Published in final edited form as:

Genet Epidemiol. 2011 September ; 35(6): 536–548. doi:10.1002/gepi.20604.

Bayesian Hierarchical Mixture Modelling to Assign Copy Number from a targeted CNV array

Niall Cardin^{1,3}, Chris Holmes^{1,2}, The Wellcome Trust Case Control Consortium, Peter Donnelly^{2,1}, and Jonathan Marchini^{1,2}

¹Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX13TG, UK

²The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX37BN, UK

³University of California, San Francisco, Diller Bldg, Room 262, 1450 3rd Street, San Francisco, CA 94158, USA

Abstract

Accurate assignment of copy number at known copy number variant (CNV) loci is important for both increasing understanding of the structural evolution of genomes as well as for carrying out association studies of copy number with disease. As with calling SNP genotypes, the task can be framed as a clustering problem but for a number of reasons assigning copy number is much more challenging. CNV-assays have lower signal to noise ratios than SNP assays, often display heavy tailed and asymmetric intensity distributions, contain outlying observations and may exhibit systematic technical differences among different cohorts. In addition, the number of copy-number classes at a CNV in the population may be unknown a priori. Due to these complications automatic and robust assignment of copy number from array data remains a challenging problem. We have developed a copy number assignment algorithm, CNVCALL, for a targeted CNV array, such as that used by the Wellcome Trust Case Control Consortium's recent CNV association study. We use a Bayesian hierarchical mixture model that robustly identifies both the number of different copy number classes at a specific locus as well as relative copy number for each individual in the sample. This approach is fully automated which is a critical requirement when analysing large numbers of CNVs. We illustrate the methods performance using real data from the WTCCC's CNV association study and using simulated data.

Introduction

It has become clear that copy number variants (CNVs) are a common and important part of human genetic variation. Early surveys discovered substantial variation between individuals [1, 2] with later studies confirming and further detailing the extent of copy number variation (CNV) in humans [3, 4]. More recently increasingly high resolution experiments directly assessing copy number and covering the whole genome have been performed, [5, 6]. As well as representing a substantial portion of human variation, it is also becoming clear that CNVs play a role in human disease. More than 25 segmental duplications are now thought to mediate a range of conditions, reviewed in [7], and CNVs have been implicated in susceptibility to some common diseases [8, 9, 10, 11, 12].

Whilst analysis of SNP data is now mature in human populations the analysis of CNV data is still in an early stage and ascertaining accurate copy number remains a significant challenge. CNVs preferentially occur in repetitive regions of the genome [4] and this can cause difficulty in designing good assays. Although catalogues of variation are beginning to emerge [5, 6] these databases are considerably less developed and comprehensive than SNP variation databases such as dbSNP [13], and databases of population variation such as the

Hapmap [14, 15]. Also, current array based technology has a relatively poor signal to noise ratio [16]. This, and the unexplained complex patterns in intensity distributions frequently observed, make assigning copy number robustly from array data a very challenging task [17].

The aim of this paper is to describe a method of providing robust assignment of copy number from array data in large samples at locations where CNVs have previously been identified. Currently such studies are most commonly performed using hybridisation arrays with multiple oligonucleotide probes for each region of interest. Each probe generates a measure of the amount of matching DNA present in the sample. With some CNV specific technologies, for example the Agilent or Roche Nimblegen arrays, a test and reference sample are assayed on the same chip and a summary of their measurements is taken to produce a *relative* copy number measurement. The log of the ratio of the two measurements is often used, though other possibilities exist and can have substantial advantages [17]. Probes on SNP chips can also be used to produce a univariate measure of copy number, though at SNP probes it may be preferable to also make use of the allelic information.

We now give an informal description of the CNVCALL approach. Full details are given in the methods section below. In order to give an outline of the method, we now assume that the data has been pre-processed leaving a univariate measure for each sample in the study at a specific locus. Example processed data from 6 CNVs from the WTCCC study are shown in Figure 1. We model the data as a mixture of Student t-distributions, each of which represent the noise in the measurements at individuals with a specific copy number, [18], and use Bayesian methods to select the number of components in the mixture. We estimate the degrees of freedom of the t-distributions from the data. We add in an explicit outlier class [19] to model outliers and this increases robustness and helps protect against overestimation of the number of mixture components that can occur if outliers are not accounted for. Where samples are collected from different sources, or use different cell types, there may be systematic differences between the collections and in an association study it is critical to account for these differences. We do this by placing a hierarchical structure on the parameters of the t-distributions, so that measurements in the same CNV class in different sample collections will have similar but not identical distributions. Many CNVs also show clusters of measurements displaying substantial skew and other patterns that are hard to model and this can artificially increase the number of classes fitted to the data. To protect against this effect we also merge adjacent clusters of individuals into a single class when the resulting distribution is unimodal.

It is worth drawing a distinction between two related tasks in assignment of copy number. The first is the task of *detecting* CNVs in an individual from whole genome chip data. The second involves *assigning genotypes* at known CNV loci in many individuals from a targeted array, such as that used in [17, 20]. Many methods have been proposed for CNV detection ([21, 22, 23, 24, 25, 26]). These methods typically use per-sample analysis and, typically HMM based, segmentation algorithms. Detection methods require high density (usually whole genome) chips with SNP and/or copy number probes; as such they cannot be used in the setting of a targeted CNV probe array, such as the Agilent array used in the WTCCC CNV study. These tasks are clearly related but assignment of genotypes at targeted CNVs presents distinct challenges and requires a different framework for success. In a targeted array the number of probes in a region may be small, and due to the lack of local genetic background information it is hard to choose appropriate intensity signals for each copy number, as this can vary by region. Rather than comparing probe intensities with a surrounding region the signal arises as a comparison between the individuals in the study.

The program CNVtools [27] can be used to assign copy number from a targeted CNV probe array. Both CNVCALL and CNVtools were employed to provide copy number assignments for the Wellcome Trust Case Control Consortium (hereafter WTCCC) genome wide CNV association experiment [17]. Of these two methods, in practice, only CNVCALL provides the ability for fully automated copy number calling in real data. Extensive manual intervention is required to effectively use CNVtools on a large number of CNVs. Firstly, CNVtools has multiple modes in which it can be run:

1. Fixed means and variances, where the means and variances in each collection must be the same
2. Different means, fixed variances, where the means are free to vary between collections, but the variances in each collection must be the same
3. Different means and variances, where the means and variances in each collection are free to vary between collections.

Unfortunately there is no single approach from these that performs well for every CNV, and the appropriate setting must be chosen manually after visual inspection of an initial run. Similarly CNVtools is not designed to automatically identify the number of classes in the population, and although a routine using BIC has been introduced, the number of classes must still be chosen manually at many CNVs. In the WTCCC CNV study extensive manual curation of 4,000 CNVs was required in order to fix the number of copy number class by eye [17]. Mixture models typically give highly multi-modal likelihoods, as a result the E-M approach of CNVtools often had to be set to a user defined starting point to avoid fitting to a local mode which gave a poor fit. As a result, this is a very labor intensive approach, which could need be performed multiple times if data changes or is processed differently.

With the large variety of preprocessing steps that can be applied to the data before calling can be performed, and with the best combination varying by CNV, a large manual effort is required to choose which preprocessing procedure to apply for each CNV. Using CNVCALL, as model fitting is fully automated, it is possible to automatically fit the model under a large number of possible preprocessing procedures and use metrics of model fits to choose the most promising. This was done in the WTCCC CNV study and greatly increased the number of CNVs at which high quality calls could be obtained (as described in [17]).

Results

We present two sets of results from using the algorithm. First we present results from the WTCCC CNV association study [17], then we give results from simulated data where we know the true 'copy number'. We also highlight some of the difficulties in calling and how these are overcome where possible.

WTCCC CNV association Study

Each sample was analysed separately. DNA from the sample was fluorescently labelled and mixed with DNA from a reference pool of individuals with an alternative label, measurements were then taken as described elsewhere [17]. The ratio of these two measurements then gives a measure of the relative amount of DNA in each individual and this is the raw data that is compared across samples. Substantial systematic technical differences between individuals can be seen in the raw data and so pre-processing is necessary before the data can be analysed. Example processed data from 6 CNVs from the WTCCC study are shown in Figure 1.

In summary: a total of 3, 432 CNVs were called as polymorphic and passed the WTCCC quality control filters (QC). At these CNVs, concordance between calls of duplicate samples

was 99.7%. A further 894 (4,326 total) CNVs showed multiple modes as identified by the CNVCALL algorithm, though the signal was too poor to provide calls appropriate for association testing. Of the 3,339 CNVs that passed QC on the autosomes, 1,321 showed two distinct classes (presumably with an unobserved rare third class), 1,697 displayed 3 distinct classes, 264 had 4 clusters and 57 had 5 or more clusters.

The WTCCC data displayed a wide variety of patterns in intensity distributions (see e.g. Figure 1 of the supplementary materials). Our method has several features that work together to allow robust and automatic model fitting in the presence of such variation. In the remainder of this section, we briefly describe properties of the WTCCC data, with examples, which we use to highlight the utility of these features.

Distributions with heavy tails were observed. To improve robustness over Gaussian mixture modeling we allowed mixtures of **t-distributions**, and allow CNVCALL to choose the number of degrees of freedom to use. An example of data with heavy tails, and how the use of t-distributions can help is given in Figure 4, and the model fit appears to give a much better fit when the algorithm is not restricted to using a mixture of normal distributions

It was also common to observe outliers. Some can be seen in Figure 1 of the supplementary materials. To address this problem CNVCALL uses an **outlier class**, to which individuals can be assigned, and which is described in the Methods section. When using CNVCALL on the autosomal CNVs: of those called as polymorphic after merging (see below), the median number of measurements assigned to this outlier class was 22 out of 17,304. Such outliers could, in principle, affect model fits and model choice if an outlier class were not implemented.

At many CNVs, it was observed that systematic differences among different collections existed. The variances of components often differed among collections (data not shown) and, perhaps more critically, the cluster means can also differ, as shown in Figure 3. Calling all the individuals without allowing for these differences can cause false associations, as discussed in [28] and [27]. The use of a **hierarchical model** allows us to be robust to such effects whilst pooling information across collections.

There was great diversity in the observed technical artifacts in the WTCCC data, as is typical for quantitative array based assays. This, combined with the very large number of samples present in the WTCCC study, meant that automatically identifying the number of classes was especially challenging. Likelihood based approaches will often identify more classes in order to fit the complex shapes not well modelled by t-distributions. To overcome this, we **merge** adjacent classes that, once merged, give rise to a unimodal distribution (as described in the methods).

Merging was applied to all of the CNVs in the WTCCC study, and the number of classes was reduced in 5,741 of 10,614 (54%) CNVs. The vast majority of these (4,718, of which 2,890 were 2-class) were at CNVs where before merging the CNV had multiple classes but after merging only one class remained. Figure 5 gives an example where a difficult to model distribution on the left hand side is initially called as two classes but is then merged such that the subsequent calls reflect the more plausible 3 class explanation at this locus.

The power to detect a CNV, or specific copy number class, in a sample will be affected by the frequency of that class in the sample. Pooling the information across collections should help to increase power to detect new classes, however the null class will also absorb individuals in very rare classes, and particularly with dispersed classes it is possible that they may be merged into nearby classes. It is very difficult to quantify the degree to which this happens: firstly the signal to noise ratio has a strong effect on power to detect a rare class (as

noted in the simulation results), but secondly true classes are unknown in real data and it is unclear how to recreate the patterns that give rise to such dubious classes in simulated data. In the WTCCC data, over 200 three class CNVs had a minor allele frequency under 0.5%, and over 1,000 were under 1%. However, quality control at rare classes is difficult, and with the CNVs originally identified using only 20 European samples we might not expect to see very many rare CNVs.

It is worth noting that with a targeted array the physical size of a CNV does not have a direct effect on our ability to assign copy number, in contrast to the setting of detecting previously unknown CNVs. In the WTCCC study the aim was to place at least 10 probes in most CNVs. Although it can be harder to find enough high quality probes in a smaller CNV, there are many important determinants of probe quality other than CNV length. The ability of CNVCALL to assign accurate copy number will depend only on the quality of the summary signal eventually produced from the probes (see Methods, and discussion in [17]).

Simulated datasets

We undertook a simulation study to assess the performance of CNVCALL in various situations where we knew the correct answer. The WTCCC study shows that real data contain a variety of complex features and artifacts and we have tried to accommodate some of these effects into the simulations. We simulated 3 sets of 600 CNVs by varying the signal-to-noise ratio. These 3 sets were denoted easy, medium and difficult. Data were simulated at the probe level and then pre-processed using the same techniques used by the WTCCC study to mimic the real data scenario as much as possible. A full description of the simulation procedure is given in the Methods section. Figure 2 gives examples of the resulting CNV intensity distributions for a CNV from each of the easy, medium and difficult simulation sets. Supplementary Figure 1 gives some additional intensity plots at CNVs from the WTCCC experiment for comparison. The results of our approach applied to the simulated datasets are summarised in Table 1.

Easy Data Set—Examination of the results shows that the quality of inference is high for the data set termed “easy”. Where the number of classes present is incorrectly inferred it is nearly always underestimated. This is because some CNV classes have very few observations, these then become absorbed into the tails of other classes or the outlier class. Of the 200 CNVs simulated with 2 classes, only one was classified as having 1 class. In this example the frequency of the rarer class was 1.5% and the data were somewhat dispersed. This meant that although the model selection procedure identified two classes, these were merged as the small and wide second class gives rise to a unimodal distribution when combined with the adjacent larger class (see Supplementary Figure 2).

With more classes, inference is more challenging, 97% of the 3 class CNVs were identified as having 3 classes, the mean frequency of the rarest class at those called with fewer classes was 0.26%. Supplementary Figure 3 shows the data and model fit at the CNV with the largest of these rarest classes (0.62%), the third class can just be seen at the right of the plot. In such cases it is hard to know whether such a small collection of points simply reflects chance outliers or interesting variation. Similarly, at the data simulated with four classes, 91% of CNVs were assigned 4 classes. The mean frequency of the missing classes was 0.29%. For comparison, the mean frequency of the rarest class across all of the four class CNVs was 5.8%.

In this simulated data a higher proportion of simulated CNVs are assigned the correct number of classes when merging is not employed, however 1 CNV from both the 3 class and the 4 class data sets was called with too many classes. Plots of these CNVs can be seen in Supplementary Figures 4 and 5. The result of such over-estimation of the number of classes

could well be to produce spurious signals of association, as the data-quality issues that give rise to the effect may well differ between collections. Also, we have observed that in real data, a high proportion of CNVs displayed patterns that caused extra clusters that seemed unlikely to represent biological variation. Finally, the amount of power lost at loci where one very rare class is merged with another class will often be small.

Medium Data Set—In the medium data set the signal to noise ratio is reduced and further complications are introduced. This means that, as well as very rare classes not being identified, pairs of classes with very large signal overlap are often called as one class. These two effects work in tandem, so that rare overlapping classes are less likely to be identified than common classes of the same overlap.

Of the 200 CNVs simulated with 2 classes, 97.5% were correctly identified as such. At four of the CNVs where the number of classes was estimated as 1, the overlap between adjacent clusters was so large that these pairs of clusters were merged. At the fifth, the frequency was 0.05%; such rare classes are difficult to distinguish from noise in any scenario. As with the easy data, the data are more challenging with more classes and with three classes, 87.5% were assigned 3 classes. The majority of the 25 CNVs not assigned 3 classes were called two class after merging was applied. The 25 corresponding cluster plots were inspected, and in each case the number of classes chosen by the algorithm would also be likely choices were manual curation used instead (see Figure 7 of the supplementary material for an example). With four classes, the situation is more challenging again. Here many of the CNVs show substantially overlapping clusters whilst a smaller number simply have very rare classes. It is not easy to give a precise breakdown as the two effects are confounded. In general, where the information in the data is clear, the correct number of classes is chosen, otherwise a reduced number of classes is usually chosen (see Supplementary Figure 6).

Difficult Data Set—Many of the CNVs in this simulation set exhibit substantial overlap between CNV classes. In the WTCCC study, CNVs with this level of overlap were filtered out using quality control filters in order to avoid false positives, so the results on this simulation set should be interpreted with this in mind. Unsurprisingly, performance is substantially worse than on the easy and medium datasets. The reasons why various CNVs have a lower number of clusters inferred are just as with the other data sets, however with this data, overlapping clusters account for almost all of cases. Note that the merging procedure creates a substantial downward bias to the number of classes called. We believe that this is appropriate when performing association studies: merging clusters has the effect of reducing the resolution of the data and reducing power, however when too many classes are identified the effect is likely to produce false positives. In other contexts, users may prefer to remove the merging step and this is straightforward using the software.

Discussion

There is currently much interest in copy number variation in natural populations, in particular, there has been significant interest in trying to identify copy number loci which are associated with human disease. Unfortunately it is both very challenging and still expensive to survey copy number variation using next generation sequencing data. It is therefore important to design methods which can assign robust copy numbers from the cheaper and more mature array technologies.

To our knowledge this is our approach first method that can provide high quality data without intensive manual curation: the use of CNVtools on the WTCCC data required manual curation for over 4,000 CNVs (see supplementary online material of [17]). Our fully automated algorithm was used to analyse the data from many pre-processing different

pipelines, then metrics of quality such as cluster separation were used to identify the best pre-processing procedure for each CNV. This approach led to a substantial increase in the number of CNVs for which the WTCCC study was able to get good copy number assignments [17]. Though it is worth noting that although assignment of copy number is automatic, there exist a variety of artefacts and complex situations which could cause misinterpretation of association results. These are discussed in detail in [17].

We have taken the decision to work within the Bayesian paradigm. This has a number of advantages. Firstly we can add robustness to the calling model by placing priors on various model parameters that match previously observed patterns when copy number variation is present. Secondly, Bayesian statistics provides a very natural framework to choose between models with different numbers of parameters, even without the requirement that such models be nested. This is very important when analysing CNVs as the number of copy number classes in the population is often unknown in advance. As the WTCCC CNV study has data from many different collections it is also a very natural context to employ a Bayesian hierarchical mixture model. The Bayesian approach also leads to more interpretable results.

The model is robust enough to cope with many of the difficult features in array data, such as non-normality of the component distributions, outliers and systematic differences between collections. In addition, we have employed techniques to deal with intrinsically more difficult problems such as skewed distributions. This algorithm has been used to generate copy number assignments and test for associations in the Wellcome Trust Case Control Consortium plus CNV study [17].

Methods

Producing a one-dimensional summary—It can be useful to use multiple probes to interrogate each CNV in order to increase the signal to noise ratio. There is then a question of how best to combine information across probes, an example of the data from multiple probes is given in the supplementary materials of [17], Figure 10 (a). CNVCALL can be used to transform this probe level data into the one-dimensional signal used for calling. In different situations, different summaries will work best, although in the WTCCC study it was found that principal component analysis (PCA hereafter) usually provides cleaner signal than a simple average, [17]. This approach finds the linear combination of probes that maximise the variance of the projected data.

A technique related to linear discriminant analysis (LDA) [27] has also been implemented. This method works by making a first round of calling, with an initial probe weighting from, for example, PCA. These calls are then used to inform a new choice of probe weightings which provide an often better one-dimensional summary for the next round of calling. The choice of probe summary can greatly affect the quality of results, for a fuller discussion of this see [17].

We found that the absolute range of probe measurements was not strongly informative about the number of classes. At some CNVs, measurements were very similar in absolute terms yet there was high signal to noise and clear polymorphism present, whilst at others there was substantial variation between individuals but this appeared to be noise and the histograms showed no evidence for polymorphism. Because of this, it proved easier to design a prior structure when the data was scaled to a fixed range. We centred the pre-processed one-dimensional signal on zero, and scaled the central 99.8% (to account for outliers) of this scaled signal to the interval $[-0.05, 0.05]$.

Bayesian Hierarchical Mixture Model

From this point we will assume that data pre-processing has produced a single one-dimensional summary of the probe information for each individual. Consider this summary at a single CNV. We use X_{ij} to denote the 1-dimensional summary of the intensity data for the j^{th} individual in the i^{th} collection. The i^{th} collection contains N_i individuals and there are C collections. We assume that each individual has an integer number of copies of the DNA in this region, and use Z_{ij} to denote the *relative* copy number of the j^{th} individual in collection i compared to the reference samples. We arbitrarily set the minimum possible copy number to zero, and include an explicit *NULL* class to capture outlying individual measurements. This gives $Z_{ij} \in \{0, \dots, K-1, \text{NULL}\}$, where K denotes the (unknown) number of different CNV classes in the sample.

The posterior distribution is given by:

$$P(\mathbf{Z}, \mu, \sigma, \lambda, \alpha, |\mathbf{X}|) \propto P(\mathbf{X}|\mathbf{Z}, \mu, \sigma)P(\mathbf{Z}|\lambda)P(\lambda|\alpha)P(\alpha)P(\mu)P(\sigma)$$

where μ and σ^2 are the vectors of means and variances of the classes respectively. The vector, λ , denotes the proportion of individuals in each class (the class proportions), and α is parameter controlling the prior probability of different sets of class proportions.

We model the one-dimensional summary, X_{ij} , conditional upon the relative copy number, Z_{ij} , using a t -distribution as

$$X_{ij}|Z_{ij} \sim t(\mu_{iZ_{ij}}, \sigma_{iZ_{ij}}^2, df)$$

where the parameters $\mu_{iZ_{ij}}$ and $\sigma_{iZ_{ij}}^2$ denote the mean and variance of the intensity distribution for the sample with relative copy number Z_{ij} . The parameter df denotes the degrees of freedom of the t -distribution.

Given the cluster means, variances and copy number, individuals are independent so that

$$p(\mathbf{X}|\mathbf{Z}, \mu, \sigma) = \prod_{i=1}^C \prod_{j=1}^{N_i} t(\mu_{iZ_{ij}}, \sigma_{iZ_{ij}}^2, df). \quad (1)$$

Cluster Mean Prior Structure—For each collection, i , the cluster means for copy number Z , μ_{iZ} , are parameterised in terms of the position of the lowest class, $\mu_{i,0}$ and the distance between the $(Z-1)$ th and Z th classes, γ_{iZ} , so that

$$\mu_{iZ} = \mu_{i,(Z-1)} + \gamma_{iZ}.$$

Our prior on the mean of the zero cluster is $\mu_{i,0} \sim N(\mu_0, \psi^2)$ where

$$\mu_0 \sim N(-0.02, 0.005^2)$$

and

$$\psi^2 \sim \text{Inverse Gamma}(\text{scale}=1/200, \text{shape}=3).$$

This gives a mean of 0.00125 and a variance of 0.0025².

Our prior on the change in cluster mean for different copy numbers is: $\gamma_{iZ} \sim \Gamma(\text{mean} = \gamma_{\text{mean}}, \text{var} = \gamma_{\text{var}})$ where

$$\gamma_{\text{mean}} \sim \Gamma(\text{scale}=1/5000, \text{shape}=100).$$

which gives a mean of 0.02 and a variance of 0.002². Also

$$\gamma_{\text{var}} \sim \Gamma(\text{scale}=1/8000, \text{shape}=16).$$

which gives a mean of 0.002 and a variance of 0.0005².

The various hierarchical components are linked together to model the centres of the signal for different copy numbers:

$$P(\mu) = \prod_{i=1}^C P(\mu_{i,0} | \mu_0, \psi^2) P(\mu_0) P(\psi^2) \times \prod_{Z=1}^K P(\gamma_{iZ} | \gamma_{\text{mean}}, \gamma_{\text{var}}) P(\gamma_{\text{mean}}) P(\gamma_{\text{var}}) \quad (2)$$

Cluster Variance Prior Structure—For each collection, i , we model log of the cluster variances in the following way:

$$\log(\sigma_{iZ}^2) = \log(\sigma_{i(Z-1)}^2) + \varphi_{iZ}.$$

Our prior on the log variance in the zero cluster is that $\log(\sigma_{i0}^2) \sim N(\eta_0, r^2)$, and we set

$$\eta_0 \sim N(-10, 0.05^2)$$

and

$$r^2 \sim \text{Inverse Gamma}(\text{scale}=0.02, \text{shape}=2).$$

This gives a mean of 0.02 and a variance of 0.07². Our prior on the change in variance for the other copy numbers is $\varphi_{iZ} \sim N(\varphi_{\text{mean}}, \varphi_{\text{var}})$ where

$$\varphi_{\text{mean}} \sim \Gamma(\text{scale}=0.0004, \text{shape}=44.4).$$

This gives a mean of 0.02 and a variance of 0.003².

$$\varphi_{\text{var}} \sim \Gamma(\text{scale}=0.0004, \text{shape}=44.4).$$

This gives a mean of 0.02 and a variance of 0.003^2 .

The various hierarchical components are linked together to model the variance of the signal in each of the different copy number classes

$$P(\sigma^2) = \prod_{i=1}^C P(\eta_{i,0} | \eta_0, r^2) P(\eta_0) P(r^2) \times \prod_{Z=1}^K P(\varphi_{iZ} | \varphi_{\text{mean}}, \varphi_{\text{var}}) P(\varphi_{\text{mean}}) P(\varphi_{\text{var}}) \quad (3)$$

Class Proportions Prior Structure—The copy number assigned to an individual depends on the relative frequencies of these in the population and the frequency of the null class. Denote the vector of these assignment proportions by $\lambda = \{\lambda_0, \dots, \lambda_{K-1}, \lambda_{\text{NULL}}\}$. Then

$$Z_i | \lambda \sim \text{Multinomial}(1, \lambda) \quad (4)$$

Consider $\lambda' = \{\lambda_0, \dots, \lambda_{K-1}\}$, the proportions of non-null classes. We model this by

$$\lambda' \sim \text{Dirichlet}(\mathbf{w}\theta) \quad (5)$$

but then scaled so that $\sum_{l=0}^{K-1} \lambda_l = 1 - \lambda_{\text{NULL}}$. The vector $\mathbf{w}$ is reparametrised by $\mathbf{w}_i = \binom{K-1}{i} \alpha^i (1-\alpha)^{(K-1)-i}$ to reflect Hardy-Weinberg Equilibrium (HWE). This makes sense in the 2 allele case as the CNV can effectively be modelled like a SNP. But in the presence of multiple alleles this is not fully general. Although some departures from this model were observed in the WTCCC study these appeared to reflect two separate variants interrogated by the same different probes. Also the program seems to be robust to such deviations. The parameter θ controls the strength of the prior, and therefore conformity to HWE. We set $\alpha \sim \beta(13.3, 13.3)$ and $\theta = 30$.

Implementation

To increase computational efficiency, we break the range of the data up into $\sim 150 - 200$ regions and replace individual measurements by the centre of the region, or grid value, that their measurement lies within. This means that the algorithm is capable of analysing very large numbers of individuals with no loss of computational efficiency.

Ideally we would provide copy number calls having integrated out the parameters in the model. Based on experience of working with a similar model for SNP calling [29] we take the computationally attractive approach of finding the values of the parameters which maximise their posterior probability (the MAP estimate), and produce probabilistic calls conditional on these. For a fixed number of classes the basic approach is to initialise the model with a variety of starting values for the parameters, and then propose updates to these parameters; updates which decrease the posterior are never accepted, otherwise they are accepted.

As well as multiple start points reflecting a range of plausible values for the parameters, we also use two more focussed initialisations. Consider the number of individuals assigned to each grid value: when the data is of good quality at a polymorphic location this plot will have multiple modes. Visually, we place a horizontal line on a plot of these values at a height x . We set x so that given K classes we find K distinct peaks which lie above x . Multiple values of x may achieve this so we use the widest vertical interval, A , such that for all y in A exactly K contiguous regions lie above y . Once these peaks have been identified the class means are initialised as the centre of these regions and the variances and class proportions are generated conditional on these means. Secondly, the model is fitted to a large number of classes first (perhaps 8–10) and then these model fits are used to initialise models with smaller numbers of classes. Each class is deleted in turn, and the class proportions distributed amongst its neighbours (or neighbour). The class means and variances of the remaining classes remain as before as this seemed to perform well. This generates l distinct start points as initialisations for fits with $l - 1$ components. The total number of start points used, for each number of classes, in the analyses for the WTCCC study was 13.

Two basic approaches are taken to update parameters. The first is to take single or multiple parameters and suggest small perturbations to their values. A balance needs to be found between proposing updates large enough to allow the model fit to progress towards better values, but also small enough that these updates are likely to be accepted. The magnitude of the values proposed is somewhat reduced in later iterations as more subtle refinement is likely to be needed at that stage. The second approach, implemented for the cluster means, variances and class proportions (also for disease parameter when the model with association is employed) is to use probabilistic calls using the current model fit, then to propose updates to the parameters of interest. For example, when updating the class means we can propose new class means for each cluster to be the mean of the calls under the current model fit, for this class, weighted by the conditional posterior probability of assignment to each cluster. For each start point a fixed number of iterations is performed leading to a set of model fits from which we choose the fit with the highest posterior probability of the parameters.

To explore the different possible distributions for the data given membership of a specific class, we first perform the procedure above using a single distribution (the default is a t -distribution with 8 degrees of freedom). Then the best fit for each number of classes is used as a seed for model fits under a range of other distributions (currently the normal distribution and t -distributions with arbitrary degrees of freedom are implemented, though the user must specify which of these is to be explored). The marginal likelihoods for each of these model fits is then approximated using BIC and the fit with the highest BIC score is chosen.

To increase computational efficiency the first few iterations are performed on all collections combined, this allows the model to settle in a broadly correct region for most parameters. Also, when the number of classes and underlying distribution have been chosen a much greater number of iterations are performed to refine the model fit. To refine the most important parameters at the end of the model fit the final proposed updates always consist of two iterations of updating the means, variances and class proportions conditional on the calls under the current model fit as described above, as this provides fast convergence when the model fit is roughly correct.

Choosing the number of components—To choose between models with different numbers of classes, we would ideally like to calculate the marginal likelihood (or model evidence) for each model. We use a quantity related to the Bayesian Information Criterion [30] which can be derived as an approximation to the marginal likelihood. This approximation requires calculation of the maximum a posteriori (MAP) estimate of the

parameters. If there is just one collection, then the approximation to the marginal likelihood we use is:

$$M \sim L^* - P \times \log(N) \quad (6)$$

where M is the marginal likelihood, L^* is the likelihood evaluated at the MAP, P is the number of parameters in the model and N the number of observations. When there are multiple collections present, the hierarchical structure of the model means that the number of parameters in the model becomes ambiguous. There is no correct solution to how many parameters to use with hierarchical models. Using the full number of parameters in the model is an overestimate as some of these parameters are tightly related, or even fully dependent on other parameters. The approach taken here is to take each collection separately: a BIC like approximation to the marginal likelihood is calculated for each collection, and then these are averaged to get an overall estimate of the number of classes. The parameters are likely to be very similar between the collections and so the penalty for these parameters is divided amongst the collections, weighted by the proportion of individuals in that collection:

$$M \sim \sum_i L_i^* - w_i P \times \log(N_i) \quad (7)$$

where N_i is the number of individuals in the study in the i^{th} collection and $w_i = \frac{N_i}{n}$ is the fraction of all individuals in this collection. When comparing models with different numbers of classes in this way, it is only necessary to consider parameters that change with the number of classes. For a K component model fit we set $P = 3K - 1$ as: there are K parameters that describe the means of the distributions, K parameterse describing the variances, and $K - 1$ for the class proportions (as these sum to 1). We have found that this approach works well on both simulated and real data.

Merging Classes—When the data do not conform well to the model specified, this can adversely affect choice of the number of classes. This is often the case with CNV data and we apply a secondary processing step: we merge classes that, when merged, give a unimodal density. The method starts by taking the densities as fitted at the MAP estimate of the parameters for the best fitting model. Each pair of adjacent components of the mixture distribution is merged if the combined density of both components has a single mode. We found that the strategy of considering adjacent pairs from left to right works well. Carrying out this procedure one collection at a time could lead to different numbers of components in each collection so the procedure is applied to an averaged model fit in which means and variances are averaged across collections:

$$m_Z = \sum_{i=1}^C w_i m_Z^{(i)}$$

$$v_Z = \sum_{i=1}^C w_i v_Z^{(i)}$$

where m_Z is the average mean for copy number Z , v_Z the average variance, and $w_i = N_i/N$ is the proportion of individuals in collection i . Figure 8 of the supplementary materials gives a schematic of the merging procedure.

Note this taking an averaged model fit is not the same as fitting a model to the data from all collections pooled together, and should be more sensitive than that approach to multi-

modality. This is because pooling the data reduces cluster separation, as can be seen in Figure 3. Once the decision of whether or not to merge the classes is made, individual model fits for each collection are themselves merged as appropriate. This procedure was used to decide the number of classes in the WTCCC study.

Probabilistic copy number calls—The primary output of CNVCALL is a set of probabilistic diploid copy number assignments for each individual in the study. These are provided in a format very similar to the file format used by the programs CHIAMO [29] and IMPUTE [31], that allows for uncertainty in copy number. We add a new 4th column, indicating the number of copy number classes inferred. The probabilities are calculated using Bayes Rule applied at the MAP estimates of the cluster mean, variance and class proportion parameters. This gives

$$P(Z_{ij}=k|X_{ij}, \hat{\mu}_{iz_{ij}}, \hat{\sigma}_{iz_{ij}}, \hat{\lambda}_k) = \frac{P(X_{ij}|Z_{ij}=k, \hat{\mu}_{iz_{ij}}, \hat{\sigma}_{iz_{ij}}) \times \hat{\lambda}_k}{P(X_{ij}|\hat{\mu}_{iz_{ij}}, \hat{\sigma}_{iz_{ij}}, \hat{\lambda}_k)}. \quad (8)$$

It is also possible to get hard calls at any specified threshold on the maximum posterior probability of assignment but we do not recommend that this method is used.

Association Testing—Although the primary focus of this paper is on assigning copy number this naturally leads to the question of how to test for association in this setting. The approach we take for testing the probabilistic copy number assignments for association is effectively the same as that used to test imputed SNP genotype data for association. For a binary phenotype, with Φ_{ij} denoting the case/control status of the j th individual in the i th collection, we use a logistic regression model of the form

$$P(\Phi|Z, \theta) = \prod_{i=1}^C \prod_{j=1}^{N_i} p_{ij}^{\Phi_{ij}} (1 - p_{ij})^{1-\Phi_{ij}} \quad (9)$$

where

$$\theta = (\mu, \psi, \gamma) \quad \log \frac{p_{ij}}{1 - p_{ij}} = \mu + \psi C_{ij} + \gamma Z_{ij} \quad p_{ij} = \frac{e^{\mu + \psi C_{ij} + \gamma Z_{ij}}}{1 + e^{\mu + \psi C_{ij} + \gamma Z_{ij}}}. \quad (10)$$

In this model, μ is the baseline log-odds of disease for copy number 0, C_{ij} is a covariate, ψ is a parameter that controls the size of the covariate effect, γ specifies the increase in log-odds due to each copy of the allele and p_{ij} is the probability that individual j of the i th collection develops the disease. This model is multiplicative on the odds scale and additive on the log-odds scale. In the analysis of the WTCCC CNV study the sex of each sample was used as a covariate.

Within the Bayesian framework, focus centres on calculation of the a Bayes factor (BF), which is a ratio of marginal likelihoods between a model of association (M_1) and a null model of no association (M_0),

$$BF = \frac{P(Data|M_1)}{P(Data|M_0)} \quad (11)$$

where the marginal likelihoods are defined by

$$P(Data|M_1) = \int \left(\prod_{i=1}^C \prod_{j=1}^{N_i} \sum_{k=0}^{K-1} P(\Phi|Z_{ij}=k, \theta) p_{ijk} \right) P(\theta|M_1) d\theta. \quad (12)$$

Equation 12 can be approximated using a Laplace approximation and a straightforward modification of the likelihood maximisation used by frequentist methods [31].

We also need to specify the prior distributions. In the case of model M_1 that means we need to specify $P(\theta_1 | M_1) = P(\mu, \psi, \gamma | M_1)$. We place diffuse prior distributions on the parameters μ and ψ which are not of primary interest. The main parameter of interest is γ . An appropriate prior will depend on the specific situation but, as is discussed in [31], we have some prior information in the case of GWAS. We do not expect to see very large effect sizes, with most odds-ratios in the range 1 – 2, and more weight near 1 – 1.3. One such prior that allows this is $\gamma \sim N(0, 0.2)$, which is used here.

This two-stage approach of calling and then testing is attractive due to its simplicity, and is now routinely used when carrying out SNP association scans involving imputed data. In the WTCCC study this allowed the calling of copy number once across collections and then testing of each collection separately based on this fixed set of calls. We investigated the approach where calling and association testing are performed jointly, and found that it gave very similar results to the two-stage procedure (data not shown).

Simulation Study

In simulating this data we aimed to capture the patterns observed in real data by implementing approximations to processes that seem likely causes of various artefacts in the data. As such, the simulated measurements are not generated using the model that we use to perform inference. We feel that we have captured the most important difficulties although the variety of patterns observed in real data is far greater, for examples, see Figures 1 and 2 from the main text, as well as Figure 1 here.

Simulated copy numbers, Z_{ij} , were generated using the following model

$$Z_{ij}|\lambda' \sim \text{Multinomial}(1, \lambda') \quad (13)$$

$$\lambda' \sim \text{Dirichlet}(\mathbf{w}\theta) \quad (14)$$

where $\mathbf{w}$ is reparametrised by $\mathbf{w}_i = \binom{k-1}{i} \alpha^i (1-\alpha)^{(K-1)-i}$, $\alpha \sim \beta(13.3, 13.3)$ and $\theta = 30$. We simulated 200 CNVs with 2, 3 or 4 copy number classes. At each CNV we simulated 2000 individuals in each of 3 collections. Intensity measurements at 10 probes per CNV were then simulated. In principle, this allows us to mimic the variation that we see between probes and to investigate sensitivity of results to different probe summaries. In practice the

noise structure in the real data of the WTCCC study was very diverse and more complex than that which we have implemented here. We denote I_{ijl} as the intensity measurement of the l th probe in the j th individual of the i th collection and used the model

$$I_{ijl} \sim N(s_l Z_{ij} + P_b^l + C_b^i, (n + P_v^l + C_v^i)^2) \quad (15)$$

where s_l is a parameter that controls the cluster mean separation at probe l , P_b^l is the bias in measurements due to probe level variation in probe l , C_b^i the bias due to the differences between the collections in collection i , n controls the level of background noise, P_v^l controls the probe specific variation at probe l and C_v^i the variation in this probe that comes from samples lying in collection i . One very noticeable feature of real data was large variation in the performance of individual probes. To mimic this effect we randomly chose 7 (bad) probes to have substantially less signal than the remaining 3 (good) probes.

We simulated 3 complete sets of 600 CNVs by varying the signal-to-noise ratio. These 3 sets were denoted easy, medium and difficult. Table 2 details the parameter settings and distributions used for s_l , P_b^l , C_b^i , n , P_v^l and C_v^i . Also, in the medium and difficult data sets, as well as these marginal increases in the effect of various biases, some quite specific effects were introduced. These were motivated by effects that are observed in real data and which can be difficult to accommodate.

- **Batch Effects:** 2 Collections were sampled at random, and in each collection a subset of 2.5% of those individuals' measurements were changed. The measurements in each individual were shifted by sampling from the distribution $\Gamma(\text{mean}_{\text{batch}}, \text{variance}_{\text{batch}})$ where the mean of this distribution was 0.03 of the range of the data and the standard is 0.07 of the range of the data.
- **Time Effects:** 2 periods of time were sampled at random, where it is assumed that individual i was measured at time period i . The periods are distributed uniformly on the total number of time periods and the length of the period is 5% of the total time. The measurements of these individuals are then shifted a draw from a $N(0, sd_{\text{time}}^2)$ distribution, where the sd_{time} is 1% of the range of the data.

Null Measurements: We then chose 0.4% of the individuals in each collection at random. The measurements of these individuals were then set to uniform across the range of the data.

Once simulated, the CNV probe intensities were processed using PCA and LDA to produce the one-dimensional measurement taken as input into our calling method. The method was then used to estimate the number of CNV classes at each CNV.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to all members of the WTCCC CNV analysis group for helpful discussions. Special thanks go to Damjan Vukcevic and Richard Pearson. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators in the consortium is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. This research was, in part supported by funding and grants from the Wellcome Trust and by NIH Grant 5U01GM79164-3. PD holds a Wolfson Royal

Society Merit Award. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. *Nature genetics*. 2004; 36(9):949–951. [PubMed: 15286789]
2. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al. Large-scale copy number polymorphism in the human genome. *Science*. 2004; 305(5683): 525. [PubMed: 15273396]
3. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. Fine-scale structural variation of the human genome. *Nature genetics*. 2005; 37(7):727–732. [PubMed: 15895083]
4. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, et al. Segmental duplications and copy-number variation in the human genome. *The American Journal of Human Genetics*. 2005; 77(1):78–88.
5. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. Global variation in copy number in the human genome. *Nature*. 2006; 444(7118):444. [PubMed: 17122850]
6. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, De Bakker PI, Maller JB, Kirby A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*. 2008; 40(10):1166–1174. [PubMed: 18776908]
7. Stankiewicz P, Lupski J. Genome architecture, rearrangements, and genomic disorders. *Trends Genet*. 2002; 18:74–82. [PubMed: 11818139]
8. Yu CE, Dawson G, Munson J, DSouza I, Osterling J, Estes A, Leutenegger AL, Flodman P, Smith M, Raskind WH, et al. Presence of large deletions in kindreds with autism. *The American Journal of Human Genetics*. 2002; 71(1):100–115.
9. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. Strong association of de novo copy number mutations with autism. *Science*. 2007; 316(5823):445. [PubMed: 17363630]
10. McCarroll SA, Huett A, Kuballa P, Chlewicki SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohns disease. *Nature genetics*. 2008; 40(9):1107. [PubMed: 19165925]
11. Stefansson H, Rujescu D, Cichon S, Pietiläinen OPH, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. Large recurrent microdeletions associated with schizophrenia. *Nature*. 2008; 455(7210):232. [PubMed: 18668039]
12. The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*. 2008; 455(7210):237–241. [PubMed: 18668038]
13. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001; 29(1):308–11. [PubMed: 11125122]
14. The International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005; 437:1299–1320. [PubMed: 16255080]
15. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449:851–861. [PubMed: 17943122]
16. Ylstra B, van den IJssel P, Carvalho B, Brakenhoff RH, Meijer GA. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic acids research*. 2006; 34(2):445. [PubMed: 16439806]
17. The Wellcome Trust Case Control Consortium. Genome-wide association study of copy number variation in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*. 2009
18. Peel D, McLachlan GJ. Robust mixture modelling using the t distribution. *Statistics and Computing*. 2000; 10(4):339–348.

19. Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, Ng R, Murphy KP. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*. 2006; 22(14):e431. [PubMed: 16873504]
20. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. Origins and functional impact of copy number variation in the human genome. *Nature*. 2009
21. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics*. 2008; 40(10):1253. [PubMed: 18776909]
22. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research*. 2007; 17(11):1665. [PubMed: 17921354]
23. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic acids research*. 2007
24. Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*. 2004
25. Huang J, Wei W, Chen J, Zhang J, Liu G, Di X, Mei R, Ishikawa S, Aburatani H, Jones KW, et al. CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC bioinformatics*. 2006; 7(1):83. [PubMed: 16504045]
26. Li C, Wong W. DNA-chip analyzer (dChip). *The Analysis of Gene Expression Data*. 2003:120–141.
27. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME. A robust statistical method for case-control association testing with copy number variation. *Nature genetics*. 2008; 40(10):1245–1252. [PubMed: 18776912]
28. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*. 2005; 37(11):1243–1246. [PubMed: 16228001]
29. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. Jun.2007 447:661–678. [PubMed: 17554300]
30. Schwarz G. Estimating the Dimension of a Model. *Ann Statist*. 1978; 6:461–464.
31. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature*. 2007:906–913.

Figure 1. Examples of 6 CNVs from the WTCCC study

Each plot shows a histogram of the processed one-dimensional signal on the x-axis and this is overlaid with lines showing the model fit. The panels show CNVs with 1,2,3,4,5 and 6 (putative) copy number classes respectively. The y-axis gives the number of individuals in each bin of the histogram. Each coloured line gives the distribution of a single component from the model fit where the area under the curve gives the relative class proportions. For each of these CNVs calling was performed using the hierarchical model in all 10 collections though for simplicity both here and elsewhere only the data from one or two collections is shown, the 1958 birth cohort is shown here. These data come from the best pipeline found for each of these CNVs, see [17] for details.

Figure 2. Data Quality

Top Row The panels show three example CNV intensity distributions from the different simulated data sets (focussing on CNVs with three classes). There exists variation in these data sets and plots of other illustrative examples follow, and can be found in the supplementary online materials. *Second row* here we see the calls made by the algorithm at these CNVs.

Figure 3. Importance of allowing differences between collections

Data at a single CNV in the 1958 birth cohort (top panel) and rheumatoid arthritis (middle panel) collections from the WTCCC study. The data for these two collections combined is given in the bottom panel. Each panel shows a histogram of intensities for different sample collections after data pre-processing. The model fit for the individual collections is overlaid. Assume that, from left to right, the classes represent 0, 1 or 2 copies of the probed DNA. We see that the separation of clusters is worse when the collections are combined. A more serious problem is that due to the shifts in cluster means there will be systematic differences between the collections if calls are made using this combined data: The yellow region gives values of the processed signal such that individuals from the 1958 birth cohort are more likely to have 1 copy whilst individuals in the rheumatoid arthritis collection are more likely to have 2 copies. It is clear that if calls were made on the combined data that an excess of 1958 birth cohort individuals would be assigned copy number 2 status whilst in rheumatoid arthritis an excess of copy number 1 calls would be made. This effect would lead to a substantial inflation of test statistics measuring the strength of association with phenotype.

Figure 4. Necessity of fitting distributions with heavy tails

Left: The model fit when an underlying mixture of normals was assumed, the BIC based procedure chose 4 classes here as it is not possible to get a good fit to the right hand cluster with a single normal distribution. *Right:* when the underlying distribution is chosen automatically, a t-distribution with 3 degrees of freedom was chosen, giving this more likely two class model fit.

Figure 5. Merging

These histograms show the distribution of one-dimensional processed data in the 1958 birth cohort at a single CNV. Left: The cluster of points at the left hand edge is inconsistent with any symmetric distribution and so has been split into two classes using the BIC procedure, as shown by the coloured lines. Right: The lines here show the model fit obtained when the merging procedure is applied to the model fit on the left. The first and second (from left to right) classes are merged to form one class that now quite accurately fits the shape of the distribution. This model fit seems more likely to represent the underlying copy number variation at this locus.

Inferring the underlying number of classes

Table summarising inference on the number of classes on the simulated data. The columns are split into 3 sections which correspond to the easy, medium and hard data sets. For each of these three the columns are further split by the number of classes simulated in each data set: 2, 3 or 4. The table is then further split into two sets of rows: **(a)** gives the results after merging as been applied whereas **(b)** gives the results for the same data but when the number of classes is chosen purely by BIC. Within these sub-tables each row corresponds to the number of classes inferred by the algorithm - so that the number of CNVs which were called as having 2 classes when merging was applied from the easy simulated data set where 2 classes were simulated is 199. The numbers given in bold give the counts of correctly inferred numbers of classes. The total number of CNVs simulated in each experiment (column) is 200.

Table 1

	Easy				Medium				Hard			
	2	3	4		2	3	4		2	3	4	
(a)												
1	1	1	0	5	1	1	0	8	1	0		
2	199	5	0	195	24	10	192	66	38			
3	0	194	18	0	175	69	0	133	95			
4	0	0	182	0	0	120	0	0	67			
5	0	0	0	0	0	1	0	0	0			
(b)												
1	0	0	0	1	0	0	0	0	0	0		
2	200	6	0	198	12	0	199	19	4			
3	0	193	17	1	186	52	1	181	85			
4	0	1	182	0	2	143	0	0	109			
5	0	0	1	0	0	5	0	0	2			

Parameters of the simulation studies. The gamma distributions here are parametrised by mean and variance respectively for interpretability.

Table 2

	good s_l	bad s_l	P_b^l	C_b^i	n	P_v^l	C_v^i
Easy	6.25	0.0625	$N(0, 0.3^2)$	$N(0, 0.004^2)$	0.2	$\Gamma(1.25, 0.28^2)$	$\Gamma(0.025, 0.0016^2)$
Medium	4.00	0.400	$N(0, 0.75^2)$	$N(0, 0.01^2)$	0.5	$\Gamma(1.25, 0.28^2)$	$\Gamma(0.025, 0.0025^2)$
Difficult	3.00	0.300	$N(0, 0.75^2)$	$N(0, 0.01^2)$	0.5	$\Gamma(1.25, 0.28^2)$	$\Gamma(0.025, 0.0025^2)$