A review of software for microarray genotyping

*Philippe Lamy,[1,2] Jakob Grove[1,3] and Carsten Wiuf[1]**

[1]Bioinformatics Research Centre, C. F. Mollers Allé 8, Building 1110, DK-8000 Aarhus C, Denmark
[2]Department of Molecular Medicine, Aarhus University Hospital, Skejby, DK-8200 Aarhus N, Denmark
[3]Department of Human Genetics, Aarhus University, The Bartholin Building, Wilhelm Meyers Allé 4, DK-8000, Aarhus C, Denmark
**Correspondence to:* Tel: +45 8942 3100; Fax: +45 8942 3077; E-mail: wiuf@birc.au.dk

Date received (in revised form): 1st March 2011

Abstract

The focus of this review is software for the genotyping of microarray single nucleotide polymorphisms, in particular software for Affymetrix and Illumina arrays. Different statistical principles and ideas have been applied to the construction of genotyping algorithms — for example, likelihood versus Bayesian modelling, and whether to genotype one or all arrays at a time. The release of new arrays is generally followed by new, or updated, algorithms.

Keywords: *SNP array, genotype, calling algorithm, copy number, intensity, software*

Introduction

The use of microarrays and microarray technology in research is now more than 15 years old and has had a tremendous impact on many aspects of research. Suddenly, it became possible to profile and survey whole genomes and to compare genomes across individuals and species to an extent that was hardly possible before. The perception of the genome changed as genome-wide data became available to everyone.

This review focuses narrowly on software used for genotyping of single nucleotide polymorphisms (SNPs) in connection with SNP microarrays (or 'arrays' for short). There are an estimated ten million or more SNPs in the human genome.[1] For each of these, there are three possible genotypes (assuming diploidy), AA, BB (homozygous) and AB (heterozygous), where A and B denote the two possible alleles. The first commercial SNP array was released in 1996 by Affymetrix (Santa Clara, CA) and targeted about 1,500 human SNPs,[2] a tiny fraction of all SNPs. Since then, many different manufacturers have developed microarrays for genome-wide genotyping, including Affymetrix, Agilent (Santa Clara, CA), Illumina (San Diego, CA) and Nimblegen (Madison, WI), with arrays designed for many different organisms.

SNP arrays have found uses in many research areas and contexts — for example, association mapping,[3] linkage disequilibrium mapping,[4] phasing,[5] inference on demography and ancestry,[6] evolution[7] and loss-of-heterozygosity analysis in cancer.[8] Early usage of SNP arrays sought to estimate loss of heterozygosity in cancer by comparing DNA from germline and tumour cells.[9] In addition, SNP arrays have been used to estimate copy numbers in cancers[10] (similar to the use of comparative genomic hybridisation [CGH] arrays) and copy number variants (CNVs) in populations.[11] The newest arrays from Affymetrix and Illumina both contain probes for CNVs and copy number polymorphisms (CNPs).

Today, SNP microarrays are able to genotype more than a million SNPs simultaneously (Table 1). This large number of SNPs poses a number of statistical, as well as computational, problems and has attracted the attention of many statisticians and bioinformaticians. Interestingly, the problems themselves

Table 1. The arrays that are currently available for the human genome from Affymetrix and Illumina. For Affymetrix, #Arrays reflects the physical number of arrays to use to obtain genotypes for all SNPs. For Illumina, #Samples gives the number of samples that can be run using the same BeadChip

	#Arrays	#SNPs	Software
Affymetrix			
GeneChip Human Mapping 10K 2.0 Array	1	10,204	MPAM
GeneChip Human Mapping 100K Set	2	116,204	DM
GeneChip Human Mapping 500K Array Set	2	500,568	BRLMM
Genome-Wide Human SNP Array 5.0	1	500,568[a]	BRLMM-P
Genome-Wide Human SNP Array 6.0	1	906,600[b]	Birdseed
Illumina	**#Samples**	**#Markers**	**Software**
HumanCytoSNP-12 DNA Analysis BeadChip	12	299,140	[d]
Human660W-Quad v1 DNA Analysis BeadChip	4	657,366	[d]
HumanOmniExpress BeadChip	12	731,442	[d]
Human1M-Duo DNA Analysis BeadChip	2	1,199,187	[d]
HumanOmni1-Quad BeadChip	4	1,140,419	[d]
HumanOmni1S-8 BeadChip	8	1,200,000[c]	[d]
HumanOmni2.5-Quad BeadChip	4	2,450,000[c]	[d]

[a]Additional 420,000 non-polymorphic probes for copy number analysis.
[b]Additional 946,000 non-polymorphic probes for copy number analysis.
[c]Also includes probes for CNVs.
[d]The BeadStudio and the GenomeStudio applications can handle all Illumina's arrays.

have led to many new developments in statistics and have fostered what we might term 'informatics of large datasets'. There are a number of statistical issues that are shared between microarrays, irrespective of the platform, chemistry and design principles. These include:

(i) Normalisation of raw intensities
(ii) Background correction and outlier detection
(iii) Genotyping

The statistical methods applied at each step are, to some extent, transferable between platforms and array types, in particular the parts relating to (i) and (ii). Normalisation of array intensities is important in order to make comparisons across arrays.[12,13] Background correction and outlier detection (individual 'bad' SNPs, as well as 'bad' arrays) are essential for correct interpretation of the data[12,13] (ie to reduce the number of false and missing calls).

A general review of SNP array platforms and their history and use is given by LaFramboise.[14]

Software

We focus on the most commonly used platforms, Affymetrix and Illumina, and do not discuss software for normalisation and background correction.

Problem formulation

Both platforms represent a SNP by a number of probes (Affymetrix) or beads (Illumina) for each allele. The probes/beads have different affinities, depending on the DNA sequence they target, and thus produce signals of various strengths (see Figure 1). The newest arrays, Human SNP Array 6.0 from Affymetrix and HumanOmni2.5-Quad BeadChip from Illumina, use three probes and around 20 beads for each allele, respectively. Earlier Affymetrix arrays additionally used mismatch probes; probes that were designed to capture non-specific binding. A first step in many algorithms for genotyping is to summarise the probe intensities for each allele and SNP, and in a second step to make a call based on the summarised intensities.

SNP calling software for Affymetrix SNP arrays

Following its release of new SNP arrays (called GeneChips), Affymetrix has developed

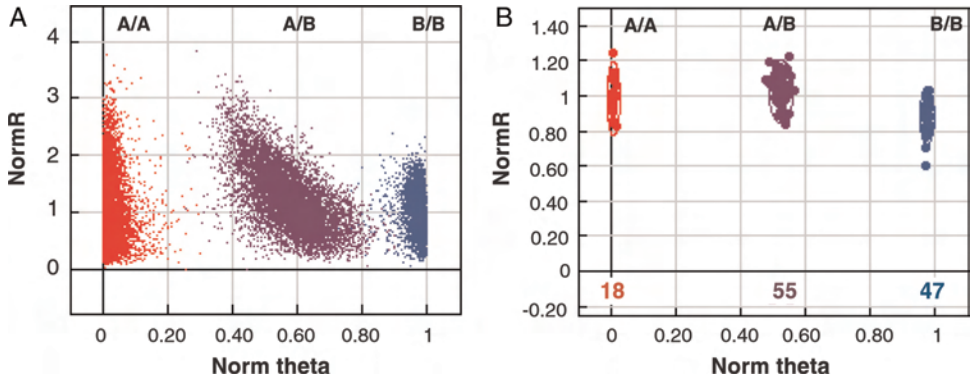

Figure 1. Normalised and summarised allele intensities from the Illumina BeadChip array. The intensities are shown in transformed polar coordinates: the theta-coordinate represents the angle from the x-axis (the angle from the x-axis to the vector [A, B] of the two allele intensities), and the R-coordinate represents the copy number (the length of the vector). (A) Intensities for a single nucleotide polymorphism (SNP) from 120 arrays, clearly separating the intensities into three groups (A/A, A/B, B/B). (B) Data from 317,000 SNPs (from the same 120 arrays). This plot clearly indicates that signal strength varies considerably with the SNP, a factor that must be taken into account when genotyping individual SNPs and deriving copy numbers. The figure is reproduced with the permission of Gunderson *et al.*[15]

accompanying software that takes into account the properties of the new arrays. The first program, Modified Partitioning Around Medoids (MPAM)[16] and Dynamic Model (DM),[17] were able to genotype one SNP on one chip at a time. The next generation of software, Robust Linear Model with Mahalanobis Distance Classifier (RLMM),[18] BRLMM[19] (which adds a Bayesian step to RLMM), BRLMM-P[20] (which uses perfect match probes only) and Birdseed,[21] increased accuracy and performance using a multi-chip approach.

One SNP, one chip

For the first SNP arrays, Affymetrix designed software modules (MPAM, DM) to genotype individual SNPs, one array at a time. The DM software[17] was introduced with the release of the 100K GeneChip and is based on statistical modelling of quartets. A quartet consists of match and mismatch probes for the two alleles. This software does not require any normalisation step and does not summarise the probe intensities. A score is assigned for each quartet and the Wilcoxon signed-ranked test is used to give a call.

Multi-chip approach

Several groups[3,18,22–28] have designed SNP calling algorithms using a multi-chip approach; the first algorithm was RLMM.[18] This approach requires pre-processing steps, such as array normalisation, in order to compare data across arrays and summation of the probe intensities for each allele. For each SNP, the two allele intensities are then clustered into three clouds, representing the different genotypes across many chips (Figure 2).

Affymetrix designed the BRLMM algorithm for the 500K SNP arrays.[19] This algorithm was a significant improvement over the DM algorithm used for the previous arrays. The BRLMM algorithm is an extension of the RLMM software and it uses a Bayesian step to define cluster centres and variances of SNP intensities. Briefly, after normalisation and allelic summation, genotypes are clustered using a Bayesian prior on cluster centres and variances and a pre-clustering made by the DM algorithm. The prior is based on a random set of SNPs, with a minimum number of individuals in each cluster. This allows for a better definition of the genotype clusters with few (potentially no) individuals. Further, new arrays can be genotyped using predefined parameters obtained from other arrays.

For the SNP5.0 GeneChip, Affymetrix designed a new version of the BRLMM algorithm, named BRLMM-P, as the array does not have mismatch probes.[20] The DM step of BRLMM is replaced

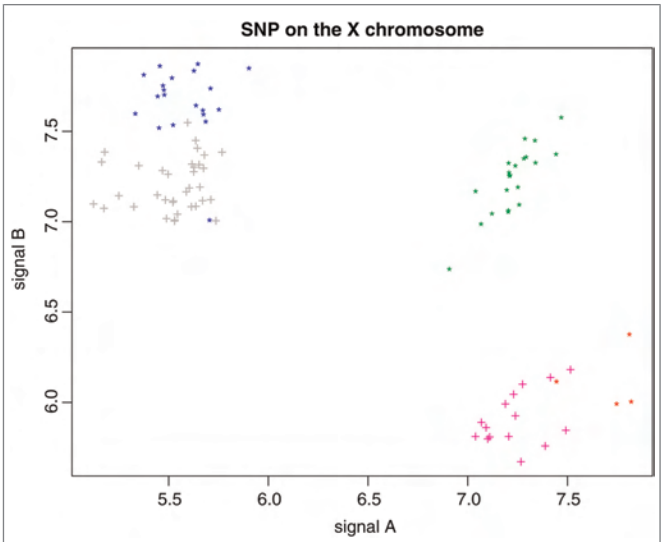

Figure 2. Normalised and summarised allele intensities from the Affymetrix GeneChip array. Each SNP is represented by a pair of intensity values (A, B) for the A and B alleles, respectively (here, on a log-scale). An X chromosome SNP is shown, clearly indicating separation into distinct genotype clusters. The plot also shows that different copy numbers can be distinguished. Males are haploid for the particular SNP (ie either AY or BY) and show up as homozygous but with reduced allele intensity. Grey: BY; blue: BB; green: AB; red: AA; and pink: AY.

by a maximum likelihood-based division into genotype cluster. Further, the prior can be a generic prior common to all SNPs or a SNP-specific prior defined using a set of training data (such as HapMap data).

For SNP6.0, the Broad Institute, in collaboration with Affymetrix, developed the Birdsuite software.[21,29] The novelty comes from relaxing the assumption that all SNPs are diploid and introducing known CNPs. Birdseed is actually composed of four applications: Canary, Birdseed, Birdseye and Fawkes. Canary can give an estimate of the copy number of known CNPs. Birdseed is a genotyping software with use restricted to diploid genomic regions. It is similar to BRLMM. Clusters are predefined using training data and then further optimised. The Birdseye software can detect rare CNVs and genotype SNPs in CNVs. Finally, Fawkes combines the output of the three previous applications to assign a comprehensive genotype (A-null, AA, AB, BB, AAB,. . .).

Other software can be used to genotype SNPs from Affymetrix GeneChips, such as Corrected Robust Model with Maximum Likelihood Distance (CRLMM),[23] Genotype calling with Empirical Likelihood (GEL),[24] SNiPer-High Density (SNiPer-HD)[25], Probe-Level Allele-specific Quantization (PLASQ),[26] MAMS[27] (combines Single-Array Multi-SNP [SAMS] with Multi-Array Single-SNP [MASS]), Chiamo[3] and JAPL.[28] Some work has been done to compare algorithms[30–32] and, generally, the performance of algorithms are compared with HapMap data in the original papers.

SNP calling software for Illumina BeadChips

Illumina has developed its own software to genotype SNPs on the BeadChip array. The software is called GenCall and has not been through the same chain of transformations as the Affymetrix software. The GenCall algorithm was implemented within the BeadStudio application (latest version v3.2.2)[33] but it is now part of the GenomeStudio application (the current version is 1.1.0). It relies on a specific normalisation occurring automatically within the Illumina GenomeStudio software and consists of several steps (including outlier removal and background estimation). The normalised intensities are then summarised, such that each SNP is assigned a pair of values corresponding to each allele. This pair represents the allele intensities in polar coordinates; the R-coordinate represents the copy number of the SNP and the theta-coordinate represents the angle from the x-axis (Figure 1). This is a multi-array approach, using information from all arrays simultaneously.

The call is made using a cluster file supplied by Illumina, based on a reference set of samples. There is an option to make the call without using the reference set, instead relying exclusively on the sampled arrays, however. This dichotomy is similar to the BRLMM (and subsequent Affymetrix software), where a call can be made with pre-defined parameters, corresponding to a reference population. Whether one should use the reference set for genotype calling depends on the number of sampled arrays, the quality of the DNA and the minimal

allele frequency (MAF) of interest, as the size of the reference set determines the MAF detectable.[34]

For SNPs with fewer than three genotype clusters, the locations and variations of the missing genotype clusters are estimated using artificial neural networks. It is also possible manually to change the call of any SNP using Illumina's visualisation tool. For CNV analysis, Illumina has developed a series of tools which are available as plug-ins to the GenomeStudio genotyping module. Software for estimation of copy numbers (cnvPartition), detection and annotation of homozygosity in single samples (Homozygosity Detector), detection and annotation of chromosomal aberrations in single samples (ChromoZone) and for calculating a likelihood score for strength of loss-of-heterozygosity (LOH Score) is available.

Other methods have been proposed for the BeadChip arrays. Teo *et al.* designed a multi-array genotype calling algorithm (Illuminus) that does not rely on a reference population.[35] By contrast, Giannoulatou *et al.* developed a method that works entirely within each sample, thereby making the performance independent of sample size and of any outside control samples.[34] Both methods rely on an expectation–maximisation (EM) algorithm. The CRLMM algorithm[23] is also available for Illumina data as a package (GenoSNP) for R/Bioconductor.[36]

Discussion and conclusions

The accuracy of genotype calling is usually reported to be above 99 per cent. This is typically the case when samples and DNA of good quality are available. Many cancer laboratories are interested in genotyping SNPs and CNPs, however, as well as estimating copy numbers, from tumour tissue. Here, there are a number of problems that have not yet fully been overcome: tumour tissue typically contains normal cells that are difficult to remove prior to analysis; also, tumour tissue tends to be heterogeneous, in the sense that different tumour cells have different copy number aberrations. These issues affect the possibility of accurately estimating genotypes and copy numbers, and significantly reduce the accuracy of calling algorithms.

We have discussed software for genotyping; however, much software also has been developed for further downstream analysis, to accommodate specific questions and needs.[37,38] Software for normalisation and background correction has likewise received much attention. These methods are also generally applicable to other types of arrays, and borrowing of ideas between array types is common.

The future of SNP arrays, in addition to many other microarray types, such as gene (RNA) expression and microRNA expression arrays, is uncertain. For the individual laboratory, the common microarray platforms are still more cost-efficient than the new platforms built on next-generation sequencing (NGS) technologies. NGS is already dominating research to an extent that few foresaw five years ago, however. In addition, it is possible to have samples sequenced through commercial organisations or scientific collaborations.

SNP and other arrays are still in use, however. They have transformed the field of genomics and sparked an intense interest among the statistics and bioinformatics communities to provide solutions to large-scale data problems. These solutions are the foundation for solving the similar large-scale data problems encountered with NGS.

Acknowledgments

The study was supported by grants from the Danish Strategic Research Council (2101-07-0059) (GEMS consortium), the Danish Cancer Society, the EC project GENICA, the Lundbeck Foundation and the John and Birthe Meyer Foundation.

References

1. Kruglyak, L. and Nickerson, D.A. (2001), 'Variation is the spice of life', *Nat. Genet.* Vol. 27, pp. 234–236.
2. Wang, D.G., Fan, J.B., Siao, C.J., Berno, A. *et al.* (1998), 'Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome', *Science* Vol. 280, pp. 1077–1082.
3. Wellcome Trust Case Control Consortium (2007), 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature* Vol. 447, pp. 661–678.
4. Pe'er, I., de Bakker, P.I., Maller, J., Yelensky, R. *et al.* (2006), 'Evaluating and improving power in whole-genome association studies using fixed marker sets', *Nat. Genet.* Vol. 38, pp. 663–667.
5. Niu, T. (2004), 'Algorithms for inferring haplotypes', *Genet. Epidemiol.* Vol. 27, pp. 334–347.

6. Price, A.L., Butler, J., Patterson, N., Capelli, C. *et al.* (2008), 'Discerning the ancestry of European Americans in genetic association studies', *PLoS Genet*. Vol. 4, p. e236.
7. Neafsey, D.E., Schaffner, S.F., Volkman, S.K., Park, D. *et al.* (2008), 'Genome-wide SNP genotyping highlights the role of natural selection in *Plasmodium falciparum* population divergence', *Genome Biol*. Vol. 9, p. R171.
8. Koed, K., Wiuf, C., Christensen, L.L., Wikman, F.P. *et al.* (2005), 'High-density single nucleotide polymorphism array defines novel stage and location dependent allelic imbalances in human bladder tumors', *Cancer Res.* Vol. 65, pp. 34–45.
9. Lindblad-Toh, K., Tanenbaum, D.M., Daly, M.J., Winchester, E. *et al.* (2000), 'Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays', *Nat. Biotechnol*. Vol. 18, pp. 1001–1005.
10. Greenman, C.D., Bignell, G., Butler, A., Edkins, S. *et al.* (2010), 'PICNIC: An algorithm to predict absolute allelic copy number variation with microarray cancer data', *Biostatistics* Vol. 11, pp. 164–175.
11. Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R. (2009), 'Copy number variation in human health, disease, and evolution', *Annu. Rev. Genomics Hum. Genet*. Vol. 10, pp. 451–481.
12. Bolstad, B.M., Irizarry, R.A., Åstrand, M. and Speed, T.P. (2003), 'A comparison of normalization methods for high density oligonucleotide array data based on variance and bias', *Bioinformatics* Vol. 19, p. 185–193.
13. Li, C. and Wong, W.H. (2001), 'Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application', *Genome Biol*. Vol. 2, pp. research0032.1–0032.11.
14. LaFramboise, T. (2009), 'Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances', *Nucleic Acids Res.* Vol. 37, pp. 4181–4193.
15. Gunderson, K.L., Kuhn, K.M., Steemers, F.J., Ng, P. *et al.* (2006), 'Genotype clustering on HumanHap300 BeadChip™', *Pharmacogenomics*, Vol. 7, pp. 641–648.
16. Liu, W., Di, X., Yang, G., Matsuzaki, H. *et al.* (2003), 'Algorithms for large-scale genotyping microarrays', *Bioinformatics* Vol. 19, pp. 2397–2403.
17. Di, X., Matsuzaki, H., Webster, T.A., Hubbell, E. *et al.* (2005), 'Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays', *Bioinformatics* Vol. 21, pp. 1958–1963.
18. Rabbee, N. and Speed, T.P. (2006), 'A genotype calling algorithm for Affymetrix SNP arrays', *Bioinformatics* Vol. 22, pp. 7–12.
19. Affymetrix, Inc. (2006), 'BRLMM: An improved genotype calling method for the mapping 500K array set', http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf (last accessed 30th April, 2011).
20. Affymetrix, Inc. (2007), 'BRLMM-P: A genotype calling method for the SNP 5.0 array', http://www.affymetrix.com/support/technical/whitepapers/brlmmp_whitepaper.pdf (last accessed 30th April, 2011).
21. Korn, J., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A. *et al.* (2008), 'Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs', *Nat. Genet*. Vol. 40, pp. 1253–1260.
22. Lamy, P., Andersen, C.L., Wikman, F.P. and Wiuf, C. (2006), 'Genotyping and annotation of Affymetrix SNP arrays', *Nucleic Acids Res.* Vol. 34, p. e100.
23. Carvalho, B., Speed, T.P. and Irizarry, R.A. (2007), 'Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data', *Biostatistics* Vol. 8, pp. 485–499.
24. Nicolae, D.L., Wu, X., Miyake, K. and Cox, N.J. (2006), 'GEL: A novel genotype calling algorithm using empirical likelihood', *Bioinformatics* Vol. 22, pp. 1942–1947.
25. Hua, J., Craig, D.W., Brun, M., Webster, J. *et al.* (2006), 'SNiPer-HD: Improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays', *Bioinformatics* Vol. 23, pp. 57–63.
26. LaFramboise, T., Weir, B.A., Zhao, X., Beroukhim, R. *et al.* (2005), 'Allele-specific amplification in cancer revealed by SNP array analysis', *PLoS Comput. Biol*. Vol. 1, p. e65.
27. Xiao, Y., Segal, M.R., Yang, Y.H. and Yeh, R.-F. (2007), 'A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays', *Bioinformatics* Vol. 23, pp. 1459–1467.
28. Plagnol, V., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2007), 'A method to address differential bias in genotyping in large-scale association studies', *PLoS Genet*. Vol. 3, p. e74.
29. McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S. *et al.* (2008), 'Integrated detection and population-genetic analysis of SNPs and copy number variation', *Nat. Genet*. Vol. 40, pp. 1166–1174.
30. Lin, S., Carvalho, B., Cutler, D., Arking, D. *et al.* (2008), 'Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays', *Genome Biol*. Vol. 9, pp. 1–12.
31. Kim, J.-H., Jung, S.-H., Hu, H.-J., Yim, S.-H. *et al.* (2010), 'Comparison of the Affymetrix SNP Array 5.0 and oligoarray platforms for defining CNV', *Genomics Informatics* Vol. 8, pp. 138–141.
32. Vens, M., Schillert, A., König, I.R. and Ziegler, A. (2009), 'Look who is calling: A comparison of genotype calling algorithms', *BMC Proc.* Vol. 3, p. S59.
33. Steemers, F.J. and Gunderson, K.L. (2007), 'Whole genome genotyping technologies on the BeadArray platform', *Biotechnol. J.* Vol. 2, pp. 41–49.
34. Giannoulatou, E., Yau, C., Colella, S., Ragoussis, J. *et al.* (2008), 'GenoSNP: A variational Bayes within-sample SNP genotyping algorithm that does not require a reference population', *Bioinformatics* Vol. 24, pp. 2209–2214.
35. Teo, Y.Y., Inouye, M., Small, K.S., Gwilliam, R. *et al.* (2007), 'A genotype calling algorithm for the Illumina BeadArray platform', *Bioinformatics* Vol. 23, pp. 2741–2746.
36. Ritchie, M.E., Carvalho, B.S., Hetrick, K.N. and Tavaré, S. (2009), 'R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips', *Bioinformatics* Vol. 25, pp. 2621–2623.
37. Aroma.affymetrix, http://groups.google.com/group/aroma-affymetrix/web/software?version=5&pli=1 (last accessed 30th April, 2011).
38. Cheng Li Lab, http://www.biostat.harvard.edu/complab/dchip/ (last accessed 30th April, 2011).