Supplemental material to A multilevel model to address batch effects in copy number estimation using SNP arrays

Robert B. Scharpf, Ingo Ruczinski, Benilton Carvalho,
Betty Doan, Aravinda Chakravarti, and Rafael A. Irizarry

Contents

A Previous Work

This paper describes the first of a three-tiered approach for the analysis of chromosomal alterations in high-throughput platforms (Scharpf et al., 2008). Briefly, first tier methods provide locus-specific estimates of copy number. Existing methods include those that provide estimates of total copy number relative to a reference (Bignell et al., 2004; Bengtsson et al., 2008), allele-specific copy number relative to a reference (Nannya et al., 2005; Huang

et al., 2006), or absolute estimates of allele-specific copy number (LaFramboise et al., 2006; Wang et al., 2007). Second tier algorithms smooth the locus-specific estimates within an individual as a function of the genomic physical position to identify alterations spanning multiple loci. This includes segmentation algorithms (Olshen et al., 2004; Hupe et al., 2004), regression-based smoothing methods (Huang et al., 2006; Rigaill et al., 2008), hidden Markov models (Colella et al., 2007; Lamy et al., 2007; Wang et al., 2007; Korn et al., 2008; Scharpf et al., 2008), or a combination. For instance, Rigaill et al. employs an iterative approach that involves segmentation (Hupe et al., 2004) and regression. A critical choice governing the suitability of a smoothing algorithm is whether cell contamination is thought to have occurred. Specifically, a mixture of cell populations can give rise to non-integer copy numbers. While hidden Markov models (HMMs) can jointly model the genotype and copy number information to identify copy-neutral regions of homozygosity in addition to copy number gains and losses (Colella et al., 2007; Wang et al., 2007; Scharpf et al., 2008), HMMs typically assume integer copy number states. Continuous state HMMs or HMMs that estimate the fraction of contaminated cells (Lamy et al., 2007) may represent viable alternatives. As segmentation algorithms can theoretically detect any non-integer change in the copy number, nonparametric methods are often preferable when there is evidence of two or more cell populations. Finally, third tier methods assess the contribution of chromosomal alterations to phenotypes in studies involving many individuals (Purcell et al., 2007; Barnes et al., 2008).

Considerations for locus-level copy number estimation. A common approach for locus-level estimation of copy number is to estimate the ratio or log ratio of the intensities at each loci relative to a reference (Bignell et al., 2004; Golden Helix, 2009; Bengtsson et al., 2008). Disadvantages of this approach include (i) the explicit requirement of a reference set, (ii) a deviation from a ratio of one can represent an alteration in either the reference or the test sample making it difficult to hypothesize about a dosage effect on phenotype, and (iii) information on the allelic copy number at polymorphic loci is often ignored. Our preference

is a quantitation of the allelic copy number dosage in both normal and disease samples.

Two critical features when estimating copy number at each locus are probe- and batch-effects. Probe effects represent variation in the observed fluorescence intensities that arise as a result of characteristics of the probe, namely the sequence. Probe effects are present in virtually all hybridization-based platforms, including gene expression microarrays. Model-based approaches for normalizing gene expression data have been useful for reducing nonbiological variation in the raw intensities that arise as a results of differences at the sequence level, such as GC content (Wu et al., 2004). In contrast to probe effects, batch effects comprise systematic differences in the intensities across samples.

General framework for locus-level models. A general framework for modeling the normalized fluorescence intensities yin gene expression arrays has been recently described (Wu and Irizarry, 2007). Specifically, Wu and Irizarry decompose the observed probe-level fluorescence intensities into optical background, non-specific binding, and specific binding,

$$\text{Observed}_{gij} \quad = \quad \text{Background}_{gij} + \text{Nonspecific}_{gij} + \text{Specific}_{gij}, \tag{9}$$

for gene $g = 1, \ldots, G$, probe $i = 1, \ldots, I_g$, and array $j = 1, \ldots J$. In the context of hybridization-based technologies, each component has an error term that is approximately log-Normal. Probe and batch-effects have been observed in genotyping platforms (LaFramboise et al., 2006; Rabbee and Speed, 2006; Beroukhim et al., 2007; Carvalho et al., 2007; Wang et al., 2008; Korn et al., 2008). Existing models for copy number estimation that fit into the general framework proposed by Wu and Irizarry include the probe-level (feature-level) model proposed by LaFramboise et al. (2006) and a locus-level model proposed by Wang et al. (2008), whereby statistical summaries of the feature-level intensities are treated as the observed data.

A.1 Feature- and locus-level models

Feature-level models. LaFramboise et al. (2006) developed a probe-level allele-specific quantitation (PLASQ) algorithm that models the feature-level intensities as a linear function of copy number on the log scale. The quantile-normalized log intensity for each feature on the array is decomposed as background, specific hybridization, nonspecific hybridization, and error. An iteratively reweighted least squares approach is used to estimate the parameters in a set of normal samples where the number of copies of allele A and allele B are treated as known covariates. In a set of test samples, the parameters for background, specific-hybridization, and cross-hybridization are now assumed to be known and the allele-specific copy number is estimated via iteratively reweighted least squares.

While fundamentally sound, there are several practical drawbacks to this approach. First, a set of normal controls is not always available. Because of genome-wide batch effects (Sections 2 and 5), the use of historical controls as part of any copy number estimation algorithm has limited value. A second drawback is computational. An iterative estimation procedure embedded within a feature-level model for the observed intensities is computationally intensive. Notably, PLASQ was first developed for the Affymetrix 100k arrays. The more recent Affymetrix 5.0 and 6.0 platforms have an order of magnitude more probes. Finally, the advantage of a feature-level model for platforms that contain sets of identical probes for each locus, such as the Affymetrix 6.0, is less clear. An approach that first summarizes the normalized probe-level intensities to the level of the locus has clear practical advantages that may outweigh the benefits of modeling the probe-level variation. A more thorough comparison of these two approaches has not been explored.

Locus-level models. Locus-level models for the *summarized* intensities have been used by several algorithms (Huang et al., 2006; Wang et al., 2008). Algorithms that provide allele-specific estimates of copy number generally use a variation of the following approach. First, biallelic genotypes are called on a set of samples from a normal training set. The allele-

specific copy number is assumed to be known from the biallelic genotype calls on this training set. In particular, the number of copies of the A and B alleles, denoted as (c_A, c_B), is $(2, 0)$ for genotypes AA, $(1, 1)$ for genotypes AB, and $(0, 2)$ for genotypes BB. Secondly, a procedure is used to estimate parameters that roughly correspond to the level of background, nonspecific hybridization, and specific hybridization. Several different approaches for estimating these parameters have been proposed, including recent approaches that take into account the correlation of the summarized intensities for the A and B alleles (Wang et al., 2008). For instance, Wang et al. compute the within-genotype average for each allele at each locus, and then regress the within-genotype averages on the allele-specific copy obtained from the biallelic genotypes. The coefficients from this regression can be used to predict the locations for other copy numbers. In addition, Wang et al. describe an approach for obtaining the posterior mean copy number that can be used for classification of discrete copy number classes.

While more amenable computationally for recent arrays, existing locus-level models for copy number estimation do not accommodate batch effects that persist after preprocessing. One approach is to fit the software separately to each plate. For instance, this is an approach advocated by Birdsuite (Korn et al., 2008; McCarroll et al., 2008). In our experience, batch effects persist in the smoothed estimates returned by the Birdseye HMM and the Canary algorithms, two components of their suite of software (e.g., supplementary Figure 3). Furthermore, Birdsuite does not currently provide locus-level estimates of copy number whereby one can more effectively assess cell contamination and batch effects. Similarly, the software proposed by Wang et al. precomputes parameter estimates from training data (Wang et al., 2008).

In summary, we believe locus-level models are attractive with fewer computational drawbacks than feature-level models. Improvements are needed to account for batch effects that persist after preprocessing, as well the potential to improve locus-level estimates of the uncertainty by borrowing strength from the millions of other loci interrogated by these platforms.

A.2 Batch effects

Batch effects can occur as a result of differences between laboratories in the handling and preparation of biological samples, as well as changes in reagents and experimental conditions over time within a laboratory. Batch effects have been previously observed and described for genotyping methods. Genotype calls for most algorithms are concordant for over 99.5% of the measured SNPs in the Affymetrix SNP arrays when the performance is assessed on individuals in the HapMap study (Consortium, 2003). Nevertheless, important differences emerge as a result of batch effects. To illustrate, supplementary Figure 1 compares two approaches for genotyping Affymetrix 6.0 data where the same HapMap samples were processed at two different labs denoted as Lab A and Lab B. The plotting symbols denote the true genotypes assigned by HapMap and the ellipses denote the prediction regions for the genotype calls in the two labs. The default software for genotyping the Affymetrix 6.0 data, Birdseed, uses plots of the A versus B allele intensities to make genotype calls (left panel). For Lab A, Birdseed makes zero mistakes, but for Lab B Birdseed makes 41 mistakes. The reason for the number of mistakes is the large shift in the A and B intensities between labs. The right panel displays a plot of the log-ratio versus the total intensity that is used for genotyping by the Corrected Robust Linear Model with Maximum-likelihood based distances (CRLMM) algorithm (Carvalho et al., 2007). Because the log-ratio is less susceptible to batch effects, the *crlmm* algorithm makes fewer mistakes in Lab B (right panel). Hence, while genotyping can be made robust to batch effect, estimates of copy number that are based on the signal abundance are much more susceptible to batch effects.

B Technical considerations for Table 1

Supplementary Table 1 (identical to Table 1 in the main text) offers a comparison of two implementations of a hidden Markov model (HMM) on the Chakravarti dataset in which 26 of the 96 samples had chromosome 21 trisomy. As the true copy number variant region is the entire chromosome 21, there is no penalty for over-smoothing that one would

expect for the detection of smaller micro-deletions and amplifications. The smoothness for the two HMMs considered in this analysis is a function of the transition probabilities and the emission probabilities. For the latter, Birdseye and VanillaICE assume a Gaussian distribution of the normalized intensities. The smoothness of the state path is influenced by the locus-level uncertainty estimates, with larger standard errors tending to promote a more smooth state sequence. While we have not formally compared the locus-level estimates of uncertainty from Birdsuite to the variance estimates in *crlmm*, we expect the estimates to differ as Birdsuite does not quantile normalize to a target reference distribution and *crlmm* quantile normalizes to a reference distribution obtained from HapMap. Furthermore, *crlmm* shrinks the locus-specific variance estimates to the population average estimated from all loci. With respect to transition probabilities, both *crlmm* and Birdseye model the probability of transitioning between two states as a function of the genomic distance, $e^{-d/c}$, where d is the distance between adjacent markers and c is a constant. Larger values of c control the smoothness and recommended values are based on the expected smoothness for a specific platform. The distance between adjacent markers in the two implementations depends on which markers were used in the analysis. For the VanillaICE HMM, we performed an analysis with polymorphic markers only and an analysis with the full set of markers. For the Birdseye HMM, the distance depends on whether any of the markers were removed as outliers; Canary calls were used in place of the Birdseye predictions in regions of common copy number polymorphisms (McCarroll et al., 2008). Using the default values of c in both implementations (see software versions in Section I of the supplementary material.), the probability of transitioning between any two states was multiplied by 0.005, 0.5, or 0.0025 as discussed in the supplementary material to Korn et al. (2008).

C Bipolar dataset: comparison with Birdsuite

We also fit the Birdsuite software to each plate in the bipolar controls. The Birdsuite software uses separate algorithms for calling copy number: a HMM for discovery of de-novo copy number variant (Birdseye) and one for calling copy number in regions that are

Copy number 2		$\widehat{CN} = 1$	$\widehat{CN} = 2$	$\widehat{CN} = 3$
Birdseye / Canary	SNPs + NPs	0.0042	0.9914	0.0043
crlmm & VanillaICE	SNPs	0.0003	0.9957	0.0041
crlmm & VanillaICE	SNPS + NPs	0.0004	0.9962	0.0034
Copy number 3		$\widehat{CN} = 1$	$\widehat{CN} = 2$	$\widehat{CN} = 3$
Birdseye / Canary	SNPs + NPs	0.0006	0.0817	0.9177
crlmm & VanillaICE	SNPs	0.0000	0.0454	0.9546
crlmm & VanillaICE	SNPs + NPs	0.0000	0.1069	0.8931

Table 1: The proportion of integer copy number estimates that agree with the *true* copy number for chromosome 21 in the trisomy dataset were computed for two HMM implementations. The true copy number for loci on chromosome 21 is assumed to be 3 for the 26 trisomy samples and 2 for the 70 normal samples. The results from Birdsuite are a merge of the Birdseye HMM and Canary calls. The VanillaICE HMM was fit to the set of polymorphic markers using the adjusted prediction regions described in Section 4.4 (row 2) and has fewer false negatives that Birdsuite for 3-copy loci. The addition of the set of nonpolymorphic markers to the analysis (row 3) results in more false negatives among the trisomy subjects relative to the polymorphic set alone (0.955 versus 0.893). At 2-copy loci (the normal subjects), the specificity was 0.991 for Birdsuite and 0.996 for VanillaICE in both the full (SNPs + NPs) and the SNP-only analysis (data not shown).

believed to contain common variants (Canary). By contrast, the current implementation of our algorithm does not use external data and assumes that the typical copy number across samples within a batch is two. A consequence is that Canary can call an amplification or deletion in nearly all of the samples within a batch in a region that is believed to contain a common variant (see supplementary Figure 3), whereas our algorithm is unlikely to do so. An interesting feature of the Birdseye segmentation is that we observe strong batch effects in regions that are thought to contain common copy number variants. These regions contain groups of probes that tend to have correlated intensity profiles across samples and, as a result, smoothing via a HMM does not reduce the batch effect. While the Canary algorithm can be helpful for reducing the batch effect in such regions, batch effects often persist (supplementary Figure 3b).

D Simulation study for common copy number alterations

To assess the extent to which our estimation procedure is robust to common variants, we simulated 26 artificial datasets from the Chakravarti study. The simulated datasets differ in the ratio of trisomy to normal controls, ranging from 1.9% (1/52) to 50% (26/52) trisomies. Supplementary Figure 4 in Section H plots the average log intensities versus the log ratio for 5 randomly selected SNPs for a dataset with 50% trisomy cases (row 1) and a dataset with 1.9% trisomy cases (row 2). The triangle plotting symbol denotes a subject with 3 copies of chromosome 21. The shading of the plotting symbols is proportional to the posterior probability of altered copy number. We suggest an additional iteration of *crlmm* that updates the within-genotype centers *via* a trimmed median, excluding subjects with high posterior probabilities of altered copy number. Supplementary Figure 5 plots the median copy number across 12,579 polymorphic loci for the trisomy subjects in each of the simulated datasets after estimating within-genotype medians from (i) all subjects (solid blue), (ii) only the normal controls (black), and (iii) samples with high posterior probabilities of normal copy number (dashed blue). When approximately 30% of the subjects have chromosome 21 trisomy, our

model begins to have more difficulty in discriminating subjects that have altered copy from normal subjects, as evidenced by the vertical separation between the dashed blue and black lines. The ability to discriminate copy number estimates in the trisomy subjects from the noise level for normal copy number (orange line) diminishes with increasing proportions of subjects with altered copy number. This simulation suggests that improved classification of subjects with altered copy number will also improve the accuracy of the copy number estimates. Such classification procedures would likely require more iterations than the two-step approach currently implemented in *crlmm*, and could be evaluated using the simulation described here.

For the study of complex germline diseases, the challenge is typically to identify small copy number variants (e.g., 10kb-100kb) that may elevate risk to complex diseases, but are generally not deterministic for the disease and are present in a subset of the diseased population and possibly a smaller subset of those without disease. Large variants, those that are greater than 1Mb, are rare and can be identified with any reasonable smoothing. While we do not generally know *a priori* the prevalence or the genomic location of small copy number variants that elevate disease risk in any given population, the simulation demonstrates that we can nevertheless derive estimates of absolute copy number with small bias even when the prevalance of the variant is high (approximately 30 percent). Note that this approach is completely unsupervised with respect to disease status and does not require an assumption that those without disease have normal copy number.

For somatic cell diseases such as cancer, the frequency and size distribution of copy number variants is substantially different than that of germline diseases. A supervised approach that uses only the normal controls to derive the batch-specific prediction regions for integer copy number may be preferable to the iterative bias adjustment. As demonstrated in the simulation, the bias remains relatively flat even when 50 percent of the samples have altered copy number (the black line in supplementary Figure 5). The trade-off for using only the normal samples is that fewer observations are available to robustly estimate model

parameters and the number of SNPs for which imputation of unobserved genotype medians is required increases.

E The *crlmm* R package

The *crlmm* R package is available for download from Bioconductor (`http://www.bioconductor.org/`).

Data considerations. The version of the R package *crlmm* described here (see Section I) does not rely on reference samples to estimate model parameters. As a consequence, one must have a moderate number of samples in a given batch to use this algorithm. As few as 10 samples in a batch are possible, but the resolution to detect small changes in copy number by downstream algorithms that smooth the copy number estimates will be less for datasets with fewer samples. In particular, outliers will be more difficult to identify in smaller datasets and can be more influential on both the prediction regions and the locus-level estimates. Furthermore, the simulation of common copy number alterations described in Section D suggests that the model begins to have difficulty discriminating amplifications from the noise level of normal copy number when approximately 27% of the subjects have altered copy number (supplementary Figure 5). In small datasets, the suggested approach of using posterior probabilities to compute more robust estimates of the within-genotype median intensities may not be feasible. Note that we advise against attempts to increase the size of the dataset by post-hoc addition of normal controls from a reference dataset such as HapMap. The addition of reference samples to the analysis without appropriately acknowledging that these samples were processed in a different batch can lead to incorrect inference in both the test and reference subjects. Statistical approaches that cluster similar batches or borrow-strength across batch are a future direction of this research.

The estimation procedure described in this paper was developed in the context of germline diseases (e.g., bipolar disease) and apparently normal subjects from HapMap. While *crlmm*

provides noninteger estimates of copy number, the performance of this approach on datasets in which the DNA has been isolated from potentially a mixture of cell populations needs to be evaluated. Additionally, chromosomal aberrations are often more extreme in diseases such as cancer, involving entire chromosomes or chromosome arms. The inclusion of normal controls in each batch will be particularly important in cancer in order to establish a baseline copy number against which shifts from normal copy number can be measured. In the event that the phenotype of interest is completely confounded by batch, additional experimentation will be required to distinguish between batch- and biologically-driven variation in the copy number estimates.

Priors. The priors used to generate the results reported in this manuscript are the defaults in the version of *crlmm* indicated in Section I. In particular, we used 50 degrees of freedom for priors involving the background and signal variances to provide moderate-heavy smoothing in batches that typically have 90 - 96 samples. For applications involving more variability in the batch size, a data-driven approach to estimate the degrees of freedom may be beneficial (e.g. Smyth (2004)). Having used the same degrees of freedom to smooth the elements of the covariance matrix, our prior is conceptually similar to the inverse Wishart that has a single parameter for the degrees of freedom. However, our preference during software development is to keep the more heavily parameterized implementation that allows more flexible exploration of shrinkage properties with respect to the prediction regions that these matrices yield.

F Suggested Bioconductor software downstream of *crlmm*

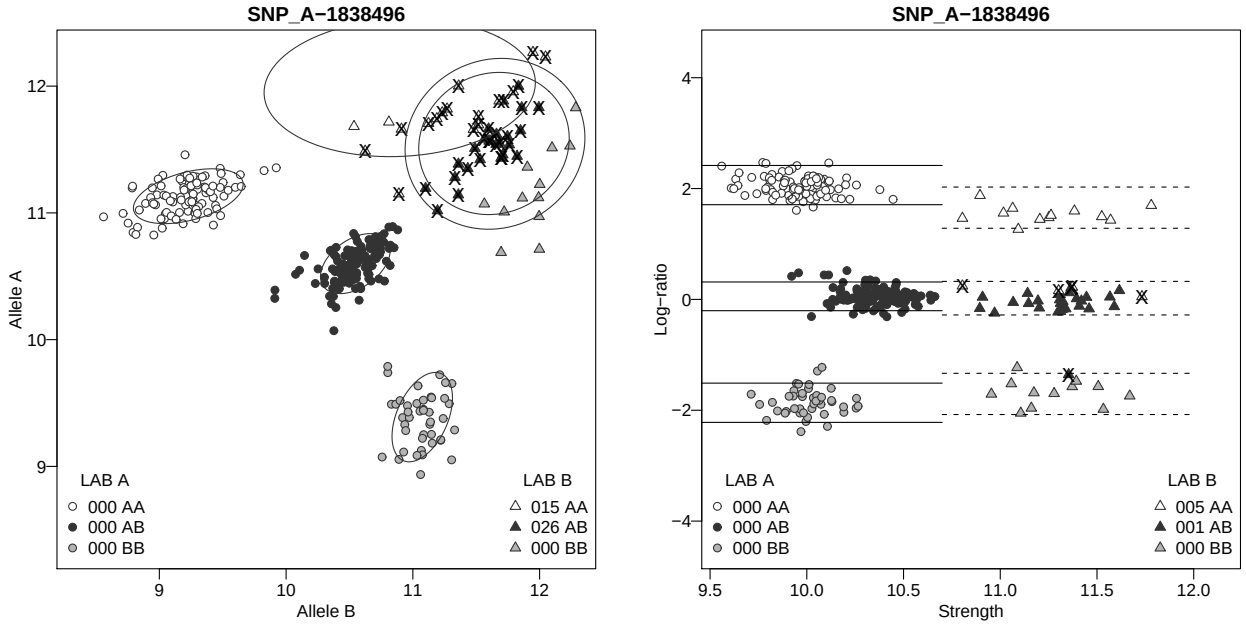
The estimates obtained from the R package *crlmm* serve as a starting point for downstream analyses that incorporate information from neighboring probes to identify regions with altered copy number. Two common downstream algorithms are segmentation approaches, such as circular binary segmentation (Olshen et al., 2004), that detect shifts in the mean

copy number and HMMs that typically assume an integer copy number state. Both of these approaches rely on estimates of copy number and can incorporate inverse variance estimates as weights. The R package *VanillaICE* provides an implementation of a HMM(Scharpf et al., 2008) and future versions will provide workflows for using *crlmm* results in conjunction with both circular binary segmentation (using *DNAcopy*) and HMMs.

G Alternative methods for pre-processing

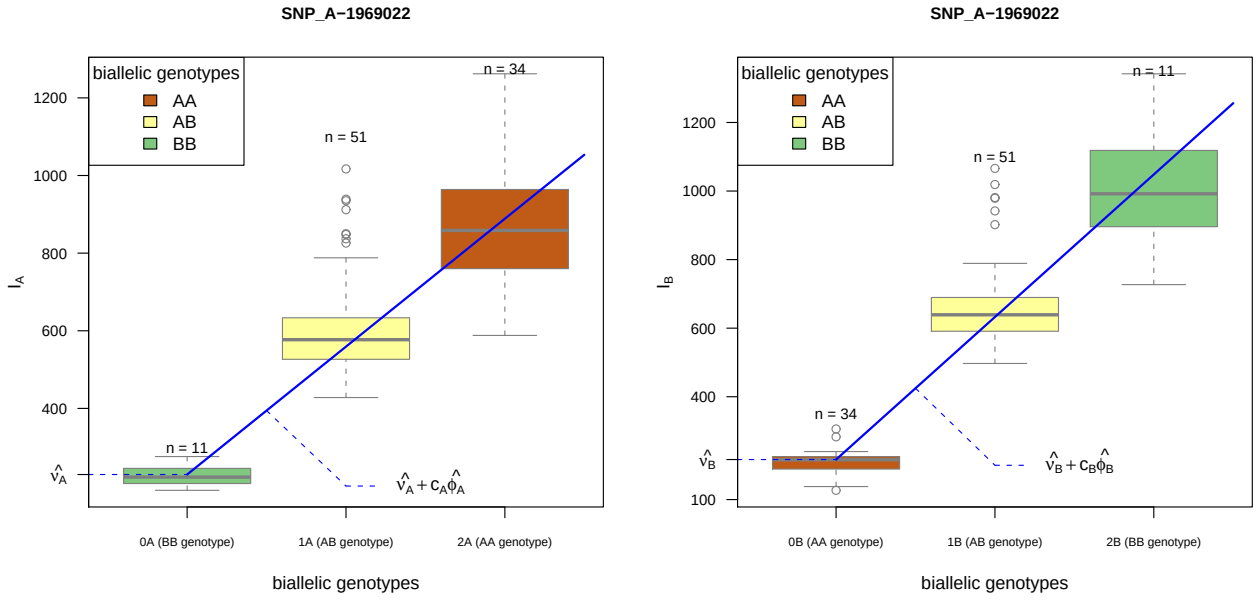
Alternative approaches to *crlmm* for preprocessing the raw intensities are available and could be explored in conjunction with the procedure described in this paper for removing batch effects and estimating copy number. For instance, Bengtsson et al. describes a preprocessing methodology in which the raw A and B intensities are adjusted for allelic cross-talk (Bengtsson et al., 2008). Quantile-normalization is then prescribed as an optional step depending on whether additional normalization is deemed necessary by the analyst. Software implementing their approach is available in the R package *aroma.affymetrix*. By contrast, *crlmm* estimates optical and nonspecific hybridization following quantile-normalization and summarization. As *aroma.affymetrix* does not adjust for batch effects, we compare the preprocessing approaches on downstream estimates of copy number using the Chakravarti dataset in which samples were processed in a single batch. Supplementary Figure 6 plots log ratios of total copy number obtained from *aroma.affymetrix* against the log ratios of total copy number obtained from *crlmm*. The log-transformed ratios from the two approaches are well correlated in the normal and trisomy samples with Pearson correlation coefficients of 0.861 and 0.854, respectively.

H Supplementary figures



(a) Birdseed (b) *crlmm*

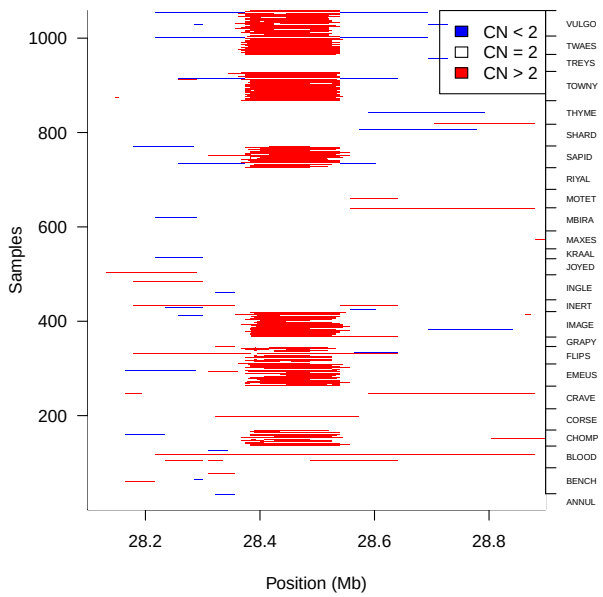
Figure 1: A set of identical samples was genotyped by two different labs. Left: A scatter plot of the A versus B allele intensities for a single SNP with plotting symbols denoting the consensus HapMap genotype. The default genotyping algorithm for this platform provided by Affymetrix, Birdseed, makes 41 mistakes in Lab B. Right: The *crlmm* algorithm uses the log ratio of the A and B allele intensities to call genotypes and makes only 6 mistakes in Lab B. As the lab-effect is mostly in the direction of the total intensity (x-axis, right panel), copy number estimates are far more susceptible to batch effects than genotype calls.



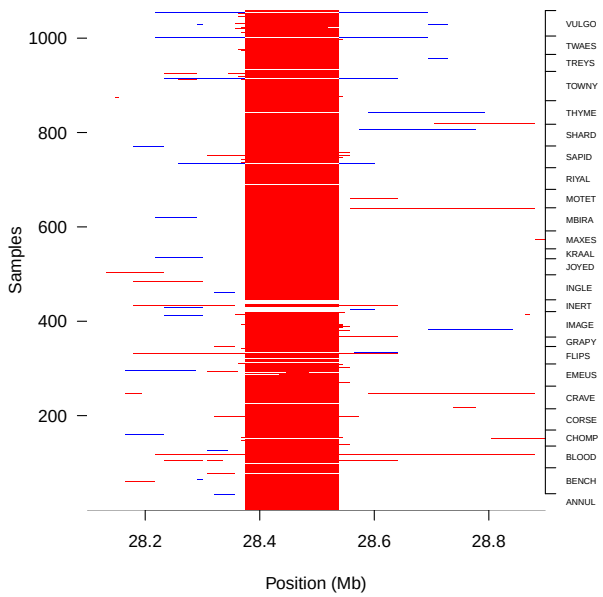
(a) A allele.

(b) B allele.

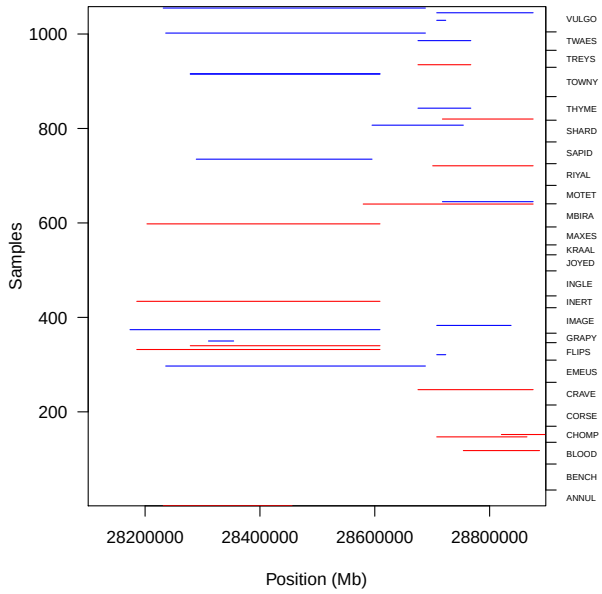
Figure 2: Boxplots of the normalized intensities stratified by the biallelic genotype for SNP_A-1969022 in the trisomy dataset. The slope and intercept parameters for allele-specific copy number (ν_A, ν_B, ϕ_A, and ϕ_B) are estimated via weighted least squares. Weights are calculated as the inverse within-genotype median absolute deviation (MAD) of the normalized intensities. As described in Section 4, the within-genotype medians for SNPs with unobserved genotypes are imputed via regression from a random sample of SNPs for which *complete* data was observed (AA, AB, and BB genotypes observed). The assumption that the relationship between allelic copy number and the median intensity is linear does not hold for all SNPs. Furthermore, the relationship becomes increasingly nonlinear for higher intensity values. We see some evidence of this nonlinearity for SNP_A-1969022. In particular, the regression line in panels (a) and (b) overestimates the observed median value for genotypes AA and BB, respectively. Appropriate modeling of the non-linear relationship at higher ranges of the intensity scale is a future direction of this work.



(a) Birdseye HMM



(b) Merged Birdseye and Canary



(c) Vanilla HMM fit to *crlmm* copy number estimates.

Figure 3: We observed plate-effects in both the Birdseye HMM predictions (a) and the merged canary predictions (b). The chi-square statistic in the 28.4 Mb region is genome-wide significant for both the Birdseye and Canary algorithms ($\chi^2_{24} > 250$, p-value $< 1.0^{-8}$). (c) An image of HMM predictions from the CRLLM copy number estimates using the default settings in the R package VanillaICE ($\chi^2_{24} = 55.84$, p-value $= 0.20$).

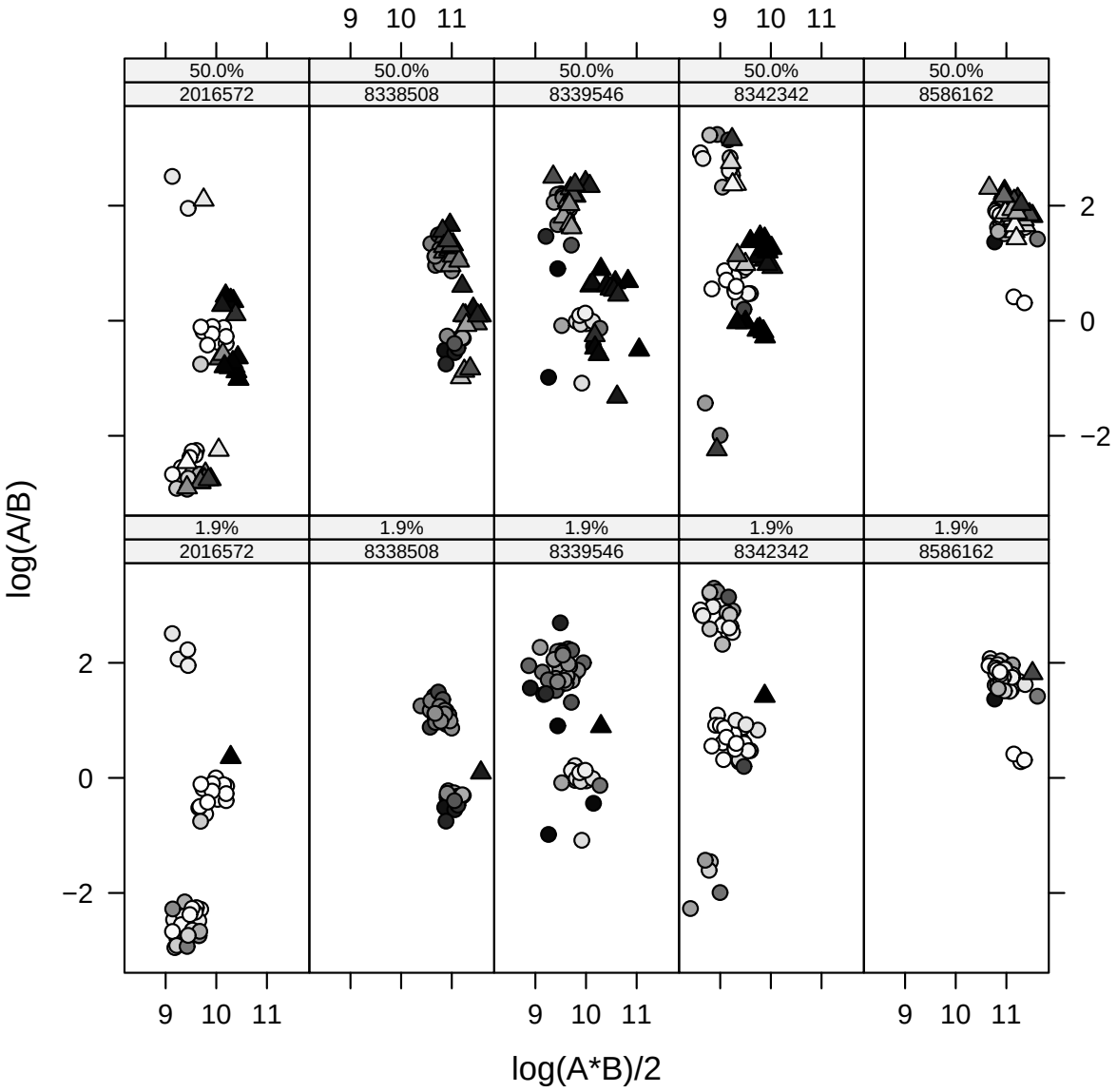


Figure 4: Five SNPs from chromosome 21 were selected at random from a dataset with a high proportion of trisomy 21 subjects (row 1: 50% (26/52)) and a dataset with a low proportion of trisomy 21 subjects (row 2: 1.9% (1/52)). Plotted in each panel is the average of the log A and log B intensities versus the log ratio. Trisomy subjects are plotted with triangles and normal controls are plotted with circles. The shading of the plotting symbols is proportional to the posterior probability of altered copy number. The proportion of black plotting symbols may be useful for evaluating departures from the assumption that the typical copy number is two.

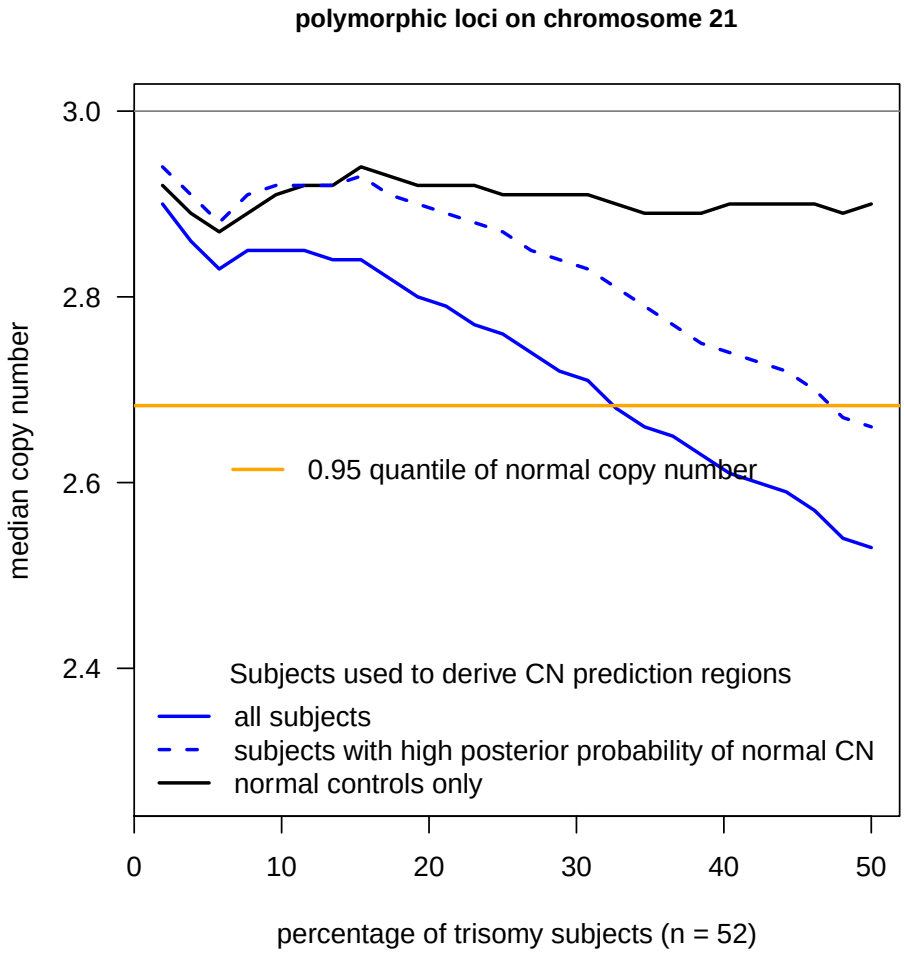
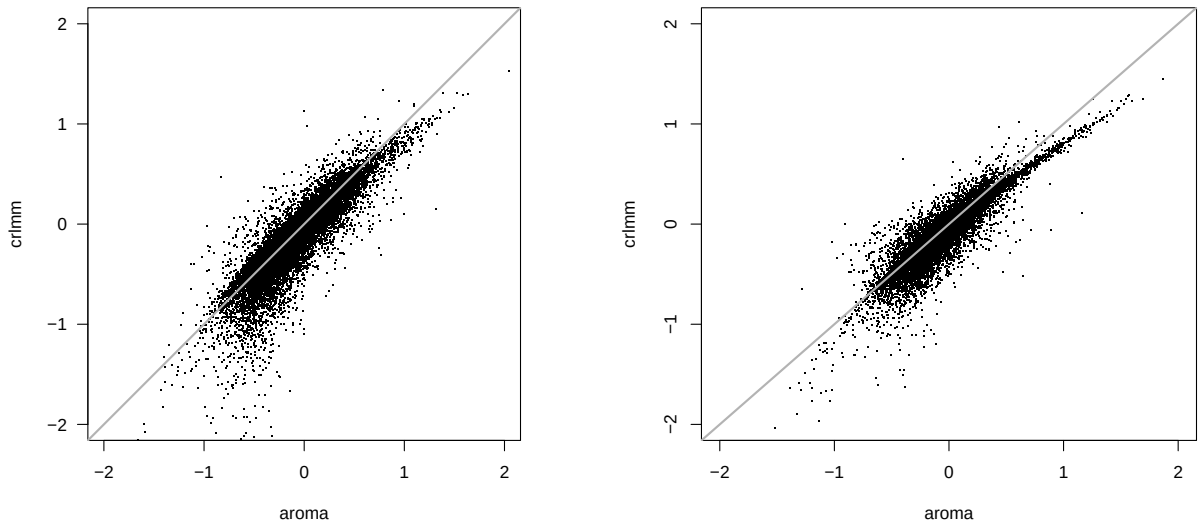


Figure 5: We assessed the impact of our assumption that the median copy number in a batch is 2 using the Chakravarti dataset. Plotted on the x-axis is the percentage of subjects with chromosome 21 trisomy. The total number of samples (normal and trisomy) included in the analysis is 52 for each of the 26 artificial datasets. For each dataset, we report the median copy number across the 12,579 polymorphic loci in the trisomy subjects. Three methods for deriving the copy number prediction regions were considered. In one approach, we estimated copy number at each locus by fitting the linear model using all of the subjects to estimate the within-genome location and scale statistics (solid blue). The second approach is iterative, whereby the initial parameter estimates are used to compute posterior probabilities of normal copy number (dashed blue). An additional iteration is used to refit the linear model after trimming samples with high posterior probabilities from the within-genotype location/scale statistics. A third approach (solid black line) provides a baseline for the experiment and is the median copy number estimated in a model that uses only the normal subjects to estimate the within-genotype location and scale.



(a) normal (b) trisomy

Figure 6: Log ratios of total copy number from a normal control (left) and a subject with chromosome 21 trisomy (right). While *aroma.affymetrix* and *crlmm* use alternative methodologies for preprocessing, the downstream estimates of locus-specific copy number are qualitatively similar for datasets without batch effects. The Pearson correlation coefficient for the normal and trisomy samples are 0.861 and 0.854, respectively.

I Computing environment and R package versions

- R version 2.11.0 Under development (unstable) (2009-11-22 r50541), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.iso885915`, `LC_NUMERIC=C`, `LC_TIME=en_US.iso885915`, `LC_COLLATE=en_US.iso885915`, `LC_MONETARY=C`, `LC_MESSAGES=en_US.iso885915`, `LC_PAPER=en_US.iso885915`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.iso885915`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: affxparser 1.19.0, aroma.affymetrix 1.3.0, aroma.apd 0.1.7, aroma.core 1.3.1, aroma.light 1.15.1, Biobase 2.7.3, crlmm 1.5.20, CrlmmCopyNumber 1.0.4, digest 0.4.2, ellipse 0.3-5, genefilter 1.29.3, IRanges 1.5.21, lattice 0.17-26, matrixStats 0.1.8, oligoClasses 1.9.24, R.cache 0.2.0, RColorBrewer 1.0-2, R.filesets 0.6.5, R.huge 0.2.0, R.methodsS3 1.0.3, R.oo 1.6.5, R.rsp 0.3.6, R.utils 1.2.4, VanillaICE 1.9.1
- Loaded via a namespace (and not attached): affyio 1.15.1, annotate 1.25.0, AnnotationDbi 1.9.2, Biostrings 2.15.11, DBI 0.2-4, grid 2.11.0, mvtnorm 0.9-8, preprocessCore 1.9.0, RSQLite 0.7-3, SNPchip 1.11.1, splines 2.11.0, survival 2.35-7, tools 2.11.0, xtable 1.5-6
- Birdsuite 1.5.3

References

Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D., and Hurles, M. E. "A robust statistical method for case-control association testing with copy number variation." *Nat Genet*, 40(10):1245–1252 (2008).

Bengtsson, H., Irizarry, R., Carvalho, B., and Speed, T. P. "Estimation and assessment of raw copy numbers at the single locus level." *Bioinformatics*, 24(6):759–767 (2008).

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J. C., Huang, J. H., Alexander, S., Du, J., Kau, T., Thomas, R. K., Shah, K., Soto, H., Perner, S., Prensner, J., Debiasi, R. M., Demichelis, F., Hatton, C., Rubin, M. A., Garraway, L. A., Nelson, S. F., Liau, L., Mischel, P. S., Cloughesy, T. F., Meyerson, M., Golub, T. A., Lander, E. S., Mellinghoff, I. K., and Sellers, W. R. "Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma." *Proc Natl Acad Sci U S A*, 104(50):20007–20012 (2007).

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. "High-resolution analysis of DNA copy number using oligonucleotide microarrays." *Genome Res*, 14(2):287–295 (2004).

Carvalho, M. A., Marsillac, S. M., Karchin, R., Manoukian, S., Grist, S., Swaby, R. F., Urmenyi, T. P., Rondinelli, E., Silva, R., Gayol, L., Baumbach, L., Sutphen, R., Pickard-Brzosowicz, J. L., Nathanson, K. L., Sali, A., Goldgar, D., Couch, F. J., Radice, P., and Monteiro, A. N. A. "Determination of cancer risk associated with germ line BRCA1 missense variants by functional analysis." *Cancer Res*, 67(4):1494–1501 (2007).

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., Bassett, A. S., Seller, A., Holmes, C. C., and Ragoussis, J. "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 (2007).

Consortium, I. H. "The International HapMap Project." *Nature*, 426(6968):789–796 (2003).

Golden Helix. *Copy Number Variation Analysis with SVS 7* (2009). Golden Helix Manual for SNP and Variation Suite.

Huang, J., Wei, W., Chen, J., Zhang, J., Liu, G., Di, X., Mei, R., Ishikawa, S., Aburatani, H., Jones, K. W., and Shapero, M. H. "CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays." *BMC Bioinformatics*, 7:83 (2006).

Hupe, P., Stransky, N., Thiery, J.-P., Radvanyi, F., and Barillot, E. "Analysis of array CGH data: from signal ratio to gain and loss of DNA regions." *Bioinformatics*, 20(18):3413–3422 (2004).

Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P. J., Darvishi, K., Lee, C., Nizzari, M. M., Gabriel, S. B., Purcell, S., Daly, M. J., and Altshuler, D. "Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs." *Nat Genet*, 40(10):1253–1260 (2008).

LaFramboise, T., Harrington, D., and Weir, B. A. "PLASQ: A Generalized Linear Model-Based Procedure to Determine Allelic Dosage in Cancer Cells from SNP Array Data." *Biostatistics* (2006).

Lamy, P., Andersen, C. L., Dyrskjot, L., Torring, N., and Wiuf, C. "A Hidden Markov Model to estimate population mixture and allelic copy-numbers in cancers using Affymetrix SNP arrays." *BMC Bioinformatics*, 8:434 (2007).

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I. W., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., and Altshuler, D. "Integrated detection and population-genetic analysis of SNPs and copy number variation." *Nat Genet*, 40(10):1166–1174 (2008).

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. "A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays." *Cancer Res*, 65(14):6071–6079 (2005).

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. "Circular binary segmentation for the analysis of array-based DNA copy number data." *Biostatistics*, 5(4):557–72 (2004).

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *Am J Hum Genet*, 81(3):559–575 (2007).

Rabbee, N. and Speed, T. P. "A genotype calling algorithm for affymetrix SNP arrays." *Bioinformatics*, 22(1):7–12 (2006).

Rigaill, G., HupÃl', P., Almeida, A., Rosa, P. L., Meyniel, J.-P., Decraene, C., and Barillot, E. "ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays." *Bioinformatics*, 24(6):768–774 (2008).

Scharpf, R. B., Parmigiani, G., Pevsner, J., and Ruczinski, I. "Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays." *Annals of Applied Statistics*, 2(2):687–713 (2008).

Smyth, G. K. "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments." *Stat.Appl.Genet.Mol.Biol.*, 3(1):Article3 (2004). PUBM: Print-Electronic; DEP: 20040212; JID: 101176023; 2004/02/12 [epublish]; ppublish.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., and Bucan, M. "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data." *Genome Res*, 17(11):1665–1674 (2007).

Wang, W., Carvalho, B., Miller, N., Pevsner, J., Chakravarti, A., and Irizarry, R. A. "Estimating genome-wide copy number using allele specific mixture models." *Journal of Computational Biology*, 15(7):857–866 (2008).

Wu, Z., Irizarry, R., Gentlemen, R., Martinez-Murillo, F., and Spencer, F. "A model-based background adjustment for oligonucleotide expression arrays." *Journal of the American Statistical Association*, 99(468):909–917 (2004).

Wu, Z. and Irizarry, R. A. "A statistical framework for the analysis of microarray probe-level data." *Annals of Applied Statistics*, 1(2):333–357 (2007).