GenoSNP: a new genotyping algorithm for the Infinium SNP genotyping platform that uses Robust Bayesian Clustering.

Eleni Giannoulatou[1,2*], Christopher Yau[1,2*], Stefano Collela[3], Jiannis Ragoussis[4] and Christopher C. Holmes [1,5]

[1] Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK
[2] Life Sciences Interface Doctoral Training Centre, University of Oxford, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK
[3] Functional Biology, Insects and Interactions, INSA-Lyon, 20 avenue Albert Einstein, 69621 Villeurbanne cedex, France
[4] Genomics Group, Wellcome Trust Centre for Human Genetics, Oxford, OX3 7BN, UK
[5] MRC Mammalian Genetics Unit, MRC Harwell, Harwell, OX11 0RD, UK

[*] These authors contributed equally to this work

Large-scale genotyping of Single Nucleotide Polymorphisms (SNPs) has enabled genome-wide association studies aiming to identify genetic variation that influences susceptibility to complex disorders. Illumina's BeadArray technology can interrogate hundreds of thousands SNPs on a single assay with sufficiently high signal-to-noise ratio. In addition to standard SNP genotyping for association studies, these high-resolution genotyping arrays are applicable to genomic profiling in order to detect loss of heterozygosity (LOH) and copy number variation (CNV) [1].
Current SNP genotyping algorithms typically call genotypes by clustering allele-specific intensity data on a SNP-by-SNP basis. We have investigated different ways of genotyping using robust Bayesian clustering and applying standard Expectation-Maximazation algorithm as well as Variational Bayes methods for fast clustering. The probe intensities of every SNP are clustered either within a sample without the need of a reference population or using a joint approach that first genotypes within a sample and subsequently genotypes on a SNP-by-SNP basis using parameters calibrated from the first pass. The within-sample approach allows borrowing information of SNPs so that accurate genotyping can be achieved (even for SNPs with low minor allele frequency) and is appropriate for small sample size studies. On the other hand, the motivation for a population-based strategy is that probe intensities vary on a SNP-by-SNP basis. Our method accounts for dye-specific and bead-specific effects on the Infinium assay of the BeadArray technology. We have compared all the available methods and we report differences in call error and accuracy, calibration as well as computational time. Our algorithm is efficient and exhibits high concordance with current methods and $> 99\%$ call accuracy on HapMap samples. Since array-based genotyping technology is moving towards assaying millions of SNPs simultaneously, fast and accurate algorithms are needed. This will enable the development of methods that detect any variation in the genome (including CNVs) and that would be optimal on population-based disease association studies.

References

[1] Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35 6:2013-2025