6-29-2009

A MULTILEVEL MODEL TO ADDRESS BATCH EFFECTS IN COPY NUMBER USING SNP ARRAYS

Robert B. Scharpf
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, rscharpf@jhsph.edu

Ingo Ruczinski
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Benilton Carvalho
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Betty Doan
The Johns Hopkins University, School of Medicine, McKusick-Nathans Institute of Genetic Medicine

Aravinda Chakravarti
The Johns Hopkins University, School of Medicine Genetics

See next page for additional authors

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

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

Abstract

Submicroscopic changes in chromosomal DNA copy number dosage are common and have been implicated in many heritable diseases and cancers. Recent high-throughput technologies have a resolution that permits the detection of segmental changes in DNA copy number that span thousands of basepairs across the genome. Genome-wide association studies (GWAS) may simultaneously screen for copy number-phenotype and SNP-phenotype associations as part of the analytic strategy. However, genome-wide array analyses are particularly susceptible to batch effects as the logistics of preparing DNA and processing thousands of arrays often involves multiple laboratories and technicians, or changes over calendar time to the reagents and laboratory equipment. Failure to adjust for batch effects can lead to incorrect inference and requires inefficient *post-hoc* quality control procedures that exclude regions that are associated with batch. Our work extends previous model-based approaches for copy number estimation by explicitly modeling batch effects and using shrinkage to improve locus-specific estimates of copy number uncertainty. Key features of this approach include the use of diallelic genotype calls from experimental data to estimate batch- and locus-specific parameters of background and signal without the requirement of training data. We illustrate these ideas using a study of bipolar disease and a study of chromosome 21 trisomy. The former has batch effects that dominate much of the observed variation

in quantile-normalized intensities, while the latter illustrates the robustness of our approach to datasets where as many as 25% of the samples have altered copy number. Locus-specific estimates of copy number can be plotted on the copy-number scale to investigate mosaicism and guide the choice of appropriate downstream approaches for smoothing the copy number as a function of physical position. The software is open source and implemented in the R package *CRLMM* available at Bioconductor (`http:www.bioconductor.org`).

1 Introduction

Segmental changes in DNA copy number arise through genomic rearrangements that cause insertions or deletions of genomic fragments. Such rearrangements are thought to arise most commonly via non-allelic homologous recombination in regions that contain low copy repeats (Gu et al., 2008), and can occur in the germline during meiosis as well as during mitosis in somatic cells. Many of the genomic rearrangements that effect DNA copy number are likely to be neutral with respect to phenotype. For instance, an extensive list of deletions and amplifications have been catalogued in apparently normal HapMap individuals (Redon et al., 2006; Kidd et al., 2008). However, genomic rearrangements that occur in regions that disrupt gene function or effect the copy number of genes that are dosage sensitive can effect phenotypes. See Lupski for a recent review (Lupski, 2009). Alterations of DNA copy number are implicated in many diseases, including autism spectrum disorders (Szatmari et al., 2007; Marshall et al., 2008), bipolar disease (Zhang et al., 2008), autoimmune disorders such as type I diabetes (McKinney et al., 2008), and cancer (Ma et al., 2009; Cappuzzo et al., 2009; Woo et al., 2009). For other heritable diseases such as schizophrenia, the role of recurrent copy number variants in disease remains elusive (Sutrala et al., 2007; Need et al., 2009).

Copy number variants (CNV) spanning regions of the genome greater than one megabase (Mb) are detectable by cytogenetic techniques such as spectral karyotyping and fluorescence in situ hybridization (FISH). However, many changes to DNA copy number are thought

to involve smaller segments of the genome that are below the level of resolution attainable by cytogenetic methods. High throughput genotyping arrays enable the measurement of genotype and copy number across the genome. The resolution for detecting CNV in current platforms is on the order of thousands of basepairs, and can therefore be used to identify segmental changes that are not detectable by spectral karyotyping (resolution: 5 - 10 Mb) or array comparative genomic hybridization (resolution: 100 kb). Screening for alterations in copy number has identified genomic regions known to be involved in disease, such as the neurexins in autism (Szatmari et al., 2007), as well as novel targets that suggest a role of less well understood pathways in disease etiology. High throughput genotyping platforms provide a useful genomic screen whereby loci exhibiting patterns of variation between normal and disease individuals can be identified and followed. While most genotype calling algorithms are highly concordant for the vast majority of SNPs, copy number estimation is more sensitive to differences in the preprocessing and normalization steps, as well as to technological artifacts that can effect the observed intensities in a systematic way.

This paper is organized as follows. In Section 2, we discuss several key features for developing locus-level estimates of copy number. The relevance of other hybridization-based technologies, such as gene expression microarrays, and recent adaptations to the problem of copy number estimation are reviewed. Section 3 motivates the need for extending these methods. Specifically, the need to adjust for batch effects in model-based approaches for copy number estimation and the importance of shrinkage. Section 4 defines a theoretical framework for copy number in hybridization-based platforms and the challenges of adapting this model to high-throughput arrays. Section 5 describes an estimation algorithm that is motivated by many of the fundamental features of standard approaches, such as maximum likelihood and empirical Bayes. In Section 6, we illustrate the main innovations of our approach using two experimental datasets and compare our results with software recommended by the array manufacturer. Concluding remarks are provided in Section 7.

2 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).

A common approach employed by tier 1 methods for copy number estimation 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 in the test sample, making it more 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. Robust-to-outlier methods for normalizing the intensities across arrays include quantile-normalization, whereby each sample is normalized to the same reference distribution (Bolstad et al., 2003).

A general framework for modeling the observed fluorescence intensities in 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, nonspecific binding, and specific binding,

$$\text{Observed}_{gij} \quad = \quad \text{Background}_{gij} + \text{Nonspecific}_{gij} + \text{Specific}_{gij}, \tag{1}$$

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 also 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. The decision to model the intensities at the feature-level or at the locus-level has important practical and computational implications that we discuss further.

A feature-level model. 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 3 and 6), 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, diallelic 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 diallelic 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 diallelic 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 (see Sections 6 and 7). Furthermore, Birdsuite does not currently provide locus-level estimates of copy number whereby one can more effectively assess cell contamination and batch effects. The software proposed by Wang

et al. has a similar drawback of PLASQ in recommending a training dataset for precomputing parameter estimates.

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.

3 Motivation

Our present work is motivated by the observation of large batch effects in several genome wide data sets and the need for improved estimates of copy number uncertainty. The former has the potential of confounding copy number-phenotype association analyses; improvements to the latter can be utilized by downstream algorithms that smooth locus-level estimates as a function of the physical position.

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. Nevertheless, important differences emerge as a result of batch effects. To illustrate, 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 num-

ber 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.

[Figure 1 about here.]

Batch effects can be addressed in several ways. One approach is to consider batch effects as part of the quality control step in the analysis of genome wide arrays. For instance, Zhang et al. (2008) excluded regions for which copy number alterations detected by the Birdsuite software were associated with batch in their analysis of the Bipolar data. This approach is sensible if a relatively small number of loci are affected by batch. In such instances, smoothing the locus-level estimates using a HMM or a segmentation procedure may reduce the impact of batch effects on downstream analyses. An alternative approach is to apply a *post-hoc* correction to the signal intensities that effectively gives each batch the same mean signal intensity, as in the GISTIC algorithm (Beroukhim et al., 2007). A third approach is to adjust for batch-effects as part of the estimation procedure for copy number. For instance, the Golden Helix software for copy number estimation provides a correction for log ratios from a principal components analysis (PCA) on the raw intensity ratios (Golden Helix, 2009). In our experience, large studies involving arrays processed over an extended period of time have batch effects that are genome-wide in scale. In these instances, quality control and *post-hoc* procedures provide inefficient protection for false positives.

One way to explore and diagnose batch effects is to perform PCA on the locus-level summaries after preprocessing, as suggested by Golden Helix. A related approach that involves less computation is simply to group samples that are processed at the same time and by the same lab. For instance, in the bipolar dataset we follow a similar practice as Korn

et al. in defining batch as the 96 well chemistry plate on which the samples were stored prior to hybridization to the SNP array. Here, we focused on the set of European ancestry controls for bipolar disease. Figures 2(a) and 2(b) provide complementary views of the batch effect at one locus on chromosome 15. The boxplots in Figure 2(b) show the distribution of the total intensities for SNP_A-4251622 by chemistry plate. The two plates that are highlighted in Figure 2(a) are also highlighted in the boxplots. A F-statistic from a one way analysis of variance (ANOVA) for the total intensities by plate at this locus is approximately 14. Figure 3 plots the distribution of all F-statistics for chromosome 15, demonstrating that moderate to strong batch-effects persist after quantile normalization at most loci. The batch effects we observed on chromosome 15 were typical of the other autosomes in this dataset (data not shown).

[Figure 2(a) - 2(f) about here.]

[Figure 3 about here.]

Shrinkage. Shrinkage of the variance estimates is likely to be useful for several reasons. First, the technology used to estimate the amount of DNA hybridized to the array affects the measured fluorescence of many probes in similar ways. Secondly, many SNPs have a low minor allele frequency or *unobserved* diallelic genotypes that complicate our estimation procedure. Third, shrinkage reduces the sensitivity of our approach to extreme values, such as variance estimates near zero. By borrowing strength from the millions of measurements at other loci, we can improve locus-specific estimates of the uncertainty. These estimates can then be propagated to higher level analyses that smooth copy number estimates as a function of the genomic physical position. For instance, the emission probabilities of an HMM can incorporate locus-specific estimates of the uncertainty.

4 Model

Batch effects appear to an unavoidable feature of studies involving a large number of arrays, one that copy number estimation algorithms should take into account. Because batch can be easily identified and visualized (e.g., Figures 2(a), 2(b), and 3), we argue that batch can effects can be successfully modeled. Here, we introduce a model for copy number estimation based loosely on an approach described by Wang et al. (2008). Our method differs from Wang et al. in several important ways. First, we model batch as a fixed effect. More generally, one can think of batch as a variable selection problem to be inferred at each locus (see Section 7). For the purpose of this paper, we treat batch as known. Secondly, we avoid using training data to estimate model parameters. Instead, we estimate model parameters using an algorithm that relies only on the experimental data (Section 5). Third, we provide prediction regions for loci with low minor allele-frequencies and potentially unobserved genotypes. In addition, we provide a solution for estimating copy number for nonpolymorphic probes in the most recent generation of genotyping platforms. Fourth, we shrink locus-level estimates of the variance and correlation across alleles that are often very noisy though a hierarchical model. Fifth, we propose an iterative estimation procedure that improves estimates of copy number at loci where many of the subjects in the experimental data have altered copy number. Finally, software for fitting this model for the Affymetrix 6.0 and Illumina platforms is publicly available from the Bioconductor website.

4.1 A multilevel model.

Polymorphic loci. We propose a multilevel model for the locus-level intensities that is motivated by past work with other hybridization based technologies. Specifically, we model the observed intensity I of the A and B alleles at locus i, sample j, and batch p as follows:

$$\begin{aligned}\begin{bmatrix} I_{A,ijp} \\ I_{B,ijp} \end{bmatrix} &= \left[\begin{pmatrix} \text{Optical}_{A,ip} + \text{Nonspecific}_{A,ip} \\ \text{Optical}_{B,ip} + \text{Nonspecific}_{B,ip} \end{pmatrix} \times \begin{pmatrix} \delta_{A,ijp} \\ \delta_{B,ijp} \end{pmatrix}\right] + \begin{bmatrix} \text{Specific}_{A,ijp} & \times & \varepsilon_{A,ijp} \\ \text{Specific}_{B,ijp} & & \varepsilon_{B,ijp} \end{bmatrix} \\ &\equiv \begin{bmatrix} \nu_{A,ip} & \times & \delta_{A,ip} \\ \nu_{B,ip} & & \delta_{B,ip} \end{bmatrix} + \begin{bmatrix} \phi_{A,ip} c_{A,ijp} & \times & \varepsilon_{A,ijp} \\ \phi_{B,ip} c_{B,ijp} & & \varepsilon_{B,ijp} \end{bmatrix}. \end{aligned} \tag{2}$$

The average fluorescence arising from optical background and nonspecific hybridization are collectively parametrized by ν and referred to as background. The slope, ϕ, in model (2) provides an estimate of the change in the average intensity at a given locus per each integer increase in the allelic copy number. Both the background and slope are allowed to depend on the SNP i and the batch p. See Figure 4 for an illustration of these parameters in the context of an A versus B intensity scatterplot of a single SNP.

[Figure 4 about here.]

The errors δ and ε account for array to array variation within a batch of the background and slope terms, respectively. These terms are each approximately log-normal and assumed to be independent across loci and independent of each other:

$$\log\left(\delta_{k,ijp}\right) \sim N\left(0, \tau^2_{k,ip}\right) \quad \text{and} \quad \log\left(\varepsilon_{k,ijp}\right) \sim N\left(0, \sigma^2_{k,ip}\right) \text{ for } k \in \{A, B\}.$$

Note, however, that the error terms are not independent across alleles. In particular, Figure 4 suggests that the correlation of the A and B intensities is most pronounced for samples with allelic copy numbers greater than 0. The correlation reflects cross-hybridization of the probe for the A allele to the B allele target sequence (and vice versa). As in Wang et al., we assume that the joint distribution of the log intensities conditional on the allelic copy number is approximately bivariate normal:

$$\begin{bmatrix} \log_2(I_{A,ijp}) \\ \log_2(I_{B,ijp}) \end{bmatrix} \Bigg| \begin{matrix} C_{A,ijp} = c_A \\ C_{B,ijp} = c_B \end{matrix} \sim \mathrm{N}\left(\begin{bmatrix} \log_2(\nu_{A,ip} + c_A\phi_{A,ip}) \\ \log_2(\nu_{B,ip} + c_B\phi_{B,ip}) \end{bmatrix}, \ \boldsymbol{\Sigma}_{ip} \right). \tag{3}$$

Note that the covariance of the A and B intensities is allowed to depend on both the SNP i and batch p, an important feature when modeling probe effects in high throughput hybridization-based technologies. The diagonal elements of $\boldsymbol{\Sigma}$ are as follows:

$$(\boldsymbol{\Sigma}_{i,p})_{11} = \tau^2_{A,ip}\mathrm{I}_{[c_A=0]} + \sigma^2_{A,ip}\mathrm{I}_{[c_A>0]} \ \text{ and } \ (\boldsymbol{\Sigma}_{i,p})_{22} = \tau^2_{B,ip}\mathrm{I}_{[c_B=0]} + \sigma^2_{B,ip}\mathrm{I}_{[c_B>0]}. \tag{4}$$

The correlation of the A and B intensities, $\rho_{i,p}$, is SNP- and batch-specific.

At the next level of the model specification, prior distributions are selected for $\boldsymbol{\Sigma}$. A commonly used prior is an inverse Wishart. However, we view this prior as too restrictive as a single degree of freedom is required for the variances. As $\boldsymbol{\Sigma}$ is a 2×2 matrix, we have considerable flexibility for exploring different priors for the standard deviations and correlation. We use independent inverse chi-squared priors with d_A and d_B degrees of freedom for the background,

$$\frac{1}{\tau^2_{A,ip}} \propto \frac{1}{d_A t^2_{A,p}}\chi^2_{A,d_A} \quad \text{and} \quad \frac{1}{\tau^2_{B,ip}} \propto \frac{1}{d_B t^2_{B,p}}\chi^2_{B,d_B}, \tag{5}$$

and slope variances,

$$\frac{1}{\sigma^2_{A,ip}} \propto \frac{1}{d_A s^2_{A,p}}\chi^2_{A,d_A} \quad \text{and} \quad \frac{1}{\sigma^2_{B,ip}} \propto \frac{1}{d_B s^2_{B,p}}\chi^2_{B,d_B}. \tag{6}$$

The terms $t^2_{A,p}$, $t^2_{B,p}$, $s^2_{A,p}$, and $s^2_{B,p}$ in equations (5) and (6) correspond to the typical variance of the background and slope terms, respectively. Note that these values are the same for all loci and depend only on the batch p. For the correlation structure, we use the following

prior:

$$\rho_{i,p} \sim \text{Beta}\,(\alpha, \beta)\,, \quad \text{where}$$

α and β are estimated empirically and place more mass at typical values. The motivation for an informative prior on the correlation is that cross-hybridization of the A and B alleles gives rise to positive correlations. In our experience, negative correlations (after conditioning on the allelic copy number) are spurious and usually occur when an insufficient number of observations are available to estimate the correlation.

Nonpolymorphic loci. For nonpolymorphic probes, only one allele is interrogated at each locus. We generically denote this allele as T. Again, we propose a theoretical model for the observed intensity for allele T at locus i, sample j and batch p as a convolution of fluorescence from optical background and non-specific binding of other probes, ν_T, and fluorescence arising from specific hybridization of the probe to the target sequence, ϕ_T. Explicitly,

$$\begin{aligned} I_{T,ijp} &= \nu_{T,ip}\delta_{T,ijp} + c_{T,ijp}\phi_{T,ip}\varepsilon_{T,ijp}, \text{ where} \\ \log(\delta_{T,ijp}) &\sim N(0, \tau_{T,ijp}) \quad \text{and} \\ \log(\varepsilon_{T,ijp}) &\sim N(0, \sigma_{T,ijp}). \end{aligned} \tag{7}$$

Again, the background and signal parameters are allowed to depend on both the nonpolymorphic locus i and batch p. The error terms corresponding to background and signal account for array to array variation within a batch and are assumed to be log-normal, independent across loci, and independent of each other. Inverse chi-squared priors for δ_T and ε_T variances complete the specification of the hierarchical model:

$$\frac{1}{\tau^2_{T,ip}} \propto \frac{1}{d_T t^2_{T,p}} \chi^2_{T,d_T} \quad \text{and} \quad \frac{1}{\sigma^2_{T,ip}} \propto \frac{1}{d_T s^2_{T,p}} \chi^2_{T,d_T}.$$

Challenges. There are several challenges to fitting models (2) and (7). First, the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\phi}$ can not be reliably estimated *a priori* from training data because of batch effects. Therefore, $\nu_A,\ \nu_B,\ \nu_T,\ \phi_A,\ \phi_B,\ \phi_T,\ c_A,\ c_B$, and c_T are allowed to depend on both the locus and batch and must be estimated from the experimental dataset. Secondly, the error terms $\delta_A,\ \delta_B,\ \delta_T,\ \varepsilon_A,\ \varepsilon_B$, and ε_T that capture within-batch variation of the background and signal intensities across arrays are not Gaussian. In principle, these parameters can be estimated using maximum likelihood or empirical Bayes. However, least squares and method of moments approaches to parameter estimation are well known to be biased, particularly when the variance of these parameters is large. The standard approach is a generalized linear model with an exponential link function, as employed by LaFramboise et al. (2006). Such an approach requires an iterative estimation procedure that we view as impractical for platforms that interrogate millions of loci. Third, for polymorphic loci the covariance matrix is a function of the allelic copy number. Fourth, outliers are common and robust-to-outlier approaches are needed. Again, least squares and method of moments are not robust to outliers. Taken together, the size of current genotyping platforms, the inevitability of batch effects in studies involving a large number of arrays, errors that are non-Gaussian, and the need for robustness has led us to develop an ad-hoc approach motivated by the fundamental features of the standard approaches.

5 Copy number estimation algorithm.

We prescribe a general strategy for copy number estimation that (i) develops naïve estimates of the allelic copy number that are taken to be known, (ii) uses a linear model to estimate batch- and locus-specific parameters for the background and slope terms and (iii) updates

the naïve estimates of allelic copy number. Robust-to-outlier procedures for preprocessing and copy number estimation are emphasized. Several problems remain after steps (i)-(iii), including unobserved genotypes at many polymorphic loci and variance estimates that are based on a small number of observations. We propose solutions to each of these problems that take advantage of the large number of observations available from other loci.

Our approach assumes that systematic artifacts that affect the overall location and scale of the intensities across arrays have been removed. For this, we quantile-normalize the arrays to a target reference distribution and then summarize these values to the level of the locus. For example, the Affymetrix 6.0 platform has 3 and occasionally 4 identical probes for each allele at polymorphic loci and one probe for each nonpolymorphic locus. For the polymorphic loci, we quantile normalize the raw intensities and then summarize the normalized intensities by taking the median. (The median is typically more robust to outliers than a trimmed mean.) For the nonpolymorphic probes, the Affymetrix 6.0 platform has only one probe per target sequence and we use the quantile-normalized intensities directly.

5.1 Allele-specific copy number

The parameters for $\nu_k,\ \phi_k,$ and c_k in model (2) are unknown for allele $k \in \{A, B\}$. As a first step, we genotype all of the samples on the array using the CRLMM software (Carvalho et al., 2007, 2009), obtaining genotype calls of AA, AB, and BB for the polymorphic loci. The genotype calls provide a naïve estimate of the allele-specific copy number — an integer value of 0, 1, or 2 for each allele. We denote the naïve estimates for the A and B alleles by $c_A^\star$ and $c_B^\star$, respectively. We use quantile-based estimators, the median and the median absolute deviation (MAD), to obtain robust estimates of the mean ($\hat{\mu}_{k,ip}^{GT}$) and variance ($\hat{\xi}_{k,ip}^{GT}$) on the intensity scale for genotype GT. For example, $\hat{\mu}_{B,ip}^{\text{AA}}$ is computed as the median of intensities $I_{B,ijp}$ for samples j with genotype AA. Inverse chi-squared priors with degrees of freedom d_ξ,

$$\frac{1}{\xi^{GT}_{k,ip}} \propto \frac{1}{d_\xi s^2_{\xi_k,p}} \chi^2_{\xi_k,d_\xi}, \tag{8}$$

are used to shrink locus-specific estimates of the variance to a typical value that is allowed to be batch-specific, $s^2_{\xi_k,p}$. Note that the within-genotype centers are approximately normal regardless of the distribution of $\boldsymbol{I}$. We (and others) have observed that the relationship of the within-genotype centers is approximately linear with the integer copy number (Huang et al., 2006; Wang et al., 2008). Using the naïve estimate of the integer copy number in the design matrix, we use weighted least squares regression to estimate ν_{A} and ϕ_{A}:

$$\frac{\mathbf{1}}{\hat{\boldsymbol{\xi}}_{\boldsymbol{A,ip}}} \times \begin{bmatrix} \hat{\mu}^{BB}_{\mathrm{A},ip} \\ \hat{\mu}^{AB}_{\mathrm{A},ip} \\ \hat{\mu}^{AA}_{\mathrm{A},ip} \end{bmatrix} = \mathrm{diag}\left(\frac{\mathbf{1}}{\hat{\boldsymbol{\xi}}_{\boldsymbol{A,ip}}}\right) \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \times \begin{bmatrix} \nu_{\mathrm{A}ip} \\ \phi_{\mathrm{A}ip} \end{bmatrix} + \boldsymbol{m}_{A,ijp}. \tag{9}$$

The errors $\boldsymbol{m}_{A,ijp}$ are approximately independent multi-Gaussian. We repeat the procedure for the B-allele to obtain batch- and locus-specific estimates of ν_{B} and ϕ_{B}. The naïve estimates of allele-specific copy number are updated by subtracting the estimated background from the observed intensity and scaling by the slope coefficient. Specifically,

$$\hat{c}_{A,ijp} = \max\left\{\frac{1}{\hat{\phi}_{A,ip}}\left(I_{A,ijp} - \hat{\nu}_{A,ip}\right),\ 0\right\} \text{ and} \tag{10}$$

$$\hat{c}_{B,ijp} = \max\left\{\frac{1}{\hat{\phi}_{B,ip}}\left(I_{B,ijp} - \hat{\nu}_{B,ip}\right),\ 0\right\}. \tag{11}$$

As discussed in Section 7, the assumption that the median intensity is linear with copy number appears reasonable for a limited range. We have observed departures from linearity for larger copy numbers as the fluorescence becomes more saturated. In practice, we constrain $\hat{c}_{A,ijp} + \hat{c}_{B,ijp} \leq 6$. The above prescription for copy number estimation is predicated on the assumption that at any given locus the majority of samples have normal copy number. In

Section 6, we explore the robustness of this approach to misspecification of the initial values for allele-specific copy number.

5.2 Unobserved genotypes and nonpolymorphic loci

For many loci, the minor allele is rare and one or more of the three possible diallelic genotypes are not observed. For SNPs with genotype GT not observed, we impute $\mu_{\mathrm{A}}^{\mathrm{GT}}$ and $\mu_{\mathrm{B}}^{\mathrm{GT}}$ via regression. For example, to impute $\mu_{\mathrm{A}}^{\mathrm{AA}}$ for SNPs with genotype AA unobserved, we regress $\hat{\mu}_{\mathrm{A}}^{\mathrm{AA}}$ on $\hat{\mu}_{\mathrm{A}}^{\mathrm{AB}}$ and $\hat{\mu}_{\mathrm{A}}^{\mathrm{BB}}$ using a set of SNPs for which all three genotypes were observed. With estimates of the coefficients for $\hat{\mu}_{\mathrm{A}}^{\mathrm{AB}}$ and $\hat{\mu}_{\mathrm{A}}^{\mathrm{BB}}$, we predict the value of $\mu_{\mathrm{A}}^{\mathrm{AA}}$ from the observed $\hat{\mu}_{\mathrm{A}}^{\mathrm{AB}}$ and $\hat{\mu}_{\mathrm{A}}^{\mathrm{BB}}$ at this locus. We repeat the procedure for the B allele to impute $\mu_{\mathrm{B}}^{\mathrm{AA}}$. For a polymorphic locus with two genotypes not unobserved, there is no information regarding the slope parameters ϕ_A or ϕ_B. Again, we impute the unobserved within-genotype medians via regression using SNPs with all three genotypes observed. The variance terms for the unobserved genotype GT, $\tilde{\xi}_A^{GT}$ and $\tilde{\xi}_B^{GT}$, are obtained from the prior in equation (8).

For nonpolymorphic loci, the parameters for background, ν_T, and slope, ϕ_T, in Model (7) are more difficult to estimate as there are no genotype clusters to guide their estimation. For each nonpolymorphic probe, we assume that the median of the observed intensities across samples in the batch corresponds to normal copy number. One approach is to impute terms for the background and slope using the nonpolymorphic loci on chromosome X and chromosome Y. This approach requires that there are both men and women in each batch. An alternative approach, one that is currently implemented in the R package *crlmm*, is to impute ϕ_T from polymorphic loci at which all three genotypes were observed. Briefly, for SNPs with three diallelic genotypes observed, we fit a linear model using $\hat{\mu}_A^{AA}$ and $\hat{\mu}_B^{BB}$ as the explanatory variables and the corresponding slopes, $\hat{\phi}_A$ and $\hat{\phi}_B$, as the response variables. The slope parameter for nonpolymorphic loci, $\hat{\phi}_T$, is predicted using the observed $\hat{\mu}_T$ at nonpolymorphic loci and the corresponding coefficient estimated from the polymorphic loci. Note that the background fluorescence, ν_T, is determined by the relationship $\hat{\mu}_T - 2\phi_T$.

The variance for non-normal copy number are obtained from the prior. Transforming the nonpolymorphic intensities to the copy number scale is achieved by

$$c_{T,ijp} \;=\; \max\left\{\frac{1}{\hat{\phi}_{T,ip}}\left(I_{T,ijp}-\hat{\nu}_{ip}\right),\ 0\right\}. \tag{12}$$

Contamination. In many applications, DNA is isolated from a mixture of two or more cell types that may have harbor different somatic alterations. As the DNA in the cell populations may differ, noninteger copy numbers are plausible. Hidden Markov models that assume integer copy number states are not appropriate. The transformations in equations (10), (11), and (12) allow one to plot $\hat{c}_A + \hat{c}_B$ and $\hat{c}_T$ as a function of the physical position to assess contamination. When contamination is likely to have occurred, a variety of nonparametric segmentation approaches are available that can be used to identify noninteger copy number gain and loss.

5.3 Uncertainty.

Estimates of the uncertainty are important for downstream algorithms that smooth estimates of the copy number as a function of the physical position. As mentioned previously, a critical choice governing the suitability of a smoothing algorithm is the presence of a mixture of cell populations that can result in noninteger copy number. In the absence of cell contamination, we advocate a HMM that can be fit directly to bivariate normal scatterplots of the log A and log B intensities, an approach originally developed elsewhere (Korn et al., 2008). When cell contamination is likely, we prefer nonparametric segmentation algorithms that can identify any noninteger shift in copy number. Because most segmentation algorithms do not readily incorporate estimates of uncertainty, our focus in this section is improving estimates of the uncertainty for the prediction regions of allele-specific copy number. For instance, see the prediction regions in Figure 4.

Integer copy number. As HMMs can incorporate locus-specific estimates of the location and scale in the emission probabilities, HMMs for detecting copy number alterations can therefore be applied directly to the bivariate normal scatterplots without first transforming the intensities to the copy number scale (Korn et al., 2008). Conditional on the allelic copy number, the logarithm of the I_{A} and I_{B} intensities is approximately bivariate normal with a mean and covariance that is locus- and batch-specific, as in model (3). Again, our procedure utilizes naïve estimates of the allelic copy number from the diallelic genotype calls to provide an estimate of $\mathbf{\Sigma}$. To illustrate this approach, we describe the estimation of $\mathbf{\Sigma}$ for a SNP with diallelic genotype AA, $(c_A^\star = 2, c_B^\star = 0)$. From equation (4), the diagonal elements of $\mathbf{\Sigma}_{ip}$ are

$$(\mathbf{\Sigma}_{ip})_{11} = \sigma^2_{A,ip} \quad \text{and} \quad (\mathbf{\Sigma}_{ip})_{22} = \tau^2_{B,ip}.$$

The background variance $\tau^2_{B,ip}$ is estimated as the MAD of the log intensities for the B allele across all subjects with genotype AA at locus i. The signal variance $\sigma^2_{A,ip}$ is estimated as the MAD of the log intensities for the A allele across all subjects with genotype AA at locus i. Implicitly, we assume that the variance of δ_A is small relative to the variance of ε_A such that $\mathrm{Var}\left\{\log(I_{A,ijp}|c_A^\star > 0\right\} \approx \hat{\sigma}^2_{A,ip}$. The assumption that the variance is constant for $c_A^\star$ greater than zero appears reasonable on the log-scale. Similarly, an initial estimate for the correlation of the log intensities for the A and B alleles, ρ_{ip}, is estimated empirically among subjects with genotype AA. The within-genotype empirical estimates for the variance terms and the correlation parameter provide an initial estimate of $\mathbf{\Sigma}$. These estimates can be very noisy when based on a small number of observations. Therefore, we shrink the initial estimates of $\mathbf{\Sigma}$ using inverse chi-squared priors as described in Section 4. Specifically, shrinkage estimates

for the background and signal variances for $(c_A^\star = 2, c_B^\star = 0)$ are obtained by

$$\begin{aligned}\tilde{\sigma}^2_{A,ip} &= \frac{(N_{AA,ip}-1)\hat{\sigma}^2_{A,ip} + d_A s^2_{A,p}}{N_{AA,ip}-1+d_A} \text{ and} \\ \tilde{\tau}^2_{B,ip} &= \frac{(N_{AA,ip}-1)\hat{\tau}^2_{B,ip} + d_B t^2_{B,p}}{N_{AA,ip}-1+d_B}.\end{aligned}$$

The count $N_{GT,ip}$ denotes the number of subjects with genotype GT at locus i in batch p. The degrees of freedom for the priors, d_A and d_B, can be estimated as described in Lönnstedt and Speed (2001). Typical values of the variance, denoted by t and s, are estimated across all loci and allowed to depend on the batch. In addition to the variances, we also shrink the empirical estimate of the correlation $\rho_{i,p}$. As motivated in section 4, we suggest a Beta prior that puts most of the mass on typical values. The resulting covariance matrix, $\widetilde{\mathbf{\Sigma}}_{i,p}$, can be used to plot prediction regions for any $(c_A^\star > 0, c_B^\star = 0)$. For instance, see the ellipses for the genotype AA in Figure 4. The covariance matrix for $(c_A^\star = 0, c_B^\star > 0)$, $(c_A^\star > 0, c_B^\star > 0)$ and $(c_A^\star = 0, c_B^\star = 0)$ are obtained using a similar procedure. By scaling $\widetilde{\mathbf{\Sigma}}_{i,p}$ by a sample-specific estimate of the variance across all loci, the variance estimate can incorporate information on the overall noise of the sample relative to other samples.

5.4 Common copy number variants

Our approach for estimating copy number uses robust estimates of the within-genotype location and scale of the intensities. In particular, we use medians for the location and median absolute deviations for the variance to prevent outliers from influencing our estimates of ν and ϕ. However, many regions of the genome appear to contain common variants in apparently normal individuals (McCarroll et al., 2008; Kidd et al., 2008), and many diseases may have regions that are commonly altered. For genomic locations where a large number of subjects contain a copy number alteration, estimates of ν and ϕ can be biased. We propose an update for the background and slope parameters that provides additional robustness to regions with a large number of alterations. An important feature of this procedure is that

we do not require any *a priori* knowledge of the genomic locations of the common variants as these are often not well characterized or highly variable for diseases such as cancer.

The general strategy is to estimate the parameters ν_A, ϕ_A, ν_B, ϕ_B, ν_T, and ϕ_T as described previously. For each loci, we determine which samples are least likely to have a *normal* copy number. Specifically, we calculate the posterior probability of belonging to a prediction region corresponding to an aggregate copy number of 0, 1, 2, or 3. For SNPs, we assume that for a fixed total copy number, any of the integer (c_A, c_B) combinations are equally likely. At each locus, we tabulate the frequency for which the posterior probability of an amplification or deletion is greater than the posterior probability of normal copy number. If the frequency is uneven for amplifications and deletions, we recompute the within-genotype location and scale parameters after trimming the tail of the distribution that has a greater frequency; otherwise, we trim both tails and recompute. See Section 6 for an application.

6 Results

We illustrate our approach for copy number estimation using two datasets that were assayed on the Affymetrix 6.0 genotyping platform: a GWAS for bipolar disease (dbGaP accession number phs000017.v3.p1) and a dataset containing 26 individuals with chromosome 21 trisomy and 70 apparently healthy controls. (Three healthy controls in the trisomy dataset were excluded because of a low signal to noise ratio.) For the bipolar disease dataset, batch is a confounder of case-control status. Our focus is on the estimation of copy number for the 1094 European ancestry controls that were processed on 29 different chemistry plates over a two month period. For the chromosome 21 trisomy data, we analyzed the 96 samples as a single batch and assess the robustness of our estimation algorithm to a dataset where as many as 27% of the samples are known to have three chromosomal copies. A difficulty in comparing our method to the software suggested on the Affymetrix website for copy number analysis is that these algorithms (Birdseye and Canary) do not provide locus-level estimates

of copy number. Rather, Birdseye provides output from a hidden Markov model for detection of de-novo CNV regions and a separate algorithm, Canary, for estimating copy number in regions that are thought to contain common copy number variants (Korn et al., 2008; McCarroll et al., 2008).

Batch effects. The European ancestry controls for bipolar disease were hybridized to Affymetrix 6.0 chips and scanned over a period of two months. As described in Section 5, we preprocessed the raw intensities using quantile normalization and summarized the intensities to the level of the locus using a median. The control samples were genotyped using the software CRLMM. Figures 2(a) and 2(b) for SNP_A-4251622 on chromosome 15 provide a visualization of the batch effect: the F-statistic from a one-way analysis of variance (ANOVA) for the sum of the quantile normalized A and B intensities across plate is 16.84 for SNP_A-4251622. We found evidence of batch effects at most loci on chromosome 15. For instance, Figure 3 plots the distribution of F-statistics for all of the polymorphic loci. We observed similar batch effects at nonpolymorphic loci and at polymorphic loci on other chromosomes (data not shown). Robust multi-array normalization procedures alone are not sufficient for removing batch effects in SNP microarray data.

Figures 2(c) and 2(d) plot copy number estimates without correcting for batch effects. Note that the copy number estimates (2(d)) are highly dependent on chemistry plate and boxplots of the estimates follow a similar pattern as the total quantile normalized intensity (2(b)). As much of the variation in the A versus B scatterplots is attributable to batch differences, the ellipses in Figure 2(c) that reflect our uncertainty of the prediction region are inflated. By allowing the parameters $\nu_{\mathrm{A}}, \nu_{\mathrm{B}}, \nu_T, \phi_{\mathrm{A}}, \phi_{\mathrm{B}}$, and ϕ_T to depend on batch, we obtain prediction regions that more accurately reflect the uncertainty of the copy number estimates 2(e) and are more robust to differences across batch 2(f). Our approach does not use a reference set for estimating locus-specific parameters and this has many advantages in light of large batch effects as well as biological differences in populations that preclude their

extrapolation.

We also fit the Birdsuite software separately for each plate in the Bipolar controls. The Birdsuite software uses separate algorithms for calling copy number: a HMM for discovery of de-novo CNV (Birdseye) and one for calling copy number in regions that are 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 (see Supplementary Figure 1), whereas our algorithm is unlikely to do so. Our approach may be preferable if the common variant maps do not extrapolate well to a given population, but can result in many false negatives if most of the samples are, in truth, amplified or deleted. An interesting feature of the Birdseye segmentation is that we observe strong plate 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. The Canary algorithm can be helpful for reducing the batch effect in such regions as a Tukey median polish provides an extra normalization step, but batch effects often persist (1b).

Common copy number variants. To explore the robustness of our approach to the assumption that the typical copy number is two at any given locus, we applied our algorithm to the trisomy dataset where approximately 26% of the samples have three copies of chromosome 21. As Discussed in Section 5, our algorithm uses diallelic genotype calls to develop naïve estimates of copy number and parameter estimates for prediction regions. To reduce the influence of samples that are not diallelic on the parameter estimates, we exclude samples with a high posterior probability of a copy number alteration. Boxplots of locus-specific estimates of copy number for the SNPs on chromosome 21 before and after the bias correction are plotted in Figures 6(a) and 6(b), respectively. Note that the initial estimates of copy

number are biased towards small values as locus-specific prediction regions in the A versus B scatterplots are shifted slightly towards higher values because 25% of the subjects have three copies of chromosome 21. After a second iteration of our method whereby samples that have a high posterior probability of belonging to a non-normal copy number state are excluded, the parameters $\nu_{\mathrm{A}}, \nu_{\mathrm{B}}, \nu_T, \phi_{\mathrm{A}}$, and ϕ_T are updated. Again, the set of samples that are excluded in this step is locus-specific and does not use any phenotype information (e.g., trisomy status) of the subjects. While locus-specific estimates are not available from the Birdsuite software, an overall copy number estimate for each chromosome is available from the Birdseye algorithm for the purpose of assessing mosaicism. From the overall Birdseye copy number for chromosome 21, one may incorrectly conclude that the samples were mosaic in copy number (dashed line in Figures 6(a) and 6(b)). The fact that the median copy number from our locus-level estimates are less biased than the overall copy number reported by Birdseye may reflect that we use more robust statistics for computing prediction regions (Birdseye uses a trimmed mean).

[Figure 6 about here.]

To more formally compare our approach to Birdseye, we fit a HMM to our copy number estimates using the same transition probabilities as that used by the Birdseye HMM (Korn et al., 2008). Assuming that the true copy number is 2 in the normal samples and 3 for the trisomy samples, we calculated the proportion of correct calls for each approach. Our approach maintains high sensitivity and specificity for detecting alterations despite 25% of the samples having a known copy number alteration. That is, naïve estimates of copy number provided by the diallelic genotype calls can be incorrect in 25% of the samples, but still provide unbiased estimates of copy number in regions that are commonly variant.

[Table 1 about here.]

7 Discussion

In this paper, we propose a multilevel model that provides absolute estimates of allele-specific copy number at polymorphic loci and total copy number at nonpolymorphic loci. The parameters used in our model are not pre-computed from training data, nor does our model require a reference set of normal samples. Rather, diallelic genotype calls of samples in the experimental dataset provide naïve estimates of allele-specific copy number that are used to derive robust-to-outlier parameters for the background and slope in a linear model fit on the intensity scale. Conditional on the naïve estimates of copy number, the log A and B intensities are correlated due to cross-hybridization of the probes for these alleles. As locus-specific estimates of the covariance are often based on a small number of observations, shrinking these estimates towards typical values provides additional robustness to unusually small or large variance estimates. Finally, all of these parameters are batch-specific. The copy number estimates obtained from this approach are shown to be robust to batch effects and robust to a large proportion of individuals having a copy number alteration. The resulting estimates can be plotted as a function of the physical position to assess issues such as cell contamination and serve as a starting point for segmentation- or HMM-based approaches for smoothing as a function of the physical position. Therefore, the copy number estimated provided by this algorithm are complementary to nonparametric segmentation algorithms, such as circular binary segmentation, or HMMs.

Our procedure for locus-level estimation of copy number reduces the impact of batch effects in several ways. First, we quantile normalize the raw intensities for each sample to a target reference distribution. By quantile-normalizing to a target distribution, we reduce the occurrence of batch effects that may be induced by how the samples were grouped by the software. Secondly, we do not rely on external data for training the model or a reference set for estimating ratios. In our experience, external datasets are unlikely to provide useful extrapolations due to (i) batch effects and (ii) biological differences between the external samples and the test samples. Rather, we provide an absolute estimate of the copy number

and a corresponding estimate of the uncertainty using only the experimental data. Third, we fit a multilevel model to the summarized intensities that allows locus-specific parameters for background and signal to depend on batch. Finally, we provide an iterative solution to copy number estimation that can detect copy number alterations when as many as 25% of the samples in a batch have non-normal copy number. This approach does not require prior knowledge of the locations of regions that are commonly altered.

Our model is most useful for datasets with 25 or more samples processed together in a batch. For datasets with fewer than 10-25 samples, our model-based approach will impute the within-genotype center for a large number of loci with unobserved diallelic genotypes. This extra layer of uncertainty from the imputation will provide less precise estimates of the copy number and a reduced resolution for detecting alterations of copy number. In the set of European ancestry bipolar controls, we defined batch as the chemistry plate on which these samples were hybridized and discarded samples from four plates that each had fewer than 20 samples. Note that a principal component analysis for batch effect may have suggested an alternative grouping, though in our experience the first few principal components correspond to plate. Although our model does not explicitly require a reference set of normal controls, we do assume that at any given locus the typical copy number is two. This assumption may not be reasonably for many datasets, particularly cancers. Ideally, one would have an enough normal controls processed in the same batch as the test samples so that the assumption of normal copy number is tenable for each batch (e.g., 50% normal controls). SNP-specific A versus B scatterplots are useful for identifying when model assumptions are unreasonable. In the trisomy example, we demonstrate that our approach is robust to as many as 30% of the samples in a batch having altered copy number.

Of the models previously proposed in the literature, the models of Wang et al. (2008) and Korn et al. (2008) are the most similar to ours. The Wang model provides allele-specific estimates of copy number that takes into account the correlation of A and B allele intensities. However, the Wang model is designed for an earlier version of the Affymetrix platform that

contained only SNP probes, proposes the use of training data to estimate model parameters, does not adjust for batch effects, and does not use shrinkage to improve estimates of the variance. Prediction regions obtained from their approach on the Bipolar controls would be similar to Figures 2(c) and 2(d) with the exception that model parameters would have been precomputed from a training dataset such as HapMap, and the variance estimates of prediction regions would be more sensitive to extreme values. The Korn model is similar to the Wang model with a few important differences. First, Korn et al. recommend fitting their software by batch. For any of the previously proposed models, one could 'add' a fixed effect for plate by applying the software independently to each plate. However, fitting software by plate is not always successful for removing batch effect. In part, this may occur because of a software-induced batch effect that occurs during the preprocessing. In particular, we expect that by not quantile normalizing samples in a batch to a target reference distribution, the splitting of samples by the software during the preprocessing can exacerbate batch effects. Secondly, Birdsuite does not provide locus-specific estimates of copy number, nor does their algorithm appear particularly robust to a substantial proportion of samples having an altered copy number. While a HMM fit to the bivariate normal A and B scatterplots is a nice feature, this does not facilitate checks for cell contamination. If cell contamination is thought to have occurred, one would have to explore a different approach for copy number estimation. The main innovation of the Korn model are different algorithms for detecting *de-novo* and common copy number variants. Such approaches are complementary to the work presented here as these tools each involves borrowing strength from neighboring loci to improve the locus-level estimates.

The multilevel model we propose for copy number estimation can be extended in several ways. This paper proposes modeling batch as a fixed effect. A compromise between a random- and fixed-effect for batch could be explored as a means to improve the copy number prediction regions for loci with low minor allele frequencies. More generally, one could regard batch as a problem of variable selection. Currently, our model assumes a linear relationship of

copy number and the within-genotype medians. While this relationship appears reasonable for zero to three copies of an allele, the relationship becomes more nonlinear for higher copy numbers. Alternative approaches that take into account this nonlinearity should be explored. Finally, adjusting for sequence characteristics such as GC-content and fragment-length can be helpful for reducing the variance associated with the probe-effect. We will explore methods that adjust for these factors along with batch effects in the future.

Our results provide a strong indication that a model-based approach for estimation of absolute allele-specific copy number can be effective in large studies with pronounced batch effects, and that borrowing strength across loci can be useful for improving estimates of the variance. Estimates of the copy number and the corresponding uncertainty will be useful for downstream assessments of copy number-phenotype association.

8 Acknowledgements

The bipolar dataset used for the analyses described in this manuscript was obtained from the database of Genotype and Phenotype (dbGaP) at `http://www.ncbi.nlm.nih.gov/gap`http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number (phs000017.v3.p1), and was provided through the Genetic Association Information Network (GAIN). We thank Erin M. Ramos and Lisa J. McNeil from the NIH for help navigating dbGap. RBS supported by NIH grant 1K99HG005015, CTSA grant ..., and training grant 5T32HL007024 from the National Heart, Lung, and Blood Institute. RAI was supported by NIH grants R01GM083084 from the National Institute of General Medicine and 5R01RR021967 from the National Center for Research Resource.

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).

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." *Bioinformatics (Oxford, England)*, 19(2):185–193 (2003). PUBM: Print; JID: 9808944; 0 (Molecular Probes); ppublish.

Cappuzzo, F., Marchetti, A., Skokan, M., Rossi, E., Gajapathy, S., Felicioni, L., Grammastro, M. D., Sciarrotta, M. G., Buttitta, F., Incarbone, M., Toschi, L., Finocchiaro, G., Destro, A., Terracciano, L., Roncalli, M., Alloisio, M., Santoro, A., and Varella-Garcia, M. "Increased MET Gene Copy Number Negatively Affects Survival of Surgically Resected Non-Small-Cell Lung Cancer Patients." *J Clin Oncol* (2009).

Carvalho, B., Louis, T. A., and Irizarry, R. "Quantifying uncertainty in genotype calls." Technical report, Johns Hopkins University (2009).

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).

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

Gu, W., Zhang, F., and Lupski, J. R. "Mechanisms for human genomic rearrangements." *Pathogenetics*, 1(1):4 (2008).

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).

Inc., P. "Partek Discovery SuiteTM." Technical report, Partek (2008). Version 6.3, Publisher: Partek Inc.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tuzun, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. "Mapping and sequencing of structural variation from eight human genomes." *Nature*, 453(7191):56–64 (2008).

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).

Lönnstedt, I. and Speed, T. "Replicated Microarray Data." *Statistica Sinica*, 0:000–000 (2001).

Lupski, J. R. "Genomic disorders ten years on." *Genome Med*, 1(4):42 (2009).

Ma, O., Cai, W.-W., Zender, L., Dayaram, T., Shen, J., Herron, A. J., Lowe, S. W., Man, T.-K., Lau, C. C., and Donehower, L. A. "MMP13, Birc2 (cIAP1), and Birc3 (cIAP2), amplified on chromosome 9, collaborate with p53 deficiency in mouse osteosarcoma progression." *Cancer Res*, 69(6):2559–2567 (2009).

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., Thiruvahindrapduram, B., Fiebig, A., Schreiber, S.,

Friedman, J., Ketelaars, C. E. J., Vos, Y. J., Ficicioglu, C., Kirkpatrick, S., Nicolson, R., Sloman, L., Summers, A., Gibbons, C. A., Teebi, A., Chitayat, D., Weksberg, R., Thompson, A., Vardy, C., Crosbie, V., Luscombe, S., Baatjes, R., Zwaigenbaum, L., Roberts, W., Fernandez, B., Szatmari, P., and Scherer, S. W. "Structural variation of chromosomes in autism spectrum disorder." *Am J Hum Genet*, 82(2):477–488 (2008).

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).

McKinney, C., Merriman, M. E., Chapman, P. T., Gow, P. J., Harrison, A. A., Highton, J., Jones, P. B. B., McLean, L., O'Donnell, J. L., Pokorny, V., Spellerberg, M., Stamp, L. K., Willis, J., Steer, S., and Merriman, T. R. "Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis." *Ann Rheum Dis*, 67(3):409–413 (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).

Need, A. C., Ge, D., Weale, M. E., Maia, J., Feng, S., Heinzen, E. L., Shianna, K. V., Yoon, W., Kasperavicite, D., Gennarelli, M., Strittmatter, W. J., Bonvicini, C., Rossi, G., Jayathilake, K., Cola, P. A., McEvoy, J. P., Keefe, R. S. E., Fisher, E. M. C., Jean, P. L. S., Giegling, I., Hartmann, A. M., Mller, H.-J., Ruppert, A., Fraser, G., Crombie, C., Middleton, L. T., Clair, D. S., Roses, A. D., Muglia, P., Francks, C., Rujescu, D., Meltzer, H. Y., and Goldstein, D. B. "A genome-wide investigation of SNPs and CNVs in schizophrenia." *PLoS Genet*, 5(2):e1000373 (2009).

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).

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F.,

Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. "Global variation in copy number in the human genome." *Nature*, 444(7118):444–454 (2006).

Rigaill, G., Hup, 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).

Sutrala, S. R., Goossens, D., Williams, N. M., Heyrman, L., Adolfsson, R., Norton, N., Buckland, P. R., and Del-Favero, J. "Gene copy number variation in schizophrenia." *Schizophr Res*, 96(1-3):93–99 (2007).

Szatmari, P., Paterson, A. D., Zwaigenbaum, L., Roberts, W., Brian, J., Liu, X.-Q., Vincent, J. B., Skaug, J. L., Thompson, A. P., Senman, L., Feuk, L., Qian, C., Bryson, S. E., Jones, M. B., Marshall, C. R., Scherer, S. W., Vieland, V. J., Bartlett, C., Mangin, L. V., Goedken, R., Segre, A., Pericak-Vance, M. A., Cuccaro, M. L., Gilbert, J. R., Wright, H. H., Abramson, R. K., Betancur, C., Bourgeron, T., Gillberg, C., Leboyer, M., Buxbaum, J. D., Davis, K. L., Hollander, E., Silverman, J. M., Hallmayer, J., Lotspeich, L., Sutcliffe, J. S., Haines, J. L., Folstein, S. E., Piven, J., Wassink, T. H., Sheffield, V., Geschwind, D. H., Bucan, M., Brown, W. T., Cantor, R. M., Constantino, J. N., Gilliam, T. C., Herbert, M., Lajonchere, C., Ledbetter, D. H., Lese-Martin, C., Miller, J., Nelson, S., Samango-Sprouse, C. A., Spence, S., State, M., Tanzi, R. E., Coon, H., Dawson, G., Devlin, B., Estes, A., Flodman, P., Klei, L., McMahon, W. M., Minshew, N., Munson, J., Korvatska, E., Rodier, P. M., Schellenberg, G. D., Smith, M., Spence, M. A., Stodgell, C., Tepper, P. G., Wijsman, E. M., Yu, C.-E., Roge, B., Mantoulan, C., Wittemeyer, K., Poustka, A., Felder, B., Klauck, S. M., Schuster, C., Poustka, F., Bolte, S., Feineis-Matthews, S., Herbrecht, E., Schmotzer, G., Tsiantis, J., Papanikolaou, K., Maestrini, E., Bacchelli, E., Blasi, F., Carone, S., Toma, C., Van Engeland, H., de Jonge, M., Kemner, C., Koop, F., Langemeijer, M., Hijimans, C., Staal, W. G., Baird, G., Bolton, P. F., Rutter, M. L., Weisblatt, E., Green, J., Aldred, C., Wilkinson, J.-A., Pickles, A., Le Couteur, A., Berney, T., McConachie, H., Bailey, A. J., Francis, K., Honeyman, G., Hutchinson, A., Parr, J. R., Wallace, S., Monaco, A. P., Barnby, G., Kobayashi, K., Lamb, J. A., Sousa, I., Sykes, N., Cook, E. H., Guter, S. J., Leventhal, B. L., Salt, J., Lord, C., Corsello, C., Hus, V., Weeks, D. E., Volkmar, F., Tauber, M., Fombonne, E., and Shih, A. "Mapping autism risk loci using genetic linkage and chromosomal rearrangements." *Nat Genet*, 39(3):319–28 (2007).

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).

Woo, H. G., Park, E. S., Lee, J.-S., Lee, Y.-H., Ishikawa, T., Kim, Y. J., and Thorgeirsson19297395, S. S. "Identification of Potential Driver Genes in Human Liver Carcinoma by Genomewide Screening." *Cancer Res* (2009).

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).

Zhang, D., Cheng, L., Qian, Y., Alliey-Rodriguez, N., Kelsoe, J. R., Greenwood, T., Nievergelt, C., Barrett, T. B., McKinney, R., Schork, N., Smith, E. N., Bloss, C., Nurnberger, J., Edenberg, H. J., Foroud, T., Sheftner, W., Lawson, W. B., Nwulia, E. A., Hipolito, M., Coryell, W., Rice, J., Byerley, W., McMahon, F., Schulze, T. G., Berrettini, W., Potash, J. B., Belmonte, P. L., Zandi, P. P., McInnis, M. G., Zllner, S., Craig, D., Szelinger, S., Koller, D., Christian, S. L., Liu, C., and Gershon, E. S. "Singleton deletions throughout the genome increase risk of bipolar disorder." *Mol Psychiatry* (2008).

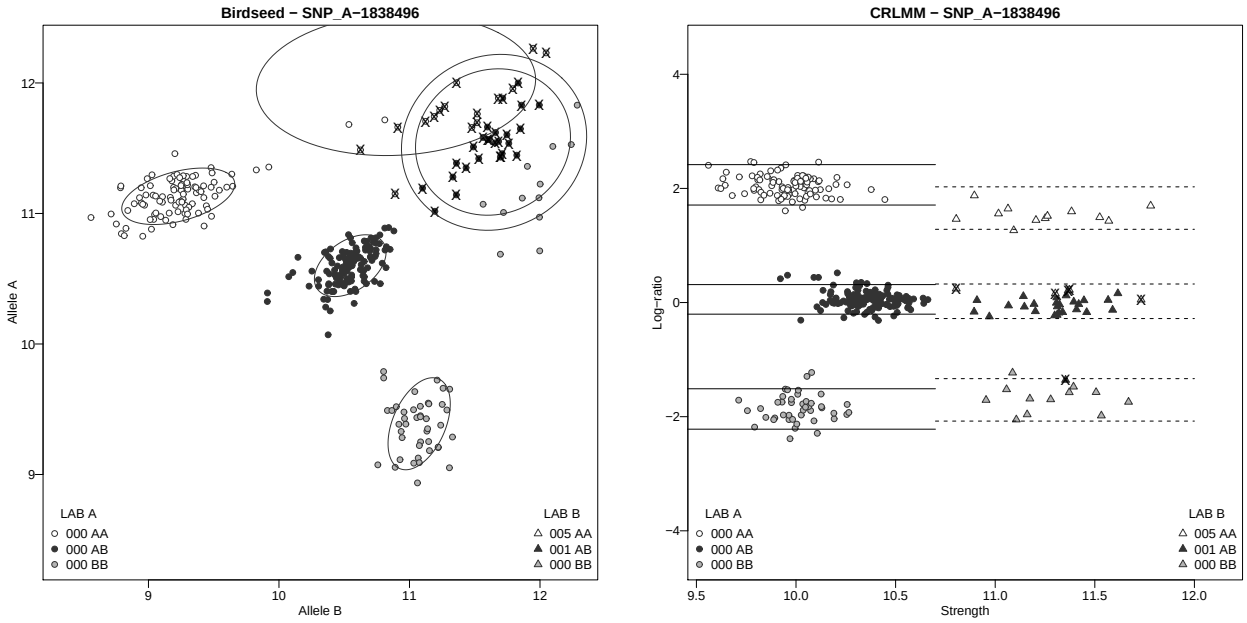


Figure 1: A set of identical samples was genotyped by two different labs. Left: A scatter plot of the A versus B allele intensities for a single SNP with plotting symbols denoting the consensus HapMap genotype. The default genotyping algorithm for this platform provided by Affymetrix, Birdseed, makes 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.

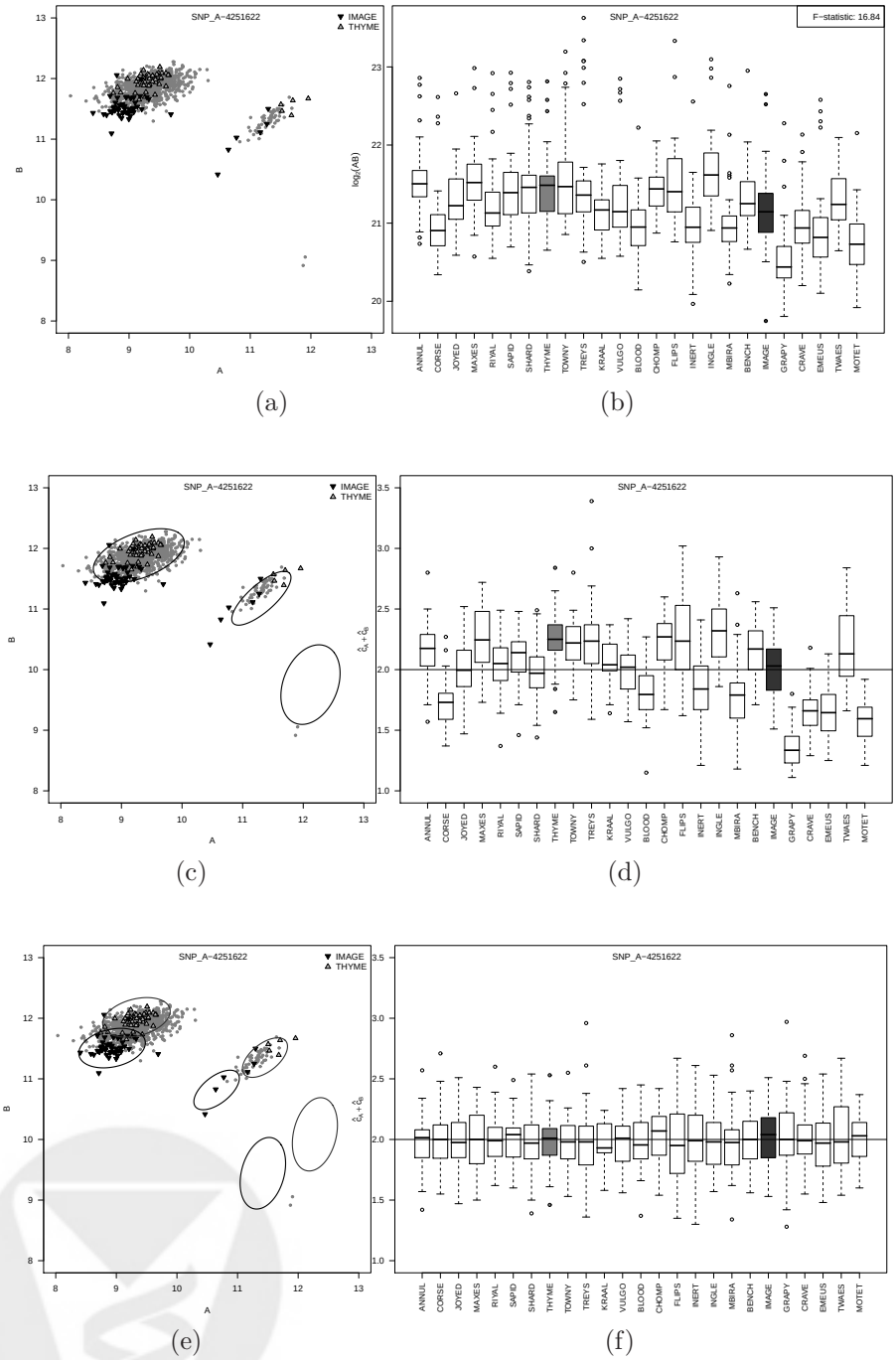


Figure 2: The European ancestry controls for Bipolar disease were run on 29 plates; we excluded 4 plates that had fewer than 20 samples. Column 1: Scatter plots of the quantile normalized intensities of the A (x-axis) and B (y-axis) alleles for SNP_A-4251622. Highlighted in the scatter plots are the samples from two of the plates (IMAGE and THYME). Column 2: Boxplots of $\log_2(A) + \log_2(B)$ (b) or copy number (d and f) stratified by plate. (c and d): Prediction regions for copy number two (c) and the corresponding copy number estimates stratified by plate (d) in a model that does not adjust for batch effects. (e and f): A multilevel model that allows the prediction regions to depend on plate improves estimates of the uncertainty (ellipses for IMAGE and THYME are shown in panel e) and provides copy number estimates that are more robust to batch differences.

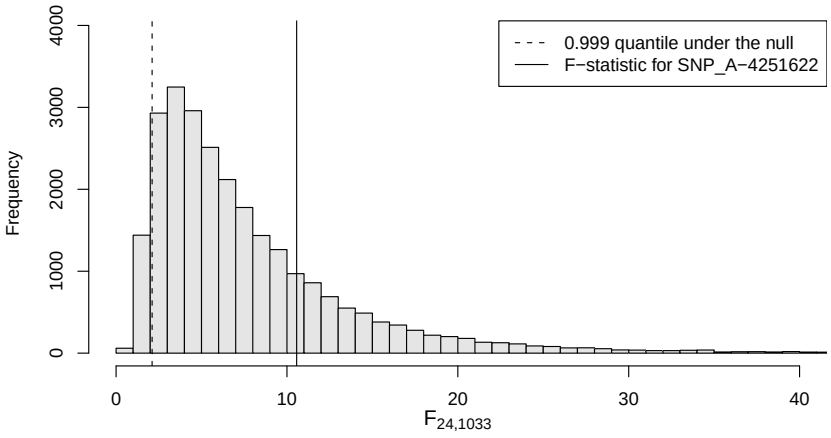
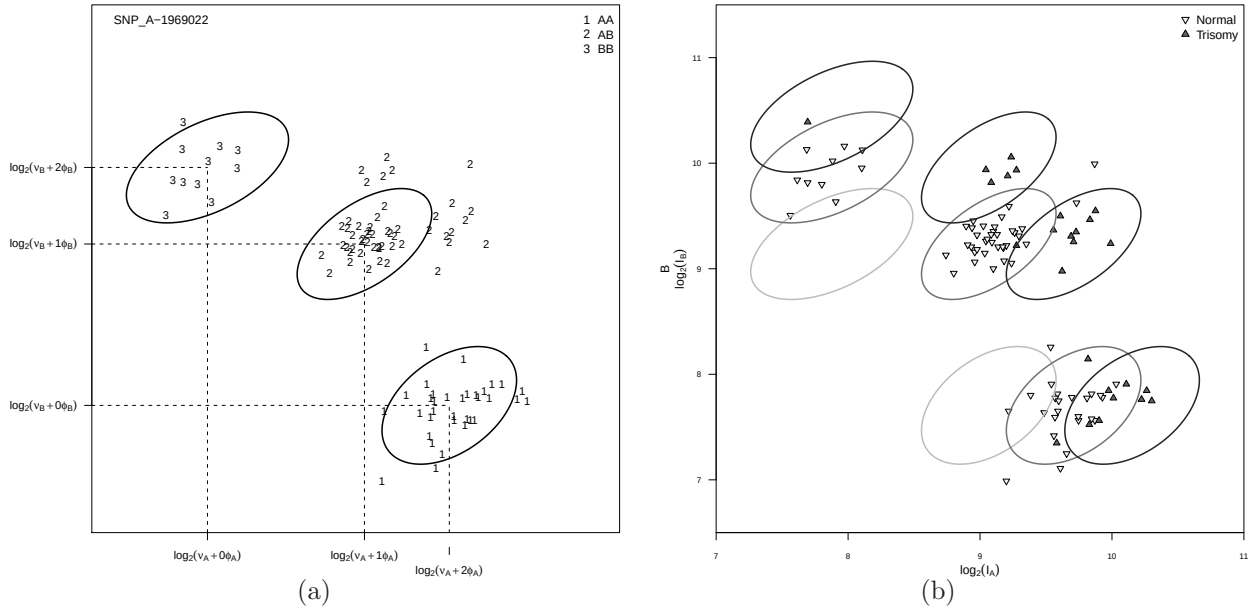


Figure 3: For each SNP on chromosome 15 of the European ancestry controls for bipolar disease, we performed an analysis of variance (ANOVA) for the quantile normalized A + B intensities by plate. After excluding four plates with fewer than 20 samples, the ANOVA provides an F-statistic with 24 and 1033 degrees of freedom for each of the 26,074 SNPs on chromosome 15.



(a) (b)

Figure 4: Scatterplots of the A and B allele intensities for SNP_A-1969022 on chromosome 21 in the trisomy dataset. (a) Our approach for copy number estimation uses naïve estimates of allele-specific copy number based on the diallelic genotype calls. A weighted linear regression is fit on the intensity scale to quantile-based estimators of the within-genotype location and scale. Estimates of ν_{A}, ν_{B}, ϕ_{A}, and ϕ_{B} are locus- and batch-specific. The ellipses demarcate a 95% confidence region for copy number 2. (b) Prediction regions for copy number 1, 2, and 3. Plotting symbols now denote the trisomy phenotype which is not known by the regression model. Note that the prediction regions are robust to incorrect diallelic genotype calls – here, 26 of the 96 subjects had chromosome 21 trisomy and, therefore, incorrect diallelic genotypes.

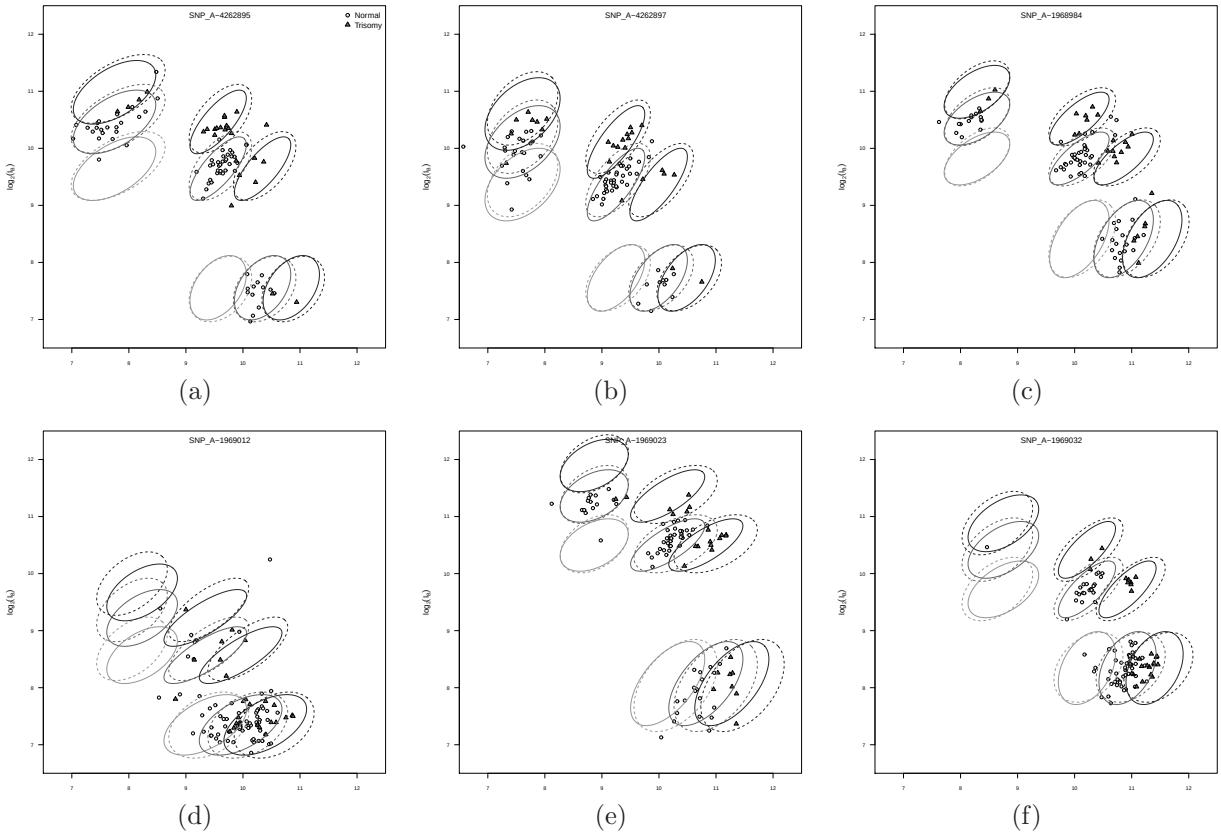
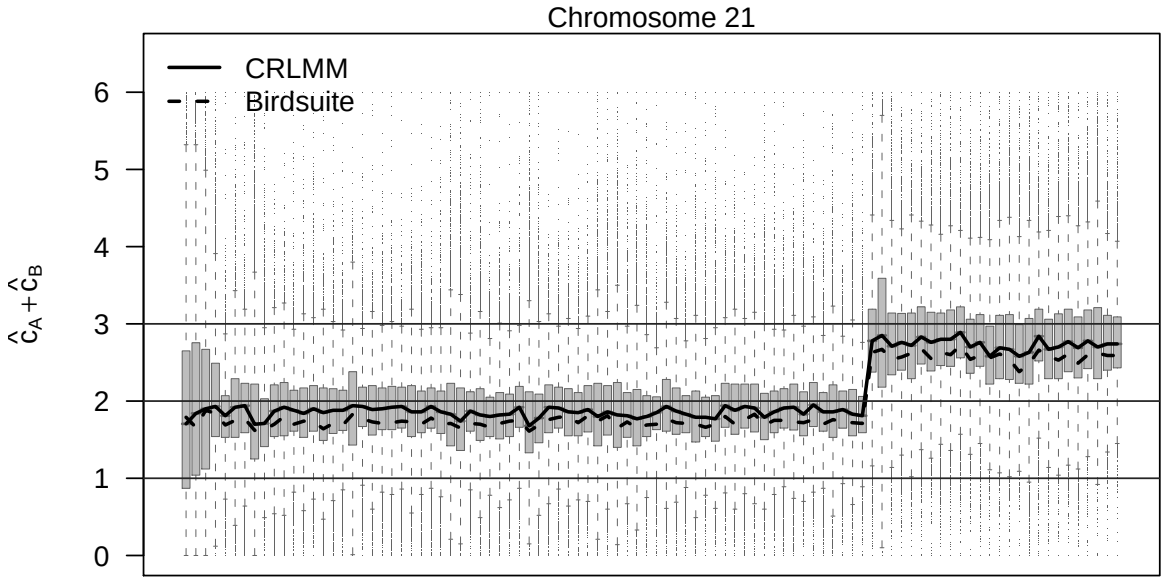
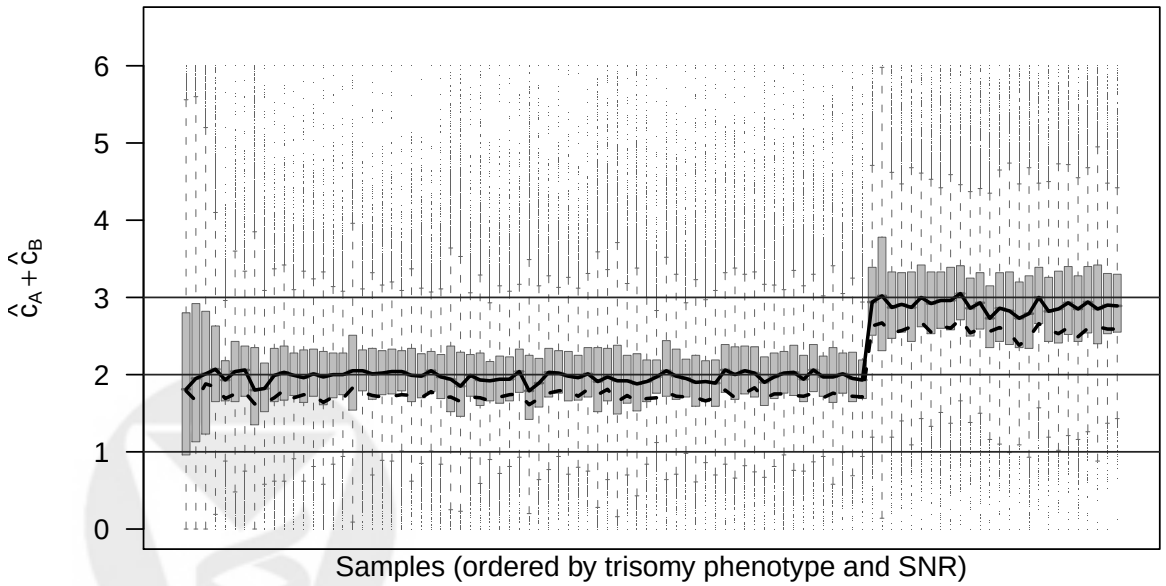


Figure 5: SNP-specific prediction regions for copy number 1, 2, and 3 before (dashed lines) and after (solid lines) bias adjustment for the trisomy study. For loci in which many individuals have a copy number alteration, the within-genotype estimates of location and scale for normal copy number are biased. The bias adjustment step involves recomputing the within-genotype centers and variance after removing samples that have a high-posterior probability of non-normal copy number. The set of samples that are removed when recomputing the location and scale is locus-specific and requires no *a priori* knowledge of common variants. For many SNPs on chromosome 21, the bias adjustment provides updated centers and variances for the ellipses that are slightly smaller than the original ellipses (a, b, c, d, e, f). Nevertheless, we slightly overestimate the A and B intensities for copy number 3. The overestimation occurs because the shift in the median intensities for copy numbers greater than 3 becomes increasingly nonlinear. (d) Finally, many SNPs simply exhibit a poor dose response in the A and B allele intensities with increasing copy number.



(a) Before bias correction.



(b) After bias correction.

Figure 6: Top: boxplots of the copy number estimates for the polymorphic probes on chromosome 21. 26 of the subjects have a chromosome 21 trisomy. Bottom: boxplots of the copy number estimates after performing a bias correction for common CNV. As described in Section 4, the bias correction does not use any phenotypic information of the samples nor does it require *a priori* specification of regions that are thought to harbor common copy number variants.

		$\%\widehat{CN} = 1$	$\%\widehat{CN} = 2$	$\%\widehat{CN} = 3$
copy number 2	Birdseye/Canary	0.0042	0.9914	0.0043
	CRLMM - HMM	0.0028	0.9926	0.0047
copy number 3	Birdseye/Canary	0.0006	0.0816	0.9177
	CRLMM - HMM	0.0003	0.0496	0.9501

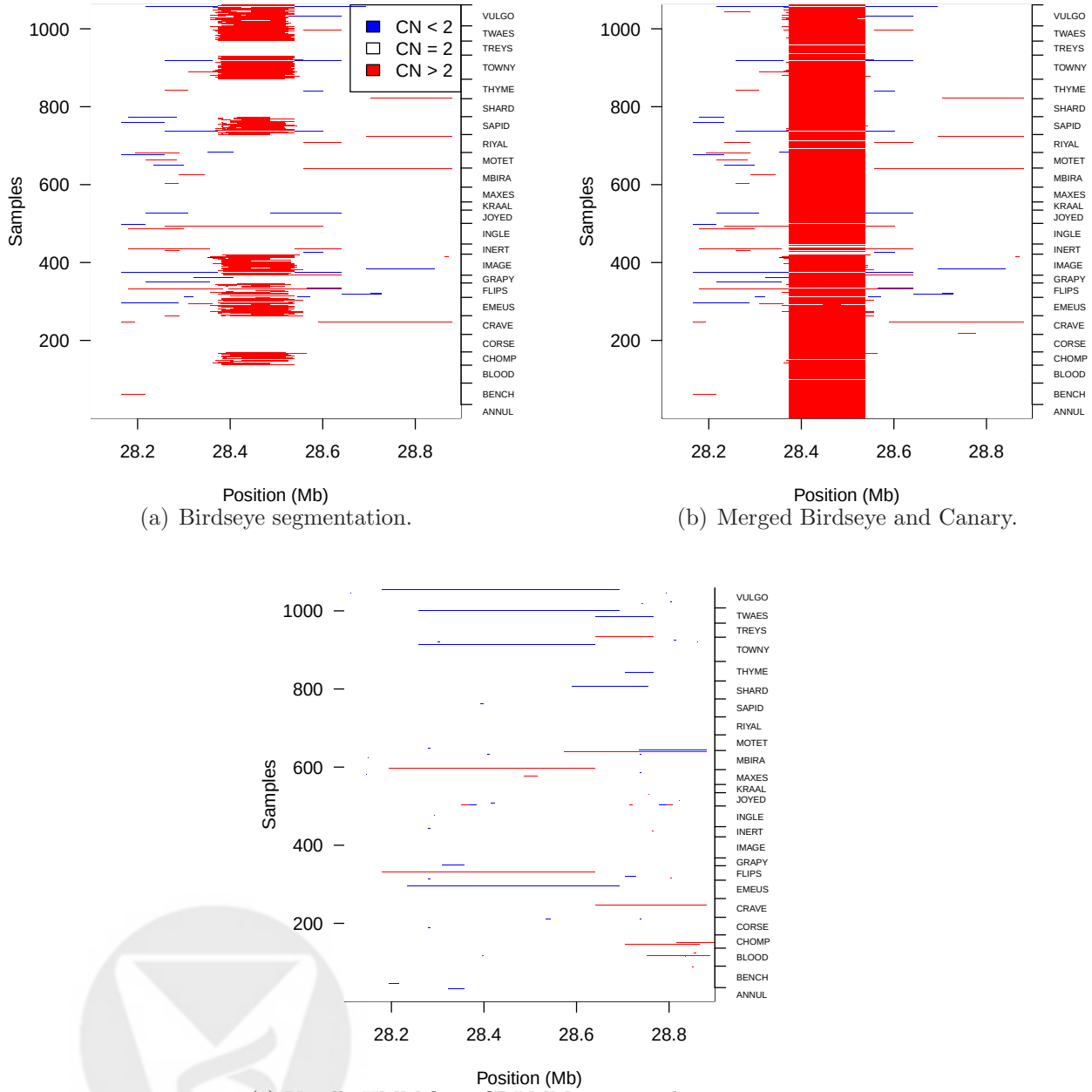
Table 1: The *true* copy number for loci on chromosome 21 is assumed to be 2 for the 70 normal samples and 3 for the 26 trisomy samples. In order to compare our method to the default software that does not provide locus-level estimates of copy number, we fit a hidden Markov model to the locus-level estimates using the same transition probabilities as used by the Birdseye HMM. Our approach decreases the (1-%) sensitivity by approximately 40% (95.01% versus 91.77%).

Supplemental Materials

Software:

- R version 2.10.0
- R packages: Biobase 2.5.3, crlmm 1.3.6, ellipse 0.3-5, genomewidesnp6Crlmm 1.0.4, RColorBrewer 1.0-2, xtable 1.5-5, affyio 1.13.3, annotate 1.23.0, AnnotationDbi 1.7.0, Biostrings 2.13.10, DBI 0.2-4, genefilter 1.25.2, IRanges 1.3.26, mvtnorm 0.9-7, oligoClasses 1.7.4, preprocessCore 1.7.4, RSQLite 0.7-1, splines 2.10.0, survival 2.35-4
- Birdsuite 1.5.3

Supplemental Figures



(a) Birdseye segmentation.

(b) Merged Birdseye and Canary.

(c) Vanilla HMM fit to CRLMM copy number estimates.

Figure 1: We observed plate-effects in both the Birdseye HMM predictions (a) and the merged canary predictions (b). The F-statistic in the 28.4 Mb region is genome-wide significant for both the Birdseye and Canary algorithms (F-statistic > 13, p-value $< 1.0^{-8}$). (c) An image of HMM predictions from the CRLLM copy number estimates using the same transition probabilities as in the Birdseye HMM (F-statistic $= 0.86$, P-value $= 0.66$).