

Published in final edited form as:

Genomics. 2008 December ; 92(6): 452–456. doi:10.1016/j.ygeno.2008.08.007.

High Fidelity of Whole-Genome Amplified DNA on High-Density Single Nucleotide Polymorphism Arrays

Jinchuan Xing, W. Scott Watkins, Yuhua Zhang, David J. Witherspoon, and Lynn B. Jorde^{*}
Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT, 84112

Abstract

Current microarray technology allows researchers to genotype a large number of SNPs with relatively small amounts of DNA. Nevertheless, researchers and clinicians still frequently face the problem of acquiring enough high-quality DNA for analysis. Whole-genome amplification (WGA) methods offer a solution for this problem, and earlier studies have shown that WGA samples perform reasonably well in small-scale genetic analyses (e.g. Affymetrix 10K array). To determine the performance of WGA products on a large-scale genotyping array, we compared the Affymetrix 250K array genotyping results of genomic DNA and their WGA products from four individuals. Our results indicate that WGA product performs well on the 250K array compared to genomic DNA, especially when using the BRLMM calling algorithm. WGA samples have high call rates (97.5% on average, compared to 99.4% for genomic DNA) and excellent concordance rates with their corresponding genomic DNA samples (98.7% on average). In addition, no apparent systematic genomic amplification bias can be detected. This study demonstrates that, although there is a slight decrease in the total call rates, WGA methods provide a reliable approach for increasing the amount of DNA samples for use with a common SNP genotyping array.

Keywords

whole genome amplification; WGA; single nucleotide polymorphism array

Introduction

Single nucleotide polymorphisms (SNPs) are the most common genetic variants in the human genome, and it is believed that more than 10 million common SNPs with minor-allele frequency >1% exist in the genome [1;2]. As such, SNPs play an important role in genome-wide association and population genetic studies [3]. With the recent development of microarray-based whole-genome SNP genotyping technology, it is now possible to genotype one million SNPs on a single microarray (e.g., Affymetrix Human SNP Array 6.0 and Illumina Human1M BeadChip). These high density arrays have dramatically reduced the amount of DNA required for large-scale genotyping. Nevertheless, insufficient quantities of DNA remain a challenge for some studies, especially when the source of DNA is limited or has low quality (e.g., tumor tissue samples, mouth wash, archival samples, etc).

^{*}Corresponding author: Lynn B. Jorde, Department of Human Genetics, Eccles Institute, of Human Genetics, University of Utah, Salt Lake City UT 84112. Tel: (801) 581-4566, Fax: (801) 585-9148. E-mail: lbj@genetics.utah.edu .

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

To increase the amount of usable DNA in low-quantity samples, a number of *in vitro* whole-genome amplification (WGA) approaches have been proposed in recent years (see [4] for review). The multiple displacement amplification with Φ 29 polymerase method, described in 2002 [5], has been shown to have a number of advantages over other methods. It has a high degree of fidelity, large amplification products, and more even amplification across the genome [4]. Several earlier studies have shown that WGA products perform reasonably well in small-scale genetic analyses including Affymetrix 10K SNP array [6;7]. Here, we investigate the performance of the WGA products on an Affymetrix 250K mapping array.

Materials and Methods

Genomic DNA extraction and whole-genome amplification

Genomic DNA used for this study was extracted from blood samples of six unrelated donors using standard phenol/chloroform extraction procedures or a Puregene DNA extraction kit (Qiagen, Valencia, CA, USA) more than 10 years ago and have been stored in TE buffer at 4°C. Whole genome amplification was performed on four of these samples using a REPLI-g mini kit (Qiagen) following the protocol in the manufacturer's manual. 10 ng of purified genomic DNA was used as template, and the amplification product was normalized to a concentration of 50 ng/ul prior to the microarray experiment.

Genotyping

High-throughput microarray genotyping of approximately 262,000 SNPs was performed using one array (version *NspI*) from the Affymetrix GeneChip® Human Mapping 500K Array set (Affymetrix, Santa Clara, CA, USA). The recommended protocol as described in the Affymetrix manual was followed. Briefly, 250 ng of genomic DNA or WGA product was digested with *NspI* (New England Biolabs, Beverly, MA, USA) and ligated to an *NspI* adapter (Affymetrix) using T4 DNA ligase (New England Biolabs, Ipswich, MA, USA). Samples were then amplified by PCR using TITANIUM *Taq* polymerase (Clontech, Mountain View, CA, USA) on an MJ Tetrad PTC-225 machine (Bio-Rad, Hercules, CA, USA). PCR products were pooled and purified using the Clontech purification kit and subjected to fragmentation using DNaseI (Affymetrix). The resulting DNA fragments were biotin-labeled with terminal deoxynucleotidyl transferase (Affymetrix). Samples were then injected into microarray cartridges and hybridized in a GeneChip® Hybridization Oven 640 (Affymetrix), followed by washing and staining in a GeneChip® Fluidics Station 450 (Affymetrix). Mapping array images were obtained using the GeneChip Scanner 3000 7G (Affymetrix).

Genotype calling and data analysis

Genotypes for each experiment were first called with the Affymetrix Dynamic Model algorithm (DM, [8]) to assess the quality of the experiment, and all samples had call rates higher than the 93% threshold recommended by Affymetrix. All samples were then called together using the BRLMM algorithm [9] with default parameters. Analyses and statistical tests were performed using MATLAB (ver. r2008a). Genotypes of all samples are available as supplemental file on our website (<http://jorde-lab.genetics.utah.edu/WGA/wga.brlmm.calls.txt>).

Results

SNP Call Rate

DNA samples from six unrelated individuals were analyzed. Four of the six samples (WGA1 to WGA4) were whole-genome amplified, and both template genomic DNA (gDNA) and whole-genome amplified (WGA) products were hybridized on the Affymetrix Human Mapping 500K Arrays version *NspI*, which assays about 262,000 SNPs. The remaining two

samples (Tech1 and Tech2) were analyzed as technical duplicates for which two identical array experiments were performed on genomic DNA from each sample.

High-quality call rates were obtained for all 12 chips, and the initial DM call rates of all samples passed the 93% threshold recommended by Affymetrix (Table 1). The sex of each sample was correctly inferred as male. All 12 CEL files were then used for the BRLMM analysis. The BRLMM call rates of the four gDNA samples varied between 99.69% and 99.03%, with a mean of 99.37% and a standard deviation (SD) of 0.28%. This result is comparable with the two technical duplicates, which had an average call rate of 98.90% and a SD of 0.49%. The four WGA samples have slightly lower call rates compared to their corresponding gDNA samples (Table 1), with an average call rate of 97.46% (1.91% lower than gDNA samples) and a SD of 0.37%. There is no SNP that failed (i.e., over the default No Call threshold) in all 12 samples.

To investigate the possible allele amplification bias of the WGA method, we selected SNPs that failed in WGA samples ("No Call" SNPs; Table 2) but were called in corresponding gDNA samples. When the proportion of calls in this set was compared with all SNPs on the array, we found that the proportion of each genotype (AA, AB and BB) in No Call SNPs is significantly different from the proportion of each genotype in all SNPs on the array (Table 2, chi-squared test, $p < 10^{-4}$ for each sample). There is a systematic decrease in AA, AB calls and an increase in BB calls for the SNPs that failed in WGA experiments. On average, SNPs that failed in WGA experiments showed a 3% decrease in AA calls, 11.3% decrease in AB calls and 11.3% increase in BB calls compared to all SNPs in the gDNA samples.

Concordance Rate

To study the fidelity of WGA products on the SNP array, we calculated SNP concordance rates between gDNA and WGA samples and between the two technical duplicates. In this analysis, only SNPs that have genotypes in both experiments were considered. The two technical duplicates have excellent concordance rates of 99.55% and 99.51%, respectively (Table 3). WGA samples have very high concordance rates with gDNA samples as well: all four samples have concordance rates greater than 98% (Table 3, mean = 98.65%, SD = 0.24%). Among all SNPs that are discordant in any WGA/gDNA comparisons (9925 total), the vast majority of the discordant calls (98.8%) are SNPs called as heterozygotes (AB) on one array but homozygotes (AA or BB) on the other array. Only 1.2% of the discordant calls are different homozygous calls (AA vs. BB) between the two experiments.

To assess the genotyping quality of discordant calls, we examined the confidence score (C-score) of these SNPs. A C-score is the ratio of the Mahalanobis distances between the observed allele signal value and two nearest typical genotype values (see [9] for detail). C-score, which varies between 0 and 1, is a measure of the genotyping quality; a smaller C-score indicates higher genotyping quality. A SNP with a C-score > 0.5 is assigned as a No Call by default. We found that the C-score distribution of the discordant calls is significantly different between gDNA and WGA samples (Figure 1A, two sample Kolmogorov-Smirnov test, $p < 10^{-12}$). The higher C-scores of WGA samples show that there is lower genotype-calling confidence for discordant SNPs in the WGA samples. Interestingly, a similar distribution difference is observed in the discordant calls between technical duplicates: as might be expected, the sample with a lower call rate in each pair has higher C-scores (Figure 1B, two sample Kolmogorov-Smirnov test, $p < 10^{-12}$). Among discordant SNPs, WGA samples showed a significantly higher proportion of heterozygotes ($> 75\%$ in each sample) compared to their corresponding gDNA samples (chi-squared test, $p < 10^{-12}$; Table 3). When the discordant calls between our two technical duplicates are examined, the experiment with a lower call rate in each duplicate also showed significantly higher proportions of heterozygotes (chi-squared test, $p < 10^{-12}$).

Therefore, discordant calls for WGA samples appear similar to discordant calls for low call-rate genomic DNA samples.

Genomic Distribution of No Call and Discordant SNPs

We next explore the possibility of systematic SNP calling failure in WGA samples. There is little overlap between No Call SNPs among WGA samples. Out of a total of 21,044 SNPs that failed in any of the four WGA samples, only 144 (0.68%) failed in all four experiments. These SNPs are dispersed across chromosomes, and no apparent clustering can be observed. Similarly, only 1.23% of all discordant SNPs (169 out of 13,712) are discordant in all four WGA/gDNA comparisons. To further investigate the distribution of No Call and discordant SNPs, we plotted the number of total SNPs analyzed (~262,200 total), the number of SNPs that failed in at least three WGA experiments (4,759 SNPs total) and the number of SNPs that have discordant calls between any WGA/gDNA comparisons (9925 total) in 1Mb non-overlapping windows across each chromosome (Supplemental Figure 1). The SNP distribution on chromosome 1 is shown in Figure 2 as an example. The No Call SNPs and discordant SNPs appears to be distributed across the genome and no apparent clustering of these SNPs are observed. As expected, the number of SNPs failed in at least three WGA experiments showed a small but significant correlation with the overall SNP density in each bin (Spearman's rank correlation $\rho = 0.16$, $p < 10^{-12}$). A very similar trend is observed ($\rho = 0.21$, $p < 10^{-12}$) for discordant SNPs. Therefore, the No Call and discordant SNPs are distributed across the whole genome as a function of SNP coverage on the microarray.

Sensitivity vs. Specificity of Different BRLMM Calling Confidence Thresholds

In the BRLMM calling process, a confidence score (C-score) threshold is used to determine the genotype calls. To investigate the effect of different thresholds on the overall call rate and the concordance rate between gDNA and WGA samples, we chose different C-score threshold varies from 0.5 to 0.05 and calculated the corresponding average call rate and concordance rate for the four WGA samples. As expected, using a more stringent threshold increases the concordance rates between samples but decreases the call rates (Figure 3). We also analyzed the two technical duplicates in the same procedure, and a very similar trend was observed (Figure 3, red curves). This result suggests that although WGA samples have a lower overall call rate, their C-score distributions are similar to those of gDNA samples.

Discussion

In this study, we investigated the performance of whole-genome amplified DNA on a commonly used Affymetrix 250K array. Because the DNA samples used in this study are more than 10 years old and have been refrigerated for most of this time, our findings should be applicable to DNA samples stored by most standard methods in common laboratories. The average WGA product call rate (97.5%) in this study is generally higher than those of previous studies using Affymetrix 10K SNP arrays [6;7] and Affymetrix 500K SNP arrays [10], in which call rates of WGA products varied between ~63% to ~97%. The higher call rate in this study is partly due to different quality of the starting DNA (e.g. some ~20 years old plasma and serum samples were used in [10]), and partly due to the different genotype-calling algorithms. The BRLMM calling algorithm used in this study has been shown to provide significant improvement over DM algorithm on the 500K array [9]. The average concordance rate (98.7%) in our samples is comparable to previous studies, in which concordance rates range from ~81% to ~99% [6;7;10]. It is noteworthy that, although our WGA samples showed lower call rate compared to their corresponding genomic DNA samples, call rates of WGA samples are not always lower than genomic DNAs. Among the 276 blood samples analyzed using the same protocol in our laboratory that have passed the 93% DM call rate threshold, 19 (7%) has BRLMM call rates below the 97.5% average call rate of the four WGA samples in this study.

In addition, a recent study showed that imputation might be used to further improve the performance of WGA products [11].

When the genomic distribution of No Call SNPs and discordant SNPs were examined, we did not find any apparent clustering patterns. Only 144 SNPs failed in all four WGA experiments (0.68% of SNPs that failed in any of the WGA experiment) and a total of 169 SNPs (1.23% of all discordant SNPs) are discordant in all four comparisons. In general, the genomic distribution of No Call and discordant SNPs correlates with the SNP coverage of the genome on the array. The random distribution and the lack of overlap of failed SNPs in different samples suggest that most of the failures are due to variability among experiments rather than systematic WGA amplification problems. Among SNPs that have discordant genotypes between gDNA and WGA samples, we found an increase in heterozygotes in WGA samples. This result is different from a previous study on the Affymetrix 10K array [6] in which most of the discrepancies are “loss of heterozygosity” (AB to AA/BB). This difference may due to the different microarray design (10K vs. 250K) and the genotyping-calling algorithm used to generate genotypes [6]. The BRLMM is known to improve the performance of heterozygotes [9] and other studies using the Affymetrix 500K array did not report a deficit of heterozygotes [10;11].

It should be noted that the performance of the BRLMM calling algorithm is thought to increase when more CEL files are analyzed together [9]. In the description of BRLMM, at least six CEL files are required for the run, and it is recommended to run more than 200 samples at a time to obtain optimal results. In our study, we analyzed 12 CEL files from 6 unrelated individuals. It is conceivable that increasing the sample size could result in better concordance rates. To test this hypothesis, we analyzed the current 12 samples with another 336 CEL files that were generated using the same protocol and compared the resulting call rates and concordance rates. In WGA samples, we observed a slight decrease in call rate (on average 2.31%) but no appreciable increase in the average concordance rate (0.13%). Therefore, despite the small sample size, our study closely represents the general quality of genotype-calling with the BRLMM algorithm.

In summary, we found that WGA products are capable of producing high-quality results on the Affymetrix 250K SNP array. Using the BRLMM calling algorithm, WGA products have high call rates and excellent concordance rates with their corresponding genomic DNA samples. In addition, no apparent amplification bias across the genome can be detected for WGA products. These results suggest that WGA is a promising solution for researchers and clinicians who want to perform large scale array-based genotyping but with limited amounts of DNA samples.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Diane Dunn and Edward Meenen for their technical support in the microarray hybridization and scanning process and Dr. Elizabeth E. Marchani for her useful comments. This work was supported by grants from the National Science Foundation (BCS-0218370), and National Institutes of Health (GM-59290 and HL-070048).

References

1. Kruglyak L, Nickerson DA. Variation is the spice of life. *Nat Genet* 2001;27:234–236. [PubMed: 11242096]
2. Reich DE, Gabriel SB, Altshuler D. Quality and completeness of SNP databases. *Nat Genet* 2003;33:457–458. [PubMed: 12652301]

3. International HapMap Consortium. The International HapMap Project. *Nature* 2003;426:789–796. [PubMed: 14685227]
4. 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]
5. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, Lasken RS. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* 2002;99:5261–5266. [PubMed: 11959976]
6. Tzvetkov MV, Becker C, Kulle B, Nurnberg P, Brockmoller J, Wojnowski L. Genome-wide single-nucleotide polymorphism arrays demonstrate high fidelity of multiple displacement-based whole-genome amplification. *Electrophoresis* 2005;26:710–715. [PubMed: 15690424]
7. Paez JG, Lin M, Beroukhi R, Lee JC, Zhao X, Richter DJ, Gabriel S, Herman P, Sasaki H, Altshuler D, Li C, Meyerson M, Sellers WR. Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res* 2004;32:e71. [PubMed: 15150323]
8. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen MM, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S. Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* 2005;21:1958–1963. [PubMed: 15657097]
9. Affymetrix, BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500K array set. 2006. http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf
10. Croft DT Jr, Jordan RM, Patney HL, Shriver CD, Vernalis MN, Orchard TJ, Ellsworth DL. Performance of whole-genome amplified DNA isolated from serum and plasma on high-density single nucleotide polymorphism arrays. *J Mol Diagn* 2008;10:249–257. [PubMed: 18403606]
11. Teo YY, Inouye M, Small KS, Fry AE, Potter SC, Dunstan SJ, Seielstad M, Barroso I, Wareham NJ, Rockett KA, Kwiatkowski DP, Deloukas P. Whole genome-amplified DNA: insights and imputation. *Nat Methods* 2008;5:279–280. [PubMed: 18376389]

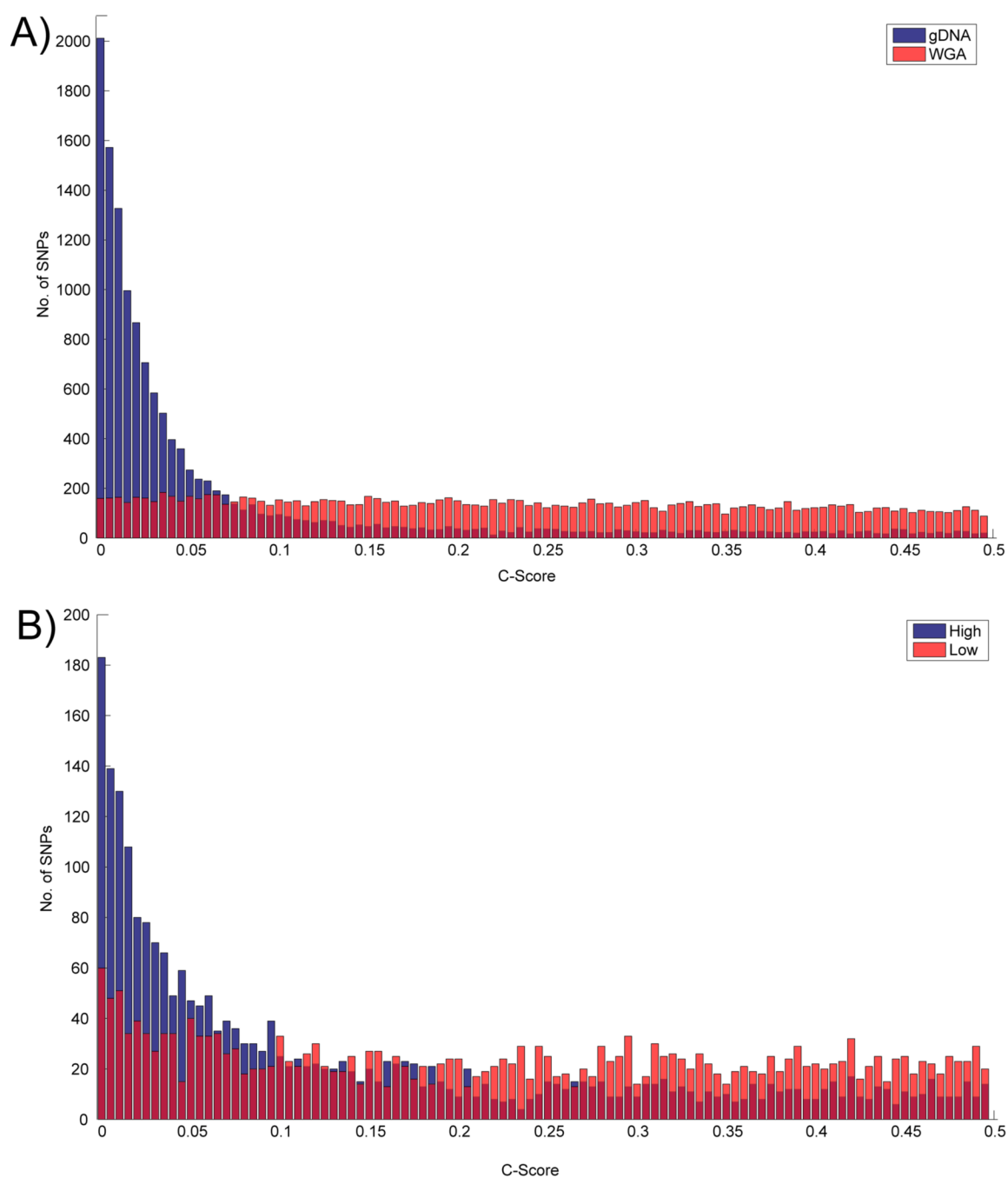


Figure 1. C-score distribution of discordant SNPs

Histograms of C-scores of discordant SNPs are shown. C-scores are shown on the x-axis and the number of SNPs is shown on the y-axis. **A)** All discordant SNPs between the four gDNA/WGA sample pairs. C-scores of the gDNA samples are shown in blue and C-scores of the WGA samples are shown in red. **B)** Discordant SNPs between the two technical duplicates. C-scores of the experiments with higher call rate in the duplicate are shown in blue and C-scores of the experiments with lower call rate in the duplicate are shown in red.

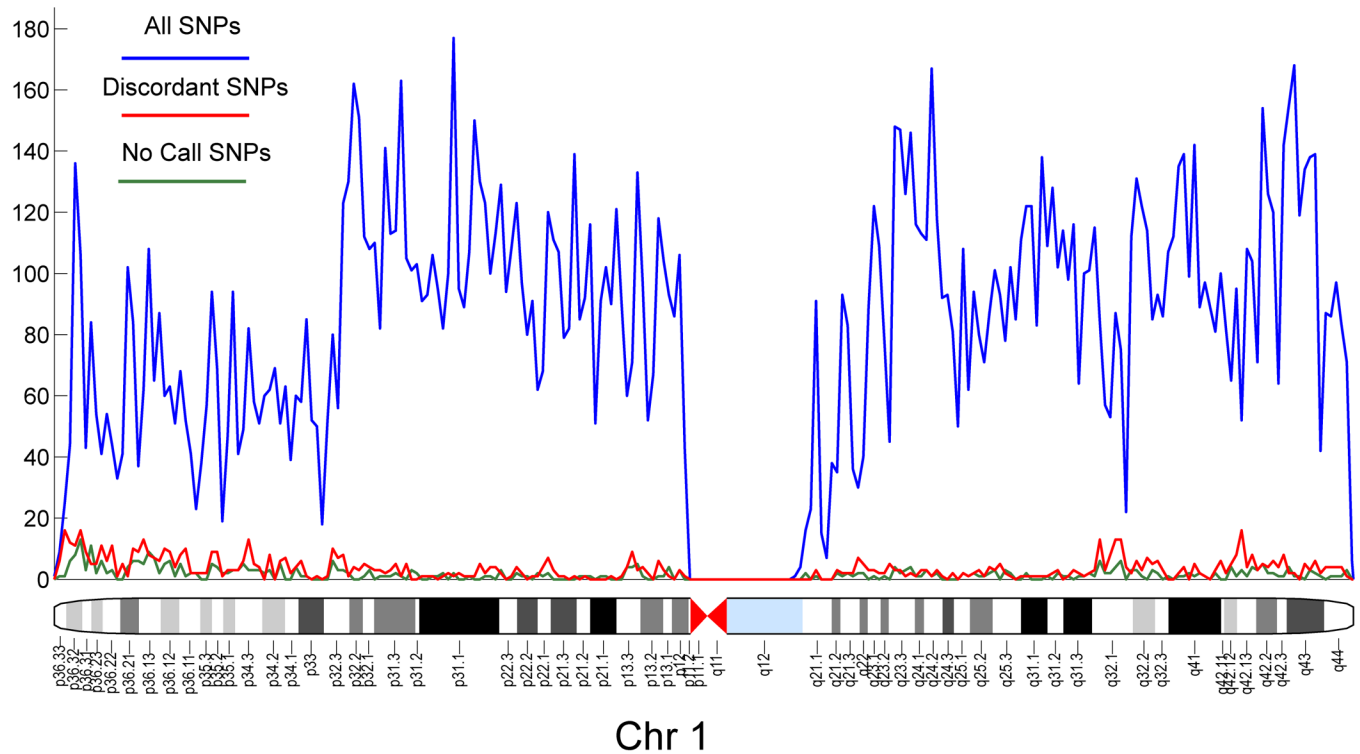


Figure 2. Distribution of No Call SNPs and discordant SNPs on chromosome 1

The number of SNPs in 1Mb bins on chromosome 1 is shown. Blue, total number of SNPs on the 250K array; green, number of SNPs that are No Call in at least three WGA samples; red, number of SNPs that are discordant in any of the four gDNA/WGA pairs.

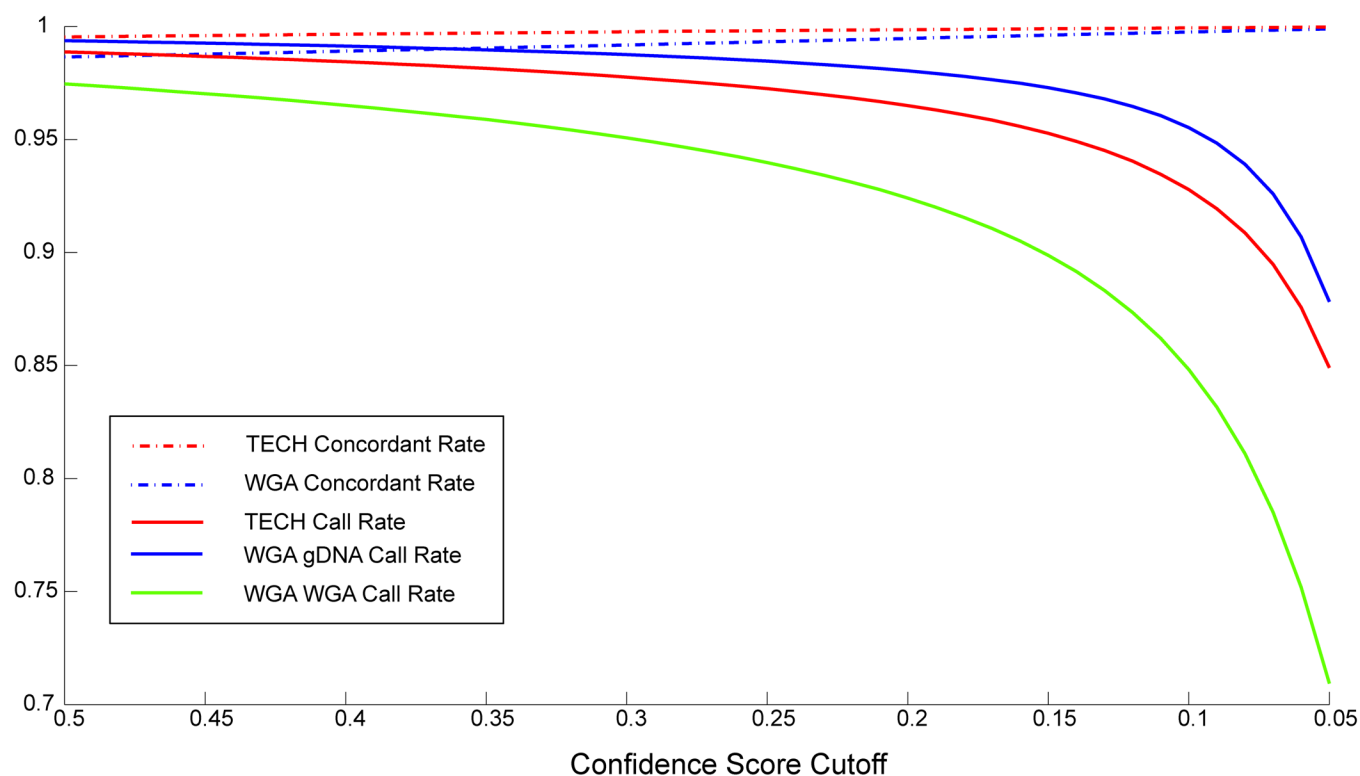


Figure 3. Performance of BRLMM at different C-score thresholds

Call Rates and Concordant rates for all six sample pairs are plotted against C-score thresholds varies from 0.5 to 0.05. Average concordance rates for the two technical duplicate pairs and the four gDNA/WGA pairs are shown in red and blue dashed lines respectively. Average call rates for four technical duplicate samples, four gDNA samples and four WGA samples are shown in red, blue and green solid lines respectively.

Table 1

SNP Call Rates on the 250K Array

Whole-genome Amplification		DM Call Rate (%)	Homozygotes (%)	Heterozygotes (%)	BRLMM Call Rate (%)	Homozygotes (%)	Heterozygotes (%)
Genomic DNA	WGA1	97.6	75.5	24.5	99.3	74.3	25.7
	WGA2	95.8	75.6	24.4	99.0	73.0	27.0
	WGA3	98.7	73.6	26.4	99.7	73.0	27.0
	WGA4	97.7	74.6	25.5	99.5	73.1	26.9
	Mean	97.5	74.8	25.2	99.4	73.4	26.6
	SD	1.2	0.9	0.9	0.3	0.6	0.6
WGA DNA	WGA1	94.3	75.7	24.3	97.4	73.4	26.6
	WGA2	95.8	74.3	25.7	98.0	72.3	27.7
	WGA3	95.0	74.9	25.1	97.1	71.9	28.1
	WGA4	95.3	74.1	25.9	97.4	72.1	28.0
	Mean	95.1	74.8	25.2	97.5	72.4	27.6
	SD	0.6	0.7	0.7	0.4	0.7	0.7
Technical Duplicates							
	Tech1A	97.1	74.1	26.0	99.3	72.8	27.2
	Tech2A	96.5	73.7	26.3	98.8	72.5	27.5
	Tech1B	94.7	75.3	24.7	98.3	72.7	27.4
	Tech2B	97.8	73.9	26.1	99.3	73.2	26.9
	Mean	96.5	74.2	25.8	98.9	72.8	27.2
	SD	1.4	0.7	0.7	0.5	0.3	0.3

Table 2
No Call SNPs in the Whole-genome Amplified Samples

Genomic	WGA1	WGA2	WGA3	WGA4	Mean	SD
AA (%)	37.7	37.1	37.1	37.2	37.3	0.3
AB (%)	25.7	27.0	27.0	26.9	26.6	0.6
BB (%)	36.6	35.9	35.9	36.0	36.1	0.3
Total SNPs *	260400	259728	261452	260889		
WGA Product						
AA -> No Call (%)	37.1	35.1	36.0	36.6	36.2	0.9
AB -> No Call (%)	21.3	24.4	25.4	23.4	23.6	1.7
BB -> No Call (%)	41.6	40.4	38.6	40.0	40.2	1.2
Total No Call SNPs **	6495	4994	7477	6546		

* Total Number of called SNPs in each gDNA sample.
** Total Number of No Call SNPs in each WGA sample that have genotype calls in their corresponding gDNA samples

Table 3

Distribution of discordant SNPs

Whole-genome Amplification	WGA1	WGA2	WGA3	WGA4
Overall Concordance Rate	98.79%	98.89%	98.37%	98.56%
Total Discordant SNPs:				
Het -> Homo *	3069	2823	4149	3671
Homo -> Homo *	659	663	712	658
Homo -> Het *	16	23	73	63
Heterozygotes in gDNA	2394	2137	3364	2950
Heterozygotes in WGA	21.47%	23.49%	17.16%	17.92%
	78.01%	75.70%	81.08%	80.36%
Technical Duplicates				
	Tech1A/1B	Tech2A/2B		
Overall Concordance Rate	99.55%	99.51%		
Total Discordant SNPs:				
Het -> Homo *	1146	1255		
Homo -> Homo *	401	248		
Homo -> Het *	4	6		
Heterozygotes in High **	741	1001		
Heterozygotes in Low **	34.99%	19.76%		
	64.66%	79.76%		

* In whole genome amplification experiments, the genotype changes are calculated based on the genomic DNA results. In technical duplicates, the number of genotype changes is calculated based on the experiments with higher call rates. Homo: Homozygotes, Het: Heterozygotes.

** Heterozygotes in High is the percentage of heterozygotes of the two experiments with higher call rates in technical duplicates and Heterozygotes in Low is the percentage of heterozygotes of the two experiments with lower call rates.