Computers in Biology and Medicine 61 (2015) 48–55

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Allele frequency calibration for SNP based genotyping of DNA pools: A regression based local–global error fusion method

Ashfaqur Rahman [a,*], Andrew Hellicar [a], Daniel Smith [a], John M Henshall [b]

[a] Digital Productivity Flagship, CSIRO, Hobart, Tasmania, Australia
[b] Agriculture Flagship, CSIRO, Armidale, NSW, Australia

ARTICLE INFO

Article history:
Received 29 July 2014
Accepted 17 March 2015

Keywords:
Allele frequency calibration
DNA pooling
SNP genotyping
Microarray
Machine learning

ABSTRACT

Background: The costs associated with developing high density microarray technologies are prohibitive for genotyping animals when there is low economic value associated with a single animal (e.g. prawns). DNA pooling is an attempt to address this issue by combining multiple DNA samples prior to genotyping. Instead of genotyping the DNA samples of the individuals, a mixture of DNA samples (i.e. the pool) from the individuals is genotyped only once. This greatly reduces the cost of genotyping. Pooled samples are subject to greater genotyping inaccuracies than individual samples. Wrong genotyping will lead to wrong biological conclusions. It is thus required to calibrate the resulting genotypes (allele frequencies).
Methods: We present a regression based approach to translate raw array output to allele frequency. During training, few pools and the individuals that constitute the pools are genotyped. Given the genotypes of individuals that constitute the pool, we compute the true allele frequency. We then train a regression algorithm to produce a mapping between the raw array outputs to the true allele frequency. We test the algorithm using pool samples withheld from the training set. During prediction, we use this map to genotype pools with no prior knowledge of the individuals constituting the pools.
Results and discussion: After data quality control we have available a dataset comprised of 912 pools. We estimate allele frequency using three approaches: the raw data, a commonly used piecewise linear transformation, and the proposed local–global learner fusion method. The resulting RMS errors for the three approaches are 0.135, 0.120, and 0.080 respectively.
Crown Copyright © 2015 Published by Elsevier Ltd. All rights reserved.

1. Introduction

Genotyping refers to the process of determining the differences in the genetic make-up (i.e. genotype) of the individuals of a species. SNPs (Single Nucleotide Polymorphisms) are one of the most common forms of genetic variation. SNP genotyping refers to the measurement of genetic variation of SNPs. A SNP refers to a DNA sequence variation at a specific locus (a specific location on gene) where a Single Nucleotide (A, T, C or G) in the genome differs between members of a species. These genetic variations within humans are commonly utilized in DNA fingerprinting (used in forensic science for identification of individuals by their DNA profile) or to provide genetic linkage to a disease to assist in drug development or identify an individual's susceptibility to the disease. Whilst there are a number of technology platforms available for SNP genotyping studies, the choice of platform is largely dependent upon the available budget, the number of samples being genotyped and the required coverage of the gene or genome for the study at hand.

The real time PCR technology, TaqMan (PE Applied Biosystems), will genotype a single SNP per assay. Furthermore, the assay requires a highly specialized, labor intensive design of allele-specific oligonucleotides (ASO) probes for each SNP in order to produce optimal hybridization between alleles [14]. Hence, TaqMan is only attractive in studies with a very small number of polymorphic markers, given cost and labor scale linearly with the number of SNP. At the opposite end of the spectra, the Illumina and Affymetrix multiplex platforms genotype a significant number of SNPs per assay. The Illumina [5] and Affymetrix [8] platforms are comprised of an array of beads or chips, respectively, with a set of SNP specific oligonucleotides probes placed upon each element. Such platforms are suited to genome wide association studies, particularly for humans, given a chip such as the Illumina Human Omni5 Beadarray can genotype over four million polymorphisms in the human genome. One of the main issues with high density array technologies, however, is the level of prior knowledge that is required to design an array for species where gene or genome coverage of polymorphic markers have yet to be determined. In this case, a low to medium density SNP platform with multiplexing

* Corresponding author.
E-mail address: ashfaqur.rahman@csiro.au (A. Rahman).

http://dx.doi.org/10.1016/j.compbiomed.2015.03.020
0010-4825/Crown Copyright © 2015 Published by Elsevier Ltd. All rights reserved.

is more a cost effective compromise. The Sequenom iPLEX platform [4] genotypes between tens and hundreds of SNP per assay and is a low cost technology for custom, low density studies. The platform utilizes a reaction where primers adjacent to SNP loci are extended with a terminator neuclotide (ddNTP) from a set of ddNTP of different masses. The platform then uses MALDI-TOF (matrix assisted laser desorption and ionization-time of flight) mass spectrometry to detect alleles by exploiting differences in the mass of reaction products. The Sequenom iPLEX platform is utilized in this paper to investigate calibration strategies for low cost genotyping technology.

In addition to the coverage aspects of the study design, the number of samples to genotype is of critical importance. In many SNP based association studies, sufficient samples need to be genotyped to obtain the statistical power necessary to achieve meaningful results. This is often an expensive and labor intensive exercise. DNA pooling is a practical attempt to improve the efficiency of SNP genotyping by combining multiple DNA samples prior to genotyping [13]. Each pool of N samples is genotyped as a single sample reducing the cost of assays by up to a factor of N. For genotyping individuals, platforms commonly use algorithms to convert their continuous intensity measurement of a SNP into a discrete call of one of three possible pairs of allele. Such calls are not informative for DNA pools given that the intensity measurement represents one of $2N+1$ possible allele calls. Consequently, when DNA pools are used, the output intensities must be acquired to compute a quantitative genotype for each SNP. This quantitative genotype is known as its allele frequency.

Despite the improvement in cost and efficiency, a major shortcoming of pooled DNA samples is that allele frequency estimates are more sensitive to measurement error than the calls of individual samples [13,10,2]. The sources of error associated with DNA pooling include pool construction, biochemical reactions and allele frequency estimation [1]. Pool construction error is associated with the process of trying to obtain equal concentrations of DNA from the samples and then mix these in equal volumes to form the pool. For the biochemical reaction associated with PCR amplification of target DNA, one allele might be more efficiently amplified than the other allele. This differential amplification is a major source of error as its causes the signal that represents the more efficiently amplified allele to be higher than its expected value, thereby inflating its estimate in the pooled DNA sample. Finally there is an analytical error associated with attempting to model the effects of differential amplification in allele frequency estimation. Whilst individual samples are subject to the same measurement errors as pooled samples, this noise is not signifi cant enough to change the allele call of individual samples. The continuity of allele frequency estimates, however, makes the pooled samples far more susceptible to being distorted by noise.

A number of works in the current literature are aimed at solving this problem. In [15] the coefficient of preferential amplification/hybridization (CPA) is used to quantify the degree of bias. The ratio of average peak intensities between two alleles was used as the bias factor. It was found that bias introduced through preferential hybridization was adequately modeled using lognormal distributions. This results in reduced error of allele frequency estimation for the human genome. A general linear model that accounts for the nested structure of the data was used in [9] for SNP genotyping. It is not required to know the CPA in [9]. It thus avoids the need for individual SNP genotyping to determine allelic ratio of hybridization. This offers scaling up to arrays with many thousands of SNPs. The piecewise linear interpolation of pooled alleles was used in [11] to correct for the bias in pooled DNA data of the human genome.

In this paper, a novel calibration method is proposed to reduce the measurement error associated with a low cost SNP genotyping approach. More precisely, local and global errors are treated differently with our proposed method. Local errors relate to issues that are specific to each SNP. For instance, differential amplification of alleles during PCR and signal to noise ratio issues associated with allele detection using mass spectrometry. The global error relates to errors that are common to all of the SNPs in the assay. Biochemical artifacts are independent of any specific SNP and are one kind of global error [11]. We designed two separate regression algorithms to calibrate allele frequency estimates: (i) one for a specific SNP to deal with local errors, and (ii) one across all the remaining SNPs to deal with global errors. The estimates from this stage are combined into a single allele frequency estimates by a fusion predictor. Experimental results demonstrate the effectiveness of the proposed method by producing accurate estimates of the allele frequencies. Such solutions will be useful for new applications or studies where there is little prior genomic knowledge available and where study budget is limited.

2. Proposed calibration method

Given the DNA sample, the genotyping process (e.g. from iPLEX platform) generates a (x, y) pair (Fig. 1) for each SNP, where x and y are the average peak intensities of the two alleles. The (x, y) pair for an individual (Fig. 1(a)) can fall in either of three clusters. The (x, y) pair for a pooled DNA sample (Fig. 1(b)) can fall into any of the $2N+1$ clusters if constructed from N individuals. Because of the PCR process and different errors, a change in (x, y) values is

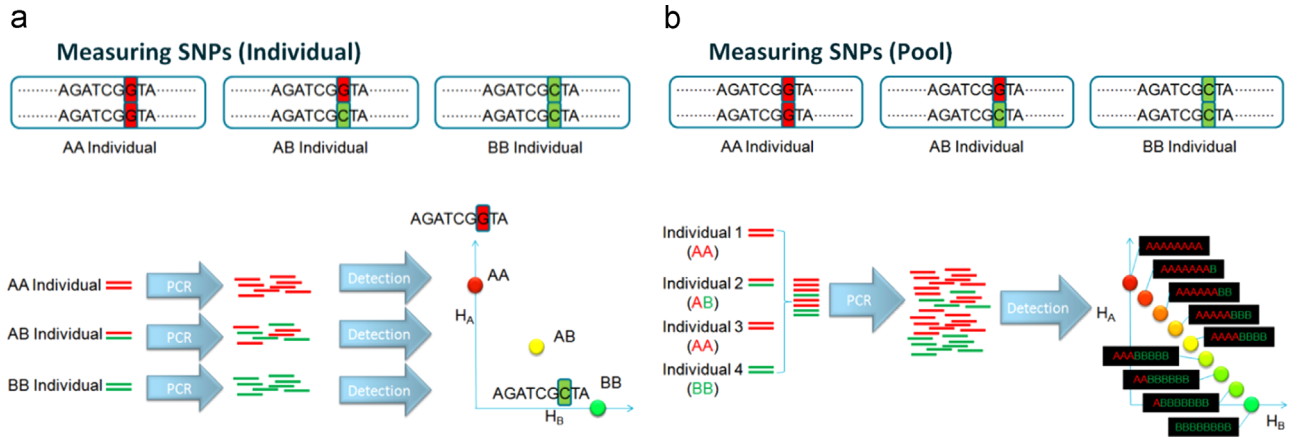

Fig. 1. Genotyping of (a) individual and (b) pooled DNA samples.

Download English Version:

https://daneshyari.com/en/article/505011

Download Persian Version:

https://daneshyari.com/article/505011

Daneshyari.com