Published in final edited form as:

Cancer Epidemiol Biomarkers Prev. 2008 December ; 17(12): 3499–3508. doi:
10.1158/1055-9965.EPI-08-0482.

WHOLE GENOME AMPLIFICATION ENABLES ACCURATE GENOTYPING FOR MICROARRAY-BASED HIGH DENSITY SNP ARRAY

Farzana Jasmine¹, Habibul Ahsan^{1,2,3,4}, Irene L. Andrulis⁵, Esther M. John⁶, Jenny Chang-Claude⁷, and Muhammad G. Kibriya¹

¹Department of Health Studies, The University of Chicago, Chicago, IL 60637

²Department of Human Genetics, The University of Chicago, Chicago, IL 60637

³Department of Medicine, The University of Chicago, Chicago, IL 60637

⁴Department of Cancer Research Center, The University of Chicago, Chicago, IL 60637

⁵Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario M5G 1X5, Cancer Care Ontario, Department of Molecular Genetics, University of Toronto, Ontario

⁶Northern California Cancer Center, Fremont CA and Stanford School of Medicine, Department of Health Research and Policy, Stanford, CA

⁷German Cancer Research Center, Heidelberg, Germany

Abstract

In large scale genome-wide association studies based on high density single nucleotide polymorphism (SNP) genotyping array, the quantity and quality of available genomic DNA (gDNA) is a practical problem. We examined the feasibility of utilizing the Multiple Displacement Amplification (MDA) method of whole genome amplification (WGA) for such a platform. The Affymetrix Early Access Mendel Nsp 250K GeneChip was used for genotyping 224,940 SNPs/sample for 28 DNA samples. We compared the call concordance using 14 gDNA samples and their corresponding 14 WGA samples. The overall mean genotype call rates in gDNA and corresponding WGA samples were comparable [97.07% (95% CI 96.17–97.97) vs. 97.77% (95% CI 97.26–98.28), $p=0.154$]. Reproducibility of the platform, calculated as concordance in duplicate samples, was 99.45%. Overall genotypes for 97.74% (95% CI 97.03–98.44) of SNPs were concordant between gDNA and WGA samples. Restricting the analysis to well performing SNPs (successful genotyping in gDNA and WGA in >90% samples), on average, 99.11% (95% CI 98.80–99.42) of the SNPs were concordant and overall a SNP showed a discordant call in 0.92% (95% CI 0.90–0.94) of paired samples. In a pair of gDNA and WGA DNA, similar concordance was reproducible on Illumina's Infinium platform as well. Although copy number analysis revealed a total of seven small telomeric regions in six chromosomes with loss of copy number, the estimated genome representation was 99.29%. In conclusion, our study confirms that high density oligonucleotide array-based genotyping can yield reproducible data and MDA-WGA DNA products can be effectively used for genome-wide SNP genotyping analysis.

To whom reprint requests and correspondence should be addressed: Muhammad G. Kibriya (kibriya@uchicago.edu), or Habibul Ahsan (habib@uchicago.edu), Department of Health Studies, The University of Chicago, 5841 S. Maryland Avenue, MC 2007, Chicago, IL 60637. Phone: (773) 834 9956; Fax: (773) 834 0139.

Keywords

Whole genome amplification (WGA); microarray; Affymetrix GeneChip; genotyping; copy number

Introduction

Quantity and quality of source DNA is a major concern for genome wide studies using rapidly evolving array-based high-throughput genotyping technologies. Currently available gene-chip platforms can interrogate up to one million single nucleotide polymorphisms (SNPs) from one DNA sample which naturally requires sufficient quantity and quality of input genomic DNA (gDNA). The source of such DNA is quickly exhausted as stored samples are used up. Obtaining gDNA from blood for very high throughput genotyping involves cost, storage space, time and skill for DNA extraction (1). DNA from other non invasive sources, e.g., buccal mucous cell, can be obtained as well but the amount and the quality will be variable (2). One option to solve the problem is to immortalize peripheral blood lymphocyte cells by insertion of the viral genome (3). However, the process is labor-intensive, costly and time consuming and most importantly viable cells must be available to obtain DNA. Whole genome amplification (WGA) is rapidly becoming a popular option to sustain source of DNA for large scale genotyping studies (4–7). WGA by the Multiple Displacement Amplification (MDA) method generates a large amount of high quality DNA from a very small amount of input DNA. The usefulness of the WGA method depends on its ability and fidelity to reproduce the entire genome with least amplification bias because any such bias potentially resulting in errors in high density SNP genotyping will have a major impact on power to detect linkage or associations (6). We carried out a study to assess the concordance rate and the reproducibility of the genotyping calls obtained from gDNA samples and corresponding WGA DNA samples in microarray-based high throughput SNP typing assay.

Material and Method

We used a total of 28 samples - 14 gDNA samples from healthy individuals and their corresponding 14 WGA DNA samples and genotyped 224940 SNPs/sample using 28 early access Affymetrix Mendel Nsp Array GeneChips. The mean and median inter-SNP distance of the SNP array chip was 11.19 kb and 4.82 kb, respectively. Of the 14 gDNA samples, 12 were collected from healthy controls from three centers (4 samples / center) and the remaining two were duplicate reference gDNA samples. The three centers were: (1) Northern California site of the Breast Cancer Family Registry (BCFR) – Northern California Cancer Center (NCCC), (2) Ontario site of the BCFR- Cancer Care, Ontario (8), and (3) the German Cancer Research Center participating in the German Breast Cancer Study (GBCS) (9). In addition to the 28 early access Affymetrix Mendel NSP Array Gene-Chips, we also tested the reference gDNA sample and the corresponding WGA DNA sample on Illumina's Infinium 610 quad SNP-chip interrogating 620,901 markers, including 592,532 SNPs and 28,369 non-polymorphic copy-number variation probes. All the DNA samples were extracted from blood, except for the 4 samples from NCCC which were obtained from lymphoblastoid cell line. These 14 gDNA samples were amplified using the Qiagen Repli-g Midi kit to obtain their corresponding WGA DNA samples, as described below.

Whole Genome Amplification

We used a MDA based WGA kit from Qiagen – Repli-g Midi kit. The manufacturer's protocol (10) was followed which includes alkali (KOH) denaturation of gDNA samples prior to amplification. This method of amplification utilizes isothermal genome amplification by Phi 29 DNA polymerase capable of replication of up to 100kb without dissociating from the

genomic DNA template. This DNA Polymerase has a 3' to 5' exonuclease proofreading activity to maintain high fidelity during replication and is used in the presence of exonuclease-resistant primers to achieve high yields of DNA product (10). The quality and integrity of the tested gDNA samples were assessed using an Agilent 2100 BioAnalyzer. The gDNA size varied between 1,500 bp to more than 10,000 bp and the 260/280 ratio measured in Nano Drop ND-1000 UV spectrophotometer was between 1.8 and 1.9. We used 2.5 µl of gDNA in TE buffer at 10ng/µl concentration for WGA reaction. DNA was incubated isothermally at 30 degree Celsius for 10 hours, followed by heat inactivation (at 65 degree Celsius) of DNA polymerase for 3 minutes. After amplification the quality of the WGA product was checked on DNA 7500 chips using the Agilent 2100 BioAnalyzer (see Figure 1). Each electropherogram clearly showed uniform amplification producing smear starting from 1.5 kb extending to >10.0 kb size with clear peak around 7.0 kb.

Genotyping

Microarray-based genome-wide SNP genotyping was done using the early-access Affymetrix Mendel Nsp 250K chip. DNA samples were normalized to 50 ng/µL concentration. The Affymetrix standard protocol (11) was followed with slight modification in PCR purification step. High-speed ultracentrifugation was used instead of vacuum extraction. To compare the effect of quantity of PCR product used for hybridization on genotype call rate, 60 µg (as suggested by Affymetrix) and 90 µg of purified PCR products were used for fragmentation. For the WGA samples, 60 µg of purified PCR products from 6 samples (2 from each center) and 90 µg from the other 6 samples (2 from each center) were used. Scanning was performed in high-resolution Affymetrix GeneChip scanner 3000 7G. The electronic data were saved as DAT and CEL files. The CAB files for the images were used to transfer the data into GCOS 1.4 for subsequent use of the data in G-TYPE v4.0 software. Using V3 annotation for the early-access chips, a total of 224,940 SNPs were genotyped per sample. The SNPs are approximately evenly distributed within the whole genome; mean and median inter-SNP distance was 11.19 kb and 4.815 kb, respectively. The Affymetrix BRLMM algorithm was used to generate the genotype calls.

Statistical Analysis

The completeness of genotyping was determined for each of the 224,940 SNPs for the 14 gDNA samples (Com_gDNA) and 14 WGA DNA samples (Com_WGA). For example, if 14 out of 14 gDNA samples could be genotyped for a given SNP, then the "Com_gDNA" was 100% for that particular SNP. But the same SNP might have been genotyped in 12/14 or 86% of the corresponding WGA DNA samples, so "Com_WGA" would be 86%.

We report the concordance between gDNA and WGA DNA samples in two ways: (1) concordance by sample pair (i.e., the mean proportion of SNPs concordant between paired gDNA and WGA DNA samples) and (2) concordance by individual SNPs (i.e., the mean proportion of paired samples for which a particular SNP was concordant). Reproducibility of the platform was calculated as the proportion of SNPs concordant between two duplicate reference gDNA samples. For a given SNP, the discordant rate was calculated as the proportion of informative pairs across the paired samples that were discordant for that particular SNP. By informative data we mean the number of paired observations for which a genotype call for a given SNP could be made both in gDNA and WGA DNA samples. For example, if a given SNP could be genotyped in all 14 gDNA samples but in only 12 WGA DNA samples, then we excluded the 2 pairs (where we have only genotype result for gDNA but not for WGA DNA) and included only the data for the 12 informative pairs, where that given SNP could be genotyped in both gDNA and WGA DNA samples to calculate the discordance. If among these 12 informative pairs, we found discrepancy of genotype call in one pair, then we calculated the discordance rate to be 1 in 12 or 8.3%.

For both Copy Number (CN) and Loss of Heterozygosity (LOH) analyses, we took the gDNA samples as reference against which the paired analysis was done for WGA samples. For CN analysis background correction was done with adjustment for fragment length and probe sequence, but no normalization was done. The Log_2 ratio of the signal intensity was used for calculation of the CN. For detection of CN change regions, the Hidden Markov Model (HMM) (12) was used with maximum probability of 0.995, genomic decay of 10,000,000 and $\sigma = 2$. Maximum probability specifies the probability of retaining the same state between neighboring observations. The genomic decay describes how quickly (expressed in base pairs) the HMM retention of state will decay towards the initial probability. Sigma specifies the Gaussian bandwidth of the distribution from which observations are drawn. Higher value of sigma would expect more noise, but may not detect smaller regions. Smaller values will result in more regions. The reported regions contain at least 10 probe sets overlapping the regions in 7 out of 14 samples. The CN change regions were mapped to cytoband regions and the length was calculated from the start and the end regions. Total number of the SNPs within a CN change region and the average CN was reported.

The paired LOH regions were calculated assuming the maximum probability of 0.99, genomic decay of 10000000 and genotype error = 0.01. The reported regions overlap at least in 7 out of 14 samples. The length of the region was calculated from the start and the end regions.

Results

The overall mean (95% confidence interval) call rate in 14 gDNA samples was 97.07% (95% CI 96.17–97.97) and in corresponding WGA samples was 97.77% (95% CI 97.26–98.28) ($p = 0.154$). Center-specific call rates, proportion of different genotypes (AA, AB or BB) and raw intensity data are presented in Table 1. There was no significant difference in call rates or raw intensity for either gDNA or WGA DNA samples across different centers. We also did not observe any significant difference in genotype call rates using 60 μg or 90 μg of amplified PCR product of WGA samples for hybridization (97.82%, 95% CI 97.01–98.63 vs. 97.71%, 95% CI 96.85–98.56, $p = 0.826$). Reproducibility of the Affymetrix platform, as measured by concordance of SNP genotyping in duplicate reference samples, was found to be 99.45%.

The overall mean (95% CI) completeness of genotyping in gDNA samples (Com_gDNA) was 97.11% (95% CI 97.08–97.14) and that of WGA samples (Com_WGA) was 97.80% (95% CI 97.77–97.92) ($p < 0.001$). Genotype completeness by chromosome in gDNA and WGA samples is presented in Figure 2A. For both the gDNA and WGA samples, the completeness of genotype call was highest in SNPs of X-chromosome (marked as chromosome 23 in the figure). Among the autosomes, completeness was more consistent across the chromosomes for gDNA samples as compared to WGA samples. In particular, for the WGA samples, with reference to chromosome 10, which had median size and is also free from regions with copy number bias, as shown in the latter section of the results, the genotyping completeness was lower (ANOVA, $p < 0.001$) for the SNPs in chromosomes 16, 17, 19, 20 and 22.

Concordance by Sample Pair

Considering all the 224,940 SNPs genotyped per sample, the overall concordance between genotype calls from gDNA and WGA DNA samples was 97.74% (95% CI 97.03–98.44) without significant differences (mean $\pm$ standard deviation) between control samples and samples from different centers (control: 98.34% $\pm$ 0.528, center 1: 97.51% $\pm$ 1.092, center 2: 97.00% $\pm$ 1.408, center 3: 98.41% $\pm$ 1.256, ANOVA $p = 0.387$). In other words, in each gDNA-WGA pair sample, there were on average 2.26% (95% CI 1.55–2.97) SNPs with discordant genotypes. In the next step, we restricted the analysis to well performing SNPs – i.e., SNPs that could be successfully genotyped in $>90\%$ samples or in other words, SNPs with Com_gDNA & Com_WGA $>90\%$. There was a total of 191,251 well performing SNPs and

the concordance rate improved to 99.11% (95% CI 98.80 – 99.42), indicating that more than 99.0% of the well performing SNPs (i.e., genotyped in >90 % samples) show the same genotypes in gDNA and WGA samples. Figure 2B shows the concordance of these 191,251 SNPs for combined sample pairs and for samples from different centers. No significant difference in concordance was noted among the centers (control: 99.31% $\pm$ 0.341, center 1: 99.05% $\pm$ 0.48, center 2: 98.75% $\pm$ 0.65, center 3: 99.43% $\pm$ 0.48, ANOVA, $p=0.343$).

To explore the characteristics of this small proportion of SNPs producing discordant calls for each of the comparisons, we further analyzed the concordance by individual SNPs among the 14 paired samples.

Concordance by SNPs

Among the total 224,940 genotyped SNPs, only 99 SNPs (0.04%) showed discordant calls in all 14 paired observations and a total of 2721 SNPs (1.2%) showed discordant calls in ≥ 7 paired observations. Figure 2C illustrates the discordant rate as a function of completeness of genotyping in WGA samples. A similar result was obtained for completeness of genotyping in gDNA samples [data not shown]. These results clearly demonstrate that the well performing SNPs (high Com_gDNA &/or high Com_WGA) had the least discordance. The finding suggests that WGA can be used efficiently with minimum error for good performing SNPs.

In the next step, we examined whether there is chromosomal bias for discordant calls. Recognizing the fact that the discordant rate is significantly influenced by completeness of genotyping for a particular SNP, for interrogating chromosomal bias we included the 191,251 well performing SNPs which had both the Com_gDNA and Com_WGA >90%. For practical purpose, for a genome-wide gene mapping study, one should filter the data on the SNP call rate or the completeness of the SNP genotyping. Figure 2D shows the mean (95% CI) discordant rate for the SNPs by chromosome. The data show that the overall mean discordant rate was only 0.92% (95% CI 0.90–0.94), i.e., overall a well performing SNP had a discordant genotype result in less than 1% sample pairs. However, there was a significant difference in the discordant rate by SNPs (proportion of discordant sample pairs) for some of the chromosomes (ANOVA, $p<0.001$). Compared to chromosome 1, the lowest discordance (0.44%, 95% CI 0.37–0.51) was observed for the SNPs in chromosome-X (shown as chromosome 23 in the graph) and significantly higher discordant rates were found among SNPs in chromosome 16 (1.08%, 95% CI 0.95–1.22), chromosome 19 (1.47%, 95% CI 1.22–1.73), chromosome 20 (1.11%, 95% CI 0.96–1.26) and in chromosome 22 (1.36%, 95% CI 1.11–1.62).

This apparent chromosomal bias for discordance and the effect of completeness of genotyping on discordant rates led us to look for copy number changes in WGA samples.

Copy Number Analysis

For copy number analysis, we took gDNA samples as the reference for the corresponding WGA samples. Figure 3 shows the regions of copy number changes in the WGA samples compared to the corresponding gDNA samples. The blue regions indicate loss of copy number. It is noted that in most of the chromosomes, the loss of copy number was detected in the telomeric regions. There were a total of 7 regions in 6 chromosomes (see Figure 3). The smallest region was 1.9 Mb in chromosome 16 p13.3 region and the largest was 4.7 Mb in chromosome 9 q34.2–q34.3 region. The seven regions with loss of copy number represent 21509539 bp. For the currently assembled human genome size of 3021400000 bp (13), these data give an estimated genome coverage of $100 \times (3021400000 - 21509539) / 3021400000 = 99.29\%$. It may be noted that all these seven regions were previously reported to have copy number variation (CNV) by different investigators (14–20). These reported variation IDs in the Database for Genomic

Variants (21) are also presented in figure 3. Figure 4A shows the copy number changes in chromosome 9 of all the 14 WGA samples in our study. The upper panel indicates the copy number loss regions marked by blue, the middle panel shows the plot of estimated copy number in reference to the gDNA samples, and the lower panel represents the heat map where blue indicates loss of copy number, gray normal copy number and the red gain in copy number. Figure 4B shows the data of chromosome 9q34.2 and 9q34.3 regions (the same 135M to 140M bp region, which are shown in figure 4A to have loss of copy number in our study) from the Database of Genomic Variants. The browser view clearly indicates that other investigators detected CNVs in that region. However, it may be noted that our study indicates defect in WGA in these regions, and this paired analysis (where g-DNA is used as reference for corresponding WGA sample), does not confirm CNV. We noted the cytoband(s) of those regions with the copy number changes (loss or gain) in all the chromosomes. SNPs in cytoband regions with loss of copy number were marked as Group-1 (n=1401) and those with normal copy number as Group-2 (n=223,593). In the next step, we further analyzed the SNPs in chromosomes in respect to the copy number changes. The overall completeness of genotyping for Group-1 SNPs was significantly lower than Group-2 [90.56% (95% CI 89.85% – 91.27%) vs. 97.84% (95% CI 97.82% – 97.87%), $p < 0.001$]. Figure 5A shows the completeness of genotyping of Group-1 and Group-2 SNPs by chromosomes. Figure 5B shows that the discordance of Group-1 SNPs was significantly higher than that of Group-2 SNPs in all the chromosomes [8.66% (95% CI 7.73% – 9.61%) vs. 2.71% (95% CI 2.68% – 2.75%), $p < 0.001$]. Therefore, it is clear that the few small regions with copy number loss in the WGA samples affect both the completeness of the genotyping call (SNP performance) and the discordant rate (inaccurate calls) and these areas are situated mainly in the telomeric regions.

LOH Analysis

We also examined paired LOH regions for the WGA samples compared to the corresponding gDNA samples. We detected five LOH regions - chromosome 2q21.1 (15505 bp), chromosome 5q13.1 (8540 bp), chromosome 8q11.22 (156577 bp), chromosome 13q31.1 (17973 bp) and chromosome 20q13.31 (39596 bp). There was no overlap between these LOH regions and the copy number change regions, indicating copy neutral LOH. Also, none of these regions was near the telomere. There was a total of only 30 SNPs (0.013% of total genotyped SNPs) covering these five very small genomic regions (total 238,191 bp, accounting for 0.00078% of whole genome) with LOH. As opposed to the usual LOH seen in tumor DNA, these were copy neutral LOH and therefore, as expected, SNPs in these regions were discordant in 22.53% (95% CI 9.93 – 35.13%) sample pairs. In fact, this also represents error of amplification.

Cross-check on Illumina's Infinium platform

The genotype call rate for the reference gDNA on Illumina's Infinium platform was 99.74% and that of corresponding WGA DNA was 99.68% with concordance rate of 99.998%. A total of 38,655 SNPs were common between the Illumina's 610 quad chip and Affymetrix early access Mendel Nsp GeneChip. Of these 38,655 common SNPs between the two platforms, for the reference gDNA sample, a total of 38,136 SNPs could be successfully genotyped on both platforms. Genotype calls for only 143 (0.375%) SNPs were discordant between the two platforms. In other words, the genotype calls by the two platforms were concordant for 99.625% of the SNPs.

Discussion

The WGA is a promising solution to eliminate the practical problem in the limitation of the source of DNA needed for genome-wide scans. In order to fulfill the purpose, WGA must satisfy some basic requirements. First, the amplification process should be highly accurate to avoid undue errors. Second, amplification should not produce a bias in the distribution of the

DNA products. Questions of amplification-induced error and template bias generated by the WGA process have been addressed elsewhere through small and large scale SNP detection methodologies (1,22–26). Third, a high amplification factor is required so that WGA generates a useful amount of DNA from small starting samples. Finally, the WGA method should be applicable to a wide array of genomic platforms (24).

Different methods of WGA have been used so far in different studies by different investigators. Three main methods have been used for WGA: (1) Multiple Displacement Amplification (MDA) (22,27), (2) Primer extension pre-amplification (PEP) (28), and (3) Degenerate Oligonucleotide-primed PCR (DOP) (5,29). Besides the methods of amplification, other critical issues include amount of DNA input (30,31), amplified DNA yield (24) and the level of bias (32). Pinard et al compared the yield of WGA product using the different amplification methods from 25ng of gDNA as starting material: the MDA based REPLI-g method generated 2100 fold amplification, GenomiPhi 640 fold, PEP 120 fold and DOP 92 fold (24). The sharp contrast between the yields derived from the two MDA based methods (REPLI-g and GenomiPhi) may be attributed to the use of KOH alkali denaturation prior to the amplification process which opens priming sites more efficiently than the thermal denaturation used in the GenomiPhi protocol (24).

There is evidence that the level of error introduced during WGA reaction appears to be a function of amount of starting material. In this connection, Dean (22) and Lovmar (33) have evaluated the genotyping performance of MDA WGA using a range of gDNA inputs and both the authors focused attention in their evaluation of genotyping performance of WGA DNA derived from 3 ng of gDNA. Bergen et al carried out extensive investigation on the effect of gDNA mass (1,10, 25, 50, 100 and 200 ng) on WGA and genotyping performance (30). They found that, for optimal performance in single-plex SNP genotyping using TaqMan platform, at least 10 ng of lymphoblastoid gDNA input in WGA reaction was required; but over 100 ng of lymphoblastoid gDNA input into WGA reaction was required to obtain optimal STR genotyping performance from WGA DNA. In their work, the WGA obtained from 25 ng of gDNA input showed 99.9% completion of genotyping with 2.3% discordance. Lasken and Egholm recommended 10–100 ng of gDNA template in the MDA WGA reaction to avoid stochastic amplification (34). In our lab, for single-plex SNP genotyping using fluorescent polarization method, we have seen up to 100% completion of genotyping with 25 ng of WGA-DNA sample per well in PCR reaction from the WGA stock obtained from 25 ng of gDNA input in 50 μ L WGA reaction volume. Figure-6 shows the clustering of 84 genotype calls for rs1476413, using 25 ng of gDNA in left panel and 25 ng corresponding WGA-DNA (from stock of WGA obtained from 25 ng of g-DNA input in WGA reaction) on the right panel. SNP concordance was 100%. Among the g-DNA samples, five samples were not clustered tightly (undetermined or no call) but clearly three were heading towards GA genotype cluster and the other two were heading towards AA genotype cluster. However, in case of the corresponding WGA samples (right panel), the samples were nicely separated in three distinct genotype clusters. Sawcer et al used a total of 508 WGA samples for genotyping on the Illumina GoldenGate platform and found that the likelihood of successful genotyping from WGA DNA correlated with the starting concentration of genomic DNA used in the amplification reaction: a large proportion of samples (n=404) failed to produce genotype calls and the mean starting concentration was 5.9 ng/ μ L, whereas for the rest of samples (n=104) for which they had successful genotype calls, the concentration of the starting gDNA was 17.4 ng/ μ L (25). The present study was not designed to find out optimal gDNA input into the WGA reaction. Rather we focused on the performance of WGA DNA derived from 25ng of gDNA as input in the WGA reaction. In the context of genome-wide genotyping, only 25 ng of good quality genomic DNA as starting material for subsequent WGA reaction may be considered a good alternative to standard requirement of 250 – 500 ng of gDNA for microarray-based high throughput genotyping.

Arriola et al amplified genomic DNA at different starting amounts (0.5, 5, 10 and 50 ng) using the Phi29 based MDA method and found that the fold amplification was highest when the input DNA was low and this higher fold amplification was correlated to amplification bias in Comparative Genomic Hybridization (CGH) profiles (31).

Paez et al, used the Phi 29 polymerase based amplification method, with or without alkali denaturation prior to amplification and tested the accuracy and genome-wide coverage of the derived WGA product through both direct sequencing of around 500,000 bp and high density oligonucleotide arrays interrogating 10K SNPs with mean inter-marker distance of 210 kb on the Affymetrix platform (32). Their study showed better call rates with prior alkali denaturation. The call rate was 92.93% in genomic DNA and 92.06% in WGA samples with prior alkali denaturation. In the present study, we used 25 ng of gDNA as starting material and treated with KOH prior to WGA by the MDA method and used the Affymetrix Early Access Mendel Nsp 250K GeneChip containing 224,940 SNPs with mean and median inter-SNP distance of 11.19 kb and 4.815 kb respectively. We found that the overall call rate was 97.07% (95% CI 96.17–97.97) in genomic DNA samples and 97.77% (95% CI 97.26–98.28) in WGA samples.

In a small-scale genotyping study in which only 6 SNPs were genotyped in 172 samples, a concordance of 100% was found among gDNA and corresponding WGA DNA (35). On the other hand, when genotyping was performed on a larger number of SNPs on the Illumina linkage panel (2320 SNPs) platform (36) or using the Illumina GoldenGate method (345 SNPs) (7), the call concordance was found to vary between 98.8% and 99.7%. One study explored the utility of MDA on 10K SNP arrays, reporting good coverage and high concordance rates but reduced call rates (32). In our study, using 250K SNP chip, the overall concordance was 97.74% (95% CI 97.03–98.45) and when restricting the analysis to well performing SNPs (Com_gDNA and Com_WGA >90%), on an average, 99.11% (95%CI 98.80 – 99.42) of the SNPs were concordant and overall a SNP showed discordant call only in 0.92% (95%CI 0.90 – 0.94) of paired samples. Moreover, we used the early access chips where the SNP panel was not yet fully optimized for SNP performance. For practical purposes, in genome-wide analysis SNPs should be filtered by call rate (across the samples). Analyzing the small number of SNPs that caused discordant calls, we identified that there were very few regions with copy number loss and those were predominantly at the telomeric regions. We also looked at paired LOH regions for the WGA samples compared to the corresponding gDNA samples and found only 5 copy- neutral LOH regions (smallest region at 2q21.1 of 8540 bp and the largest one at 8q11.22 of 156577bp), none of which was located near telomeric regions. In a previous study, Paez et al. also found few chromosomal regions with loss of copy number in MDA-based WGA samples, but none of those regions were telomeric (32). To our knowledge, this is one of the first studies to examine the SNP concordance of WGA product with healthy human germ line gDNA samples on very high-density oligonucleotide based SNP chips interrogating 224,940 SNPs. Although only in one pair of samples, we also tested the performance of MDA-based WGA product on a different platform – Illumina's 610 Quad chip interrogating 592,532 SNPs, and noticed 99.998% concordance with the gDNA. Previous studies have not used such a high resolution microarray platform to address this issue. It may be noted that neither the Affymetrix nor the Illumina GoldenGate assay protocol uses further WGA step in sample processing, rather PCR amplification is used. On the other hand, Illumina's Infinium chemistry uses WGA as a part of DNA sample processing before hybridization.

The present study was limited to the use of high quality intact gDNA as input into the WGA reaction. Considering the fragment size of the degraded DNA extracted from formalin-fixed paraffin embedded (FFPE) samples, MDA based WGA may not be suitable option for Affymetrix GeneChip. However, fragmentation-PCR-based method for WGA is an appropriate choice for the FFPE samples. In a very recent publication (Epub 2008 Jun 12),

Mead et al have documented that degraded DNA amplified with MDA-based-WGA gave low call rates and concordance across all platforms at standard loading concentration; but the fragmentation-PCR-based method of WGA gave high call rate and concordance for degraded DNA (37).

In summary, our results suggest that Phi29 MDA based WGA product provides a highly accurate and reasonably comprehensive representation of the unamplified human genome, suitable for high resolution genome-wide genotyping studies using oligonucleotide-based SNP genotyping arrays.

Acknowledgments

Grant Support and Acknowledgements:

This work was partly supported by the National Cancer Institute, National Institutes of Health under RFA-CA-06-503 and through cooperative agreements with members of the Breast Cancer Family Registry and P.I.s and partly by U01 CA122171 and P30 CA 014599. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CFR.

REFERENCES

1. Hosono S, Faruqi AF, Dean FB, et al. Unbiased whole-genome amplification directly from clinical samples. *Genome Res* 2003;13:954–964. [PubMed: 12695328]
2. Harty LC, Garcia-Closas M, Rothman N, Reid YA, Tucker MA, Hartge P. Collection of buccal cell DNA using treated cards. *Cancer Epidemiol Biomarkers Prev* 2000;9:501–506. [PubMed: 10815695]
3. Packer, RJ.; Bolton, BJ. A Laboratory Handbook: Immortalization of B-Lymphocyte by Epstein-Barr Virus. Academic Press; 1998. Cell Biology.
4. Cheung VG, Nelson SF. Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proc Natl Acad Sci U S A* 1996;93:14676–14679. [PubMed: 8962113]
5. Little SE, Vuononvirta R, Reis-Filho JS, et al. Array CGH using whole genome amplification of fresh-frozen and formalin-fixed, paraffin-embedded tumor DNA. *Genomics* 2006;87:298–306. [PubMed: 16271290]
6. Montgomery GW, Campbell MJ, Dickson P, et al. Estimation of the rate of SNP genotyping errors from DNA extracted from different tissues. *Twin Res Hum Genet* 2005;8:346–352. [PubMed: 16176719]
7. Pask R, Rance HE, Barratt BJ, et al. Investigating the utility of combining phi29 whole genome amplification and highly multiplexed single nucleotide polymorphism BeadArray genotyping. *BMC Biotechnol* 2004;4:15. [PubMed: 15279678]
8. John EM, Hopper JL, Beck JC, et al. The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res* 2004;6:R375–R389. [PubMed: 15217505]
9. Chang-Claude J, Eby N, Kiechle M, Bastert G, Becher H. Breastfeeding and breast cancer risk by age 50 among women in Germany. *Cancer Causes Control* 2000;11:687–695. [PubMed: 11065005]
10. QIAGEN. REPLI-g Mini/Midi Handbook. 2005. Available from <http://www1.qiagen.com/ts/msds.asp>
11. Affymetrix. GeneChip Mendel Array Protocol Early Access Version 2.0. 2005. Available from <http://www.affymetrix.com>
12. Rabiner, LR. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition; *Proceedings of the IEEE*; 1989. p. 257-285.
13. UCSC genome browser NCBI build 35. 2004 Aug 26. Available from <http://genome.ucsc.edu>
14. Simon-Sanchez J, Scholz S, Fung HC, et al. Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum Mol Genet* 2007;16:1–14. [PubMed: 17116639]

15. Wong KK, deLeeuw RJ, Dosanjh NS, et al. A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet* 2007;80:91–104. [PubMed: 17160897]
16. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. *Hum Mol Genet* 2007;16(Spec No. 2):R168–R173. [PubMed: 17911159]
17. de Smith AJ, Tsalenko A, Sampas N, et al. Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum Mol Genet* 2007;16:2783–2794. [PubMed: 17666407]
18. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet* 2004;36:949–951. [PubMed: 15286789]
19. Zogopoulos G, Ha KC, Naqib F, et al. Germ-line DNA copy number variation frequencies in a large North American population. *Hum Genet* 2007;122:345–353. [PubMed: 17638019]
20. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature* 2006;444:444–454. [PubMed: 17122850]
21. Database of Genomic Variants. Human genome build 36. Available from <http://projects.tcag.ca/variation/cgi-bin/gbrowse/hg18>
22. Dean FB, Hosono S, Fang L, et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* 2002;99:5261–5266. [PubMed: 11959976]
23. Lovmar L, Fredriksson M, Liljedahl U, Sigurdsson S, Syvanen AC. Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA. *Nucleic Acids Res* 2003;31:e129. [PubMed: 14576329]
24. Pinard R, de Winter A, Sarkis GJ, et al. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 2006;7:216. [PubMed: 16928277]
25. Sawcer S, Ban M, Maranian M, et al. A high-density screen for linkage in multiple sclerosis. *Am J Hum Genet* 2005;77:454–467. [PubMed: 16080120]
26. Wells D, Sherlock JK, Handyside AH, Delhanty JD. Detailed chromosomal and molecular genetic analysis of single cells by whole genome amplification and comparative genomic hybridisation. *Nucleic Acids Res* 1999;27:1214–1218. [PubMed: 9927758]
27. Lage JM, Leamon JH, Pejovic T, et al. Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. *Genome Res* 2003;13:294–307. [PubMed: 12566408]
28. Zhang L, Cui X, Schmitt K, Hubert R, Navidi W, Arnheim N. Whole genome amplification from a single cell: implications for genetic analysis. *Proc Natl Acad Sci U S A* 1992;89:5847–5851. [PubMed: 1631067]
29. Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA, Tunnacliffe A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* 1992;13:718–725. [PubMed: 1639399]
30. Bergen AW, Qi Y, Haque KA, Welch RA, Chanock SJ. Effects of DNA mass on multiple displacement whole genome amplification and genotyping performance. *BMC Biotechnol* 2005;5:24. [PubMed: 16168060]
31. Arriola E, Lambros MB, Jones C, et al. Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridisation. *Lab Invest* 2007;87:75–83. [PubMed: 17170740]
32. Paez JG, Lin M, Beroukhi R, et al. Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res* 2004;32:e71. DOI: 10.1093/nar/gnh069. [PubMed: 15150323]
33. Lovmar L, Syvanen AC. Multiple displacement amplification to create a long-lasting source of DNA for genetic studies. *Hum Mutat* 2006;27:603–614. [PubMed: 16786504]
34. Lasken RS, Egholm M. Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. *Trends Biotechnol* 2003;21:531–535. [PubMed: 14624861]
35. Tranah GJ, Lescault PJ, Hunter DJ, De Vivo I. Multiple displacement amplification prior to single nucleotide polymorphism genotyping in epidemiologic studies. *Biotechnol Lett* 2003;25:1031–1036. [PubMed: 12889810]

36. Barker DL, Hansen MS, Faruqi AF, et al. Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel. *Genome Res* 2004;14:901–907. [PubMed: 15123587]
37. Mead S, Poulter M, Beck J, et al. Successful amplification of degraded DNA for use with high-throughput SNP genotyping platforms. *Hum Mutat* 2008. Epub 2008 Jun 12.

Figure 1.

Agilent 2100 BioAnalyzer electropherogram of 10 WGA DNA samples (normalized to 50 ng/ul concentration) overlaid on ladder marker peaks (shown in red). After the initial spike at 50 bp, the subsequent ladder peaks correspond to 100, 300, 500, 700, 1000, 1500, 2000, 3000, 5000, 7000, 10380 bp respectively. All the samples (in different color) show uniform smearing effect starting around 1500 bp extending to >10 kb size with clear peak around 7 kb region (11th peak).

Figure 2A:

Figure 2B:

Figure 2C:

Figure 2D:

Discordant rate by chromosome in well performing SNPs (com_gDNA & com_WGA > 90% samples)

Figure 2.

(A) X axis represents the mean (95% CI) completeness of genotype call in gDNA samples (blue solid square) and WGA samples (red open square) and Y axis represents chromosomes. Chromosome X is represented by chromosome 23. The error bars represent 95% confidence intervals.

(B) Genotype concordance rate in all samples combined and samples from different centers. The error bar represents SD.

(C) Discordant rate (on Y axis) is shown as a function of completeness of genotyping in WGA samples (on X axis). The number of SNPs in each group is also shown. The error bar represents 95% CI.

(D) For the well performing SNPs (those which could be successfully genotyped in > 90% of the cases in both gDNA and WGA samples) discordant rate is shown in different chromosomes. Total number of SNPs in each chromosome is also shown. The error bar represents 95% CI.

*Database of Genomic Variants (<http://projects.tcag.ca/variation/>)

Figure 3.

Upper panel shows graphical representation of the genome-wide copy number (CN) change regions. Blue regions indicate loss of copy number in WGA samples compared to the corresponding gDNA samples. No region was identified as gain of copy number. The lower panel shows the chromosomal location, length, average CN, the CNV variation ID# from the database of genomic variants and number of SNPs in those CN loss regions shown in the upper panel.

Figure 4A:

Figure 4B:

Figure 4.

(A) Detailed view of CN changes detected in all the 14 WGA samples in chromosome 9. The upper panel indicates the copy number loss regions marked by blue, the middle panel shows the plot of estimated copy number in reference to the gDNA samples and the lower panel represents the heat map of all the 14 WGA samples (each row represents a WGA sample) where blue indicates loss of copy number, grey normal copy number and the red gain in copy number. The 8 samples with loss of CN are highlighted in black box in the lower panel. The cytoband regions of chromosome 9 are also shown at the bottom of the lower panel.

(B) Genome browser view from the Database of Genomic Variants for the copy number loss region of chromosome 9q34.2, q34.3 region (130M to 140M) shown in figure 4A. The cytoband regions are shown in dark gray, the CNVs reported in the publicly available DGV database are shown in orange and the Insertion-Deletions (InDels) between 100 bp and 1kb size are shown in green.

Figure 5A:

Figure 5B:

Figure 5.

(A) Completeness of genotyping of group-I SNPs (those in cytobands with loss of CN) and group-II SNPs (in cytobands with normal CN) by chromosome in WGA samples.

(B) Discordant rate of group-I SNPs (those in cytobands with loss of CN in WGA samples) and group-II SNPs (in cytobands with normal CN in WGA samples) by chromosome.

Figure 6. Genotyping for rs1476413 using Fluorescent Polarization method for g-DNA samples (left) and corresponding WGA-DNA samples (right)

Genotyping from gDNA and corresponding WGA DNA for rs1476413 using Fluorescent Polarization method (a single base extension method). Clustering of 84 genotype calls, using 25 ng of gDNA in left panel and 25 ng of corresponding WGA-DNA (from stock of WGA obtained from 25 ng of g-DNA input in WGA reaction) on the right panel. SNP concordance was 100%.

Genotype calls in gDNA and WGA samples by center using Affymetrix early access MendelNsp Array GeneChips interrogating 224940 SNPs using BRLLM algorithm.

Table 1

	Center-1			Center-2			Center-3			ANOVA p-value
	mean	95%	CI	mean	95%	CI	mean	95%	CI	
gDNA samples *	n=4			n=4			n=4			
Call Rate (%)	96.67	95.01	-98.32	96.16	93.19	-99.14	97.68	95.14	-100.21	0.405
AB call (%)	28.76	27.05	-30.48	28.80	27.03	-30.58	28.42	26.66	-30.17	0.863
AA call (%)	35.09	33.68	-36.51	34.79	32.50	-37.08	35.73	33.76	-37.71	0.552
BB call (%)	32.81	30.83	-34.79	32.57	30.13	-35.00	33.52	31.23	-35.81	0.626
Raw Intensity	8.77	8.27	-9.27	8.66	8.26	-9.07	8.67	8.26	-9.08	0.831
WGA samples **	n=4			n=4			n=4			
Call Rate (%)	98.17	97.51	-98.83	97.42	96.28	-98.56	98.36	97.40	-99.33	0.111
AB call (%)	27.76	26.92	-28.6	27.48	26.74	-28.22	27.91	27.34	-28.48	0.436
AA call (%)	36.22	35.63	-36.81	36.03	35.21	-36.84	36.23	35.53	-36.94	0.773
BB call (%)	34.18	33.44	-34.92	33.90	32.85	-34.96	34.21	33.42	-35.01	0.682
Raw Intensity	8.67	8.20	-9.13	8.62	8.10	-9.14	8.76	8.27	-9.25	0.805

* genomic DNA samples

** whole genome amplified samples