Accuracy of allele frequency estimates in pool DNA analyzed by high-density Illumina Human 610-Quad microarray

P Janicki, J Liu

Citation

P Janicki, J Liu. *Accuracy of allele frequency estimates in pool DNA analyzed by high-density Illumina Human 610-Quad microarray*. The Internet Journal of Genomics and Proteomics. 2008 Volume 5 Number 1.

Abstract

Purpose: DNA pooling method has been previously proven successful for allele frequency estimation using earlier versions of Illumina Beadchip and Affymetrix GeneChip microarrays. The aim of this study was to evaluate the accuracy of the DNA pooling using the High Density Illumina Infinium microarray system.Patients & methods: The accuracy of three different methods of allele frequency estimation on the High Definition Infinium II Beadchip (Human 610-Quad) was tested using the DNA samples obtained from saliva collected from seven Caucasian subjects who had been individually genotyped. Our first approach was to calculate estimated allelic frequency (expressed as B allele frequency) from pooled DNA sample based on the correlation coefficient obtained for individual SNPs, and to correlate it with the true (real) allele frequency computed using individually genotyped samples. The second approach was to estimate SNP frequency in the pooled DNA sample as calculated automatically by the proprietary Beadstudio software, which is based on the internal calibration and normalization procedures. The third approach was to study the pool allele frequency calculated using the estimation of correlation coefficient from microarray SNP sub-pools.Results and conclusions: All analyzed methods of allelic frequency estimates from pooled DNA produced similar results in respect to the calculated correlation coefficient (R) (R ranging from 0.947 to 0.958 with R>0.99 for correlation between each three of them). All three methods of calculations seem to be sufficient and equally effective for calculation of individual allelic frequency for the given SNPs from the pooled DNA.

INTRODUCTION

Conceptualized a decade ago, genome-wide association studies (GWAS) have only recently become practical in large samples. Costs are still a limiting factor. An initial screen which measures allele frequencies between pools of cases and controls, instead of between individuals, would greatly reduce costs of GWAS. DNA pooling has been used successfully on several platforms including Illumina microarray BeadChips [1234567], but its accuracy and reliability has not been yet demonstrated on the high density Illumina Infinium system. This assay is based on allele-specific hybridization coupled with primer extension of genomic DNA by primers directly surrounding the SNP on randomly ordered bead arrays [89]. The Infinium assay has been further developed into allele-specific single base extension using two colors labeling with the Cy3 and Cy5 fluorescent dyes (Infinium II). Infinium II is a two-channel assay and data consist of two intensity values (x, y) for each SNP, with one intensity channel for each of the fluorescent dyes associated with the two alleles of the SNP. The alleles measured by the x channel (Cy5 dye) are arbitrarily, with respect to haplotypes, called the A alleles, whereas the alleles measured by the y channel (Cy3 dye) are called the B alleles. All SNP markers are present at a high redundancy on Infinium II assays and the allele specific intensities are summarized estimates from replicate markers. Current generations of high density Infinium II arrays are able to interrogate more than 1 million SNPs simultaneously.

In the present pilot project we assessed the accuracy of the B allele frequency estimates from pooled DNA samples analyzed on the High Definition Infinium II BeadChip (Human 610-Quad) (by comparing the estimates obtained using three different methods of calculations with the allele frequencies computed using individually genotyped DNA samples.

MATERIAL AND METHODS

Samples consisted of 7 unrelated individual saliva samples

(2 ml) collected into the Oragene (DNA Genotek, Canada) collection containers containing 2 ml of the proprietary preservative. The containers were kept at room temperature until DNA extraction. DNA was extracted according to the Puregene DNA purification kit manual. The concentration of DNA in the final elute was measured in duplicates using the picoGreen method and ranged from 60 to 300 ng/ l. The DNA samples were then diluted with appropriate amount of Tris buffer to obtain the final DNA concentration of 50 ng/ l in each sample. The equimolar pool was created by combining the same amount of diluted individual DNA samples producing the final pooled DNA sample with the concentration of 50 ng/ l re-confirmed again with the picoGreen method.

All DNA samples (7 individual and 1 pooled samples) were analyzed using two Illumina Human 610-Quad microarray BeadChips (each containing 4 sample ports) and processed according to the manufacturer recommendations. The final processed microarrays were read by the Laser Scaning device and results processed by the Illumina Beadstudio software (ver. 3.2).

The recorded output variables for each analyzed SNP on the BeadChip consisted of raw fluorescence readings for x and y channels (x raw and y raw), internally normalized x and y (x norm and y norm), typed genotype (AA, BB or AB), as well as calculated B allele frequency. The reported raw beadscore for x and y and for each SNP represents the composite value of up to 64 PAF estimates from individual beads (range 4-64) per SNP, as calculated by the Beadstudio software (i.e., no individual reading were available for separate beads, as this would require modification of the calculation option in the software, not available at the time of data collection).

Initial number of available SNPs to analyze on each chip was 620,901. The SNPs containing 21890 monoallelic copy number variant (CNV) sites were then removed from further analysis leaving 599,011 SNPs. From this number, further 114,854 SNPs were observed to be exclusively homozygotic in all 7 individuals and therefore these samples were also removed from further analysis leaving us with the final total number of 484,157 SNPs which were subject to all subsequent calculations.

Method 1: Estimation of allelic frequency in the pool DNA calculated using the calibration coefficient obtained for each heterozygous SNP in seven individual samples genotyped separately.

The obtained data included the composite red (x raw) and green (y raw) beadscores for each analyzed SNP, as well as the results of the genotyping for each individual SNP in the individual and pool DNA samples. The average y raw/x raw ratio (k) for the heterozygotic SNP in all individuals was calculated and used for the transformation of the x raw into calibrated x (kx) for each corresponding SNP in the pool DNA. The estimated B allele frequency in pool DNA was then calculated from the ratio calibrated y, B frequency = y raw/(y raw + kx) for each SNP in the pool DNA.

Method 2. Estimation of allelic frequency in the pool DNA calculated automatically by the Beadstudio software and based on the internal calibration and normalization procedures (no prior individual genotyping necessary)

The generation of B allele frequency was performed automatically using the self-normalization algorithm built in the BeadStudio software (Genotyping Module ver. 3.2). Illumina's standard normalization algorithm is implemented as the first step in the SNP genotyping data analysis. The intensity data (x raw and y raw) are normalized automatically when they are loaded into Illumina's BeadStudio software. The normalization algorithm is designed to adjust for channel-dependent background and global intensity differences, and to scale the data. It is important to note that the normalization process uses the information that links a bead type to a sub-bead pool. Illumina uses a 6-degree of freedom affine transformation to normalize sample intensities. The six parameters are offset_x, offset_y, theta, shear, scale_x, and scale_y. The normalization process consists of five main steps: Outlier removal; Background estimation (offset_x, offset_y); Rotational estimation (theta); Shear estimation (shear); 5. Scaling estimation (scale_x, scale_y). Briefly, within each sub-beadpool, outlier SNPs are removed if their allelic intensities are smaller than either the 5th smallest or 1st percentile as compared to all SNPs, or if their intensities are larger than the 5th largest or 99th percentile as compared to all SNPs. Background estimation occurs by uniform sampling of 400 points along each intensity axis to create a linear fitting to candidate homozygotes. The intercept of the linear fittings from both homozygotes then defines the origin. Rotation and shear of the data points by the same uniform sampling then occurs with respect to this defined origin. The final normalized intensities are then determined by mean scaling via virtual control points. This procedure at present occurs automatically within the Illumina BeadStudio

software and outputs the normalized intensities (x norm and y norm), which provide a pair of coordinates corresponding to the signals for the two alleles at each SNP.

After normalization, data should be as canonical as possible with homozygous SNPs positioned along the transformed X and Y intensity axes. To visualize the data after normalization, the genotyping data are transformed to a polar coordinate plot of normalized intensity R = Xnorm + Ynorm and allelic composition (copy angle), using the equation theta = (2/pi)*arctan2 (Ynorm,Xnorm), where Xnorm and Ynorm represent transformed normalized signals from alleles A and B for a particular locus.

The B Allele frequency in this method of calculations represents the theta value for a SNP, corrected for cluster position. Cluster positions are generated from a large set of normal individuals. Standard cluster files provided with Infinium products identify expected intensity levels of genotype classes for each SNP. Comparing sample intensities to this cluster file is usually sufficient for generating extremely high quality data for human projects. Standard cluster files for standard human Infinium products are created using a diverse set of over 100 samples from the Caucasian (CEU), Asian (CHB+JPT), and Yoruban (YRI) HapMap populations, and therefore should incorporate much of the genetic diversity in these populations.

The B Allele Frequency can also be referred to as “copy angle” or “allelic composition. B allele freq is described by the following equation.

B allele frequency = 0 if theta < tAA, or B allele freq = 0.5 * (theta - tAA) / (tAB - tAA) if theta < tAB, or B allele freq = 0.5 + 0.5 * (theta - tAB) / (tBB - tAB) if theta < tBB, or B allele freq = 1 if theta >= tBB

where:

tAA = mean theta value of all genotypes in the AA cluster plotted in polar normalized coordinates

tAB = mean theta value of all genotypes in the AB cluster plotted in polar normalized coordinates

tBB = mean theta value of all genotypes in the BB cluster plotted in polar normalized coordinates

Method 3. Estimation of allelic frequency in pool DNA calculated using the approximation of correlation coefficient from each microarray sub-pools (total of 24) by the method described by MacGregor et al. (2008). No prior individual genotyping necessary or internal automatic normalization were required for this method of calculations.

This approach is based on the observation that green (Y) channel beadscores tended to be larger than red (x) beadscores and that the magnitude of these differences varies between different sub-pools of SNP on the chip. In other words, the calibration of the beadscores should be performed not over the entire microarray content, but instead separately for each sub-pool of SNPs on the chip. Each microarray is divided into a number of sub-bead pools (25 sub-bead pools for the 610-Quad chip) and normalization of the bead intensities occurs at the sub-beadpool level to adjust for channel-dependent background and global intensity differences, and to scale the data. A sub-bead pool is a set of beads that were manufactured together and are located in roughly the same analytical location (stripe) on a BeadChip. For each sub-pool the calibration was performed by re-scaling the red (x) channel raw beadscore (x_raw) to make the mean value of the pooling allele frequency (PAF) = 0.5 (over all SNP in this sub-pool). The PAF for each SNP were then computed as the corrected red intensity divided by the total (corrected red plus green) intensity. Green channel intensities were available as the y-raw in the data output from the microarray scan.

PAF = k x_raw / (k x_raw + y_raw); where k indicates the correction factor.

The correlation factors were calculated separately for each of 24 sub-pools on microarray. The calculated correction factor was then used for computation of the allele frequency estimates for each SNP in the analyzed sub-pool on microarray.

Statistical calculations. The statistical analysis of the results included correlation analysis of the real B allele frequencies (ie. frequencies obtained from the individually genotyped samples) vs. estimated B allele frequencies for each analyzed method of calculations across all analyzed SNP on microarray. All calculations were performed using the SPSS ver. 16.01 software for Windows PC (SPSS Inc., Chicago IL).

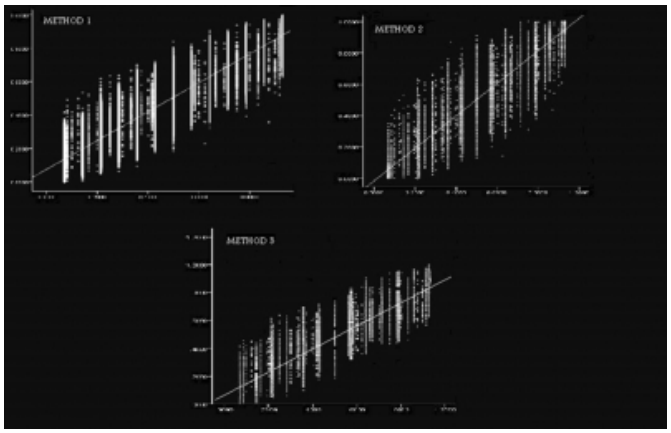
RESULTS

All analyzed methods of calculation of real B freq produced very similar results in respect to the calculated correlation coefficient (R). The graphs presenting the original plots of

calculated B allele frequency versus true (or real) B allele frequency are shown below for all three methods (Fig. 1).

Figure 1

Fig. 1 Scatter plot of calculated allele frequencies calculated with different methods versus real B allele frequencies



Horizontal axis - True (real) B allele frequency as calculated from individual genotypes in 7 samples

Vertical axis - Estimated B allele frequency in pool DNA sample

The accuracy data are presented as R, R_2 and SEE (standard error of estimation) in the Table 1 and show a linear correlation between frequency of all tested SNP in pool (calculated by 3 different methods) and real frequency for the analyzed alleles obtained from seven individual samples.

Figure 2

Table 1. The accuracy data for three different methods of calculation of B frequency in pool DNA as compared with real allelic frequencies obtained from the individually genotyped samples.

Parameter/Method	Method 1	Method2	Method 3
R	0.956	0.958	0.947
R^2	0.914	0.918	0.896
SEE	0.07	0.085	0.08

R - correlation coefficient, R^2 - squared correlation coefficient, SEE - standard error of estimation

DISCUSSION

All analyzed methods of calculations seem to be sufficient and equally effective (R>0.99 for correlation between three of them) for calculation of individual allelic frequency of analyzed SNP from the pool DNA.

It is noteworthy that the analyzed methods of calculation differ significantly in respect to the need for either previous genotyping of individuals (in order to obtain correlation coefficient data for SNP) (e.g., method 1) or labor-intensive calculations for the estimation of the correlation coefficient from the sub-pools (e.g., method 3). In contrast to method 1, methods 2 and 3 do not require previous calibration with individual samples.

It appears that the method 2 is the simplest one to employ for routine calculations as it relies completely on the automatic, internal algorithm in the BeadStudio software without any additional need for either previous individual genotyping or complicated internal calculations for correlation coefficient in sub-pools on the microarray. The attractiveness of this approach is also based on the fact that the B frequency values are obtained automatically in the BeadStudio software without any need for additional calculations, and as such could be used immediately for the comparison of B allele frequencies between different experimental pools (which usually constitutes the main purpose of the pooling method).

This conclusion was supported recently for the Illumina Human 300 genotyping by Sebastiani et al (2008) who reported that when B allele frequencies were estimated (in duplicates) using the BeadStudio software, they obtained the correlation coefficient in the range of 0.98-0.999 (depending on the number of individual DNA samples in the pool, ranging from 30 to 60 individuals).

I summary, we compared the accuracy of several different methods of estimation of B allele frequency from the pooled DNA sample. Our results show that all three analyzed methods of calculation seem to be sufficient and equally effective for estimation of allelic frequency for the given SNP from the pooled DNA sample.

ACKNOWLEDGEMENTS

We would like to thank, Dr. Willard Freeman, Rob Brucklacher and Georgina Bixler from the Functional Genomics Core Facility of the Section of Research Resources, Penn State College of Medicine for their help in the genotyping part of the study.

References

1. Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, Georgieva L, Dowzell K, Cichon S, Hillmer AM, O'Donovan MC, Williams J, Owen MJ, Kirov G: A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. BMC Medical Genomics 2008, 1: 44
2. Kirov G, Zaharieva I, Georgieva L, Moskvina V, Nikolov

I, Cichon S, Hillmer A, Toncheva D, Owen MJ, O'Donovan MC: A genome-wide association study in 574 schizophrenia trios using DNA pooling. Molecular Psychiatry 2008,; doi:10.1038/mp.2008.33.
3. Macgregor S: Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. European Journal of Human Genetics 2007, 15: 501-4.
4. Macgregor S, Visscher PM, Montgomery G: Analysis of pooled DNA samples on high density arrays without prior knowledge of differential hybridization rates. Nucleic Acids Research 2006, 34: e55.
5. Macgregor S, Zhao ZZ, Henders A, Nicholas MG, Montgomery GW, Visscher PM: Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. Nucleic Acids Research 2008, 36: e35.
6. Sebastiani P, Zhao Z, Abad-Grau MM, Riva A, Hartley SW, Sedgewick AE, Doria A, Montano M, Melista E, Terry D, Perls TT, Steinberg MH, Baldwin CT: A hierarchical and modular approach to the discovery of robust associations in genome-wide association studies from pooled DNA samples. BMC Genetics 2008, 9: 6.
7. Steer S, Abkevich V, Gutin A, Cordell HJ, Gendall KL, Merriman ME, Rodger RA, Rowley KA, Chapman P, Gow P, Harrison AA, Highton J, Jones PB, O'Donnell J, Stamp L, Fitzgerald L, Iliev D, Kouzmine A, Tran T, Skolnick MH, Timms KM, Lanchbury JS, Merriman TR: Genomic DNA pooling for whole-genome association scans in complex disease: empirical demonstration of efficacy in rheumatoid arthritis. Genes and Immunity 2007, 8: 57-68.
8. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS: A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 2005, 37:549-554.
9. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL: High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 2006, 16:1136-1148

Author Information

Piotr K. Janicki, MD, PhD, DSci
Professor, Laboratory of Perioperative Genomics, Department of Anesthesiology Pennsylvania State University College of Medicine

Jiabin Liu, MD, PhD
Resident Physician, Laboratory of Perioperative Genomics, Department of Anesthesiology Pennsylvania State University College of Medicine