Published online 12 April 2011

Nucleic Acids Research, 2011, Vol. 39, No. 12 ***e79***
doi:10.1093/nar/gkr197

cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate

Djork-Arné Clevert[1,2,*], Andreas Mitterecker[1], Andreas Mayr[1], Günter Klambauer[1], Marianne Tuefferd[3], An De Bondt[3], Willem Talloen[3], Hinrich Göhlmann[3] and Sepp Hochreiter[1,*]

[1]Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria, [2]Department of Nephrology and Internal Intensive Care, Charité University Medicine, Berlin, Germany and [3]Johnson & Johnson Pharmaceutical Research & Development, a Division of Janssen Pharmaceutica, Beerse, Belgium

Received January 13, 2011; Revised March 15, 2011; Accepted March 18, 2011

ABSTRACT

Cost-effective oligonucleotide genotyping arrays like the Affymetrix SNP 6.0 are still the predominant technique to measure DNA copy number variations (CNVs). However, CNV detection methods for microarrays overestimate both the number and the size of CNV regions and, consequently, suffer from a high false discovery rate (FDR). A high FDR means that many CNVs are wrongly detected and therefore not associated with a disease in a clinical study, though correction for multiple testing takes them into account and thereby decreases the study's discovery power. For controlling the FDR, we propose a probabilistic latent variable model, 'cn.FARMS', which is optimized by a Bayesian maximum a posteriori approach. cn.FARMS controls the FDR through the information gain of the posterior over the prior. The prior represents the null hypothesis of copy number 2 for all samples from which the posterior can only deviate by strong and consistent signals in the data. On HapMap data, cn.FARMS clearly outperformed the two most prevalent methods with respect to sensitivity and FDR. The software cn.FARMS is publicly available as a R package at http://www.bioinf.jku.at/software/cnfarms/cnfarms.html.

INTRODUCTION

Copy number varations (CNVs) are one or more kilobases long DNA regions with varying copy numbers between individuals (1). In biology and population genetics, CNVs help to understand the origin and evolution of genomes (1–3). In medicine, associations between CNVs and diseases were discovered, e.g. for systemic autoimmunity (4), HIV (5), Crohn's disease and type 1 diabetes (6), type 2 diabetes (7–9), malaria, breast and prostate cancer, multiple sclerosis and bipolar disorder (10). In most CNV studies, DNA oligonucleotide arrays like the Affymetrix Genome-wide SNP 6.0 arrays are applied. These arrays possess both high coverage and high resolution through their large number of genetic markers (the probes). They are able to detect CNVs in formalin-fixed, paraffin-embedded (FFPE) tissue samples which were stored decades ago (11,12). FFPE samples are attractive because instead of designing new studies, existing biobanks can be utilized, though the measurements are more noisy.

If analyzing CNV data from microarrays, researchers face the serious problem of high false discovery rates (FDRs), i.e. the fraction of wrongly detected or too large CNV regions. CNVs are wrongly detected because of random probe variations through measurement noise. Current array techniques strive steadily to increase the number of probes in order to obtain higher coverage and higher resolution. However, this coverage is traded off against more false discoveries, which increase proportional to the number of probes. Each falsely discovered CNV region may give a false hint for population geneticists or may generate a spurious correlation with a disease and, therefore, misguides the medical expert. More seriously, a high FDR at CNV detection decreases the discovery power of studies and the significance of discoveries after correction for multiple testing. Falsely discovered CNVs are not associated with diseases, though correction for multiple testing takes them into account and reduces the discovery power of the study. Therefore, FDR control

*To whom correspondence should be addressed. Tel: +49 30 6883 5306; Fax: +49 30 6883 5307; Email: okko@clevert.de
Correspondence may also be addressed to Sepp Hochreiter. Tel: +43 732 2468 8880; Fax: +43 732 2468 9511; Email: hochreit@bioinf.jku.at

is a highly desired feature of CNV analysis methods to avoid that the advantage of higher coverage is counteracted by correction for multiple testing. However, current CNV analysis methods do not control the FDR, as Baross *et al*. (13) write 'The frequency of false positive deletions was substantial' with different methods like dChip (14) and CNAG (15). We introduce cn.FARMS for array-based CNV analysis which is designed to control the FDR while ensuring high sensitivity.

Previous array-based CNV analysis methods

We assume that the DNA is first cut by enzymes into fragments which are then amplified by PCR. The PCR products are then mechanically fragmented into smaller pieces before being put on the array. Each CNV region is broken by enzymes into several DNA fragments each of which is targeted by several probes. This gives a copy number hierarchy probes-fragment-region which is depicted in Figure 1. The more copies of the region exist, the more fragment copies exist, the higher are the probe intensities.

As visualized in Figure 2, copy number analysis is, in principle, a three-step pipeline: (i) normalization, (ii) probe-level modeling and (iii) segmentation. We introduce this pipeline to describe previous methods in the following and to describe our cn.FARMS method in section 'MATERIALS AND METHODS' (note, that cn.FARMS neither does segmentation nor integer copy number estimation).

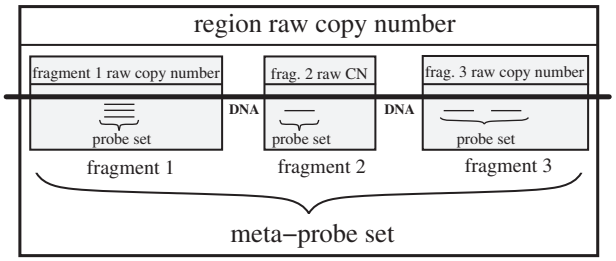

Figure 1. The copy number hierarchy probes-fragment-region. Fragment copy numbers serve as meta-probes used for 'multi-loci modeling' which yields region copy numbers. Inner boxes: the probes which target a fragment (often at a SNP position) are summarized to a raw copy number of this fragment. Note, that instead of fragments a DNA probe loci can be summarized. Outer box: the raw fragment copy numbers are the meta-probes for a DNA region and are summarized to a raw region copy number.

Normalization. Normalization is performed at two levels. It has as 'input' the raw probe intensity values and as 'output' intensity values at chromosome locations which are leveled between arrays and are allele independent. At the 'first level', normalization methods remove technical variations between arrays arising from differences in sample preparation or labeling, array production (e.g. batch effects) or scanning differences. The goal of the first level is to correct for array-wide effects. At the 'second level', alleles are combined to one intensity value at a chromosome location. Optional correction for cross-hybridization between allele A and allele B probes is performed. Cross-hybridization arise due to close sequence similarity between the probes of different alleles, therefore a probe of one allele picks up a signal of the other allele. The optional corrections for differences in PCR yield can be performed at this step or after 'single-locus modeling' (see below). After normalization, arrays have comparable, allele-independent probe intensity values, which measure the copy number of a specific target fragment or DNA probe site.

Modeling. Modeling is also performed at two levels. The 'input' is the probe intensity values which independently measure the copy number of a specific target fragment or DNA probe locus. The 'output' is an estimate for the region copy number. At the 'first level', 'single-locus modeling', the probes which measure the same fragment are combined to a raw fragment copy number ('raw' means that the copy number is still a continuous value; Figure 1). An optional intermediate level corrects for the fragment length and sequence features like the GC content to make raw fragment copy numbers comparable along the chromosome. Nannya *et al*. (15) suggested considering fragment characteristics like sequence patterns and the

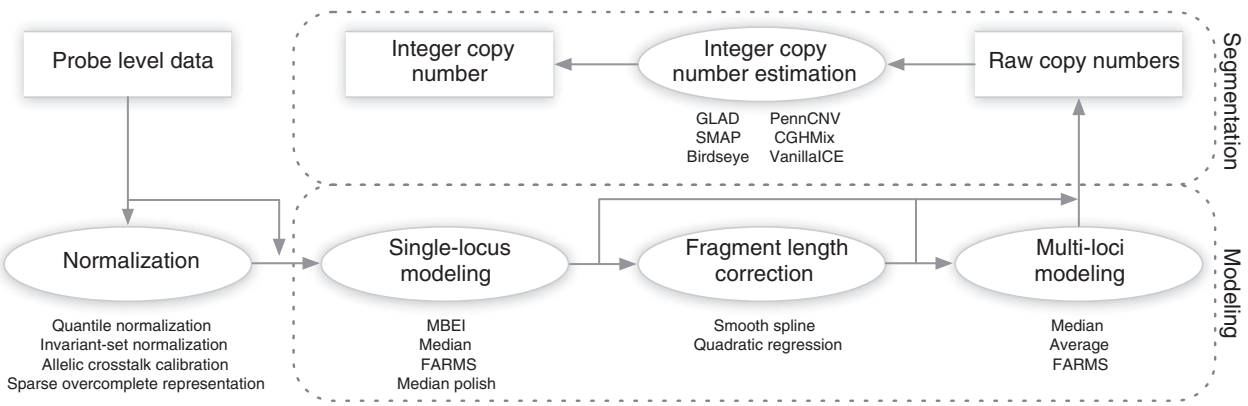

Figure 2. Copy number analysis for (Affymetrix) DNA genotyping arrays as a three-step pipeline: (i) normalization, (ii) modeling and (iii) segmentation. Modeling is divided into 'single-locus modeling' and 'multi-loci modeling' with 'fragment length correction' as an optional intermediate step. As described in subsection 'cn.FARMS: FARMS for CNV Detection', cn.FARMS' pipeline is as follows: normalization by sparse overcomplete representation, single-locus modeling by FARMS, fragment length correction and multi-loci modeling by FARMS.

length because they affect PCR amplification. For example, PCR is usually less efficient for longer fragments, which lead to fewer copies to hybridize and result in weaker probe intensities. At the 'second level', 'multi-loci modeling', the raw copy numbers of neighboring fragments or neighboring DNA probe loci are combined to a 'meta-probe set' which targets a DNA region. Raw fragment copy numbers or DNA probe loci in a region now serve as probes themselves which measure the region's copy number (Figure 1). Multi-loci modeling considerably reduces the FDRs, because raw copy numbers of neighboring fragments or neighboring DNA probe loci must agree to each other on the copy number, which reduces the likelihood of a discovery by chance. However, low FDR is traded against high resolution by the window size for multi-loci modeling, i.e. by how many raw copy numbers of neighboring fragments or neighboring DNA probe loci are combined.

Segmentation. Segmentation is also performed at two levels. It has as 'input' the continuous raw copy numbers and as 'output' integer copy numbers for segments. At the 'first level', segmentation groups together adjacent raw copy numbers with similar intensity values. At the 'second level', integer copy numbers are assigned to the regions. Neighboring regions are separated by breakpoints which indicate a change in the copy number (16). Note, that this step overlaps with the previous modeling step because in both steps single loci can be combined to regions. For example, hidden Markov models automatically assign integer copy numbers (the hidden states) and segment the DNA by runs of the same hidden state.

Using this pipeline, we next categorize existing methods for analyzing copy number variations on microarray data: (i) the first CNV analysis method has been supplied by Affymetrix with the hardware. It is called 'Chromosome Copy Number Analysis Tool' (CNAT) where version 1.0 appeared as early as 2004 but now version 4.0 (17) can be used. (a) Normalization is performed at the first level by quantile normalization (18). The second level is skipped because the alleles are separately modeled. (b) Modeling uses robust multichip average [RMA (18–20)] for allele-specific single-locus modeling. RMA is an additive model fitted by median polish. (ii) Following CNAT, the 'DNA-Chip Analyzer' (dChip) software for transcriptomic data was modified to allow for CNV analysis (21). (a) Normalization at the first level is based on the invariant set method which corresponds to normalize the arrays based on probes with known copy numbers. At the second level, allele A and B probe intensities are added. (b) Modeling is based on model-based expression index [MBEI (14)] for single-locus modeling. MBEI iteratively estimates a linear model that is the product of a raw copy number and a probe pattern by least squares. (c) Segmentation is either performed by computing the median over a region or by a hidden Markov model. (iii) One of the early CNV analysis methods is 'Copy Number Analyser for GeneChip' [CNAG (15)]. (a) Normalization starts with the second level, namely to remove allele-specific probe signals by adding allele A and B probes to give allele-independent fragment probe intensities per array. Next the arrays are normalized to have the same mean signal intensity for all autosomal probes which make fragment probes comparable between arrays. (b) Modeling skips single-locus modeling and directly corrects for fragment length and for the GC content. Both corrections are realized by a quadratic regression which predicts intensities based on GC content and fragment length. (iv) A CNV analysis software, which is broadly used, is Birdsuite's Birdseye (22). (a) Normalization is performed at the first level by quantile normalization like with CNAT. Normalization at the second level is realized by SNP genotyping through the Birdseed method via a mixture clustering. (b) Modeling and (c) Segmentation are performed together at the multi-loci level. The hidden states of a hidden Markov model (HMM) give the copy numbers and its outputs are the probe intensities for the estimated genotype. The HMM reuses the mixture distributions from Birdseed as emission probabilities for copy number 2 while emission probabilities for copy number 0 and 1 are estimated on the X chromosome using the sex information. (v) Most recently 'Copy-number estimation using Robust Multichip Analysis' [CRMA (23), CRMA_v2 (24)] has been proposed as an extension of the RMA model. (a) Normalization at the first and second level are combined by allelic cross-hybridization correction (ACC). ACC performs allele correction array-wise in the 2D space of the allele A and allele B intensity. A cone is fitted to the data such that one border of the cone is a regression line for the AA genotype and the other border for the BB genotype. Similar to the left and right line in Figure 3. The cone fitting allows estimating how much allele A cross-hybridizes at the allele B probe and vice versa. Genotype AA (allele A only) should lead to minimal intensity at the allele B probe and genotype BB (allele B only) to minimal intensity at allele A probe. The genotype AB is assumed to have the same cross-hybridization characteristics as genotypes AA and BB. Finally, the probes are normalized by scaling them to a pre-specified mean intensity value. (b) Modeling for single-locus raw copy numbers is performed via RMA. Then CRMA corrects for the GC pattern and for the fragment length where the former showed little effect and is therefore not recommended by the authors (23). Most CNV analysis methods allow using an arbitrary segmentation algorithm [for an overview see Ref. (25)].

Popular is the Gain and Loss Analysis of DNA (GLAD) model which is a local constant Gaussian regression model (26). Using a weighted maximum likelihood estimator, GLAD estimates regions with constant copy numbers. Other methods like CGHMIX (27) estimate the copy number by a mixture model incorporating spatial information. Spatial information is also utilized by segmentation with an HMM like in Birdseye and in the 'Segmental Maximum A Posteriori' approach [SMAP (28)]. Also 'PennCNV' (29) and 'vanillaICE' (30) apply an HMM to integer copy number estimation using spacial and genotype information.

However, all mentioned methods do not control the FDR and are prone to high FDRs. We will control the

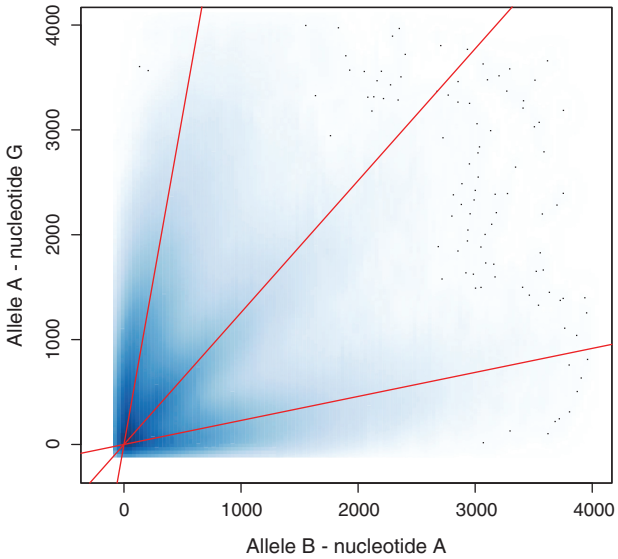

Figure 3. Sparse overcomplete representation of allele A and B probes. The smooth scatter plot for a HapMap Affymetrix 250K_NSP array sample (CEU_NA12878, G/A allele probes). The three clouds going outwards from the origin correspond to genotypes AA (upper left cloud), AB (middle cloud), and BB (lower right cloud). For the genotype AA, allele A probes show a strong signal and allele B probes shows a weak signal due to cross-hybridization (analog for genotype BB). Note, that the middle cloud is closer to the left cloud than to the right (violating CRMA's ACC assumptions). The lines are the estimates of sparse overcomplete representation. They are used to correct for cross-hybridization by moving the left cloud to be vertical, the middle cloud to be at the 45° line and the lower right cloud to be horizontal.

FDR by selecting CNVs based on high information content determined by a latent variable model.

MATERIALS AND METHODS

We propose a novel CNV detection method, called 'cn.FARMS', which is based on our FARMS ['factor analysis for robust microarray summarization' (31)] algorithm for summarizing probe sets of expression arrays. Expression array summarization estimates the expression value of a gene which is basically its mRNA copy number. The expression value of an mRNA is computed from intensity values of all probes targeting it that is the probe intensities are summarized. Since 2006, FARMS is the leading summarization method of the international 'affycomp' competition if sensitivity and specificity are considered simultaneously. We extend FARMS to cn.FARMS for detecting CNVs by moving from mRNA copy numbers to DNA copy numbers.

cn.FARMS: FARMS for CNV detection

cn.FARMS is described by the pipeline depicted in Figure 2: (i) normalization at the first and second level are combined similar as for CRMA (23). However, instead of CRMA's ACC, we propose sparse overcomplete representation in the 2D space of allele A and B intensity. Therefore, we do not only estimate the AA and the BB cross-hybridization like CRMA but also the AB cross-hybridization. The latter takes into account that hybridization and cross-hybridization may be different for the AB genotype, where for both allele probes target fragments are available and compete for hybridization. After allele correction, we follow CRMA and normalize by scaling the probes to a pre-specified mean intensity value. CNV probes which have only one allele are scaled in the same way. (ii) At the first level, 'single-locus modeling', raw fragment copy numbers are estimated by FARMS. The original FARMS was designed to summarize probes which target the same mRNA. This can readily be transfered to CNV analysis where FARMS now summarizes probes which target the same DNA fragment. Either both strands can be summarized together or separately where our default is the former. Following the suggestions in Nannya *et al.* (15), cn.FARMS performs GC and fragment length correction. At the second level, 'multi-loci modeling', the raw copy numbers of neighboring fragments or neighboring DNA probe loci are combined to a 'meta-probe set' which targets a DNA region. The raw fragment copy numbers from single-locus modeling are now themselves probes for a DNA region as depicted in Figure 1. Again, we use FARMS to summarize metaprobes and to estimate a raw copy number for the region. This modeling across samples is novel as previous methods only model along the chromosome. FARMS supplies an informative/non-informative (I/NI) call (32,33) which is used to detect CNVs. Additionally, the I/NI value gives the signal-to-noise-ratio of the estimated raw copy number. (iii) Segmentation and estimation of integer copy numbers is performed by segmentation methods like those which were mentioned at the end of the 'Introduction' section.

In our pipeline, FARMS is used for both single-locus and multi-loci CNV analysis. The more loci are combined, the more the FDR is reduced, because more metaprobes must mutually agree on the region's copy number. The window size for multi-loci modeling is a hyperparameter which trades off low FDR against high resolution. We recommend a window size of 5 as default, 3 for high resolution and 10 for low FDR. Alternatively to a fixed number of CNV or SNP sites, the cn.FARMS software allows defining a window in terms of base pairs. In this case, multi-loci modeling may use a different number of metaprobes at different DNA locations, in particular for less than two metaprobes multi-loci modeling is skipped. Note, however, that controlling the FDR is more difficult because a minimal number of metaprobes cannot be assured for each window and modeling with few metaprobes is prone to false discoveries. cn.FARMS introduces at several steps novel algorithms into the CNV detection pipeline. First, at the normalization step sparse overcomplete representation is used for allele correction. Second, FARMS is used for 'single-locus modeling'. Third, FARMS is used for 'multi-loci modeling' which supplies the raw region copy numbers. Fourth, and most importantly, I/NI calls for controlling the FDR are supplied. In the following subsections,

we describe the methods which are utilized by cn.FARMS and are novel in the CNV detection pipeline.

Sparse overcomplete representation

At the pipeline's step (i), the normalization, cn.FARMS corrects for cross-hybridization between allele A and allele B probes. We generalize the ACC method of CRMA. ACC performs a cone fitting in the 2D space of the allele A and allele B intensity, where cone borders lay at the AA and BB genotype (see left and right line in Figure 3). For each array, the probes are first divided into the allele groups A/T, A/C, A/G, T/C, T/G, C/G to each of which a cone is separately fitted. ACC assumes that cross-hybridization for the AB genotype has the same characteristics as for AA and BB genotypes. Consequently, the AB genotype regression line is supposed to be exactly between the AA and BB genotype regression line (the cone borders), that is the AB regression line divides the cone into two equal halves. However, the assumption on the AB genotype regression line is not always true as shown in Figure 3 for a HapMap Affymetrix 500K array sample. In this example, the AB regression line does not divide the cone into two equal halves, which indicates that cross-hybridization is different for the AB genotype. For the AB genotype, target fragments for both alleles are present and compete for hybridization at the probe's spots. Motivated by such examples, at the ACC step we not only estimate a regression line for the AA and BB genotype but also for the AB genotype. After correction for cross-hybridization, the AA and BB genotypes should lay on the x-axis (allele A) and y-axis (allele B), respectively, because one probe allele is supposed to be zero, while the AB genotype should be on the 45° line. This problem of fitting three lines in a 2D space is solved in the field of machine learning by sparse overcomplete representation (34,35). Data points are described by more vectors than the dimension of the space, therefore the description of a data point is not unique. Sparse overcomplete representations choose the most sparse one from the set of all possible data descriptions. A sparse description is appropriate if each data point is mainly determined by few describing vectors. For allele correction, the sparse description is justified because a data point which is given by the two allele probe intensities can be described by (i) its angle given by the genotype (AA, AB and BB–the genotype determines three main directions) and (ii) its radius given by the copy number. Thus, we represent the 2D vector of allele A and allele B probe intensity by a 3D vector where the components correspond to the genotypes AA, AB and BB. The solution of a sparse overcomplete representation is shown as the lines in Figure 3. A sparse overcomplete representation of 2D data $\boldsymbol{x}_s \in \mathbb{R}^2$ can be modeled as:

$$\boldsymbol{x}_s \;=\; \lambda_s \, \boldsymbol{z}_s \;+\; \epsilon_s \qquad (1)$$

where $\boldsymbol{z}_s \in \mathbb{R}^3$, $\lambda_s \in \mathbb{R}^{2\times 3}$ and $\epsilon_s \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{\Psi}_s)$. Here $\mathcal{N}(\mathbf{0},\ \boldsymbol{\Psi}_s)$ is the 2D Gaussian distribution with mean vector $\mathbf{0} \in \mathbb{R}^2$ and covariance matrix $\boldsymbol{\Psi}_s \in \mathbb{R}^{2\times 2}$. Sparseness is enforced by assuming a Laplacian prior for $\boldsymbol{z}_s$:

$$p(\boldsymbol{z}_s) \;=\; (2)^{-\frac{3}{2}} \prod_{l=1}^{3} \exp\left(\sqrt{2}\ |z_{sl}|\right) . \qquad (2)$$

Because the likelihood for this model is analytically intractable, we employ a variational approach according to Girolami (36). The Laplacian prior is locally approximated from below by a local Gaussian at the mode of the Laplacian. An expectation–maximization algorithm (37) is used to optimize the parameters $\boldsymbol{\lambda}_s$ and $\boldsymbol{\Psi}_s$. Using these parameters, the maximum of the $\boldsymbol{z}_s$-posterior $\hat{\boldsymbol{z}}_s$ allows back-transforming the data to $\hat{\boldsymbol{x}}_s$ by $\hat{\boldsymbol{x}}_s = \lambda_s \hat{\boldsymbol{z}}_s$.

FARMS algorithm

Overview. The main idea of the FARMS algorithm is to detect a common hidden cause in the measurements assuming independent noise. The probabilistic FARMS model:

- regards that probes measuring the same target (fragment or region) can only be positively correlated,
- estimates (meta-)probe-specific characteristics,
- automatically trades off signal against noise via the z-posterior distribution,
- can adjust the signal/noise tradeoff via the priors on the parameters and
- supplies I/NI calls (32,33).

The I/NI call measures the information gain of the posterior over the prior which can be interpreted as the negative log signal-to-noise ratio. High data information content leads to a low variance of the latent variable's posterior and a high confidence in the copy number estimate. The original FARMS applied to 30 real-life expression data sets could exclude 70–99% of all probe sets because of their low information content while never excluding a gene that was known to be biologically meaningful (32). We want to introduce this I/NI call property into the field of CNV analysis to control the FDR.

Brief review. The vector of n probes $\boldsymbol{x}$ is modeled by probe-effects $\boldsymbol{\lambda}$ and a factor z (latent variable or signal) representing the raw normalized copy number as:

$$\boldsymbol{x} \;=\; \lambda\, z \;+\; \epsilon, \qquad (3)$$

where $\boldsymbol{x},\ \boldsymbol{\lambda} \in \mathbb{R}^n$ and $z \sim \mathcal{N}(0,\ 1)$, $\epsilon \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{\Psi})$. Here $\boldsymbol{\Psi} \in \mathbb{R}^{n\times n}$ is the diagonal noise covariance matrix to address independent measurement noise. ϵ and z are assumed to be statistically independent. Given these assumptions, $\boldsymbol{x}$ is distributed according to the following Gaussian:

$$\boldsymbol{x} \;\sim\; \mathcal{N}\left(\mathbf{0}\ ,\ \boldsymbol{\lambda}\boldsymbol{\lambda}^T \;+\; \boldsymbol{\Psi}\right) . \qquad (4)$$

The covariance matrix of $\boldsymbol{x}$ is decomposed into signal $\boldsymbol{\lambda}\boldsymbol{\lambda}^T$ and noise $\boldsymbol{\Psi}$. Because $\boldsymbol{\Psi}$ is diagonal, probe correlations are attributed to the signal z via $\boldsymbol{\lambda}$. That means highly correlated probes lead to large $\boldsymbol{\lambda}$ which in turn leads to

low noise because the diagonal of the covariance matrix of $\boldsymbol{x}$ is mainly explained by $\boldsymbol{\lambda}$.

Higher intensity of the probes means more copies and vice versa, therefore noise-free probes must be positively correlated. FARMS ensures the positive correlation of probes by a prior on $\boldsymbol{\lambda}$ which enforces only positive values: $p(\lambda) = \prod_{j=1}^{n} p(\lambda_j)$, where the rectified Gaussian $p(\lambda_j)$ is given by

$$\lambda_j = \max\{y_j, 0\} \text{ with } y_j \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda) . \quad (5)$$

Further, the prior on $\boldsymbol{\lambda}$ prefers small values and, therefore, model selection tends to explain variation by noise instead by a signal. Using μ_λ and σ_λ, the prior's influence on model selection and, therefore, the signal/noise tradeoff can be adjusted. FARMS selects the model parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\Psi}$ by an expectation–maximization algorithm (37) that maximizes the parameter posterior. To ensure data consistency, negative entries in the data covariance matrix are set to zero.

I/NI calls. The I/NI call measures the information gain of the posterior hidden variable distribution compared to its prior distribution where the latter represents the null hypothesis. Therefore, the I/NI call measures the tendency to reject the null hypothesis based on the observed data.

From the model Equation (4) and the Gaussian z-prior $N(0, 1)$, we can compute the z-posterior $p(z \mid \boldsymbol{x})$ as

$$\begin{aligned} z \mid \boldsymbol{x} &\sim \mathcal{N}\left(\mu_{z|\boldsymbol{x}}, \sigma^2_{z|\boldsymbol{x}}\right) \\ \mu_{z|\boldsymbol{x}} &= (\boldsymbol{x})^T \boldsymbol{\Psi}^{-1}\lambda \left(1 + \lambda^T\boldsymbol{\Psi}^{-1}\lambda\right)^{-1} \\ \sigma^2_{z|\boldsymbol{x}} &= \left(1 + \lambda^T\boldsymbol{\Psi}^{-1}\lambda\right)^{-1} . \end{aligned} \quad (6)$$

We see that large $\boldsymbol{\lambda}$ (going with low noise $\boldsymbol{\Psi}$) leads to low variance of $z \mid \boldsymbol{x}$, which means a precise conditional z.

The variance of z is decomposed into a signal and a noise part:

$$\text{var}(z) = \frac{1}{N}\sum_{i=1}^{N} E_{z_i|\boldsymbol{x}_i}\left(z_i^2\right) = \frac{1}{N}\sum_{i=1}^{N} \mu^2_{z_i|\boldsymbol{x}_i} + \sigma^2_{z|\boldsymbol{x}}, \quad (7)$$

where the noise part $\sigma^2_{z|\boldsymbol{x}}$ is independent of $\boldsymbol{x}_i$ according to Equation (6) and serves as I/NI call in FARMS (32).

At the same time $-\log\sigma_{z|\boldsymbol{x}}$ measures the information gain between the prior and the posterior because the prior has unit variance and therefore zero entropy.

cn.FARMS: I/NI calls and FDR control

As the FARMS I/NI call also cn.FARMS' I/NI call measures the information gain of the posterior hidden variable distribution compared to its prior distribution that represents the null hypothesis. The variance across samples of the signal part of maximum posterior hidden variable z given the observation x is cn.FARMS I/NI call. This signal variance is zero for the prior. In contrast to FARMS I/NI call, cn.FARMS I/NI call also includes the signal strength. This reflects the assumption that data from null hypotheses produce only spurious signals that are low. Such spurious signals are more likely to be observed for cn.FARMS at multi-loci modeling with few metaprobes than for FARMS on expression arrays with larger probe sets.

First, we compute the signal strength S. The data $\{\boldsymbol{x}_i\}$ has been probewise standardized to variance 1 and mean zero, where $\text{std}(\boldsymbol{x}^{\text{raw}})$ is the probes' SD vector of the raw data $\boldsymbol{x}^{\text{raw}}$. We reintroduce the signal strength S as the median of $\boldsymbol{\lambda}$ scaled by std $\boldsymbol{x}^{\text{r}}$:

$$S = \text{median}(\lambda \cdot \text{std}(\boldsymbol{x}^{\text{raw}})), \quad (8)$$

where '$\cdot$' is the element-wise product.

Second, we extract the variance of the maximum a posterior hidden variable z given the observation $\boldsymbol{x}$:

$$\begin{aligned} \text{sigvar}(z) &= \frac{1}{N}\sum_{i=1}^{N} \mu^2_{z_i|\boldsymbol{x}_i} \\ &= \lambda^T\boldsymbol{\Psi}^{-1}\,\text{covar}(\boldsymbol{x})\,\boldsymbol{\Psi}^{-1}\lambda\left(1 + \lambda^T\boldsymbol{\Psi}^{-1}\lambda\right)^{-2}, \end{aligned} \quad (9)$$

which is between 0 (no signal, only noise) and 1 (only signal, no noise). Note, that sigvar(z) is one minus FARMS' I/NI call squared and corresponds to the part of the variance in the data explained by the signal.

cn.FARMS' I/NI call is signal variance multiplied by the signal strength squared:

$$\text{I/NI} = \text{sigvar}(z)\, S^2. \quad (10)$$

Note, that I/NI calls allow comparing two data sets with respect to common CNVs. In this case, the model is selected on one data set $\{\boldsymbol{x}_i\}$ and the calls are made on the other data set $\{\boldsymbol{y}_i\}$ using covar($\boldsymbol{y}$).

The I/NI call value considers both the signal strength and the information gain. If the true copy numbers vary, then probe intensities are consistent (correlated) and large (high signal-to-noise-ratio) and therefore lead to a large $\boldsymbol{\lambda}$ and a small $\boldsymbol{\Psi}$ which in turn gives a large sigvar(z) (close to 1). In contrast to these true positives, false positives come from random independent Gaussian noise variations, which are unlikely to produce consistent and large probe intensities. Thus, the larger the I/NI call, the less likely it was caused by noise. Consequently, the ratio of false positives decreases with increasing I/NI call values. A CNV is detected by an I/NI call value exceeding a detection threshold, therefore the threshold controls the FDR. The effect of the detection threshold can be seen in Figure 5, where precision–recall curves on HapMap SNP 6.0 arrays are shown for cn.FARMS. Note that the precision is 1−FDR, thus the distance of the curve to the upper limit gives the FDR. Therefore, the curve shows the FDR as a function of the threshold where indeed higher thresholds (more to the left) result in smaller FDRs. The detection threshold for a desired FDR can either be estimated at chromosome locations where CNVs are unlikely or at reference data sets.

RESULTS

We compare the new cn.FARMS algorithm with the two methods which performed best in other comparative studies on raw copy number estimation (23), namely the

dChip software for CNV analysis (21) and CRMA (23,24) (see the 'Introduction' section for a brief description of these and the following methods). In Bengtsson *et al.* (23), it was shown that both CRMA and dChip perform better than CNAG and CNAT . Therefore, these two methods are not regarded in our experiments. Other methods like Birdseye do not estimate raw copy numbers and incorporate segmentation and integer copy number estimation. The latter methods can still be applied on the output of cn.FARMS for single-locus or multi-loci modeling.

Because true copy numbers are in general not known, we use two benchmark data sets from 'The International HapMap Project' where the sex must be determined by the raw copy numbers at the X chromosome. (i) We first use the 250K Affymetrix array benchmark data set from Bengtsson *et al.* (23). Even if these arrays are outdated, they allow comparisons to other CNV analysis methods like CNAT and CNAG investigated in Bengtsson *et al.* (23). (ii) Next, this benchmark was upgraded to Affymetrix SNP 6.0 arrays, to allow further assessment on recent arrays. (iii) Finally, we assess the FDR at CNV detection on the HapMap phase 2 data set with Affymetrix SNP 6.0 arrays. To estimate the FDR, we define as true CNVs those which were multiple confirmed by other techniques and reported in Conrad *et al.* (38).

250K array benchmark

The first data set is from Bengtsson *et al.* (23). It comprises the 90 CEU founders (30 triplets of father, mother, child) from 'The International HapMap Project' (phase 2) where the children are removed to avoid biases due to inherited CNVs. For these 60 CEU founders, their DNA has been analyzed by Affymetrix Mapping250K_NSP arrays. Female NA12145 had too low copy number level on chromosome X and has been excluded (23) which leads to the final data set of 59 CEU founders. The X chromosome serves as ground truth to assess the performance of CNV detection methods because there males possess one copy and females two. At every location on the X chromosome, raw fragment copy numbers (single-loci) and raw region copy numbers (multi-loci) are used to classify the sex of the person the sample stems from. To allow multi-loci classification for dChip and CRMA_v2, adjacent raw fragment copy numbers are averaged within a region to give a raw region copy number. However, not all locations on the X chromosome can distinguish the sex based on the copy numbers. At the pseudo-autosomal regions (PAR1 and PAR2), the copy numbers of males and females match. Besides PAR1 and PAR2, there are segmental duplications on chromosome Y which match regions at chromosome X (obtained from 'Segmental duplication DB' at http://humanparalogy.gs.washington.edu/build36/). Further chromosome X has CNV regions (1,2). All loci in pseudo-autosomal, segmental duplications in Y and CNV regions are excluded in our classification task. Finally, 5557 single loci on the X chromosome for distinguishing males from females were kept which gives 327 863 ($=59\times5557$) single loci sex classification tasks. The performance of the methods is measured by receiver operating characteristic (ROC) curves. The ROC curve plots the true positive rate (sensitivity) as a function of the false positive rate (1−specificity). Methods with ROC curves at the upper left corner indicate better performance of the corresponding method—a method's ROC curve above another method's ROC curve shows that the former method performs better than the latter. The classification results are shown as ROC curves (A and B) in Figure 4. The ROC curves are summarized by the area under the ROC (AUC) in Table 1. Further, we give the false positives (males classified as females) where the numbers of false positives and false negatives are equal—that is the false positives in the largest 161 153 (number of true female loci = 29×5557) raw copy numbers. To evaluate the statistical significance of the method's differences in performance, we use McNemar's χ^2 test under the null hypothesis that the compared algorithms should have the same error rate (39). The results show that cn.FARMS performs significantly better than dChip and CRMA_v2 and has much fewer false discoveries—confirming that cn.FARMS yields low FDRs.

SNP 6.0 array benchmark

Because the Affymetrix 250K arrays are outdated, we perform the same benchmark test as in the previous

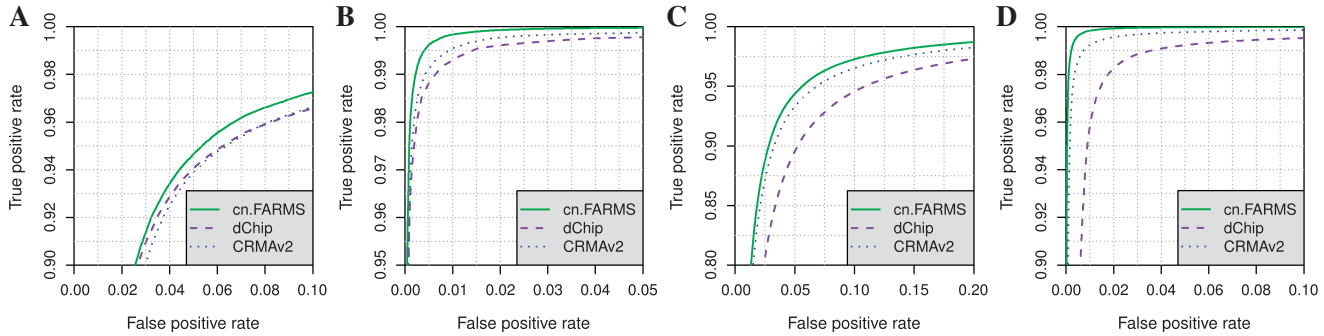

Figure 4. ROC curves for cn.FARMS, CRMA_v2 and dChip at the sex classification task for 59 HapMap CEU founders based on the X chromosome copy numbers. The panels show (**A**) single-locus and (**B**) three-loci modeling of Affymetrix Mapping250K_NSP arrays. While panels show (**C**) single-locus and (**D**) three-loci modeling of Affymetrix SNP 6.0 arrays. ROC curves more at the upper left indicate better performing methods (AUC values for Affymetrix Mapping250K_NSP and Affymetrix SNP 6.0 are given in Table 1). cn.FARMS performs better than CRMA_v2 and dChip.

Table 1. AUC values for cn.FARMS, CRMA_v2 and dChip at the sex classification task for 59 HapMap CEU founders based on the X chromosome copy numbers measured by Affymetrix 250K and Affymetrix SNP 6.0 arrays

Loci	Criteria	Affymetrix Mapping250K_NSP			Affymetrix SNP 6.0		
		cn.FARMS	CRMA_v2	dChip	cn.FARMS	CRMA_v2	dChip
1	AUC	**0.9852**	0.9820	0.9819	**0.9838**	0.9807	0.9721
	FP	**8472**	9106	9018	**56145**	68593	77438
	p-value	–	1.8e-65	3.1e-26	–	1e-1160	1e-6949
2	AUC	**0.9983**	0.9974	0.9969	**0.9983**	0.9963	0.9894
	FP	**1375**	1449	1611	**9777**	11705	18039
	p-value	–	2.7e-4	2.5e-12	–	1e-317	1e-3713
3	AUC	**0.9998**	0.9995	0.9992	**0.9998**	0.9990	0.9953
	FP	**240**	366	440	**1573**	3462	6625
	p-value	–	2.6e-38	7.2e-58	–	1e-896	1e-3455
4	AUC	**1.0000**	0.9999	0.9998	**0.9999**	0.9995	0.9976
	FP	**49**	95	153	**366**	1338	2985
	p-value	–	2.8e-10	1.9e-48	–	1e-594	1e-2023

The first column gives the number of combined loci, where '1' means single-locus modeling. The second column gives (i) area under the receiver operating curve given in Figure 4 ('AUC'), (ii) false positives ('FP' – females are classified as males) and (iii) the *P*-value of McNemar's χ^2 test for difference to the cn.FARMS ('*p*-value'). False positives are counted in the lowest 166 710 and 1 075 680 (number of true male loci) raw copy numbers for Affymetrix 250K and Affymetrix SNP 6.0 arrays, respectively. The last six columns give the values for the according array types and methods, where significant better performance is indicated by boldface numbers. cn.FARMS clearly outperforms CRMA_v2 and dChip.

subsection but now with up-to-date Affymetrix SNP 6.0 arrays. The SNP 6.0 data set comprises again the same 59 CEU founders as for the 250K array benchmark. Note, in contrast to Affymetrix 250K arrays, for Affymetrix SNP 6.0 arrays we model single SNP and CNV loci instead of fragments because for Affymetrix SNP 6.0 the fragment that is targeted by a probe is ambiguous as both Sty and Nsp fragments can hybridize to a probe. We again excluded regions which are pseudo-autosomal, have segmental duplications or have been reported as CNV regions and kept 35 856 single loci for the classification task which sums up to 2 115 504 ($=59\times35\,856$) sex classification tasks. ROC curves (C and D) in Figure 4 show the results for which the AUCs and McNemar significance tests are given in Table 1. Again we report the false positives (females classified as males) while equalizing the numbers of false positives and false negatives. By doing this, we give the number of false positives in the largest 1 039 824 raw copy numbers, which is the number of true female loci $= 29\times35\,856$. Again cn.FARMS significantly outperforms CRMA_v2 and dChip and has fewer false discoveries. The absolute improvement in terms of the AUC values seem to be marginal. However, for single locus modeling, we obtained *P*-values of 1.8e-65 and 3.1e-26 by the McNemar test for 250K, even going down to 1e-1160 and 1e-6949 for SNP 6.0 arrays. Clearly, these *P*-values indicate significant performance improvement of cn.FARMS over its competitors. For 250K arrays, cn.FARMS has 8472 false positives and the second best method (dChip) has 9018, which is about 6.5% more false positives. For SNP 6.0 arrays, cn.FARMS has 56 145 false positives and the second best method (CRMA) has 68 593, which is about 21% more false positives. For 250K arrays and multi-loci modeling with 4 loci, the number of 49 false positives almost doubles if we look at the next best method with 95 false positives. For SNP 6.0 arrays and multi-loci modeling with 4 loci, the number of 366 false positives increases by a factor of 3.5 if we look at the next best method with 1338 false positives.

CNV Detection on HapMap

In this subsection, we want verify that cn.FARMS can indeed control the FDR. In the previous two subsections, we classified male/female based on raw copy numbers at X chromosome locations. The majority of loci have a CNV as half of the samples are male with copy number one and the other half are female with copy number two. Therefore, false discoveries can only appear at the few pseudo-autosomal or CNV regions. In CNV association studies, however, false discoveries are much more likely because true CNVs are rather rare. Therefore, we define rare true CNV regions in this experiment where we use again 'The International HapMap Project' phase 2 data set with Affymetrix SNP 6.0 arrays. The goal is now to identify true rare CNV regions with a low FDR.

We define as 'true CNV regions' those regions which were detected and verified by different biotechnologies in Conrad *et al.* (38). In Conrad *et al.* (38), first, CNV candidate regions were identified by NimbleGen tiling arrays with 2.1 million long oligonucleotide probes covering the genome with a median probe spacing of 56 bp. From the identified CNVs, random control samples were selected and successfully verified by quantitative PCR. The CNV regions identified by NimbleGen tiling arrays served to design CNV-typing Agilent CGH arrays comprising 105 000 long oligonucleotide probes. With these Agilent arrays, 4978 CNVs were detected on 450 HapMap phase 3 samples and then completed by 59 CNV regions from McCarroll *et al.* (40). The third platform, Illumina Infinium genotyping (Human660W), found CNVs of which 87% were already genotyped by the Agilent CGH arrays. Almost all CNVs from Conrad *et al.* (38) were confirmed by at least two different platforms (NimbleGen tiling arrays, Agilent CGH or Illumina

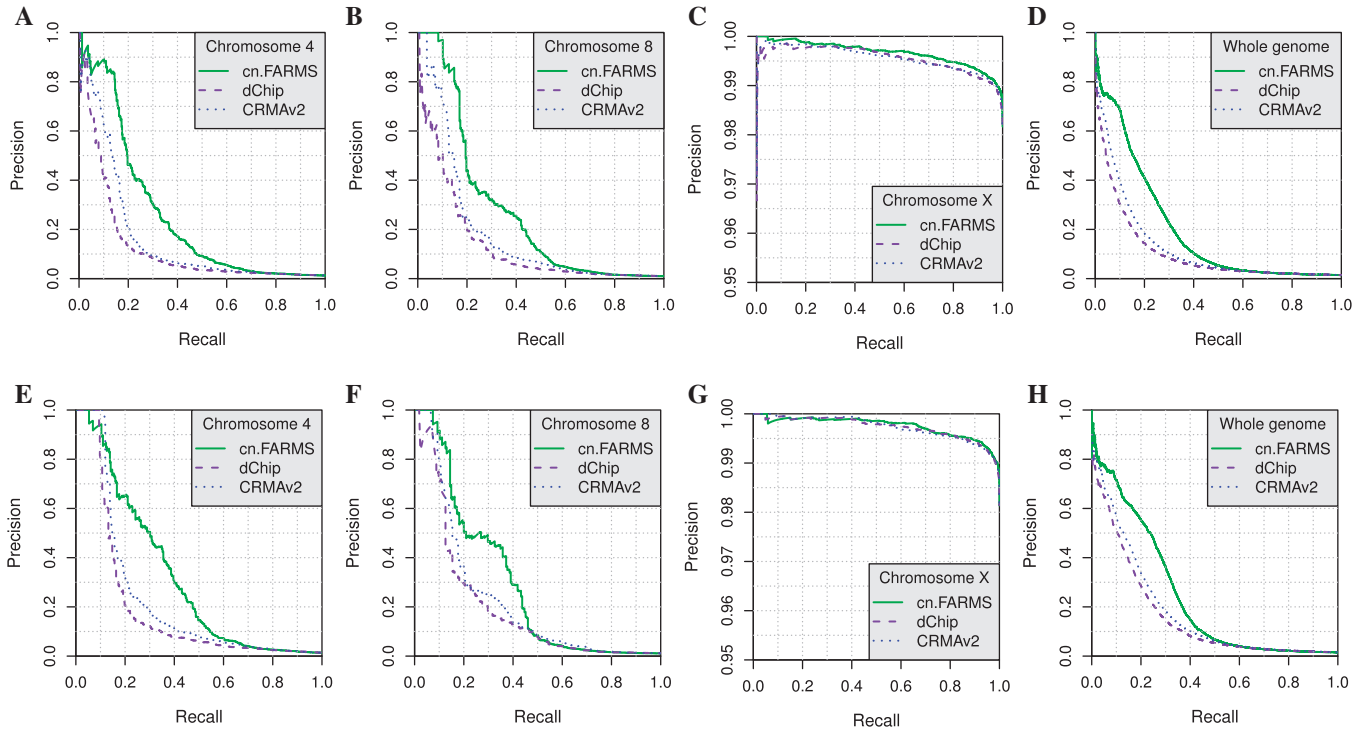

Figure 5. Precision-recall curves (PRCs) on HapMap SNP 6.0 arrays for cn.FARMS, CRMA_v2, and dChip at detecting previously multiple confirmed CNVs reported in Conrad *et al.* (38). cn.FARMS detection criteria is the I/NI call, whereas CRMA_v2 and dChip use the variance of raw copy numbers. A PRC more in the upper-right hand corner indicates better performance. Note, that precision is (1−FDR) thus the FDR is the distance of the curve to the upper limit. Panels (**A–D**) give the PRC for chromosome 4, 8, chromosome X and the whole genome for 3 loci. Panels (**E–H**) show the same for 5 loci. cn.FARMS (solid green) has a clear advantage over dChip (dashed purple) and CRMA_v2 (dotted blue). cn.FARMS has a considerable lower FDR compared to the other methods.

Human660W). Of these 5037 CNV regions, we only selected CNV regions from the 60 CEU HapMap phase 2 samples (CEU trios without children). Finally, we obtained 2515 true CNV regions as reference for our experiment.

For detecting CNV regions, cn.FARMS uses its I/NI calls. However for CRMA_v2 and dChip, we have to define a CNV calling criterion. We tested different criteria of which the variance of the raw copy numbers on the samples gave the best results. This variance calling criterion is like I/NI call independent of the test statistic, thus correction for multiple testing is still valid (33,41).

Using the true CNVs, we can assess the FDR. Instead of reporting the FDR for a fixed classification threshold, we present the CNV detection results as precision–recall curves (PRCs). PRCs plot the precision (which is 1−FDR) as a function of the true positive rate (recall or sensitivity). Thus, a PRC that is more in the upper-right hand corner performs better. A larger *y*-value of the PRC means a lower FDR for a given sensitivity. Figure 5 shows the PRC plots where cn.FARMS has indeed lower FDRs compared to the other methods. The corresponding areas under the precision–recall curves are listed in Table 2. A larger value means that the method has lower FDR averaged over different given recall values. We observed that for some chromosomes increasing the window size also increases the FDR because of the reduced resolution and an overestimate of CNV regions. cn.FARMS performed significantly better than CRMA_v2 and dChip. The significance was obtained by a one sample *t*-test under the hypothesis that the differences between values of the area under the PRC for two methods have a mean equal to 0. The Gaussian assumption of the *t*-test was verified beforehand by a Shapiro–Wilk test. The *P*-values of the *t*-test were smaller than 3.9e-7 for CRMA_v2 and smaller than 7.1e-8 for dChip. Figure 6 shows CNV calling plots across chromosome 4 for 3-loci and 5-loci regions. The *y*-axis gives cn.FARMS' I/NI call and for both CRMA_v2 and dChip the raw copy number variance across samples. Calling values are scaled such that the maximum is one. Local calling densities are encoded by blue color shades. True CNVs (reported in 38) are marked as light-rose bars and calls at these loci by red circles. A perfect calling method would call all true CNVs (red circles at 1) and would not call others (dark blue background at 0). True positives (true CNVs) are better separated from true negatives by cn.FARMS as the smaller variance of true negatives, which is indicated by dark blue density at the bottom. The red arrows, e.g. at positions 65 or 85 Mb in the upper cn.FARMS panel, indicate verified CNVs which were detected by one method, in this case cn.FARMS, but not by both others. cn.FARMS identifies true CNVs with a lower FDR than CRMA_v2 and dChip.

Table 2. Area under the PRCs on HapMap SNP 6.0 arrays for cn.FARMS, CRMA_v2, and dChip at detecting previously multiple confirmed CNVs reported in Conrad *et al.* (38)

Method	Chr	Area under the PRC for combined loci of				Chr	Area under the PRC for combined loci of				Chr	Area under the PRC for combined loci of				Chr	Area under the PRC for combined loci of			
		3	4	5	7		3	4	5	7		3	4	5	7		3	4	5	7
cn.FARMS	**1**	0.20	0.23	0.23	0.25	**7**	0.16	0.19	0.19	0.22	**13**	0.22	0.26	0.26	0.26	**19**	0.17	0.21	0.23	0.26
CRMA_v2		0.16	0.19	0.20	0.23		0.10	0.12	0.13	0.16		0.10	0.14	0.18	0.23		0.11	0.13	0.15	0.19
dChip		0.14	0.18	0.19	0.22		0.09	0.11	0.12	0.15		0.08	0.11	0.16	0.20		0.09	0.11	0.13	0.16
cn.FARMS	**2**	0.19	0.21	0.22	0.25	**8**	0.27	0.28	0.29	0.27	**14**	0.06	0.06	0.06	0.06	**20**	0.24	0.26	0.25	0.31
CRMA_v2		0.12	0.14	0.16	0.21		0.18	0.20	0.22	0.26		0.06	0.06	0.06	0.07		0.12	0.18	0.19	0.25
dChip		0.10	0.13	0.15	0.20		0.13	0.16	0.19	0.21		0.05	0.06	0.06	0.06		0.11	0.14	0.15	0.21
cn.FARMS	**3**	0.27	0.31	0.34	0.39	**9**	0.14	0.13	0.12	0.12	**15**	0.11	0.12	0.14	0.16	**21**	0.05	0.11	0.12	0.19
CRMA_v2		0.16	0.20	0.23	0.29		0.09	0.08	0.10	0.09		0.07	0.09	0.11	0.14		0.04	0.05	0.10	0.10
dChip		0.13	0.16	0.20	0.25		0.06	0.07	0.07	0.07		0.06	0.08	0.09	0.12		0.03	0.04	0.06	0.07
cn.FARMS	**4**	0.25	0.30	0.31	0.34	**10**	0.14	0.17	0.20	0.23	**16**	0.21	0.31	0.36	0.35	**22**	0.54	0.61	0.64	0.70
CRMA_v2		0.16	0.20	0.22	0.26		0.09	0.12	0.15	0.19		0.14	0.21	0.25	0.33		0.41	0.50	0.54	0.62
dChip		0.12	0.16	0.19	0.21		0.08	0.11	0.15	0.18		0.12	0.19	0.23	0.32		0.37	0.44	0.49	0.58
cn.FARMS	**5**	0.19	0.22	0.22	0.24	**11**	0.24	0.25	0.25	0.27	**17**	0.17	0.22	0.22	0.26	**X**	0.99	0.99	0.99	0.99
CRMA_v2		0.11	0.15	0.17	0.21		0.14	0.16	0.19	0.24		0.16	0.22	0.22	0.26		0.99	0.99	0.99	0.99
dChip		0.09	0.11	0.13	0.16		0.08	0.11	0.15	0.19		0.14	0.20	0.21	0.26		0.99	0.99	0.99	0.99
cn.FARMS	**6**	0.23	0.25	0.25	0.27	**12**	0.26	0.32	0.33	0.38	**18**	0.11	0.14	0.15	0.16	**all**	0.20	0.22	0.24	0.26
CRMA_v2		0.16	0.20	0.22	0.25		0.16	0.22	0.25	0.30		0.05	0.06	0.10	0.11		0.13	0.16	0.18	0.21
dChip		0.13	0.16	0.20	0.23		0.12	0.17	0.21	0.26		0.03	0.04	0.05	0.08		0.11	0.14	0.16	0.19

A larger value means that the method has lower FDR averaged over different given recall values. 'Chr' gives the chromosome; 'Area under the PRC for combined loci of' reports the area under the PRCs for different number of combined loci. Note, that large windows can increase the FDR again because CNV regions are overestimated. cn.FARMS clearly outperforms the other methods.

Computational complexity

Finally, we give the computation time for cn.FARMS, dChip and CRMA_v2. The required computation time can be an important factor for choosing an appropriate method because for many samples and large arrays (e.g. Affymetrix SNP 6.0 comprises 6.6 million probes), CNV analysis can take some hours. Table 3 shows the computational times for the compared methods. cn.FARMS requires less time than other methods. cn.FARMS's low computational load is due to the fact that FARMS's update rules both for single and multi-loci modeling are based on an EM algorithm which converges in only a few iterations.

DISCUSSION

Variation across samples versus variation across the chromosome

cn.FARMS identifies regions in the genome that have variable copy numbers across samples. If a CNV is found, it is straightforward to select the samples which caused the variation. In a next step (not considered here), integer copy numbers will be assigned by segmentation methods which find deviations along the chromosome. Thus, segmentation methods serve as a second filter which are able to sort out wrongly detected CNVs stemming from few high variable (noisy) or outlier samples. High variable samples inject locally variation across samples which may be detected by cn.FARMS as a CNV. However, if segmentation methods scan along a chromosome of a high variable sample, the local variation may be considered as being in the range of copy number two. Concluding, cn.FARMS finds variations across samples and segmentation finds variations across the chromosome—only locations having variations in both directions are finally considered as CNV regions.

Affymetrix Mapping250K_NSP to SNP 6.0 arrays

Affymetrix Mapping250K and 500K arrays contain only SNP probes which are allele A or allele B, strand or antistrand, prefect match or mismatch, shifted or not. In contrast to these arrays, Affymetrix SNP 6.0 arrays have, besides single CNV probes, for each SNP and allele three identical probes on one strand. One may think that single locus modeling is superfluous for SNP 6.0 arrays, but we observed that for SNP loci it still improves the results. Though the probes are identical, their fixed array location leads to consistent intensity differences which are captured by single locus modeling.

cn.FARMS for other platforms

Of course, cn.FARMS is not limited to the Affymetrix platform and can be applied to other platforms like Illumina bead arrays or Agilent arrays. The concept remains the same: do genomically adjacent measurements agree on copy numbers? If they contain variation, then the more they agree to each other, the more confident cn.FARMS is in its copy number estimates.

Combining array types and platforms

cn.FARMS can integrate a mixture of arrays or a mixture of platforms if normalization is done carefully to make

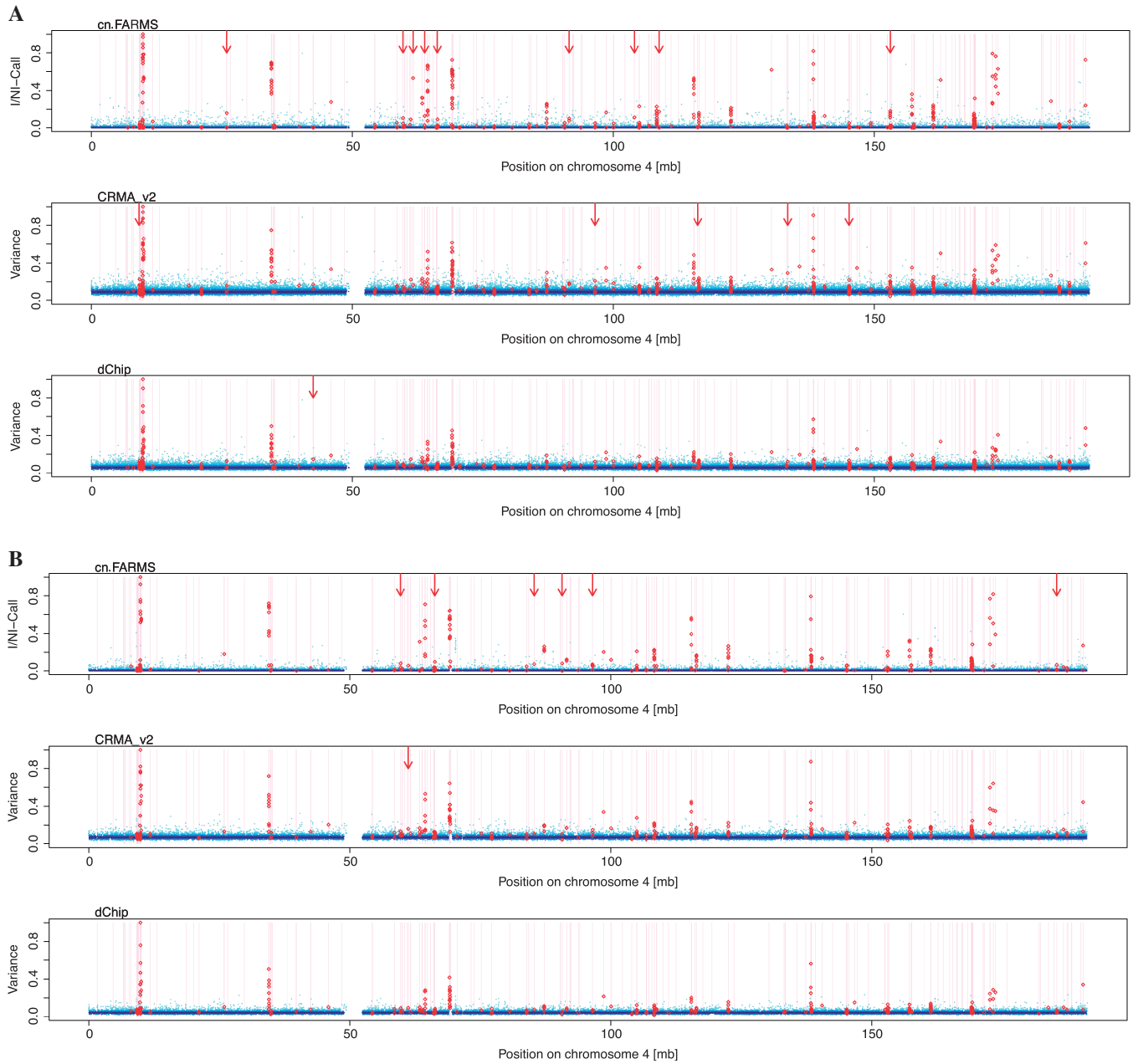

Figure 6. (**A**) CNV calling plots across chromosome 4 for 3 loci regions (each point in the plot summarizes 3 loci). The *y*-axis gives the I/NI call estimated by cn.FARMS and for both CRMA_v2 and dChip it gives the variance. Calling values are scaled such that the maximum is one. Local calling densities are encode by blue color shades. True CNVs [reported in Conrad *et al.* (38)] are marked as light rose bars and calls at these loci by red circles. A perfect calling method would call all true CNVs (red circles at 1) and does not call others (dark blue background at 0). cn.FARMS separates called true positives (true CNVs) from true negatives better than other methods which can be seen at less variance in true negatives indicated by dark blue density at the bottom. The red arrows, e.g. at positions 65 or 85 Mb in the upper cn.FARMS panel, indicate verified CNVs which were detected by one method, in this case cn.FARMS, but not by both others. cn.FARMS identifies true CNVs with a lower FDR than CRMA_v2 and dChip. (**B**) The same plot for 5 loci (each point in the plot summarizes 5 loci). The FDR is further reduced, as can be seen by the lower variance of non-call values at the bottom. Again, cn.FARMS identifies true CNVs with a lower FDR than CRMA_v2 and dChip.

single arrays comparable. We created metaprobe sets for the Affymetrix 500K where the metaprobes for one set are from both the 250K_NSP and the 250K_STY array. Metaprobe sets can in principle consist of metaprobes from different platforms like from Affymetrix and Illumina. The combination of meta-probes across array types or platforms has the advantage that it increases the resolution and coverage, but, on the other hand, it may introduce between array type or between platform variations. It may even be possible to combine array metaprobes with metaprobes obtained from next-generation sequencing (NGS). To provide these NGS metaprobes, we currently work on adapting the idea of cn.FARMS to NGS data by a mixture of Poissons model.

Table 3. Computational time (in seconds) to process 60 Mapping250K_NSP arrays, respectively SNP 6.0. cn.FARMS, is faster than CRMA and dChip

time (s)	cn.FARMS	CRMA_v2	dChip
60 250K	**1055**	3363	2493
60 SNP 6.0	**3657**	11 850	6210

CONCLUSION

We introduced a novel method for detecting CNVs called 'cn.FARMS' which controls the FDR. In experiments, cn.FARMS outperformed its competitors both with respect to FDR and sensitivity, i.e. has fewer false positives while detecting more true CNVs. The reduced FDR increases the discovery power of studies and avoids that researchers are misguided by spurious correlations between CNVs and diseases.

ACKNOWLEDGEMENTS

The authors thank Dr Ulrich Bodenhofer for helpful discussion and comments.

FUNDING

Funding for open access charge: Funds from the Institute of Bioinformatics, Johannes Kepler University Linz.

Conflict of interest statement. None declared.

REFERENCES

1. 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. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
2. Conrad,D.F. and Hurles,M.E. (2007) The population genetics of structural variation. *Nat. Genet*., **39**, S30–S36.
3. Jakobsson,M., Scholz,S.W., Scheet,P., Gibbs,J.R., VanLiere,J.M., Fung,H.-C., Szpiech,Z.A., Degnan,J.H., Wang,K., Guerreiro,R. *et al*. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**, 998–1003.
4. Fanciulli,M., Norsworthy,P.J., Petretto,E., Dong,R., Harper,L., Kamesh,L., Heward,J.M., Gough,S.C., deSmith,A., Blakemore,A.I. *et al*. (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet*., **39**, 721–723.
5. Gonzalez,E., Kulkarni,H., Bolivar,H., Mangano,A., Sanchez,R., Catano,G., Nibbs,R.J., Freedman,B.I., Quinones,M.P., Bamshad,M.J. *et al*. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, **307**, 1434–1440.
6. Wellcome-Trust-Case-Control-Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.
7. Scott,L.J., Mohlke,K.L., Bonnycastle,L.L., Willer,C.J., Li,Y., Duren,W.L., Erdos,M.R., Stringham,H.M., Chines,P.S., Jackson,A.U. *et al*. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341–1345.
8. Zeggini,E., Weedon,M.N., Lindgren,C.M., Frayling,T.M., Elliott,K., Lango,H., Timpson,N., Perry,J., Rayner,N., Freathy,R. *et al*. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, **316**, 1336–1341.
9. Frayling,T.M. (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat. Rev. Genet*., **8**, 657–662.
10. Estivill,X. and Armengol,L. (2007) Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet*., **3**, e190.
11. Jacobs,S., Thompson,E.R., Nannya,Y., Yamamoto,G., Pillai,R., Ogawa,S., Bailey,D.K. and Campbell,I.G. (2007) Genome-wide, high-resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using microarrays. *Cancer Res*., **67**, 2544–2551.
12. Tuefferd,M., Bondt,A.D., Wyngaert,I.V.D., Talloen,W., Verbeke,T., Carvalho,B., Clevert,D.-A., Alifano,M., Raghavan,N., Amaratunga,D. *et al*. (2008) Genome-wide copy number alterations detection in fresh frozen and matched FFPE samples using SNP 6.0 arrays. *Genes Chromosomes Cancer*, **47**, 957–964.
13. Baross,A., Delaney,A., Li,I.H., Nayar,T., Flibotte,S., Qian,H., Chan,S., Asano,J., Ally,A., Cao,M. *et al*. (2007) Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data. *BMC Bioinformatics*, **8**, 368.
14. Li,C. and Wong,W. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
15. Nannya,Y., Sanada,M., Nakazaki,K., Hosoya,N., Wang,L., Hangaishi,A., Kurokawa,M., Chiba,S., Bailey,D.K., Kennedy,G.C. *et al*. (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*., **65**, 6071–6079.
16. Van de Wiel,M.A., Picard,F., van Wieringen,W.N. and Ylstra,B. (2010) Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief. Bioinformatics*, **12**, 10–21.
17. Affymetrix. (2007) CNAT 4.0: Copy number and loss of heterozygosity estimation algorithms for the GeneChip human mapping 10/50/100/250/500K array set. *Technical report*. Affymetrix Inc.
18. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
19. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y., Antonellis,K., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
20. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*., **31**, 1–8.
21. Lin,M., Wei,L.-J., Sellers,W.R., Lieberfarb,M., Wong,W.H. and Li,C. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*, **20**, 1233–1240.
22. Korn,J.M., Kuruvilla,F.G., McCarroll,S.A., Wysoker,A., Nemesh,J., Cawley,S., Hubbell,E., Veitch,J., Collins,P.J., Darvishi,K. *et al*. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet*., **40**, 1253–1260.
23. Bengtsson,H., Irizarry,R., Carvalho,B. and Speed,T.P. (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**, 759–767.
24. Bengtsson,H., Wirapati,P. and Speed,T.P. (2009) A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics*, **25**, 2149–2156.
25. Dellinger,A.E., Saw,S.-M., Goh,L.K., Seielstad,M., Young,T.L. and Li,Y.-J. (2010) Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res*., **38**, e105.
26. Hupe,P., Stransky,N., Thiery,J.P., Radvanyi,F. and Barillot,E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA region. *Bioinformatics*, **20**, 3413–3422.

27. Broët,P. and Richardson,S. (2006) Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics*, **22**, 911–918.
28. Andersson,R., Bruder,C.E.G., Piotrowski,A., Menzel,U., Nord,H., Sandgren,J., Hvidsten,T.R., deStähl,T.D., Dumanski,J.P. and Komorowski,J. (2008) A segmental maximum a posteriori approach to genome-wide copy number profiling. *Bioinformatics*, **24**, 751–758.
29. Wang,K., Li,M., Hadley,D., Liu,R., Glessner,J., Grant,S., Hakonarson,H. and Bucan,M. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.
30. Scharpf,R.B., Parmigiani,G., Pevsner,J. and Ruczinski,I. (2008) Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *Ann. Appl. Stat.*, **2**, 687–713.
31. Hochreiter,S., Clevert,D.-A. and Obermayer,K. (2006) A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**, 943–949.
32. Talloen,W., Clevert,D.-A., Hochreiter,S., Amaratunga,D., Bijnens,L., Kass,S. and Göhlmann,H.W.H. (2007) I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data. *Bioinformatics*, **23**, 2897–2902.
33. Talloen,W., Hochreiter,S., Bijnens,L., Kasim,A., Shkedy,Z. and Amaratunga,D. (2010) Filtering data from high-throughput experiments based on measurement reliability. *Proc. Natl Acad. Sci. USA*, **107**, 173–174.
34. Olshausen,B.A. and Field,D.J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607–609.
35. Lewicki,M.S. and Sejnowski,T.J. (2000) Learning overcomplete representations. *Neural Comput.*, **12**, 337–365.
36. Girolami,M. (2001) A variational method for learning sparse and overcomplete representations. *Neural Comput.*, **13**, 2517–2532.
37. Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, **39**, 1–22.
38. Conrad,D.F., Pinto,D., Redon,R., Feuk,L., Gokcumen,O., Zhang,Y., Aerts,J., Andrews,T.D., Barnes,C. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.
39. Dietterich,T.G. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, **10**, 1895–1923.
40. McCarroll,S.A., Kuruvilla,F.G., Korn,J.M., Cawley,S., Nemesh,J., Wysoker,A., Shapero,M.H., deBakker,P.I.W., Maller,J.B., Kirby,A. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.
41. Bourgon,R., Gentleman,R. and Huber,W. (2010) Independent filtering increases detection power for high-throughput experiment. *Proc. Natl Acad. Sci. USA*, **107**, 9546–9551.