window sizes until a local maximum *t*-score was found. The resulting high confidence lists of CNVs were then compared as appropriate for each analysis. (**B**) In this set of comparisons against a common reference set, we investigated the effect of WGA on array noise (i.e., the distribution of log_2 ratios) and the ability to resolve CNVs. To this end, each unamplified and amplified sample was independently compared against the Affy48 reference set, log_2 ratios calculated and detected CNVs were compared. (**C**) To assess the nature of bias induced by WGA, this data set directly compared matched pre- and post-WGA samples. Since matched samples were used, all CNVs detected in this analysis are due to the amplification technique. (**D**) This set of comparisons examined the ability of pair-wise analysis of amplified samples to reciprocate CNVs detected in unamplified samples. Three pair-wise comparisons were conducted using both unamplified and amplified material and the observed CNVs were compared.

48 samples representing five HapMap CEPH trios, five HapMap Yoruban trios, three other non-HapMap trios, and nine unrelated HapMap Asian samples. To analyse sample pairs on the NimbleGen platform, we used qspline normalized data and log_2 intensity ratios provided by NimbleGen for each probe on the array.

To identify significant deviations in the log_2 ratio data from both platforms, the following *t*-score optimization algorithm was used. First, log_2 ratios were sorted by genome coordinate and moving windows representing a number of adjacent probes were subjected to a *t*-test against the rest of the data outside of the window on the same chromosome. This was done across the entire genome for all window sizes from 3 to 30 probe sets for the Affymetrix and NimbleGen data. To establish a comparison-specific false-positive threshold, the order of log_2 ratios was then randomized and moving window *t*-tests were recalculated. Two *t*-score thresholds, one for amplifications and one for deletions, were then defined at which no amplifications or deletions were identified in the randomized data. These thresholds were then applied to the *t*-scores derived from the original data and regions with *t*-scores exceeding these thresholds were identified. To identify apparent variants impacting regions larger than our largest moving window size, *t*-scores were optimized for aberrations encompassing more than 27 probe sets using larger and larger windows until a local maximum *t*-score was found. As no CNVs met the false positive thresholds set for the NimbleGen data, a 50 probe window was used to detect statistically significant CNVs and a comparison-specific false positive threshold was not applied.

Sequence analysis of recurrent whole genome amplification-induced artifacts

In the analysis of recurrent WGA-induced artifacts, several sets of genomic coordinates were defined based on the human genome reference sequence Build 36/hg18 (released March, 2006) downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/). To define a set of regions that were consistently over- or under-amplified by the whole genome amplification technique, we analysed apparent variants arising from our comparison of matched pre- and post-WGA samples for overlapping genomic coordinates across all three comparisons and defined minimal overlapping regions (Supplementary Tables 1 and 2). These minimal overlapping regions were defined as the smallest region overlapped by a WGA-induced variant in all three comparisons. To define a subset of recurrently under-amplified chromosome ends, the first or last 2.5% of the reference genome sequence of any chromosome was recorded if it was impacted by a region consistently under-amplified by the WGA technique. To serve as reference sets representing the remainder of the human genome, random sets of coordinates were generated with equivalent size distributions for the regions consistently over- or under-amplified by the whole genome amplification technique and for the subset of recurrently biased regions affecting chromosome ends. In these reference sets, 10 random segments were generated with sizes corresponding to each entry in the list of regions affected by WGA-induced bias (i.e. 1900 amplifications and 750 deletions). The GC and repeat content of each entry in the above sets of coordinates were calculated in the following manner. For each set, the genomic sequence for each coordinate was downloaded from the Ensembl database (http://www.ensembl.org). To calculate the GC content of the sequence, the number of Gs and Cs in the sequence was counted and that number divided by the total length of the sequence. To calculate the repeat content of the sequence, the coordinates of the UCSC Genome Browser 'Simple Repeats' track generated by Tandem Repeats Finder (23) was used to identify base pairs belonging to repeat sequences. The number of these base pairs was then divided by the total length of the sequence to give the percentage of repeat sequence in the region. As most of the sets were not normally distributed in GC or repeat content as found by the Jarque-Bera test, the two-sample Kolmogorov-Smirnov test (KS test) was used to test whether these sets differed in their distribution of these two parameters.

RESULTS

Array noise and CNV in samples pre- and post-WGA

To establish a base line for array noise and CNV detection prior to amplification, each unamplified DNA sample was compared to the Affy48 reference set (Methods; Figure 1b) and candidate CNVs were identified. This comparison versus the Affy48 set was then repeated using amplified samples. As a measure of array noise, we quantified the distribution of $\log_2$ ratios resulting from these comparisons by calculating the mean, standard deviation (SD), and interquartile range (IQR) (Table 1, Figure 2). As expected due to normalization by CNAT4, the mean $\log_2$ ratios from both unamplified and amplified samples were very close to zero. The SDs and IQRs of $\log_2$ ratios from amplified samples were nearly twice those of the unamplified samples suggesting an increase in array noise using WGA material.

To compare the CNVs detected pre- and post-WGA, we counted apparent CNVs with p-values more significant than each comparison's false-positive detection limit (Table 1, Figure 3). The analysis of unamplified samples detected 13 candidate CNVs, 11 of which overlapped the coordinates of genomic variants listed in the Database of Genomic Variants (http://projects.tcag.ca) (5) (Table 2). In contrast, the analysis of the amplified samples identified 1572 apparent CNVs, an approximately 100-fold increase

Table 1. Distribution of $\log_2$ ratios from comparison of unamplified and amplified samples versus a common reference set of 48 individuals

Sample compared versus Affy48	Mean[a]	SD[b]	IQR[c]	Apparent amplifications		Apparent deletions	
				Count	$P<$	Count	$P<$
Sample 1 - Unamplified	0.0002517	0.3079	0.3428	2	1.99×10^{-8}	3	1.65×10^{-9}
Amplified	0.001971	0.3790	0.4793	322	9.76×10^{-7}	368	9.39×10^{-9}
Sample 2 - Unamplified	0.002710	0.2602	0.3152	2	3.70×10^{-7}	2	1.00×10^{-16}
Amplified	−0.0001297	0.4188	0.5412	254	8.91×10^{-7}	157	8.33×10^{-9}
Sample 3 - Unamplified	0.003530	0.2584	0.3176	3	5.42×10^{-10}	1	1.00×10^{-16}
Amplified	−0.0004284	0.4076	0.5178	295	7.45×10^{-7}	176	1.36×10^{-8}

[a] Mean value of $\log_2$ ratios resulting from each comparison. A site with with equivalent copy number in both samples would return a $\log_2$ ratio of 0.
[b] Standard deviation of $\log_2$ ratios resulting from from each comparison. These values are interpreted as a measure of data noise from each comparison.
[c] Interquartile range of $\log_2$ ratios resulting from from each comparison. These values are interpreted as a measure of data noise from each comparison.

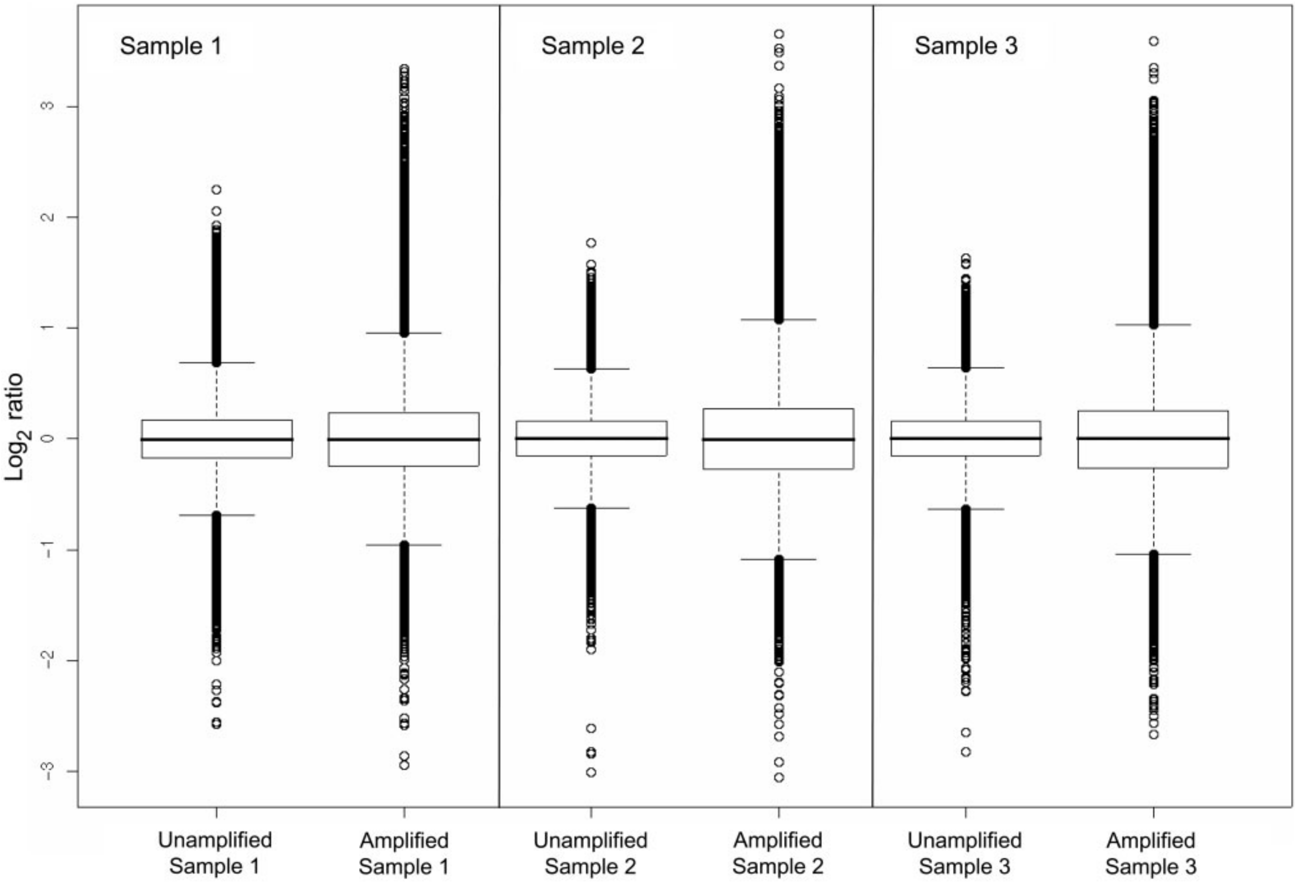

Figure 2. Boxplots comparing the spread of $\log_2$ ratios in unamplified and amplified samples. The $\log_2$ ratios resulting from comparison of each sample against the Affy48 reference set were plotted using a standard box and whisker plot displaying a five number summary: maximum value or Q3 + 1.5 × IQR, Q3, mean, Q1, and minimum value or Q1 – 1.5 × IQR. Outliers, defined as values that fall more than 1.5 × IQR above Q3 or below Q1, are displayed as individual data points. Due to normalization as part of the CNAT4 analysis pipeline, the mean $\log_2$ ratio from each sample is close to zero. However, the IQR, as well as the maximum and minimum values, were further from the mean in the amplified samples relative to the unamplified samples. The increased spread of data distribution is likely due to increased array noise and the detection of amplification biases induced by WGA.

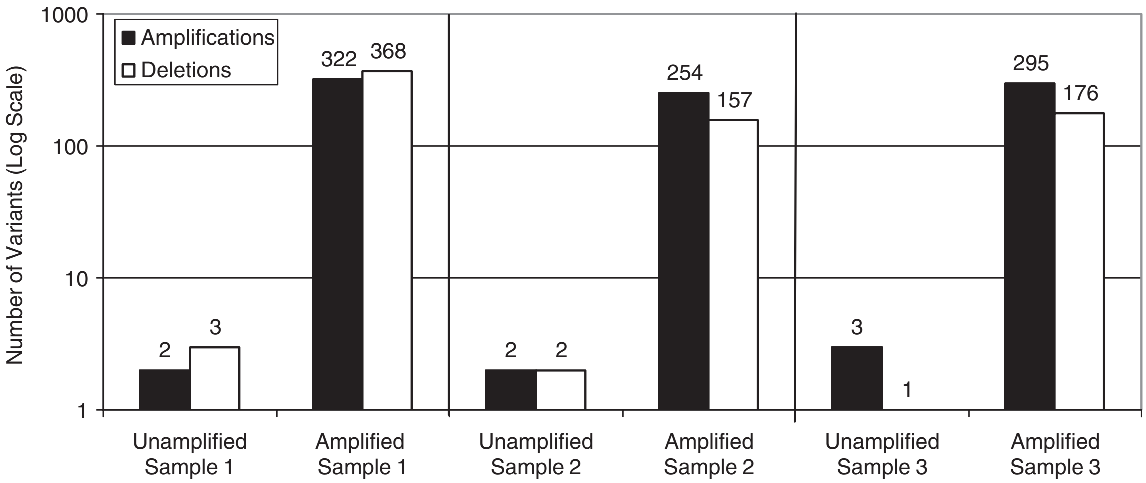

Figure 3. Apparent CNVs in unamplified and amplified samples. The number of variants detected in unamplified and amplified samples from comparison against the Affy48 reference set were counted. The amplified samples appear to contain hundreds of CNVs not seen in the unamplified samples suggesting that WGA over- and under-represents of specific regions of the genome.

Table 2. Apparent amplifications and deletions detected prior to amplification through comparison with a reference set of 48 individuals

Sample compared versus Affy48	Genome coordinates of variant (NCBI Build 36/hg18/Mar 2006)	Size (bp)	CN within variant	CN outside variant	SNP count	*P*-value	Variation locus[a]
Amplifications							
Sample 1	chr7:48424572–48431182	6610	2.88184	2.04848	11	1.99×10^{-8}	–
	chr14:19381928–19492423	110495	2.93812	2.03610	28	4.85×10^{-13}	Locus 2636
Sample 2	chr2:113809804–113849256	39452	2.28770	2.04023	12	3.70×10^{-7}	Locus 0397
	chr17:41569489–41709662	140173	3.07396	2.03694	41	2.31×10^{-12}	Locus 3029
Sample 3	chr9:29695281–29706655	11374	2.19958	2.04042	4	$<1.00 \times 10^{-16}$	–
	chr14:19309086–19459561	150475	2.65807	2.03481	25	5.42×10^{-10}	Locus 2639
	chr15:19163125–20077554	914429	2.66995	2.04165	72	$<1.00 \times 10^{-16}$	Locus 2748
Deletions							
Sample 1	chr7:142030227–142210594	180367	1.54593	2.04848	27	1.61×10^{-10}	Locus 1656
	chr14:21451264–22044096	592832	1.51299	2.03610	161	$<1.00 \times 10^{-16}$	Loci 2644 and 2645
	chr22:33661041–33725126	64085	1.75349	2.06794	21	1.65×10^{-9}	Locus 3489
Sample 2	chr2:50682535–50865587	183052	1.44974	2.04023	40	$<1.00 \times 10^{-16}$	Locus 0329
	chr14:21792331–22040096	247765	1.38419	2.02893	60	$<1.00 \times 10^{-16}$	Locus 2645
Sample 3	chr14:21800768–21932862	132094	1.53811	2.03481	32	$<1.00 \times 10^{-16}$	Locus 2645

[a] From the database of genomic variants (http://projects.tcag.ca/variation/).

Table 3. Distribution of $\log_2$ ratios from pair-wise comparison of experimental replicates of unamplified and amplified samples

Sample		Mean	SD	IQR
Sample 1 -	Unamplified	0.005517	0.2579	0.3223
	Amplified	0.002538	0.2840	0.3544
Sample 2 -	Unamplified	0.008175	0.2658	0.3299
	Amplified	0.0003263	0.3264	0.4153
Sample 3 -	Unamplified	0.0064235	0.2585	0.3187
	Amplified	0.001687	0.2842	0.3517

in the number of apparently significant amplifications and deletions versus the unamplified samples (Table 1). These artifactual CNVs are likely the result of WGA-induced biases.

To assess experimental variation prior to amplification, each unamplified and amplified sample was subjected to a pair-wise comparison against an experimental replicate of itself (Table 3). The lack of fluctuation in mean, SD and IQR in the $\log_2$ ratios from unamplified replicates suggests a high degree of reproducibility of the array method used. Similarly, while still elevated relative to unamplified samples, there is no major fluctuation in these values between amplified replicates further supporting the notion that the WGA method behaves consistently. However, the values obtained from unamplified samples versus values obtained from amplified samples, using the Affy48 reference set, showed a substantial decrease in SDs and IQRs. This indicates that amplified samples produce different signal intensity distributions than unamplified samples, suggesting that comparison of amplified to unamplified data sets is potentially problematic.

CNVs induced by whole genome amplification

To identify apparent CNVs arising from non-uniform amplification bias in the WGA technique, data from paired pre- and post-WGA samples were directly compared to each other (Figure 1b). Our analysis identified apparent WGA-induced over- and under-amplifications in each of the three comparisons of amplified versus unamplified material. In sample 1, we detected 502 amplifications (*P*-value threshold of detection, $P < 1.68 \times 10^{-6}$) and 580 deletions ($P < 1.71 \times 10^{-8}$). In sample 2, we detected 467 amplifications ($P < 1.68 \times 10^{-6}$) and 202 deletions ($P < 1.64 \times 10^{-8}$). In sample 3, we detected 546 amplifications ($P < 1.68 \times 10^{-6}$) and 259 deletions ($P < 3.45 \times 10^{-8}$). Our analysis also revealed a set of 265 recurrent apparent WGA-associated aberrations that were detected in all three comparisons. This set consisted of 190 over-amplifications (Supplementary Table 1) and 75 under-amplifications (Supplementary Table 2). 39 of these regions overlapped one of the 92 regions of bias (31 of 62 over-amplifications, 8 of 30 under-amplifications) identified by three previous studies (16,17,20). 110 of the regions we identified overlapped genomic regions with known CNVs (2) (64 over-amplifications, 46 under-amplifications) but there was no correlation between regions susceptible to WGA-associated bias and known CNVs ($P = 1.00$). In a set of 2650 random genomic coordinates with the same size distribution as the WGA-induced artifacts, 36.26% overlapped a known CNV, a proportion near the 41.51% overlap observed with the set of WGA-induced biases.

The minimal overlapping regions (see Methods) of WGA-induced over-amplifications ranged from 2207 bp to 357 399 bp with a median size of 58 961 bp, an IQR of 66 524 bp and encompassed 13.6 Mbp of the reference human genome sequence. These recurrently over-amplified sites were distributed throughout the genome and had a statistically significant increase in GC content relative to a set of 1900 random genomic segments with identical size distribution ($P = 8.36 \times 10^{-40}$). These over-amplified sites were also enriched for repeat sequences relative to the set of 1900 random genomic segments ($P = 1.76 \times 10^{-6}$). These results are compatible with the notion that over-amplification by the WGA technique is related to the GC and repeat content of the underlying sequence.

The minimal overlapping regions of the recurrent WGA-induced under-amplifications ranged from 5206 bp to 1.93 Mbp with a median size of 75 698 bp, an IQR of 64 619 and encompassed 8.37 Mb of the reference human genome sequence. These regions of under-amplification appeared to fall into two groups: those near chromosome ends and those distributed throughout the genome. Comparison of the 54 under-amplified sites distributed throughout the genome with a set of 540 random genomic segments with identical size distribution found no statistically significant difference in GC content ($P = 0.0796$) or repeat sequences ($P = 0.1901$). However, the under-amplifications were greatly depleted for GC-rich regions compared to the over-amplifications ($P = 1.93 \times 10^{-5}$) which supports the notion that WGA amplification efficiency is related to the GC content of the underlying sequence. A plot of GC content versus copy number shows a trend of increasing amplification magnitude (i.e. increasing copy number) with increasing GC content (Figure 4).

Of the 39 chromosome ends (see Methods) assayed by probe sets, 15 contained regions of under-amplification (Table 4). Only three chromosome ends contained over-amplifications, suggesting that under-representation of chromosome ends is a consistent result of whole genome amplification. The set of chromosome end under-amplifications impacted 2.547 Mbp of the reference human genome sequence and the GC content was statistically greater than that of a set of 150 random genomic segments with identical size distribution ($P = 1.12 \times 10^{-6}$). However, there was no statistical difference in GC content been the under-amplified chromosome ends and the 25 appropriately amplified chromosome ends ($P = 0.8215$). This suggests that amplification bias due to GC content does not play a role in under-amplification of specific sub-telomeric regions. Under-amplified chromosome ends were enriched for repetitive sequences (see Methods) relative to both a set of 150 random genomic segments with identical size distribution ($P = 1.52 \times 10^{-9}$) and the 25 assayed chromosome ends that were not under-amplified ($P = 0.0022$) suggesting that increased repeat content of specific chromosome ends may result in their under-amplification.

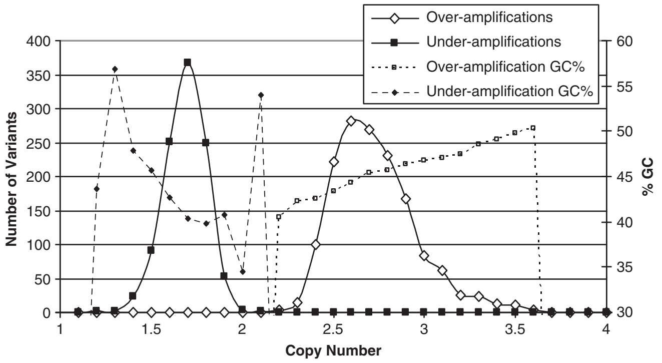

Figure 4. Copy number distribution and GC content of WGA-induced CNVs. The number of variants and percentage GC content were plotted against copy number magnitude for all of the CNVs detected by comparisons of each pre- and post-WGA sample pair. There appears to be a direct relationship between the magnitude of over-amplification and increased GC content.

To assess WGA-induced CNV artifacts using a second array platform, we compared pre- and post-amplification sample pairs in three comparative genome hybridization (CGH) experiments using the NimbleGen 385k array. The $\log_2$ ratios from these experiments were widely distributed (average SD = 0.378, average IQR = 0.457) and while several thousand CNVs were detected, none were identified with p-values passing the stringent false positive thresholds set by our algorithm due to the high level of noise in this data ($P < 3.51 \times 10^{-7}$ for over-amplifications, $P < 3.30 \times 10^{-11}$ for under-amplifications). Analysis of this data using a 50 probe moving window without filtering for false positives detected 2116 WGA-induced CNVs (466 over-amplifications, 1650 under-amplifications) of which 141 occurred in all three comparisons (29 over-amplifications, 112 under-amplifications). Despite their relatively large size (average = 1.06 Mb, median = 0.36 Mb, SD = 4.10 Mb), only 28 of these overlapped recurrent artifacts detected by the Affymetrix comparisons (17 of 190 over-amplifications, 11 of 75 under-amplifications). This amount of overlap is similar to that seen with a random set of 2116 random genomic coordinates with the same size distribution as the CNVs detected by the NimbleGen platform of which 65 overlapped a WGA-induced CNV detected by the Affymetrix platform. These results suggest that these are artifacts resulting from the difficulty in distinguishing real CNVs from background noise when co-hybridizing amplified and unamplified samples even when a large moving window of 50 probes is used.

Table 4. Regions of recurrent WGA under-amplification within chromosome ends

Genome coordinates (Build 36/hg18/Mar 2006)	Size (Mbp)	% GC content	Mbp from nearest chromosome end
P-terminal end			
chr1:3058506–3129776	0.071	57.113	3.059
chr1:5857077–5871605	0.015	57.168	5.857
chr2:554079–613259	0.059	45.934	0.554
chr2:1841469–1968296	0.127	45.876	1.841
chr5:487981–738504	0.251	56.251	0.488
chr5:2187888–2267721	0.080	49.395	2.188
chr5:2836714–2884070	0.047	41.89	2.837
chr5:3160861–3195828	0.035	46.205	3.161
chr8:791584–850907	0.059	47.539	0.792
chr8:1816651–1946694	0.130	49.183	1.817
chr10:2593122–2624375	0.031	37.102	2.593
chr19:373238–892603	0.519	59.541	0.373
q-terminal end			
chr6:170198708–170308225	0.110	51.929	0.592
chr7:158582043–158739710	0.158	45.905	0.082
chr10:134327710–134332916	0.005	49.165	1.042
chr12:130611957–130673802	0.062	51.924	1.676
chr13:112193014–112294946	0.102	42.808	1.848
chr13:113053814–113215730	0.162	50.548	0.927
chr15:99580062–99745948	0.166	47.27	0.593
chr16:87408466–87706274	0.298	59.068	1.121
chr20:60967459–61027216	0.060	49.085	1.409

Use of amplified material for pair-wise copy number comparisons

To assess the use of WGA material in pair-wise comparisons, each sample was compared to the other samples one-by-one and relative differences in copy number in the three samples assessed using: (i) unamplified samples versus unamplified samples, (ii) amplified samples versus unamplified samples, and (iii) amplified samples versus amplified samples (Figure 1d). An example of the output from one such set of comparisons is illustrated in Figure 5.

The unamplified versus unamplified comparisons identified 21 apparent differences in copy number among the three samples (Tables 5 and 6). These pair-wise comparisons identified 5 of 13 apparent differences expected from the individual comparisons of samples to the Affy48 reference set. Twelve of these apparent differences, including the five differences expected from comparison with the Affy48 set, overlap variants listed in the Database of Genomic Variants (http://projects.tcag.ca). The amplified versus unamplified comparisons identified 3207 apparent differences in copy number among the three samples (Table 5). Only seven of these apparent differences were detected by both unamplified/amplified and amplified/unamplified comparisons suggesting that systematic WGA-induced variants and random WGA-reaction variability mask real events.

The amplified versus amplified comparisons identified 275 apparent differences in copy number among the three samples (Table 5). These amplified versus amplified comparisons identified 2 of the 12 apparent amplifications and 5 of the 9 apparent deletions seen in the unamplified

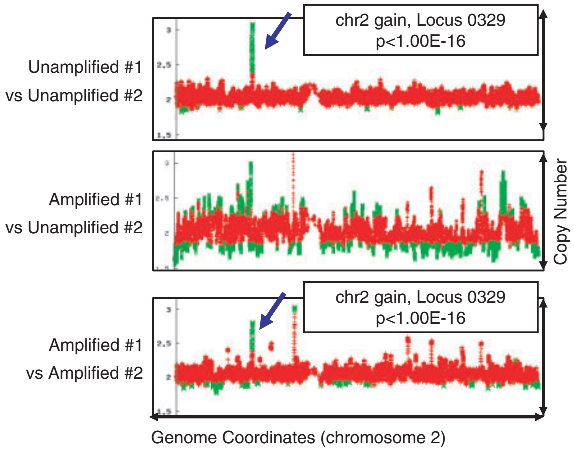

Figure 5. Example of how a pair-wise comparison of amplified material can partially compensate for WGA-induced bias. Shown is the output of three copy number analyses conducted using our CNV discovery software pipeline. Copy number, calculated directly from $\log_2$ ratios of probe intensities, is plotted against genome location using a sliding window of averaged data points, in this case 60 probes. In this example, a pair-wise comparison of two unamplified samples, identified a gain of copy number ($P < 1.00 \times 10^{-16}$) in unamplified sample #1 relative to unamplified sample #2 at a locus documented to be copy number variable in the Database of Genomic Variants. Conducting the same comparison after WGA of sample #1 results in hundreds of confounding CNVs from which the known CNV is indistinguishable. However, conducting this comparison after WGA of both samples restores the ability to detect this CNV. Artifactual variants do still remain as a result of random variation in the WGA process, however they do not reach the level of significance of the real event. Therefore, when interpreting results from comparisons of WGA samples, only the top-most hits are likely to be representative of the unamplified sample.

Table 5. Apparent copy number differences identified by pair-wise comparisons of all possible combinations of unamplified and amplified samples

Samples compared	Apparent amplifications		Apparent deletions		Total apparent CNVs	CNVs in common between matched comparisons
	Count	$P<$	Count	$P<$		
Unamplified sample 1 Unamplified sample 2	4	4.26×10^{-7}	3	1.40×10^{-8}	7	–
Unamplified sample 1 Unamplified sample 3	4	3.88×10^{-8}	4	1.05×10^{-13}	8	–
Unamplified sample 2 Unamplified sample 3	4	1.09×10^{-10}	2	3.44×10^{-15}	6	–
Amplified sample 1 Unamplified sample 2	369	1.26×10^{-6}	367	7.77×10^{-9}	736	2
Unamplified sample 1 Amplified sample 2	69	1.05×10^{-6}	358	7.04×10^{-9}	427	
Amplified sample 1 Unamplified sample 3	471	1.81×10^{-6}	498	1.28×10^{-8}	969	1
Unamplified sample 1 Amplified sample 3	110	1.60×10^{-6}	536	1.53×10^{-8}	646	
Amplified sample 2 Unamplified sample 3	183	1.07×10^{-6}	49	5.64×10^{-8}	232	4
Unamplified sample 2 Amplified sample 3	67	1.28×10^{-6}	130	3.31×10^{-8}	197	
Amplified sample 1 Amplified sample 2	21	2.03×10^{-6}	49	1.71×10^{-8}	70	–
Amplified sample 1 Amplified sample 3	18	9.67×10^{-7}	82	2.69×10^{-8}	100	–
Amplified sample 2 Amplified sample 3	44	1.82×10^{-6}	61	8.23×10^{-8}	105	–

Table 6. Copy number variants detected by pair-wise comparisons of unamplified and amplified sample sets

Sample comparison	Relative CN difference	Detected by pairwise comparison of unamplified samples			Detected by pairwise comparison of amplified samples			Variation locus[a]
		Coordinates (Build 36)	$P\leq$	Rank	Coordinates (Build 36)	$P\leq$	Rank	
1 versus 2	Increase	chr2:50775422–51014967	1.00×10^{-16}	1	chr2:50828689–50960764	1.15×10^{-9}	1 of 21	0329[b]
		chr14:19272965–19489991	1.38×10^{-10}	2	–			2636
		chr3:21942154–21975950	3.91×10^{-7}	3	–			–
		chr16:22640088–22688093	4.26×10^{-7}	4	–			2893
	Decrease	chr17:41569489–41708649	1.00×10^{-16}	1	chr17:41587072–41709662	1.00×10^{-16}	1 of 48	3029
		chr9:11936421–11997006	5.09×10^{-11}	2	–			1901
		chr10:95243220–95304377	1.40×10^{-8}	3	–			–
1 versus 3	Increase	chr8:124654695–124656225	1.00×10^{-16}	1	–			–
		chr13:43692360–43696382	3.99×10^{-13}	2	–			–
		chr18:20691186–20697540	4.86×10^{-13}	3	–			–
		chr14:19402695–19502641	3.88×10^{-8}	4	–			2636
	Decrease	chr14:21715523–22040167	1.00×10^{-16}	1	chr14:21531617–22057862	1.00×10^{-16}	1 of 82	2644/5
		chr10:54588936–54590136	1.00×10^{-16}	1	–			–
		chr17:76310141–76321112	1.00×10^{-16}	1	–			–
		chr15:19876834–20005562	1.05×10^{-13}	4	chr15:19877365–20077554	2.11×10^{-10}	37 of 82	2748
2 versus 3	Increase	chr17:41572099–41708649	1.00×10^{-16}	1	chr17:41522422–41647903	8.47×10^{-13}	1 of 44	3029
		chr15:84684853–84693981	1.00×10^{-16}	1	–			2830
		chr15:98087203–98095507	1.11×10^{-11}	3	–			2860
		chr16:77105899–77109454	1.09×10^{-10}	4	–			–
	Decrease	chr15:18711364–20079140	1.00×10^{-16}	1	chr15:19313868–20329239	1.00×10^{-16}	1 of 61	2748
		chr2:50870615–51020480	3.44×10^{-15}	2	chr2:50828689–51018056	1.00×10^{-16}	1 of 61	–

[a] From the database of genomic variants (http://projects.tcag.ca/variation/).

[b] This CNV locus is overlapped only by the coordinates expected from comparison versus the Affy48 reference set.

comparisons (Table 6), suggesting that pair-wise comparisons of material where both samples have been subjected to WGA can partially compensate for reproducible WGA-induced bias (Figure 5). The most significant deletion identified by each unamplified comparison was recapitulated as the most significant deletion identified by the corresponding amplified comparison (Table 6). This was also true of the most significant amplification in two of the three comparisons (Table 6). The list of variants detected at lower levels of significance than these top scoring events may still contain real CNVs although it is difficult to isolate these from the remaining artifactual events resulting from random experimental variation without independent validation of each one.

Validation of WGA pair-wise comparisons for copy number detection

To determine the extent to which amplified pair-wise comparisons mask known, validated CNVs, DNA from the blood of three father/child pairs with previously described CNVs (9) were subjected to WGA and copy number analysis using the 250k Nsp chip of the Affymetrix 500k set. The original analysis of unamplified DNA performed using the Affymetrix Mapping 100k SNP array set (9) identified a total of 32 CNVs within the three father/child pairs of which five (two amplifications, three deletions) were validated by conventional cytogenetic analysis (Table 7).

The amplified child versus amplified father comparisons identified 63 CNVs in copy number in total within the three pairs. Analysis of amplified family pair #8379 identified 41 copy number differences (13 relative amplifications $P < 3.48\times10^{-6}$, 28 relative deletions $P < 8.38\times10^{-8}$), analysis of amplified family pair #1280 identified six copy number differences (two relative amplifications $P < 2.14\times10^{-6}$, four relative deletions $P < 1.05\times10^{-8}$), and analysis of amplified family pair #3476 identified 16 copy number differences (six relative amplifications $P < 2.07\times10^{-6}$, 10 relative deletions $P < 6.09\times10^{-9}$). These copy number differences were then ranked by *P*-value (most significant to least) and the coordinates compared to those of the validated aberrations. The amplified versus amplified comparisons identified four of the five CNVs (two amplifications, two deletions) validated by FISH (9) and each received the lowest *P*-value for its comparison (Table 7). The single validated CNV that was not detected by the amplified comparisons may have been missed due to a difference in array coverage at this site. On the 250k Nsp array, this region was covered by three probe sets (10 683 bp/probe set) compared to six probe sets (5341 bp/probe set) on the 100k array. This was also the smallest feature of the set of validated CNVs (0.03 Mb) and may reflect a decrease in detection sensitivity when using amplified comparisons. Among the top-ranked variants (i.e. those with the most significant *P*-values), six variants were identified by the 250k WGA experiment that were not detected by the original experiments. Five of these are covered by six or fewer probe sets (5743–93 452 bp/probe set, one with no probes) on the 100k array. In addition to the possibility of an increased false positive rate due to increased array noise, differences in each array's probe coverage may explain why these regions were only detected by the experiment using amplified samples.

Table 7. Copy number variants detected in MR families by pair-wise comparisons of unamplified and amplified sample sets (child versus father)

Family ID[9]	Relative CN	Validated aberrations detected by pairwise comparison of unamplified samples [9] (100k array set)				Detected by pairwise comparison of amplified samples (250k Nsp array)			Variation locus[a]
	difference	Coordinates (Build 36)	Mbp	Validation	Cyto-band	Coordinates (Build 36)	$P\leq$	Rank[b]	
8379	Increase	chr10:259695–23144645	22.88	karyotyping	10p12.2–p15.3	chr10:1000464–24070263	1.00×10^{-16}	1 of 13	many
		chr15:19208413–19943075	0.73	karyotyping	15q11.2	chr15:18850150–20335459	1.00×10^{-16}	1 of 13	2748
		–	–	–	–	chr14:21394980–21864733	1.00×10^{-16}	1 of 13	many
1280	Increase	–	–	–	–	chr9:10069844–10104307	5.54×10^{-7}	1 of 2	–
		–	–	–	–	chr13:100974064–101034679	2.14×10^{-6}	2 of 2	–
	Decrease	chr4:22943293–23102259	0.16	FISH (BAC)	4p15.2	chr4:22828003–23025619	3.64×10^{-10}	1 of 4	0794
3476	Increase	–	–	–	–	chr5:64484426–64535538	1.00×10^{-16}	1 of 6	–
		–	–	–	–	chr20:50794691–50801972	1.00×10^{-16}	1 of 6	3405
	Decrease	chr1:83242288–83274337	0.03	FISH (fosmid)	1p31.1	–	–	–	0104
		chr4:82282746–85558739	3.28	FISH (BAC)	4q21.23	chr4:82531241–92371701	1.00×10^{-16}	1 of 10	many
		–	–	–	–	chr22:46869824–46963276	1.00×10^{-16}	1 of 10	–

[a] From the database of genomic variants (http://projects.tcag.ca/variation/).
[b] Ranked by significance (*P*-value). Only variants with the lowest *P*-value scores are shown.

Genotype fidelity

To compare the fidelity of genotype calls derived from WGA product to those from corresponding unamplified samples, data from matched pairs of these sources were compared. Average genotype call rates (±1 SD) were 96.74 ± 1.14% from the unamplified samples and 93.14 ± 2.68% from the WGA samples, suggesting a modest degree of information loss following amplification. Of the SNPs which were unsuccessfully called in the amplified samples, only 2% were common to all three samples and only one of these fell within a region of WGA-induced bias (an over-amplification). Genotype concordance was 98.57 ± 0.53% between calls successfully made from both amplified and unamplified samples in each matched pair. There was very little overlap in the coordinates of SNPs with non-concordant genotypes and regions of recurrent WGA-induced bias. Of the non-concordant calls, 58.77% were called heterozygotes in the unamplified sample and homozygotes in the amplified sample (i.e. AB called as AA or BB) and 0.2% of these were located in regions of WGA-induced over-amplification while none were in regions of WGA-induced under-amplification, 40.66% were called homozygotes in the unamplified sample and heterozygotes in the amplified sample (i.e. AA or BB called as AB) of which none were located in regions of WGA-induced bias, and 0.57% were incorrectly called homozygotes (i.e. AA called as BB or BB called as AA) of which none were located in regions of WGA-induced bias. Twelve regions each containing 3–7 SNPs were identified as displaying loss of heterozygosity (LOH) in total from the three pre- and post-amplification comparisons. Three of the LOH regions had an allele-specific copy number of 3 while the others had a copy number of 2. These regions impacted a total of 58 SNPs, 0.01% of all of the SNPs assayed, and none overlapped a region recurrently over- or under-amplified by WGA. These results suggest that increased random array noise is likely a greater source of genotype non-concordance than systematic allele-specific amplification bias or polymerase error.

DISCUSSION

The ability to discover CNVs in unamplified human DNA using data generated by the Affymetrix Mapping SNP array platform has been previously demonstrated by our group and others (1–3,9). However, with small amounts of DNA, from tumour biopsies for example, amplification of the starting material prior to discovery of CNVs is often necessary to generate enough material to conduct such analyses. We aimed to assess the nature of biases that are introduced by this amplification, and to determine their impact on copy number detection and whether pair-wise comparisons could compensate for these biases. For the first time, we have used a high resolution microarray platform to explicitly define regions susceptible to WGA-induced bias, statistically assessed the sequence features underlying these biases, and demonstrated an ability to correct for these biases and resolve real CNVs. In this study, three unamplified DNA samples were used to establish a base line for array noise and CNV detection. These were compared to the same DNA samples that were amplified in duplicate using a WGA technique. The apparent CNVs we detected by comparing unamplified samples to the unamplified Affy48 reference set were likely real events, as the variants were relatively large, statistically significant, and 11 of the 13 CNVs corresponded to previously documented genomic variants (5). While our variant detection approach adjusts its threshold of significance based on the level of noise of each array, comparisons using amplified samples still identified hundreds of apparent CNVs not seen in the unamplified comparisons on the Affymetrix array platform. Since these comparisons were performed against an unamplified reference, it is likely that these artifactual apparent CNVs were the result of preferentially amplifying of regions of the genome and not due to an increased level of array noise. The data from the NimbleGen platform appeared to have a high level of noise that affected our ability to detect WGA-induced CNVs when co-hybridizing unamplified and amplified samples. Our results suggest that amplified and unamplified samples cannot be directly compared to uncover WGA-induced artifacts using the NimbleGen CGH array. However, this should not preclude the comparison of similarly amplified samples on this platform as we have shown using Affymetrix arrays that the biases are largely systematic and the noise is reduced substantially when comparing two amplified samples.

To explore the nature of this bias, we directly compared Affymetrix data from pre- and post-amplification sample pairs and observed a set of regions apparently over- or under-amplified in all three samples. These regions impacted a total of 21.97 Mb of sequence, consisted of 190 over-amplifications and 75 under-amplifications, and overlapped 39 of 92 regions of WGA-induced bias identified by other studies (16,17,20). The low amount of overlap is perhaps due to differences in genome coverage by the arrays used in these studies, particularly as there was no previous consensus on any region being susceptible to WGA-induced bias. Results reported are for DNA amplified using the QIAgen Mini kit and it is conceivable that DNA amplified using different protocols will exhibit different bias. While the lack of a correlation between regions of WGA-induced bias and known CNVs is different from a previous observation (17), we have demonstrated that the degree of overlap of the amplification biases we identified with known CNVs is only slightly greater than would be expected by chance. The amount of overlap observed is likely due to the fact that documented CNVs are generally large, 165 kb on average, and, in total, impact ~27% of the genome.

The difference in size and size distribution of the over- and under-amplifications that we identified suggests focal over-amplification of specific sequences and broader under-representation of others. We observed a direct relationship between amplification efficiency and GC-content as over-amplified regions had a statistically significant increase in GC content relative to the deletions ($P = 1.93 \times 10^{-5}$) and the magnitude of over-amplification appeared to scale directly with GC richness (Figure 4). These results are consistent with the notion that WGA-induced over-amplification bias is related to the increased binding affinity of GC-rich hexamers relative to AT rich hexamers and not a shortage of hexamers corresponding to repetitive regions in the genome. There is also the possibility that, unlike many polymerases, Phi29 polymerase is more efficient in synthesizing GC-rich sequences, thereby resulting in over-amplification of these regions. These effects likely also contribute to under-amplification of GC-poor regions distributed throughout the genome but not likely the loss of chromosome ends. The lack of a relationship between regions of WGA-induced bias and the presence of known CNVs suggests that different mechanisms account for these phenomena.

The loss of chromosome ends appears to be a consistent result of the WGA procedure as 15 of the 39 ends assayed were under-amplified in all samples compared to only three that were over-amplified. Relative to chromosome ends that were not affected by bias, the under-amplified ends were enriched for repetitive sequences ($P = 0.0022$) but did not have a statistically significant difference in GC content ($P = 0.8215$). These results suggest that the source of amplification bias at chromosome ends is different from GC-content-derived biases affecting the rest of the genome. One possible explanation is the positional effect of having fewer overlapping amplification products at the ends of linear stands of DNA than in the middle. However, if this were the case then all chromosome ends should be similarly under-amplified which they are not. Another possible explanation is that the limited quantities of hexamers corresponding to subtelomeric repeats result in fewer priming events in these regions. This may account for the loss of repetitive chromosome ends more frequently than less repetitive ends.

We found that samples subject to Phi29-based WGA can be used for accurate genotyping, albeit with some data loss. From the WGA samples, we consistently observed a decrease in the average number of genotype calls and a wider range of call rates compared to those from the unamplified samples. However, of the genotype calls that were made, over 98% were concordant between amplified and unamplified sample pairs. The less than 2% non-concordant calls were 99.43% discrepant heterozygotes (i.e. AB called as AA or BB, AA or BB called as AB), rather than incorrectly called homozygotes, and nearly none (<0.12%) were located in regions of WGA-induced bias. This discrepancy rate is very near that observed between unamplified replicates on the Affymetrix 500k array (24). It is likely that the source of genotype call non-concordance is related to the genotyping accuracy of the array in the presence of increased noise due to WGA and not truly genotype changes induced by WGA through allele-specific amplification or polymerase error.

Regardless of the source of the systematic biases induced by WGA, we have shown that pair-wise analysis of amplified samples is a viable strategy for CNV detection, albeit with an appropriate threshold of significance to filter the number of low-significance random artifacts induced by this technique. While the greater number of apparent copy number differences detected using amplified samples has the potential to mask real events, we observed that pair-wise comparisons of such samples can detect real differences between samples. On comparing amplified samples to amplified samples, the number of artifactual copy number differences is reduced by an order of magnitude relative to comparisons of amplified versus unamplified samples due to the systematic nature of the bias induced by the technique. Conceivably, the use of a large, amplified reference set would be a practical alternative to pair-wise comparisons for larger batches of amplified samples requiring a universal reference. Of the apparent copy number differences detected by the three pair-wise comparisons using unamplified material, all of the top deletions and two of the three top amplifications were identified as the most significant by the corresponding comparisons using amplified material. By applying this technique to paired child/father samples with known, validated copy number differences (9), four of the five validated differences detected by the original study using unamplified DNA were the most significant in the same comparisons using amplified DNA. The only validated CNV that was missed using WGA material was due to a difference in coverage by the array platforms used. A similar difference in coverage partially explains the presence of six high confidence CNVs detected by the WGA experiments not seen in the original study as one of these has recently been observed in the unamplified material using a higher resolution platform. Therefore, when evaluating the results from amplified comparisons, CNVs with the top ranked significance are more likely to be real CNVs in the unamplified sample.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the expert technical assistance of Susanna Chan, Jennifer Asano and Adrian Aly of the Affymetrix Array group at the Genome Sciences Centre, BC Cancer Agency. Support for this work and funding to pay the Open Access publication charges for this article was received from the BC Cancer Foundation, the Canada Foundation for Innovation, and Hoffmann-La Roche Ltd. of Canada. TJP is a Senior Graduate Trainee of the Michael Smith Foundation for Health Research (MSFHR) and the BC Cancer Foundation and has been supported by fellowship grants from Eli Lilly and the University of British Columbia. MG is a Senior Graduate Trainee of the MSFHR and Genome BC and is supported by a fellowship from the Natural Sciences and Engineering Research Council. MAM is a senior scholar of the MSFHR and a Terry Fox Young Investigator.

Conflict of interest statement. None declared.

REFERENCES

1. McCarroll,S.A., Hadnott,T.N., Perry,G.H., Sabeti,P.C., Zody,M.C., Barrett,J.C., Dallaire,S., Gabriel,S.B., Lee,C., Daly,M.J. *et al.* (2006) Common deletion polymorphisms in the human genome. *Nat. Genet.*, **38**, 86–92.
2. Feuk,L., Carson,A.R. and Scherer,S.W. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.
3. Conrad,D.F., Andrews,T.D., Carter,N.P., Hurles,M.E. and Pritchard,J.K. (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 75–81.
4. Sharp,A.J., Locke,D.P., McGrath,S.D., Cheng,Z., Bailey,J.A., Vallente,R.U., Pertz,L.M., Clark,R.A., Schwartz,S., Segraves,R. *et al.* (2005) Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.*, **77**, 78–88.
5. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
6. Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Maner,S., Massa,H., Walker,M., Chi,M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
7. Rovelet-Lecrux,A., Hannequin,D., Raux,G., Le Meur,N., Laquerriere,A., Vital,A., Dumanchin,C., Feuillette,S., Brice,A., Vercelletto,M. *et al.* (2006) APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat. Genet.*, **38**, 24–26.
8. Zhang,X., Snijders,A., Segraves,R., Zhang,X., Niebuhr,A., Albertson,D., Yang,H., Gray,J., Niebuhr,E., Bolund,L. *et al.* (2005) High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative genomic hybridization. *Am. J. Hum. Genet.*, **76**, 312–326.
9. Friedman,J.M., Baross,A., Delaney,A.D., Ally,A., Arbour,L., Armstrong,L., Asano,J., Bailey,D.K., Barber,S., Birch,P. *et al.* (2006) Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation. *Am. J. Hum. Genet.*, **79**, 500–513.
10. Tonon,G., Wong,K.K., Maulik,G., Brennan,C., Feng,B., Zhang,Y., Khatry,D.B., Protopopov,A., You,M.J., Aguirre,A.J. *et al.* (2005) High-resolution genomic profiles of human lung cancer. *Proc. Natl Acad. Sci. USA*, **102**, 9625–9630.
11. Zhao,X., Li,C., Paez,J.G., Chin,K., Janne,P.A., Chen,T.H., Girard,L., Minna,J., Christiani,D., Leo,C. *et al.* (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.*, **64**, 3060–3071.
12. Hughes,S., Arneson,N., Done,S. and Squire,J. (2005) The use of whole genome amplification in the study of human disease. *Prog. Biophys. Mol. Biol.*, **88**, 173–189.
13. Dean,F.B., Hosono,S., Fang,L., Wu,X., Faruqi,A.F., Bray-Ward,P., Sun,Z., Zong,Q., Du,Y., Du,J. *et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA*, **99**, 5261–5266.
14. Spits,C., Le Caignec,C., De Rycke,M., Van Haute,L., Van Steirteghem,A., Liebaers,I. and Sermon,K. (2006) Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.*, **1**, 1965–1970.
15. Corneveaux,J.J., Kruer,M.C., Hu-Lince,D., Ramsey,K.E., Zismann,V.L., Stephan,D.A., Craig,D.W. and Huentelman,M.J. (2007) SNP-based chromosomal copy number ascertainment following multiple displacement whole-genome amplification. *Biotechniques*, **42**, 77–83.
16. Paez,J.G., Lin,L.M., Beroukhim,R., Lee,J.C., Zhao,X., Richter,D.J., Gabriel,S., Herman,P., Sasaki,H., Altshuler,D., Li,C. *et al.* (2004) Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.*, **32**, e71.
17. Arriola,E., Lambros,M.B., Jones,C., Dexter,T., Mackay,A., Tan,D.S., Tamber,N., Fenwick,K., Ashworth,A., Dowsett,M. *et al.* (2007) Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridisation. *Lab. Invest.*, **87**, 75–83.
18. Lage,J.M., Leamon,J.H., Pejovic,T., Hamann,S., Lacey,M., Dillon,D., Segraves,R., Vossbrinck,B., Gonzalez,A., Pinkel,D. *et al.* (2003) Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. *Genome Res.*, **13**, 294–307.
19. Tzvetkov,M.V., Becker,C., Kulle,B., Nurnberg,P., Brockmoller,J. and Wojnowski,L. (2005) Genome-wide single-nucleotide polymorphism arrays demonstrate high fidelity of multiple displacement-based whole-genome amplification. *Electrophoresis*, **26**, 710–715.
20. Bredel,M., Bredel,C., Juric,D., Kim,Y., Vogel,H., Harsh,G.R., Recht,L.D., Pollack,J.R. and Sikic,B.I. (2005) Amplification of whole tumor genomes and gene-by-gene mapping of genomic aberrations from limited sources of fresh-frozen and paraffin-embedded DNA. *J. Mol. Diagn.*, **7**, 171–182.
21. Esteban,J.A., Salas,M. and Blanco,L. (1993) Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J. Biol. Chem.*, **268**, 2719–2726.
22. Pinard,R., de Winter,A., Sarkis,G.J., Gerstein,M.B., Tartaro,K.R., Plant,R.N., Egholm,M., Rothberg,J.M. and Leamon,J.H. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, **7**, 216.
23. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
24. Affymetrix (2006) BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500K Array Set. *Technical Report*, White Paper. Santa Clara, CA: Affymetrix, Inc.