Harvard University
Harvard University Biostatistics Working Paper Series

Year 2006 *Paper* 44

PLASQ: A Generalized Linear Model-Based Procedure to Determine Allelic Dosage ini Cancer Cells from SNP Array Data

Thomas LaFramboise* David P. Harrington†
Barbara A. Weir‡

*Dana-Farber Cancer Institute, tlafram@broad.mit.edu
†Dana-Farber Cancer Institute and Harvard School of Public Health, dph@hsph.harvard.edu
‡Dana-Farber Cancer Institute

http://biostats.bepress.com/harvardbiostat/paper44

PLASQ: A Generalized Linear Model-Based Procedure to Determine Allelic Dosage in Cancer Cells from SNP Array Data

THOMAS LAFRAMBOISE*

Department of Medical Oncology, Dana-Farber Cancer Institute, MA, 02115, USA

*To whom correspondence should be addressed: tlafram@broad.mit.edu,

Tel: (617) 324-0812, Fax: (617) 582-7880

DAVID HARRINGTON

Department of Biostatistics, Harvard School of Public Health, MA, 02115, USA

BARBARA A. WEIR

Department of Medical Oncology, Dana-Farber Cancer Institute, MA, 02115, USA

Abstract

Human cancer is largely driven by the acquisition of mutations. One class of such mutations is copy number polymorphisms, comprised of deviations from the normal diploid two copies of each autosomal chromosome per cell. We describe a probe-level allele-specific quantitation (PLASQ) procedure to determine copy number contributions from each of the parental chromosomes in cancer cells from SNP microarray data. Our approach is

based upon a generalized linear model that takes advantage of a novel classification of probes on the array. As a result of this classification, we are able to fit the model to the data using an expectation-maximization algorithm designed for the purpose. We demonstrate a strong model fit to data from a variety of cell types. In normal diploid samples, PLASQ is able to genotype with very high accuracy. Moreover, we are able to provide a generalized genotype in cancer samples (e.g. CCCCT at an amplified SNP). Our approach is illustrated on a variety of lung cancer cell lines and tumors, and a number of events are validated by independent computational and experimental means. An R software package containing the methods is freely available.

1. Introduction

Over the course of the past decade, high throughput probe-based microarray technology has become a vital tool in genomic research. These microarrays contain thousands of unique nucleotide probe sequences, each designed to hybridize to a "target" nucleic acid molecule. When a DNA or RNA sample is properly prepared and applied to the array, specialized equipment can produce a measure of the intensity of hybridization between each probe and its target in the sample. The underlying principle is that the hybridization intensity depends upon the amount of target DNA or RNA in the sample, as well as the affinity between target and probe. Extensive processing and analysis of these raw intensity measures gives estimates of some characteristic of the target sequences in the sample. The subject of this paper is the analysis of data from a specific array type, the single nucleotide polymorphism (SNP) array.

The GeneChip Mapping 100K Set (Affymetrix, 2004) is a pair of arrays able to interrogate over 100 000 human SNPs. Herein, we shall refer to this pair simply as the SNP array. The original aim of the SNP array was to identify which of the two SNP alleles — arbitrarily labeled allele A and allele B — occurs for each chromosome copy (maternal and paternal) at each SNP in an individual's genome. Thus, the individual can be genotyped at a SNP as either homozygous AA, homozygous BB, or heterozygous AB. More recently, it has been demonstrated that these arrays may be used to identify loss-of-heterozygosity (LOH) (Lindblad-Toh *et al.*, 2000; Lin *et al.*, 2004), as well as to produce a measure of genomic copy number at each SNP (Bignell *et al.*, 2004; Zhao *et al.*, 2005), in cancer samples. Regions of LOH are loci at which one of the two parental copies of a chromosome is deleted. Typically, one may use SNP array data to detect

LOH at SNPs where the cancer cell is homozygous, but its matched normal (same individual) counterpart is heterozygous. In copy number inference, the goal is to identify chromosomal regions in which the number of copies deviates from the normal diploid two. These lesions include amplifications (copy number greater than two), heterozygous deletions (copy number one), and homozygous deletions (copy number zero).

The SNP array is designed so that each probe is a sequence of length 25 bases, and is a member of a probe set comprised of 40 unique sequences. Within a probe set, half of all probes are "perfect match" (PM) probes. All PM probes within the set are perfectly complementary to some 25-base subsegment of the same target DNA fragment. Additionally, every PM probe has a corresponding "mismatch" (MM) probe that is identical to its PM counterpart, save that the central (13^{th}) base is altered so as not to be perfectly complementary to the target sequence. The PM probes are complementary to either the A or B allele of the SNP, and thus the SNP array probes have been typically classified as either PM_A, PM_B, MM_A, or MM_B. In fact, the probes on the array may be grouped as quartets comprised of one of each of these four classes, with each quartet interrogating the same 25-base subsequence of the target genomic DNA fragment.

In this paper, we provide a generalization of the three applications — genotyping, LOH detection, and copy number inference — of SNP arrays. Specifically, we present a probe level allele-specific quantitation (PLASQ) procedure to infer allele-specific copy number (ASCN) and parent-specific copy number (PSCN). The ASCN is a generalization of both genotype and copy number at a SNP, in that all sample SNPs are assigned a genotype, regardless of copy number. Thus, ASCNs for normal (diploid) regions are simply the usual AA, AB, or BB. However, a

SNP in an amplified region may have ASCN $AAAAB$; a SNP in a heterozygously deleted region may have ASCN B. PSCN, on the other hand, refers to the contributions to copy number of each of the two parental chromosomes. Within this framework, for example, we may more precisely identify LOH as a region in which the PSCNs are $(c, 0)$ for some positive integer c.

Our PLASQ procedure is rooted in a generalized linear model for the behavior of probe intensities, exploiting a novel classification of the SNP array probes that is fundamentally different from the usual PM_A, MM_A, PM_B, MM_B classification. An earlier version of the procedure (LaFramboise *et al.*, 2005) — also termed PLASQ — used a simpler general linear model, and its performance with regard to genotyping and copy number determination was inferior to the version we present here. In the present work we analyze statistical properties (which were not discussed in our earlier paper) of this updated model, demonstrating the improvements in fit and performance. In light of these improvements, our intent is that the current PLASQ replace the version described in our previous work.

After specifying our model in Section 2, its fitting, via an expectation-maximization (EM) (Dempster *et al.*, 1977) algorithm that takes advantage of the inherently discrete nature of the quantity being measured, is detailed in Section 3. In Section 4, we apply our approach to a variety of cell types, demonstrating the ability to: a) very accurately genotype over 100 000 SNPs in normal samples as either AA, AB, or BB; b) determine copy number, genome-wide, at a very high resolution in cancer samples; c) reveal the contributions of each of the two parental chromosomes to the amplifications and deletions in these aberrant samples; and d) infer ASCNs at each of the SNPs on the array. We provide statistical justification for the suitability of our model, and our *in silico* results are validated using a variety

of independent *in silico* and *in vitro* methods. We conclude in Section 5 with a discussion of the relevance of our results in cancer genomics research.

2. Array design and model specification

Studies employing SNP arrays have focused almost exclusively on the PM_A, MM_A, PM_B, and MM_B probe classification. However, another classification is relevant. A PM/MM pair may either be centered precisely so that the middle (13^{th}) base of the PM probe is complementary to the SNP site, or may be offset (by between 1 and 4 bases in either direction). The three dichotomizations of the probe set therefore leave us with eight probe types: PM_A^c, MM_A^c, PM_B^c, MM_B^c, PM_A^o, MM_A^o, PM_B^o, and MM_B^o, where the superscript denotes centered (c) or offset (o). Our method focuses on the nucleotide-level affinities between each probe and the two target DNA sequences (corresponding to the two SNP alleles). We can count the number of bases at which each probe mismatches each of the target alleles; indeed, this information is encoded in the .CDF (Chip Definition File) provided by the manufacturer. Each probe mismatches each of the two target alleles by either 0, 1, or 2 bases, and the eight probe classes completely determine these counts. See Supplementary Figure 1 for a specific example of a probe set.

Our model is motivated by the following set of principles. First, the relationship between the target quantity and probe intensity is approximately linear (with an additive term) on a log-log scale, as demonstrated in studies involving known quantities of RNA (Irizarry *et al.*, 2003) and genomic DNA (Huang *et al.*, 2004). Second, the authors in Irizarry *et al.* (2003) justified, via spike-in studies, a multiplicative stochastic error term on the standard (non-log) scale, as evidenced by larger probe variance at higher intensity levels. Third, within

a probe set, each probe is complementary to a subsegment of either the forward or reverse strand in the target DNA fragment. This "forward" or "reverse" distinction is referred to as the probe's orientation, and empirical evidence indicates differences in hybridization intensities between the orientations. Finally, it is reasonable that, aside from orientation, the main factor determining probe/target hybridization affinity within the same probe set would be the number of bases that the probe mismatches the target. More specifically, we reasonably assume that the hybridization affinity of a target for a probe is a decreasing function of the number of bases at which the probe is not complementary to the target. The exception to this assumption arises in differences in the hybridization affinities of the A and B target fragments. Since the A and B difference represents the only potential significant difference in GC content between the probes in a set, we have accommodated target-allele-specific differences in hybridization affinity in our model.

In an array with J probe sets/SNPs (so $J > 100\,000$ in our case) let $C_A^{(ij)}$ and $C_B^{(ij)}$ denote the number of copies of the alleles A and B, respectively, in the i^{th} sample at the j^{th} SNP site ($j = 1, \ldots, J$). The model we propose for the normalized, log-transformed intensity $Y^{(ijk)}$ of probe k in the probe set for SNP j in an array interrogating sample i is

$$Y^{(ijk)} = \log(\gamma_{O_{jk}}^{(j)} + \alpha_{A_{jk}O_{jk}}^{(j)} C_A^{(ij)} + \beta_{B_{jk}O_{jk}}^{(j)} C_B^{(ij)}) + e^{(ijk)}. \tag{2.1}$$

Here $O_{jk} = F$ (forward) or R (reverse) denotes the orientation of the probe, A_{jk}, $B_{jk} = 0$, 1, or 2 indicate the number of bases at which the probe mismatches the A and B allele targets, respectively, and $\gamma_F^{(j)}, \gamma_R^{(j)}$ represent the unwanted background contributions of optical noise and non-specific binding to the forward

and reverse orientation probe intensities, respectively. One may think of these last terms as representing the signal from a probe whose target is completely absent. The independent, normally distributed, mean zero error terms $e^{(ijk)}$ are meant to capture additional sources of variation. They are assumed to have standard deviation $\sigma_F^{(j)}$ when $O_{jk} = F$ and $\sigma_R^{(j)}$ when $O_{jk} = R$. The distributions of these error terms are the same for any fixed values of j and O_{jk}, but are allowed to vary for different probe sets and different orientations within the same probe sets. Finally, we have found in practice that hybridization intensities between probes and targets that mismatch at two bases are indistinguishable from background noise, and thus we fix

$$\alpha_{2F}^{(j)} = \beta_{2F}^{(j)} = \alpha_{2R}^{(j)} = \beta_{2R}^{(j)} = 0.$$

Thus, the parameters of interest for each probe set/SNP j are $\gamma_F^{(j)}$, $\gamma_R^{(j)}$, $\alpha_{0F}^{(j)}$, $\alpha_{0R}^{(j)}$, $\alpha_{1F}^{(j)}$, $\alpha_{1R}^{(j)}$, $\beta_{0F}^{(j)}$, $\beta_{0R}^{(j)}$, $\beta_{1F}^{(j)}$, and $\beta_{1R}^{(j)}$.

3. Model fitting and copy number inference

Equation (2.1) above models mean log-transformed probe intensity as a log-linear function of copy number. There are some complications to fitting the model. First, the log transformation on the right side of the equation precludes the use of ordinary least squares. However, the model is a generalized linear model (McCullagh and Nedler, 1989) with an exponential link, and thus we fit the model using iteratively reweighted least squares (IRLS). A more severe obstacle to model fitting is the fact that we usually know neither parameter nor covariate values $C_A^{(ij)}$ and $C_B^{(ij)}$ *a priori*. We do know that, in a normal sample, each SNP is in one of

three states — AA, AB, or BB. This implies three different covariate combinations, and therefore an EM algorithm is a natural approach to fitting the model to diploid data. The first step is to quantile normalize (Bolstad *et al.*, 2003) the raw probe intensity data from these normal references together with those from the test cancer samples we wish to analyze. This step ensures that the results are comparable across arrays by removing differences (such as overall brightness) unrelated to the underlying molecular biology. After next estimating the model parameters from the normal references using the EM procedure, we fit to data from cancer samples (again using IRLS), which yields raw ASCNs at each SNP site. Further processing produces our final ASCN and PSCN calls. In this section, we describe each of these steps.

3.1 *Model calibration on normal samples*

For SNP arrays, normal samples provide a convenient basis for model fitting and testing, as the pairwise ASCN sums $C_A^{(ij)} + C_B^{(ij)}$ are known to be two. We exploit this fact to find estimates $\hat{\gamma}_F^{(j)}$, $\hat{\gamma}_R^{(j)}$, $\hat{\alpha}_{0F}^{(j)}$, $\hat{\alpha}_{0R}^{(j)}$, $\hat{\alpha}_{1F}^{(j)}$, $\hat{\alpha}_{1R}^{(j)}$, $\hat{\beta}_{0F}^{(j)}$, $\hat{\beta}_{0R}^{(j)}$, $\hat{\beta}_{1F}^{(j)}$, and $\hat{\beta}_{1R}^{(j)}$ of $\gamma_F^{(j)}$, $\gamma_R^{(j)}$, $\alpha_{0F}^{(j)}$, $\alpha_{0R}^{(j)}$, $\alpha_{1F}^{(j)}$, $\alpha_{1R}^{(j)}$, $\beta_{0F}^{(j)}$, $\beta_{0R}^{(j)}$, $\beta_{1F}^{(j)}$, and $\beta_{1R}^{(j)}$, respectively. Model (2.1) may be fit to (normalized) probe intensities using an EM algorithm, and the genotyping inferences automatically result. Details of this procedure are given in the Appendix.

3.2 *Parent- and allele-specific copy numbers in tumor samples*

Supplementary Figure 2 gives a diagrammtic overview of the procedure to obtain ASCNs and PSCNs from (normalized) probe-level data from tumor sample i_0. We assume that parameters have been estimated as above from a battery of normal samples, and we replace the parameters in the model with these estimates at each

SNP. Our model becomes

$$Y^{(i_0jk)} = \log(\hat{\gamma}^{(j)}_{O_{jk}} + \hat{\alpha}^{(j)}_{A_{jk}O_{jk}} C_A^{(i_0j)} + \hat{\beta}^{(j)}_{B_{jk}O_{jk}} C_B^{(i_0j)}) + e^{(i_0jk)}. \tag{3.1}$$

We may now obtain raw ASCN inferences $(C_{A\ \mathrm{raw}}^{(i_0j)}, C_{B\ \mathrm{raw}}^{(i_0j)})$ via IRLS as applied to model (3.1). In effect, we are treating the covariates $C_A^{(i_0j)}$ and $C_B^{(i_0j)}$ as parameters to be estimated. The ASCN inferences at this stage are "raw" because we have not yet taken advantage of the fact that total copy number is locally constant; that is, chromosomal copy number aberrations occur in discrete segments, typically spanning many consecutive SNP sites. We may therefore apply a smoothing or break point procedure to the pairwise sums of the raw ASCNs, mapped to their genomic locations. For our study, we have employed the GLAD algorithm (Hupé *et al.*, 2004) because of its sensitivity, specificity, and computational efficiency. GLAD attempts to detect chromosomal segments with constant total copy number using an adapted weights smoothing (Polzehl and Spokoiny, 2000) breakpoint-detection algorithm. Our inferred total copy number $T^{(i_0s)}$ for a GLAD-determined segment s is the rounded median of the pairwise raw ASCN sums in the segment.

Next, we infer PSCNs in each segment s from inferred total copy number $T^{(i_0s)}$ and raw ASCNs as follows. First, if the inferred total copy number is 0 or 1, then our PSCN calls are obviously (major chromosome, minor chromosome)= $(0,0)$ or $(1,0)$, respectively. If not, we next decide whether LOH has occurred. When a matched normal sample is available, this is easily determined by querying for homozygosity SNPs that are heterozygous in the matched normal. In the absence of a matched normal sample, we make use of the fact that the average heterozygosity rate for SNPs on the array is approximately 30% (Affymetrix,

2004). Therefore, we may think of the number of homozygous SNPs in a segement with m SNPs as an approximate Binomial$(m, 0.7)$ variable. Making a Bonferroni correction for the number S of segments, we call LOH for segments in which the number of homozygous SNPs is greater than the $1 - 0.05/S$ quantile in the Binomial$(m, 0.7)$ distribution (here a SNP j is assumed to be homozygous when the rounded minimum$(C_{A\ \mathrm{raw}}^{(i_0 j)}, C_{B\ \mathrm{raw}}^{(i_0 j)})$ is less than one). If LOH is deemed to have occurred, our PSCNs for the segment are $(T^{(i_0 s)}, 0)$. Otherwise we ignore homozygous SNPs, as they are noninformative with regard to PSCN, and our PSCN call is $(T^{(i_0 s)} - \nu, \nu)$ where

$$\nu = T^{(i_0 s)} \times \frac{\sum \mathrm{minimum}(C_{A\ raw}^{(i_0 j)}, C_{B\ raw}^{(i_0 j)})}{\sum (C_{A\ \mathrm{raw}}^{(i_0 j)} + C_{B\ \mathrm{raw}}^{(i_0 j)})}$$

rounded to the nearest integer. Both sums in this expression are taken over all heterozygous SNPs j in segment s.

Finally, we determine ASCNs from PSCNs and raw ASCNs at each SNP j. If the SNP is heterozygous, then the ASCNs are the same as the PSCNs, with the copy number of the major SNP allele (as determined by raw ASCNs) identical to that of the major parental chromosome segment. If the SNP is homozygous, the allele with the higher raw ASCN is assigned ASCN $T^{(i_0 s)}$, and the other 0.

4. Application to normal and cancer data

4.1 *Data sets*

The SNP array data are encoded in a pair of .cel files (one for each chip type) for each sample. We employed data from 21 normal samples in our study. These data include 24 .cel files from Zhao *et al.* (2005) that corresponded to all of the

normal samples in that study, as well as 18 .cel files (corresponding to samples NA6985, NA6991, NA6993, NA12707, NA12716, NA12717, NA12801, NA12812, and NA12813) that were generated as part of the International HapMap Project (http://www.hapmap.org). The latter samples, which we refer to as the HapMap data set, are available for download at the Affymetrix web site (http://www.affymetrix.com). For cancer samples, we used .cel files from 12 lung tumors and cell lines (see Tables 2 and 3) that were generated in Zhao *et al.* (2005).

4.2 *Application to normal samples*

To validate the assumptions of our model, we first fit it to the HapMap data set. We examined the residuals from the model to check the assumption of normally distributed error terms. Note that, although the error terms are assumed to be identically distributed within same-orientation subsets of a probe set, their variances are allowed to differ across probe sets and orientations. We therefore constructed a normal quantile-quantile (q-q) plot (Figure 1a) of the standardized residuals, with the understanding that the model implies a standard normal distribution for these across all probe sets. For clarity, we randomly selected 10 000 such residuals to plot. To demonstrate the necessity of the log-log transformation, we also plotted the standardized residuals resulting from fitting the linear model

$$\tilde{Y}^{(ijk)} = \gamma^{(j)}_{O_{jk}} + \alpha^{(j)}_{A_{jk}O_{jk}} C_A^{(ij)} + \beta^{(j)}_{B_{jk}O_{jk}} C_B^{(ij)} + e^{(ijk)}, \tag{4.1}$$

where $\tilde{Y}^{(ijk)}$ now denotes the normalized, but untransformed, probe intensity. We note that the model in LaFramboise *et al.* (2005) was similar to (4.1), but even simpler — it did not allow for different coefficients for the C_A and C_B terms, and thus forced $\alpha^{(j)}_{A_{jk}O_{jk}} = \beta^{(j)}_{B_{jk}O_{jk}}$ for each $j = 1, \ldots, J$ and $k = 1, \ldots, 40$. We fit

(4.1) using the EM algorithm as with model (2.1), except that the M-step involves ordinary least squares rather than IRLS. The resulting q-q plot (Figure 1b) clearly shows a severe departure from normality. This demonstrates the improvement of our new generalized linear model-based approach over the previous work.

As mentioned above, probes on SNP arrays have traditionally been classified in PM/MM or A allele/B allele terms. The advantage of our approach — classifying probes by base mismatch count — can be seen in Figure 2. The first scatterplot shows the mean MM_A^o intensity versus the mean MM_A^c intensity across 10 782 HapMap sample SNPs. Each point represents one orientation (F or R) of one SNP for one sample. The means are taken over all MM_A^c (x-axis) or MM_A^o (y-axis) probe intensities for the given orientation/SNP/sample. Each point is colored according to HapMap genotype. Although the traditional classification treats these two probe types as being equivalent measures, there is clearly a separation of the three genotypes visible in the plot. As expected, the centered probes generally have a greater affinity for the B target than the offset probes, and both types have roughly the same affinity for the A target. This effect is even more dramatic when the background γ term is subtracted, as shown in Figure 2b. These figures show that the practice of ignoring MM probes, as some approaches do, in fact discards relevant information. Moreover, if we construct a similar plot for MM_A^c versus MM_B^c (Figure 2c), no separation of the genotypes is discernible, even though the traditional classification would treat these two intensities as being measures of separate quantities.

Many of the SNPs in the HapMap data set have been independently genotyped, using a variety of genotyping platforms. Of these, 1198 were genotyped by at least two different HapMap centers. Calls that were concurrent among at least

two different centers may be considered as being very close to ground truth, and we employed these as the "gold standard" data set against which we compared our PLASQ method. As shown in Table 1, our method performs quite well. The rate of agreement between PLASQ and the HapMap concordant calls is similar to the HapMap Project's concordance rate, and our No Call rate is considerably lower. We should note that the 16 sample SNPs in the Table for which PLASQ called AA and the HapMap effort called BB are all from the same two SNP loci. Close inspection of the raw array data from these SNPs reveals a strong AA signal (data not shown). Thus, we suspect that this is simply a case of an error being made by Affymetrix when the "A" and "B" labels were assigned to the nucleotide residues. In any case, the results in the Table clearly indicate that the model captures the relevant aspects of the data, and underscore the validity of our EM fitting approach.

4.3 *Application to lung cancer*

We applied our PLASQ method to SNP array data from 12 lung cancer samples, using the 12 diploid samples from the same study as normal references on which to train the model. Figure 3 shows an example of a genome-wide view of PSCN for one of these samples, the cell line H2087. Note that LOH is clearly identifiable as a region comprised of only the major chromosome (all green). For example, all of one copy of chromosome 13 appears to be lost, though the total copy number remains at two. This phenomenon is referred to as copy-neutral LOH.

To assess the accuracy of our method, we compared our results to PCR-based copy number estimates. A total of 16 deletions and 10 amplifications in our 12 lung cancer samples were previously PCR-measured in Zhao *et al.* (2005). These PCR measurements quantified only total (not allele-specific) copy number, so we

have developed an experimental method to measure copy number on an allele-specific basis. This quantitative PCR-based method is described in LaFramboise *et al.* (2005). Tables 2 and 3 compare the PLASQ results with the PCR results. As quantitative real-time PCR is a very sensitive technique, the putative homozygous deletions in Table 2 are most likely valid. Our PLASQ procedure is able to identify each deletion, and in fact they are almost always apparent at the raw ASCN level. The estimates for amplifications are, however, not as concordant. Although PLASQ detects each amplification, the results tend to be lower than the PCR-based estimates in the higher-copy-number alleles. This is quite possibly due to well-known saturation effects in oligonucleotide arrays (Naef *et al.*, 2003), and is difficult to mitigate. On the other hand, it is possible that our allele-specific quantitative PCR technique may not be a precise measure. In any case, an argument could be made that these errors are of little consequence, as the aim in these studies is to identify amplifications, deletions, and the haplotypes involved, all of which PLASQ can clearly uncover.

5. Discussion

Human cancer is driven by the acquisition of genomic changes in the cell. One extremely important class of such changes is amplifications and deletions — deviations from the normal two copies of each chromosome in a cell. Regions of amplification may harbor cancer-causing oncogenes, while deletions often contain tumor suppressor genes. The localization of such alterations is therefore a central goal in cancer research. We have presented a procedure, PLASQ, for determining the copy numbers of SNP alleles and parental chromosomes in cancer cells from SNP array data. Our SNP allele copy number result is particularly of interest in

LOH determination, since existing methods often mistakenly call LOH where in fact allelic balance (due to amplification of one allele) has occurred, resulting in apparent (though false) homozygosity. We avoid these false LOH calls by taking into accont the contribution to copy number from both alleles. Two recent papers (Ishikawa *et al.*, 2005; Nannya *et al.*, 2005) have been published that aim to determine parent-specific copy number. However, their approaches require additional SNP array data from matched normal cells, which are often unavailable. Moreover, both methods ignore MM probes, and thereby discard half of the information available in SNP arrays. As we have shown, MM probes are in fact informative.

Finally, we should mention two potential weaknesses of our approach. First, we are assuming a diploid copy number two in autosomal chromosomes of normal cells. Recent studies (Iafrate *et al.*, 2004; Sebat *et al.*, 2004) have uncovered copy number polymorphisms in normal cells. Given that our approach (and all others that we are aware of) compares signal intensities to normal references, this could in theory present a problem. In practice, however, we feel that this problem is mitigated by the fact that we use a sizable collection of normal reference samples, and that polymorphic genomic regions common to most normal reference samples are likely rare, small in length, or both. A second concern is our practice of fitting the model to normal samples and then applying the result to data from tumors. We are implicitly assuming that the model parameters are appropriate outside of the range of covariates with which they were estimated. Although this is indeed a concern (and may be partially to blame for copy number underestimation of high-level amplifications), we would argue that the results shown in Tables 2 and 3 demonstrate the value of the model, even for aberrant copy numbers.

All procedures described herein are available in an R (R Development Core Team, 2006) package, freely downloadable at
http://genome.dfci.harvard.edu/∼tlaframb/PLASQ

Acknowledgements

We wish to acknowledge the contributions of the referees and editors, whose insightful comments resulted in a much-improved paper. We also thank Matthew Meyerson for support and guidance during the early development of this work. David Harrington was supported by NIAID grant 2R01 AI052817.

APPENDIX

We describe in detail the EM approach to fitting model (2.1) to probe-level SNP array data from normal samples.

Notation. Fix an arbitrarily chosen SNP j_0. Suppose that we have N normal samples. For $i = 1, \ldots, N$ and $l = 0, 1, 2$, let Z_{ij_0l} denote the (unobserved) indicator variable $I(C_A^{(ij_0)} = l)$. Model (2.1) may be rewritten, using this notation, as

$$Y^{(ij_0k)} = \log(\gamma^{(j_0)}_{O_{j_0k}} + \alpha^{(j_0)}_{A_{j_0k}O_{j_0k}}(Z_{ij_01} + 2Z_{ij_02}) + \beta^{(j_0)}_{B_{j_0k}O_{j_0k}}(2Z_{ij_00} + Z_{ij_01})) + e^{(ij_0k)}. \tag{A.1}$$

We think of the Z_{ij_0l} as missing data, whose values provide the genotypes of our samples. Let $\phi(x \mid \mu, \tau)$ denote the density function of the normal distribution with mean μ and variance τ^2, and let $\mathbf{Y}^{(ij_0)}$ denote the data vector $(Y^{(ij_0k)})_{k=1,\ldots,40}$ from probe set j_0 for sample i. For $l = 0, 1, 2$, let p_{j_0l} denote the (unknown) proportion of samples for which $C_A^{(ij_0)}$ is l at the SNP j_0. We consider the p_{j_0l} to be part of the

set $\mathbf{\Psi}$ of parameters (which also includes the α, β, γ, and σ model parameters) to be estimated during the M-step. It follows from (A.1) that the density function for $\mathbf{Y}^{(ij_0)}$ is

$$f_{\mathbf{Y}^{(ij_0)}}(\mathbf{y}) = \sum_{l=0}^{2} p_{j_0 l} f_{j_0 l}(\mathbf{y})$$

where

$$\begin{aligned}
f_{j_0 0}(\mathbf{y}) &= \prod_{k=1}^{40} \phi(y_k \mid \log(\gamma^{(j_0)}_{O_{j_0 k}} + 2\beta^{(j_0)}_{B_{j_0 k} O_{j_0 k}}), \sigma^{(j_0)}_{O_{j_0 k}}) \\
f_{j_0 1}(\mathbf{y}) &= \prod_{k=1}^{40} \phi(y_k \mid \log(\gamma^{(j_0)}_{O_{j_0 k}} + \alpha^{(j_0)}_{A_{j_0 k} O_{j_0 k}} + \beta^{(j_0)}_{B_{j_0 k} O_{j_0 k}}), \sigma^{(j_0)}_{O_{j_0 k}}) \\
f_{j_0 2}(\mathbf{y}) &= \prod_{k=1}^{40} \phi(y_k \mid \log(\gamma^{(j_0)}_{O_{j_0 k}} + 2\alpha^{(j_0)}_{A_{j_0 k} O_{j_0 k}}), \sigma^{(j_0)}_{O_{j_0 k}})
\end{aligned}$$

We refer to the vector $(\mathbf{Y}^{ij_0}, \mathbf{Z}_{ij_0}) = (Y^{(ij_0 k)}, Z_{ij_0 l})_{k=1,\ldots,40;\ l=0,1,2}$ as a the complete data vector. The complete data density is

$$\begin{aligned}
f^{\text{comp}}_{(\mathbf{Y}^{(ij_0)}, \mathbf{Z}_{ij_0})}(\mathbf{y}, \mathbf{z}) &= (p_{j_0 0} f_{j_0 0}(\mathbf{y}))^{z_1} (p_{j_0 1} f_{j_0 1}(\mathbf{y}))^{z_2} (p_{j_0 2} f_{j_0 2}(\mathbf{y}))^{z_3} \\
&= p^{z_1}_{j_0 0} p^{z_2}_{j_0 1} p^{z_3}_{j_0 2} g_{j_0}(\mathbf{y}, \mathbf{z}), \qquad \text{(A.2)}
\end{aligned}$$

where

$$g_{j_0}(\mathbf{y}, \mathbf{z}) = \prod_{k=1}^{40} \phi(y_k \mid \log(\gamma^{(j_0)}_{O_{j_0 k}} + \alpha^{(j_0)}_{A_{j_0 k} O_{j_0 k}}(z_2 + 2z_3) + \beta^{(j_0)}_{B_{j_0 k} O_{j_0 k}}(2z_1 + z_2)), \sigma^{(j_0)}_{O_{j_0 k}})$$

and $\mathbf{z} = (z_1, z_2, z_3)$.

Initialization. We have found our procedure to be somewhat sensitive to starting values for the missing data. Therefore, rather than randomly assigning these values as a first step, we use a reasonable yet crude t-test approach to provide initial values $z^{(0)}_{ij_0 l}$ of the expectations of the $Z_{ij_0 l}$. For each i, a one-sided t-test is per-

formed for the null hypothesis that the mean of the (normalized, log-transformed) PM_A probe intensities is larger than that of the PM_B probes. Let P denote the resulting P-value. If $P \leq 0.5$, we assign initial probabilities $(z_{ij_00}^{(0)}, z_{ij_01}^{(0)}, z_{ij_02}^{(0)}) = (\frac{P}{2}, P, 1 - \frac{3P}{2})$. If $P > 0.5$, we assign $(z_{ij_00}^{(0)}, z_{ij_01}^{(0)}, z_{ij_02}^{(0)}) = (\frac{3P}{2} - \frac{1}{2}, 1 - P, \frac{1-P}{2})$.

M-step. For the m^{th} M-step, we consider the complete data log likelihood, assuming the current expectations $z_{ij_00}^{(m-1)}$, $z_{ij_01}^{(m-1)}$, and $z_{ij_02}^{(m-1)}$ for the values of the missing data along with the observed data $\mathbf{Y}^{(ij_0)} = \mathbf{y}^{(ij_0)}$. By the factorization in expression (A.2), this log likelihood can be written as

$$\log \mathcal{L}^{(m)} = \sum_{i=1}^{N} \sum_{l=0}^{2} z_{ij_0l}^{(m-1)} \log p_{j_0l} + \sum_{i=1}^{N} \log g_{j_0}(\mathbf{y}^{(ij_0)}, \mathbf{z}_{ij_0}^{(m-1)}).$$

On the right side of this equation, the p_{j_0l} appear only in the first term, while the α, β, γ, and σ parameters appear only in the second term. Thus, we may maximize each term separately. It is easy to see that the first expression, subject to the constraint $p_{j_00} + p_{j_01} + p_{j_02} = 1$, is maximized at the values

$$\hat{p}_{j_0l}^{(m)} = \sum_{i=1}^{N} z_{ij_0l}^{(m-1)} / N.$$

The maximum likelihood estimates for the model parameters may be computed using iteratively reweighted least squares, as applied to the model (A.1) with the Z_{ij_0l} replaced by $z_{ij_0l}^{(m-1)}$ and the $\mathbf{Y}^{(ij_0)}$ by $\mathbf{y}^{(ij_0)}$.

E-step. We find the expected values $z_{ij_0l}^{(m)}$ of the Z_{ij_0l} based on the m^{th} M-step parameter estimates $\hat{\mathbf{\Psi}}^{(m)}$. Given that the value of Z_{ij_0l} is either 0 or 1, we have

$$z_{ij_0l}^{(m)} = E_{\hat{\mathbf{\Psi}}^{(m)}}[Z_{ij_0l} | \mathbf{Y}^{(ij_0)} = \mathbf{y}^{(ij_0)}] = P_{\hat{\mathbf{\Psi}}^{(m)}}[Z_{ij_0l} = 1 | \mathbf{Y}^{(ij_0)} = \mathbf{y}^{(ij_0)}].$$

By Bayes' Theorem and (A.2), we have

$$z_{ij_0l}^{(m)} = \hat{p}_{j_0l}^{(m)} f_{j_0l}(\mathbf{y}^{(ij_0)}) / f_{\mathbf{Y}^{(ij_0)}}(\mathbf{y}^{(ij_0)}),$$

where the density functions use $\hat{\mathbf{\Psi}}^{(m)}$ for their parameter values.

The E- and M-steps are alternated repeatedly until the changes in the estimates are very small, say after m_0 steps. In this way, we obtain two important results. First, model parameter estimates are produced, which can be used in (2.1) to fit to SNP data from any sample, producing raw allele-specific copy number estimates at the SNP as demonstrated in Section 3.2. Second, the $z_{ij_0l}^{(m_0)}$ may be used to infer genotypes for the normal samples. If a call is desired for sample i, a simple rule would be:

$$\text{genotype}_i = \begin{cases} \text{homozygous } AA & \text{if} \quad \arg\max_l(z_{ij_0l}^{(m_0)}) = 2 \\ \text{heterozygous } AB & \text{if} \quad \arg\max_l(z_{ij_0l}^{(m_0)}) = 1 \\ \text{homozygous } BB & \text{if} \quad \arg\max_l(z_{ij_0l}^{(m_0)}) = 0 \end{cases}$$

This scheme automatically provides a way to measure uncertainty in the genotype calls. The researcher may set a threshold for the value of $\max_l(z_{ij_0l}^{(m)})$, below which the call is considered uncertain and a "No Call" determination is given. In practice, we have found 99% to be a suitable such threshold.

FIGURES

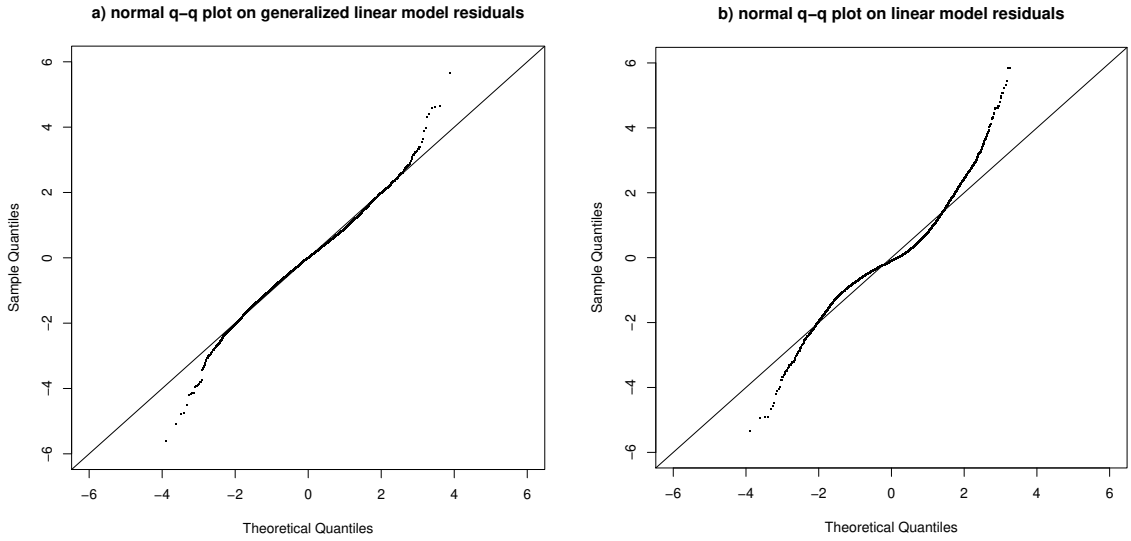

Figure 1: Normal quantile-quantile plots comparing standardized residuals to the standard Gaussian distribution. a) 10 000 randomly-selected residuals from the generalized linear model (2.1) fit to SNP array data from HapMap samples. b) 10 000 randomly-selected residuals from the linear model (4.1) fit to SNP array data from HapMap samples.

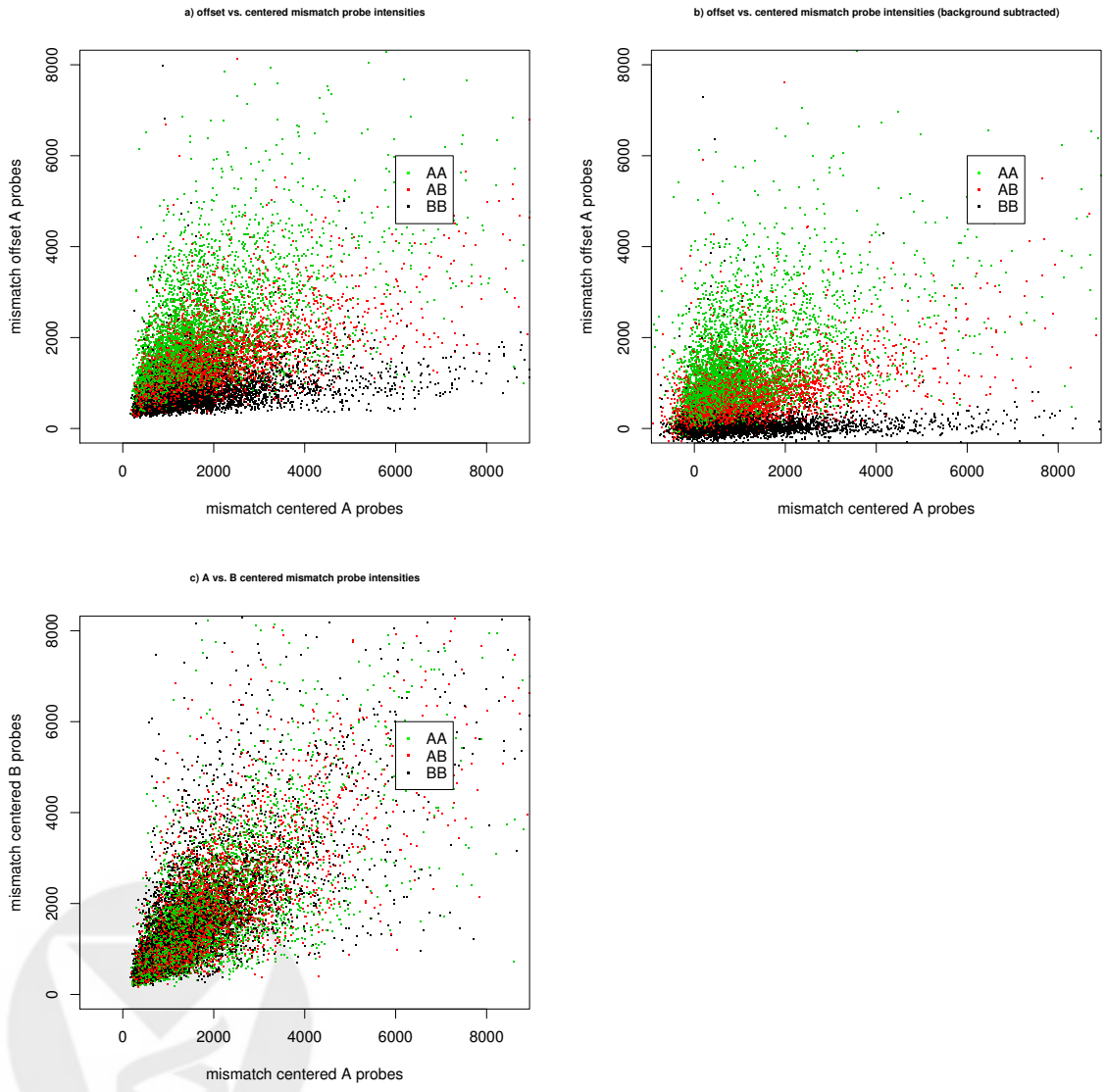

Figure 2: Scatterplots of mean intensities of probe types across 10 782 sample SNPs. a) MM_A^o probes vs. MM_A^c probes. Although traditionally considered to be of the same type, these probes clearly behave differently with different genotypes. b) The differences from a) are even more pronounced when background is subtracted. c) No such difference is apparent in MM_B^c vs. MM_A^c, even though these are traditionally considered to be measuring different alleles.

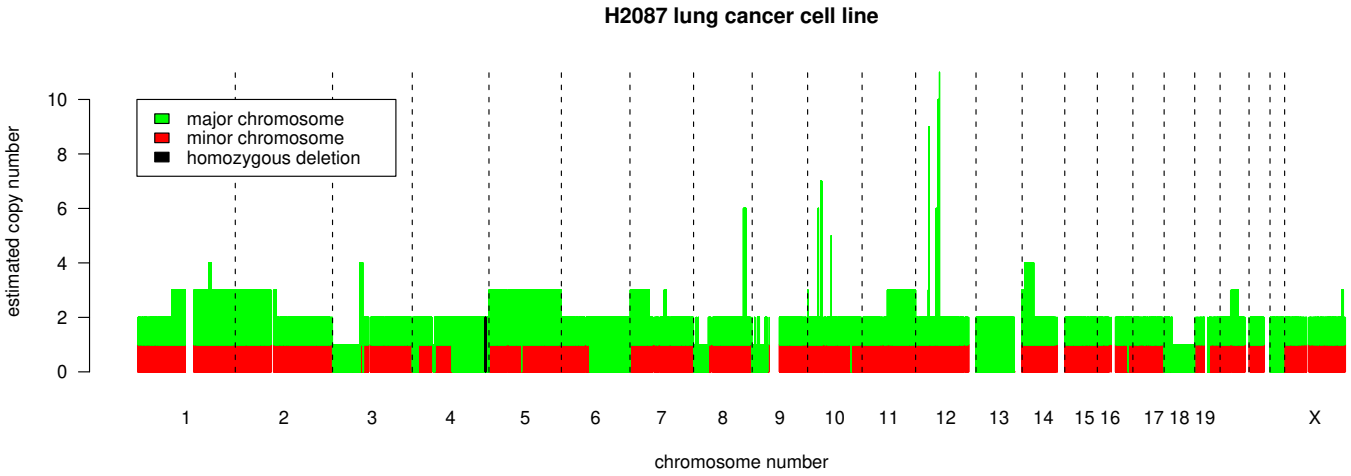

Figure 3: Parent-specific copy number of the H2087 cell line, as determined by the PLASQ procedure. Various types of genomic lesions are apparent in this view. For example, there are high-level amplifications on chromosome 12, copy-neutral LOH on chromosome 13, heterozygous deletion of the p arm of chromosome 3, and a focal homozygous deletion (thin black bar) on chromsome 4.

Tables

	PLASQ AA	PLASQ AB	PLASQ BB	PLASQ No Call	totals
HapMap AA	3787 (35.12%)	4 (0.04%)	1 (0.01%)	2 (0.02%)	3794 (35.19%)
HapMap AB	15 (0.14%)	3158 (29.29%)	4 (0.04%)	6 (0.05%)	3183 (29.52%)
HapMap BB	16* (0.15%)	4 (0.04%)	3595 (33.34%)	11 (0.10%)	3626 (33.63%)
HapMap No Call	44 (0.41%)	49 (0.45%)	46 (0.43%)	2 (0.02%)	141 (1.31%)
HapMap Discordant	5 (0.05%)	24 (0.22%)	9 (0.08%)	0 (0%)	38 (0.35%)
totals	3867 (35.87%)	3239 (30.04%)	3655 (33.90%)	21 (0.19%)	10,782 (100%)

Table 1: Concordance between our procedure's calls and those made by more than one center in the International HapMap Project effort. The HapMap calls are considered discordant if any two centers, neither producing a No Call for the SNP, call it differently. If all but one center produce a No Call, the SNP is placed in the table's "HapMap No Call" category. *Likely the result of mislabeled A and B alleles (see text).

SNPID (rs)	Chromosome	Position (Mb)	Sample	raw allele A ASCN	raw allele B ASCN	PLASQ allele A ASCN	PLASQ allele B ASCN	Real-time PCR copy number[a]
4133302	2	142.07	H2126	−0.05	−0.03	0	0	0.00
4133302	2	142.07	H2122	−0.04	−0.07	0	0	0.01
4133302	2	142.07	H157	0.63	−0.06	0	0	0.06
10496876	2	142.29	HCC95	−0.02	0.01	0	0	0.00
2687167	3	60.32	HCC95	−0.01	−0.05	0	0	0.00
930589	3	152.87	H2882	0.09	−0.03	0	0	0.00
930589	3	152.87	S0177T	0.13	0.01	0	0	0.02
2033554	9	8.73	S0177T	−0.12	0.15	0	0	0.01
655125	9	9.59	HCC1171	0.22	0.04	0	0	0.08
4074785	9	21.97	HCC1359	−0.04	−0.08	0	0	0.00
4074785	9	21.97	H2126	−0.04	0.00	0	0	0.00
4074785	9	21.97	H2122	0.05	−0.04	0	0	0.01
4074785	9	21.97	H2882	0.05	−0.05	0	0	0.00
4074785	9	21.97	HCC1171	−0.05	0.10	0	0	0.00
4074785	9	21.97	HCC95	0.00	0.12	0	0	0.00
1162609	9	24.58	H157	0.03	−0.03	0	0	0.03

Table 2: Comparison of raw and inferred ASCNs with PCR results for deletions.
[a]From Zhao *et al.* (2005).

SNP ID (rs)	Chromosome	Position (Mb)	Sample	raw allele A ASCN	raw allele B ASCN	PLASQ allele A ASCN	PLASQ allele B ASCN	PCR allele A copy number	PCR allele B copy number
4859257	3	183.98	S0465T	6.89	0.70	6	1	25.18	1.68
2049284	3	183.49	S0515T	−0.38	22.90	0	14	2.42	38.37
1569265	7	54.61	HCC827	10.72	0.99	11	1	135.92	1.97
2893603	8	128.04	H2122	6.61	−0.11	7	0	58.46	3.39
9283954	8	128.33	HCC827	0.28	6.69	0	6	0.06	7.58
2392827	8	128.91	H2087	5.63	0.87	5	1	6.03	1.23
10506101	12	32.60	S0515T	−0.13	10.94	0	10	0.06	7.12
1486883	12	33.80	H2087	7.55	0.10	9	0	17.32	0.03
3913094	12	57.20	H2087	9.19	0.21	10	0	4.86	0.17
448041	22	19.77	HCC1359	0.93	6.32	1	4	1.03	8.36

Table 3: Comparison of raw and inferred ASCNs with PCR results for amplifications.

References

Affymetrix (2004). *GeneChip Human Mapping 100K Set Data Sheet*. Santa Clara, CA.

Bignell, G. R., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigorova, M., Jones, K. W., Wei, W., Stratton, M. R., Futreal, P. A., Weber, B., Shapero, M. H., and Wooster, R. (2004). High- resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research* **14**, 287–295.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Huang, J., Wei, W., Zhang, J., Liu, G., Bignell, G. R., Stratton, M. R., Futreal, P. A., Wooster, R., Jones, K. W., and Shapero, M. H. (2004). Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Human Genomics* **1**, 287–299.

Hupé P., Stransky N., Thiery J. P., Radvanyi F., and Barillot E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413–3422.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics* **36**, 949–951.

Irizarry, R. A., Hobbs, B., Collin, F., Beaxer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Ishikawa, S., Komura, D., Tsuji, S., Nishimura, K., Yamamoto, S., Panda, B., Huang, J., Fukayama, M., Jones, K. W., and Aburatani, H. (2005). Allelic dosage analysis with genotyping microarrays. *Biochemical and Biophysical Research Communications* **333**, 1309–1314.

LaFramboise, T. L., Weir, B. A., Zhao, X., Beroukhim, R., Li, C., Harrington, D., Sellers, W. R., and Meyerson, M. (2005) Allele-Specific Amplification in Cancer Revealed by SNP Array Analysis. *PLoS Computational Biology*, **1**(6): e65.

Lin, M., Wei, L. J., Sellers, W. R., Lieberfarb, M., Wong, W. H., and Li, C. (2004). dChipSNP: Significance Curve and Clustering of SNP-Array-Based Loss-of-Heterozygosity Data. *Bioinformatics* **20**, 1233–1240.

Lindblad-Toh, K., Tannenbaum, D. M., Daly, M. J., Winchester. E., Lui, W. O., Villapakkam, A., Stanton, S. E., Larsson, C., Hudson, T. J., Johnson, B. E., Lander, E. S., and Meyerson M. (2000). Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nature Biotechnology* **18**, 1001–1005.

McCullagh, P. and Nedler, J. A. (1989). *Generalized Linear Models*, 2nd ed. Boca Raton, Florida: CRC Press.

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D. K., Kennedy G. C., and Ogawa, S. (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Research* **65**, 6071–6079.

Naef F., Socci, N. D., and Magnaso M. (2003). A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations. *Bioinformatics* **19**, 178–184.

Polzehl, J. and Spokoiny, S. (2000). Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society, Series B* **62**, 335–354.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528.

Zhao, X., Weir, B. A., LaFramboise, T., Lin, M., Beroukhim, R., Garraway, L., Beheshti, J., Lee, J. C., Naoki, K., Richards, W. G., Sug-

ARBAKER, D., CHEN, F., RUBIN, M. A., JANNE, P. A., GIRARD, L., MINNA, J., CHRISTIANI, D., LI, C., SELLERS, W. R., AND MEYERSON, M. (2005). Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Research* **65**, 5561–5570.