BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 19 2009, pages 2621–2623
doi:10.1093/bioinformatics/btp470

Genetics and population analysis

R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips

Matthew E. Ritchie[1,*], Benilton S. Carvalho[2], Kurt N. Hetrick[3], Simon Tavaré[4] and Rafael A. Irizarry[2,*]

[1]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville Victoria 3052, Australia, [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, North Wolfe Street, Baltimore, MD 21205, [3]Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21224, USA and [4]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

Received on May 18, 2009; revised on July 8, 2009; accepted on July 28, 2009

Advance Access publication August 6, 2009

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Illumina produces a number of microarray-based technologies for human genotyping. An Infinium BeadChip is a two-color platform that types between 10^5 and 10^6 single nucleotide polymorphisms (SNPs) per sample. Despite being widely used, there is a shortage of open source software to process the raw intensities from this platform into genotype calls. To this end, we have developed the R/Bioconductor package *crlmm* for analyzing BeadChip data. After careful preprocessing, our software applies the CRLMM algorithm to produce genotype calls, confidence scores and other quality metrics at both the SNP and sample levels. We provide access to the raw summary-level intensity data, allowing users to develop their own methods for genotype calling or copy number analysis if they wish.

Availability and Implementation: The *crlmm* Bioconductor package is available from http://www.bioconductor.org. Data packages and documentation are available from http://rafalab.jhsph.edu/software.html.

Contact: mritchie@wehi.edu.au; rafa@jhu.edu

1 INTRODUCTION

In recent years, large-scale genome-wide association studies have provided significant insight into the genetics underpinning many complex diseases (Grant and Hakonarson, 2008). High-density microarrays, which allow many single nucleotide polymorphisms (SNPs) to be genotyped simultaneously in a sample at low cost, have been the technology driving this research. Illumina Inc. (San Diego, CA, USA) is a major provider of such arrays.

Illumina BeadChips are composed of a number of rectangular strips, each containing many randomly arranged, replicated beads. For Infinium genotyping, beads are coupled with specific 50mer probes designed to be complementary to the sequence adjacent to the SNP site, and the two alleles (A, B) are discriminated using either a red or green dye (Steemers *et al*., 2006). Data are acquired by scanning each strip at different wave lengths using Illumina's scanning device followed by automatic image analysis (Galinsky, 2003). A robust summary of the intensity in each channel for each SNP assayed is reported in proprietary idat files. BeadChips of varying SNP density and sample format (single, duo, quad) are available for human genotyping. Some contain non-polymorphic probes for assessing copy number variation.

Many algorithms that take summarized alleles A and B signals as inputs to produce genotypes (AA, AB, BB) have been developed for Affymetrix SNP arrays (Carvalho *et al*., 2007; Hua *et al*., 2007; Rabbee and Speed, 2006; Xiao *et al*., 2007). A smaller number of Illumina-specific methods (Giannoulatou *et al*., 2008; Teo *et al*., 2007) including Illumina's GenCall algorithm in BeadStudio/GenomeStudio are also available. Software for the analysis of Illumina data such as *beadarray* (Dunning *et al*., 2007), *beadarraySNP* and *lumi* (Du *et al*., 2008) is available in R/Bioconductor (Gentleman *et al*., 2004); however, current packages do not deal specifically with Infinium BeadChip data.

In this article, we present the *crlmm* package for Illumina genotyping. Our software extracts summarized intensities, performs normalization and applies the CRLMM algorithm (Carvalho *et al*., 2007) to remove chip- and SNP-specific biases and call genotypes.

2 METHODS

To begin, summarized data are read from idat files (two per array, one for each channel) using the function `readIdatFiles`. Binary idat files are a convenient starting point, as they are routinely output by the scanning software, provide a compact representation of the data and have a consistent format (unlike output from Illumina's BeadStudio/GenomeStudio software, which is exported at the user's discretion, meaning the raw signals needed for the analysis are not always available). Access to the raw data allows for low-level plotting to help visualize trends and biases that may be present (Fig. 1A and B). It also allows alternative genotyping algorithms, which require data on the raw scale, to be applied.

Next, the allele A (X Raw) and allele B (Y Raw) signals are normalized between channels and samples simultaneously using strip-level quantile normalization. The between-channel aspect of the normalization [also recommended in Oosting *et al*., (2007) and Staaf *et al*., (2008)] aims to remove any dye-bias effects, while the strip-level component corrects

*To whom correspondence should be addressed.

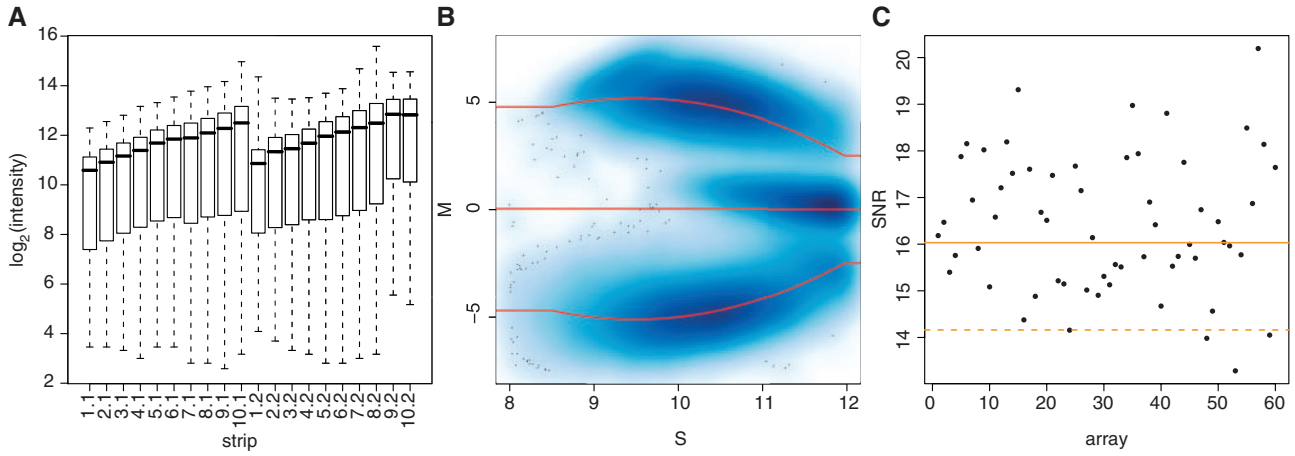

Fig. 1. (**A**) Plot of the $\log_2$ alleleB (green) intensity by strip (labelled by Row.Column position), which steadily increases from rows 1 to 10 down the BeadChip. This effect is less prominent in the allele A (red) channel (data not shown). The source of this trend is related to the way post-hybridization reagents are applied to the BeadChip and scan order, and its presence motivates strip-level normalization. (**B**) A smoothed scatter plot of M versus S for a typical array, where darker regions indicate a higher density of points. This plot shows intensity-dependent effects in M which vary for the AA and BB genotypes, and motivate the three-component mixture model in CRLMM. The curves represent the smoothing splines that model this effect. (**C**) SNR for 60 arrays, with the median (solid line) and median–median absolute deviation (dashed line) SNR values plotted. Lower scores correspond to poorer separation between the genotype clouds depicted in (B). This metric can be used to flag low-quality arrays to exclude from further analysis.

for intensity gradients which can occur within BeadChips (Fig. 1A). Normalization at the strip-level has also proven useful for data from Illumina's gene expression BeadChips (Wei Shi, personal communication). By default, the strip-level quantiles are standardized against a reference distribution obtained from HapMap samples (International HapMap Consortium, 2007) run on the same platform to correct for lab and batch effects.

After normalization, the CRLMM genotyping algorithm (Carvalho *et al*., 2007; Lin *et al*., 2008) is applied. For each array, SNP-specific log-ratios ($M = \log_2 \text{alleleA} - \log_2 \text{alleleB}$) and average intensities [$S = (\log_2 \text{alleleA} + \log_2 \text{alleleB})/2$] are calculated. As noted for Affymetrix data (Carvalho *et al*., 2007), S appears to have an effect on M. The effect appears to be a smooth function of S, but only applies to the AA and BB intensities (Fig. 1B). To remove this effect, we fit a three-component mixture model with a spline used to model the smooth function. This model is fitted per array via the expectation-maximization (EM) algorithm using a random sample of data-points. Due to the different chemistry used for Illumina genotyping, the fragment length covariate described in Carvalho *et al*. (2007) can be ignored.

Next, a two-level hierarchical model is applied. SNP-specific means and standard deviations (SDs) are obtained for each genotype via supervised learning using HapMap data. Independent genotype calls (available from http://www.hapmap.org/) provide the true states for samples that have been genotyped using the respective BeadChip platform. Normalized signals from these arrays are then used to estimate robustly the genotype means and SDs. The intensity-dependent splines from the EM (which explain the between-SNP variation) and the SNP-specific genotype means and SDs (obtained from training data) are combined in the model. New genotype calls are assigned by choosing the class that minimizes the negative log likelihood.

CRLMM produces a number of metrics for quality assessment (Lin *et al*., 2008). Confidence scores for each call are provided using the log-likelihood ratio test from the hierarchical model. The sample-specific SNR (signal-to-noise ratio) assesses the separation of the three genotypes within an array. Lower SNR values indicate poorer quality, and this metric can be used to exclude samples from further analysis (Fig. 1C). SNP-specific quality is measured as the minimum distance between the heterozygote centre and either of the two homozygous centres.

The preprocessing and genotyping steps above are performed by the `crlmmIllumina` function. All code is written in R (R Development Core Team, 2009) and existing *Biobase* classes are used to store the data.

The software requires chip-specific data packages (available at http://rafalab.jhsph.edu/software.html) that store basic SNP annotation information and various parameters used by CRLMM. We also provide the *hapmap370k* data package, which contains idats from 40 HapMap samples hybridized to HumanHap 370 K Duo BeadChips, and a user guide that provides example R code to analyse these samples (see Supplementary Material). A 64-bit Linux system takes ~90 s and uses up to 1.2 GB of RAM to read these data, while normalization and genotyping takes a further 470 s and uses up to 3.3 GB of RAM. This equates to processing around 600 SNPs per second.

3 DISCUSSION

The *crlmm* package provides bioinformaticians with an additional tool outside of Illumina's proprietary software for analysing Infinium BeadChip data. Our software also facilitates the analysis of Affymetrix SNP chips and the use of a consistent algorithm and framework to process both the platforms allows data from different studies to be combined more easily.

Implementation in R/Bioconductor gives users the opportunity to exploit other tools that have been adapted for Illumina data. For example, if raw bead-level data were available, the BASH spatial artefact detection method (Cairns *et al*., 2008) in the *beadarray* package could be applied. Once summarized, the data could be further processed using *crlmm*.

The CRLMM algorithm can be applied to new versions of Illumina BeadChips for humans and other species, provided that the necessary training data and prior information on genotype calls are available. Future work will benchmark the performance of our method with other genotyping algorithms tailored to suit Illumina data, such as Illuminus (Teo *et al*., 2007) and Illumina's own

algorithms in BeadStudio/GenomeStudio. Furthermore, tools for copy number analysis are being developed in the *crlmm* package.

ACKNOWLEDGEMENTS

We thank Illumina for providing access to HapMap datasets for each platform and for technical support on their products; Keith Baggerly (MD Anderson Cancer Center) for sharing R code to read idat files; Kimberly Doheny and Elizabeth Pugh (CIDR) for providing test HapMap data; and Stephen Wilcox and Melinda Ziino (AGRF) for providing example idat files for testing purposes.

Funding: Isaac Newton Trust and NHMRC Program (grant 406657); NHMRC IRIISS (grant 361646); Victorian State Government OIS grant (to M.E.R.); National Institutes of Health grants R01GM083084, R01RR021967 and P41HG004059 (to B.S.C., R.A.I.); Cancer Research UK (S.T.). Funding for open access charge: National Institutes of Health (grant P41HG004059).

Conflict of Interest: none declared.

REFERENCES

Cairns,J. *et al*. (2008) BASH: a tool for managing BeadArray spatial artefacts. *Bioinformatics*, **24**, 2921–2922.

Carvalho,B. *et al*. (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, **8**, 485–499.

Du,P. *et al*. (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, **24**, 1547–1548.

Dunning,M.J. *et al*. (2007) beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, **23**, 2183–2184.

Galinsky,V.L. (2003) Automatic registration of microarray images. II. Hexagonal grid. *Bioinformatics*, **19**, 1832–1836.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Giannoulatou,E. *et al*. (2008) GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. *Bioinformatics*, **24**, 2209–2214.

Grant,S.F. and Hakonarson,H. (2008) Microarray technology and applications in the arena of genome-wide association. *Clin. Chem.*, **54**, 1116–1124.

Hua,J. *et al*. (2007) SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. *Bioinformatics*, **23**, 57–63.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Lin,S. *et al*. (2008) Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays. *Genome Biol.*, **9**, R63.

Oosting,J. *et al*. (2007) High-resolution copy number analysis of paraffin-embedded archival tissue using SNP BeadArrays. *Genome Res.*, **17**, 368–376.

R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at http://www.R-project.org.

Rabbee,N. and Speed,T.P. (2006) A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics*, **22**, 7–12.

Staaf,J. *et al*. (2008) Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics*, **9**, 409.

Steemers,F.J. *et al*. (2006) Whole-genome genotyping with the single-base extension assay. *Nat. Methods*, **3**, 31–33.

Teo,Y.Y. *et al*. (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, **23**, 2741–2746.

Xiao,Y. *et al*. (2007) A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics*, **23**, 1459–1467.