Poster K-4

Genotyping and Annotation of Affymetrix SNP Arrays

Authors:
Philippe Lamy (*Bioinformatics Research Center, University of Aarhus*)
Carsten Wiuf (*Bioinformatics Research Center, University of Aarhus*)
Claus L. Andersen (*Molecular Diagnostic Laboratory, Aarhus University Hospital, Skejby*)
Friedrik P. Wikman (*Molecular Diagnostic Laboratory, Aarhus University Hospital, Skejby*)

Short Abstract: We have developed a new method for genotyping Affymetrix SNP array. The method is based on using multiple arrays at the same time to determine the genotypes and on a model that relates intensities of individual SNPs to each other. It is therefore possible to annotate SNPs that experimentally perform badly.

Long Abstract:

Introduction

Affymetrix SNP arrays have become popular and are widely used. Originally, an array with 1500 SNP was released, later the 10k SNP array followed and quite recently arrays with up to 500k SNPs have been made available. The array technique is based on genomic hybridization to synthetic high density oligonucleotide microarrays. Each of the two alleles of a SNP is represented by 10 or 14 oligonucleotides (together called a probe set) and hybridization (probe) intensities are measured for all probes in a probe set. Affymetrix has developed software (GDAS) for genotyping SNPs based on the intensities and subsequently the derived genotypes can be used in the further analysis of the data. GDAS genotypes SNPs arraywise, one SNP at a time. For the larger arrays, Affymetrix has developed a new dynamic model-based algorithm (DM) that also is based on arraywise genotyping. Here we present an alternative algorithm (PBG – Pool Based Genotyping) for genotyping Affymetrix SNP arrays. If allele intensities (probe intensities combined into one value for each allele) are plotted for a typical SNP, three distinct clusters are generated that correspond well to the three possible genotypes. Naturally, this suggests that the genotype of a SNP could be derived from the distribution of allele intensities by choosing the genotype of the cloud that statistically (in some sense) is closest to the observed allele intensities. PBG builds on this observation. In addition, we base PBG on a model that allow identification and annotation of SNPs that either are difficult to genotype correctly for experimental reasons or have probes that are not suited for copy number analysis.

Material and Methods

For this study we used 113 samples collected at Aarhus University Hospital, Skejby. The GeneChip®Mapping 10k Early Access Array was applied to all 113 samples. This array has 10126 SNPs. Of these 9600 (9430 autosomal and 170 X chromosomal) mapped to a unique position in the genome (using the April 2003 genome assembly (hg15), http://www.genome.ucsc.edu). The probe set intensities were normalized using the dChipSNP software. Then they were combined into two values by taking the logarithm of the average over all probes for the each allele, A or B.

Our model is based on a relationship between the intensities of 0 or 2 copies of one allele and a single copy of the same allele. This relationship is not SNP dependant.

Our method comprises two steps: the first step is to iteratively fit the different parameters from our model and the second is to genotype the SNPs. The two steps are run iteratively until no more (few) changes in genotypes occur. It also requires an initial clustering. In that study, we used the GDAS genotyping.

We defined different measures in order to compare the two algorithms. First, for each SNP, we tested whether the genotype assignments complied with Hardy Weinberg equilibrium. Second, we defined a weighted Euclidian distance between the observed means and the expected means for the A and the B intensities alone and for both jointly. Thus, we were able to flag alleles and/or SNPs which do not fit the model.

Results and Discussion

We have developed a new method for genotyping Affymetrix SNP arrays and compared the performance of our method (PBG) to that of Affymetrix (GDAS). PBG is based on analysing multiple arrays at the same time, in contrast to GDAS that analyses SNPs arraywise, one SNP at a time. Generally, the two methods agree (99.25% when a call is given by the two methods), but PGB appears to be able to genotype correctly with a lower no call rate (1% vs 6%) and also appears to produce more genotypes than GDAS that comply with Hardy-Weinberg equilibrium. In addition, PBG is based on a model that relates allele intensities from different SNPs to each other. We use this relationship to annotate SNPs and alleles.

Our method is based on dChipSNP normalized probe intensities. One array is selected as reference array and all other arrays are normalized relatively to the reference array. This has the advantage that new arrays (a test set) can be genotyped using fitted parameters obtained from a training set.

If the test set is normalized relatively to the reference array of the training set the fitted parameters of the training set can be used to genotype the test set. Particularly, this should be useful when genotyping only few arrays, provided the fitted parameters of the test set and the reference array is publicly available. We showed that this approach is feasible by analysing an additional 10 arrays that was not used for fitting.

Some SNPs, that do not fit the model, are flagged as ’poor’ and they can be excluded from the analysis. Flagging or annotation of ’poor’ performing SNPs is a two-sided advantage. First of all, SNPs that perform ’poor’ because of experimental reasons can be excluded from the analysis. Secondly, SNPs can be ’poor’ performing because for one or both alleles the probes do not behave in a dose-response manner and should therefore be excluded. These SNPs might still be genotyped correctly, but are not suitable for copy number analysis.

In general, genotyping and copy number analysis are separate issues ; i.e. if genotypes are used in a copy number analysis the genotypes are obtained before the copy number analysis is conducted. It would be natural to combine the two into a single analysis. We showed that the level of the A-intensity is not affected by the copy number of the B allele, and vice versa. This leads us to speculate that cross hybridization can be ignored generally in the sense that the level of the A-intensity is only affected by the copy number of the A allele, not the copy number of the B allele. Assuming a linear relationship between log-copy number and log-intensity, the intensity levels for higher allele copy numbers could be extrapolated from the observations made in this study.