Increasing the Discovery Power of -Omics Studies

Djork-Arné Clevert[1] , Andreas Mayr[1], Günter Klambauer[1], Andreas Mitterecker[1], Armand Valsesia[2], Marianne Tuefferd[3], Karl Forner[2], Willem Talloen[3], Jérôme Wojcik[2], Hinrich Göhlmann[3], and Sepp Hochreiter*[1]

[1] Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria; [2] Bioinformatics, Merck Serono SA, Geneva, Switzerland; [3] Functional Genomics, Johnson & Johnson Pharmaceutical R&D, A Division of Janssen Pharmaceutica, Beerse, Belgium

Email: Djork-Arné Clevert - okko@clevert.de; Sepp Hochreiter*- hochreit@bioinf.jku.at;

*Corresponding author

Abstract

Motivation: Current clinical and biological studies apply different biotechnologies and subsequently combine the resulting -omics data to test biological hypotheses. The plethora of -omics data and their combination generates a large number of hypotheses and apparently increases the study power. In contrary to these expectations, the wealth of -omics data may even reduce the statistical power of a study because of a large correction factor for multiple testing. Typically this loss of power at -omics data is caused by an increased false detection rate (FDR) in single measurements like falsely detected DNA copy numbers or falsely identified differentially expressed genes. The false detections are likely to fail the test because they are random and, therefore, are not related to the tested conditions. Thus, a high FDR at the detection level considerably decreases the discovery power of studies, specifically if different -omics data are involved.

Methods: A remedy for suffering from too high FDRs is to filter out putative false detections. We suggest to use probabilistic latent variable models to identify putative false detections by large noise or by measurement inconsistencies across samples. To select such a model, a Bayesian approach starts with the maximum a priori model that assumes no detection and selects the maximum a posteriori model. Hence a detection results in a deviation of the maximal posterior from the maximal prior model measured by the information gain obtained by the data. If this information gain exceeds a threshold then the selected model obtains an informative/non-informative (I/NI) call that indicates a detection. Even if the I/NI call filtering has been successfully applied, it

was not shown that correction for multiple testing after I/NI call filtering still controls the type I error rate. We prove this important property of the I/NI and show that it is independent from commonly used test statistics for null hypotheses. We apply the I/NI call to transcriptomics (gene expression), where the prior model corresponds to constant gene expression level across samples, and to genomics (copy number variation) data, where the prior model corresponds to constant DNA copy number 2 across samples.

Results: On a HapMap data set, where known CNVs have to be re-detected, I/NI call filtering was much more efficient than variance-based filtering. In particular the I/NI call filter outperforms variance-based filters on data with rare events like the CNVs in the HapMap data set. We assessed the efficiency of the I/NI call filter in reducing the FDR on two different cancer cell lines where it reduces the FDR 18 to 22 fold.

Introduction

Currently, bio-medical research moves more and more from hypothesis-driven to data-driven approaches and therefore more high-throughput technologies are utilized in one study. For example such a study may first identify genetic variations (e.g. copy number variations or single nucleotide variants) and then correlate them to the transcriptom (miRNA or mRNA) or may first extract transcriptomics variations (differentially expressed genes) and then correlate them to the proteom or to the metabolom. For these high-throughput technologies like oligonucleotide arrays or next generations sequencing, the number of markers like probes or reads is steadily increasing. Researchers demand more markers because they expect an increase of the study's power as obtained by bio-technological innovations of the last decade. These innovations helped to reveal molecular causes of various diseases like systemic autoimmunity [1], Crohn's disease and type 1 diabetes [2], type 2 diabetes [3], malaria, non-small-cell lung cancer [4], multiple sclerosis, and bipolar disorder [5]. These disease causes were found because more markers allowed to investigate more hypothetical causes from which the true ones are identified by statistical tests.

However, more markers may even result in a loss of power (one minus type II error rate) through correction for multiple testing. The more hypotheses are investigated, the larger is the number of type I errors that are false discoveries due to falsely rejected null hypotheses (p-value below a threshold). Therefore, the type I error must be controlled through correction for multiple testing by bounding (in probability) either the familywise error rate (FWER), i.e. the probability of making a type I error, or the false discovery rate, i.e. the proportion type I errors in a set of rejected hypotheses. Typically a loss of power is caused by an increased false detection rate (FDR) like falsely detected DNA copy numbers or falsely identified differentially expressed genes. The false detections are likely to fail the test because they are not caused by bio-molecular

states but by random noise and, therefore, they are not related to the tested conditions. More markers are prone to increase the FDR because of larger noise for single markers, increased dependencies between the markers, or higher sensitivity to small biological variations. For example gene expression markers that target different iso-forms show larger noise, copy number markers may be related if they target the same DNA fragment, or markers may be sensitive to the local GC content [6]. In the context of integrative analyses of -omics data, the problem of power loss due to high FDR becomes even more apparent because of the combinatorial multiplicity in generating hypotheses (e.g. each CNV may be correlated with each gene). An approach to reduce the FDR and thereby to increase the study's power is to filter out false detections in single -omics data [7, 8]. Such filters have successfully been applied to transcriptomics [7–13]. The filter removes hypotheses that are based on false detections and very likely will not obtain low p-values. Consequently for fewer hypotheses has to be corrected while the hypotheses with low p-values are kept. Thus, the discovery power of the study is increased. Concluding, a stringent filter criterion, which allows for filtering out false detections, is highly desired.

However, not every filter is appropriate to increase the study's power. Correction for multiple testing to control the type I error rate requires a test that produces independently, uniformly distributed p-values of true negatives (null hypotheses) [7, 14, 15]. This must also hold after filtering. Thus, an appropriate filter must fulfill three conditions: (a) it should be dependent on the test statistic for alternative hypotheses to enrich the remaining hypotheses with low p-values, (b) it must not introduce dependencies between hypotheses, and (c) it must be independent of the subsequent test statistic for null hypotheses in order to control the type I error rate [7]. Item (a) assures an increase of the study's power while (b) and (c) ensure control of the type I error rate. We briefly review existing filtering methods in the framework of gene expression and CNV analysis.

Gene-filtering methods. In array-based gene expression analysis filtering is typically done by removing probes (the markers) or probe sets which have a small expected signal-to-noise ratio. The variation across the conditions may be used to estimate the signal strength while fluctuations of the background intensities or non-specific binding effects using mis-match probes allow both to estimate the measurement noise. The Absent/Present call (A/P) [9] for Affymetrix arrays was one of the first methods which estimates the signal strength by a Wilcoxon's signed rank test on perfect-match and mis-match probes. Probe sets with small expected signal are filtered out. McClintick and Edenberg [10] improved the A/P call by additionally filtering out signals with low intensities as they cannot contain large signals. Then Calza et al. [11] proposed the

"Filtering Likely Uninformative Sets of Hybridizations" (FLUSH) method, which treats probes and arrays as fixed effects in a linear model. The FLUSH filter excludes probe sets that have statistically small array-effects (small signals) or large residual variance (large noise). Note, that Affymetrix's most recent arrays no longer contain mis-match probes and therefore neither permit A/P calls nor FLUSH. Later, in Talloen et al. [12] we introduced the Informative/Non-Informative (I/NI) call based on our "Factor Analysis for Robust Microarray Summarization" (FARMS) algorithm [16] (see subsection "FARMS algorithm"). Bourgon et al. [7] found that variance-based filtering can tremendously increase a study's power. To keep control of the type I error after filtering is important to increase the power. Therefore the filter must be independent under the null hypotheses of the test statistic. On the other hand, filter and test statistic must be correlated under alternative hypotheses to enrich the remaining hypotheses with low p-values and to increase the power.

CNV-filtering methods. In copy number variations analysis, false detections, that are falsely identified CNVs, result from random variations which may occur during sample preparation, during DNA fragment extraction, or during the actual measurement. CNV detection in oligonucleotide arrays first segments the data to identify CNVs and then filters the CNVs using criteria such as the length of a CNV in base pairs or the number of probes that a CNV contains [17]. Note that DNA segmentation methods apply the same model to each chromosome location and, therefore, may introduce dependencies between CNV detection values. For example the difference between segmental mean and chromosomal mean in cancer data may depend on the number and sizes of deletions in this chromosome. Therefore p-values of a test using this segmental mean differences in diseased and matched normals may be correlated. To avoid such dependencies Clevert et al. [18], adapted the I/NI call to CNV detection. The I/NI calls were determined by locally independent models without using information from a segmentation algorithm.

We successfully extended the principle of I/NI calls to CNV detection in next generation sequencing data [19]. For sequencing data an I/NI call is present, if read counts in adjacent intervals mutually agree on the copy number. The interval's copy number is determined by the maximum a posteriori mixture component of a mixture of Poissons model where each mixture represents a certain copy number.

Next section introduces the I/NI call filter and proofs some of its properties. Its first subsection briefly summarizes the FARMS algorithm and the I/NI call. Its second subsection is devoted to the properties of the I/NI call filter like its independence of certain test statistics. The last section provides an experimental evaluation of the I/NI call filter on phase 2 data of "The International HapMap Project" and on two tumor genome data sets.

Methods

We introduce the I/NI call filter and show that it is an appropriate filter which fulfills above mentioned conditions (a) to (c). We first review the FARMS algorithm and the I/NI call and show then properties of the I/NI call filter. In particular we show its independence of the test statistic under null hypotheses for permutation invariant test statistics and for the t-test statistic.

Brief Review of the I/NI Call

FARMS Algorithm

The I/NI call filter is derived from the "Factor Analysis for Robust Microarray Summarization" (FARMS) algorithm [16, 18] which is a method to summarize microarray probe set data. Its main idea is to detect a common hidden cause in the measurements assuming independent noise. The probabilistic FARMS model (1) regards that probes measuring the same target (fragment or region) can only be positively correlated, (2) estimates probe-specific characteristics, (3) automatically trades off signal against noise via the latent variable posterior distribution, (4) can adjust the signal/noise trade-off via the priors on the parameters, and (5) allows for Informative/Non-Informative calls (see next subsection).

The vector $\boldsymbol{x}$ of n probes, the probe set, is modeled by probe effects $\boldsymbol{\lambda}$ and a factor z (latent variable or signal) representing the DNA or mRNA fragment concentration. Higher concentrations will lead to higher intensities of the single probes which is modeled by

$$\boldsymbol{x} \;=\; \boldsymbol{\lambda}\, z \;+\; \boldsymbol{\epsilon}\;, \tag{1}$$

where $\boldsymbol{x}, \boldsymbol{\lambda} \in \mathbb{R}^n$ and $z \sim \mathcal{N}(0,1)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$. Here $\boldsymbol{\Psi} \in \mathbb{R}^{n\times n}$ is the diagonal noise covariance matrix resulting from the assumption of independent measurement noise. $\boldsymbol{\epsilon}$ and z are assumed to be statistically independent. The model parameters are the factor loadings $\boldsymbol{\lambda}$ and the noise variance $\boldsymbol{\Psi}$.

Given these assumptions, $\boldsymbol{x}$ is distributed according to the following Gaussian:

$$\boldsymbol{x} \;\sim\; \mathcal{N}\left(\mathbf{0}\;,\; \boldsymbol{\lambda}\boldsymbol{\lambda}^T \;+\; \boldsymbol{\Psi}\right)\;. \tag{2}$$

The covariance matrix of $\boldsymbol{x}$ is decomposed into a signal part $\boldsymbol{\lambda}\boldsymbol{\lambda}^T$ and a noise part $\boldsymbol{\Psi}$. 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 higher fragment concentration and vice versa, therefore noise-free probe measurements must be positively correlated. FARMS integrates this prior knowledge and ensures the

positive correlation of probes by enforcing non-negative components of $\boldsymbol{\lambda}$ through a rectified Gaussian [16] as prior on the components of $\boldsymbol{\lambda}$. Further, the prior prefers models with small factor loadings and, therefore, selected models tend to explain probe variation rather by noise than by a signal. The mean and variance of the prior are hyperparameters that determine how much of the data variance the maximum posterior model explains by signal and how much by noise.

FARMS selects the model parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\Psi}$ by an expectation-maximization algorithm [20] that maximizes the parameter posterior.

Informative/Non-Informative Call

The Informative/Non-Informative call (I/NI-call) has been introduced in [8,12] and measures the information gain of the maximum posterior model over the maximum prior model. The maximum prior model assumes that no signal is present in the data, that is a constant gene expression level or a constant DNA copy number 2 across samples. Therefore the I/NI call is a propensity for a detection through a signal in the data.

We measure the information gain of the maximum posterior model over the maximum prior model by the difference of the signal part of the latent variable. In contrast to information theoretic approaches which are dependent on data scaling, the focus of the I/NI call is on the signal-to-noise ratio which is independent of data scaling. To assess the signal-to-noise ratio, the variance of the latent variable is decomposed into a signal and a noise part. The I/NI call is the noise part, therefore a small I/NI call corresponds to low noise and high signal. The maximum prior model explains the data only by noise, thus the signal variance is zero and the I/NI call one. The I/NI call filter removes detections with large I/NI call values and keeps those with small values that correspond to low noise and high signal.

From the model eq. (2) and the Gaussian z-prior $\mathcal{N}(0,1)$, we can compute the z-posterior $p(z \mid \boldsymbol{x})$ after observing $\boldsymbol{x}$ as

$$
\begin{aligned}
z \mid \boldsymbol{x} \;&\sim\; \mathcal{N}\left(\mu_{z|\boldsymbol{x}} \;,\; \sigma^2_{z|\boldsymbol{x}}\right) \;, \\
\mathrm{E}_{z|\boldsymbol{x}}\left(z\right) \;&=\; \mu_{z|\boldsymbol{x}} \;=\; \left(\boldsymbol{x}\right)^T \;\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}\;\left(1 \;+\; \boldsymbol{\lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}\right)^{-1} \;, \\
\mathrm{E}_{z|\boldsymbol{x}}\left(\left(z \;-\; \mu_{z|\boldsymbol{x}}\right)^2\right) \;&=\; \sigma^2_{z|\boldsymbol{x}} \;=\; \left(1 \;+\; \boldsymbol{\lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}\right)^{-1} \;.
\end{aligned}
\tag{3}
$$

The variance $\mathrm{var}(z)$ of the latent variable z is decomposed into signal and noise part:

$$\begin{aligned} 1 \;=\; \mathrm{var}(z) \;&\approx\; \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{z_i|\boldsymbol{x}_i}\left(z_i^2\right) \\ &=\; \frac{1}{N}\sum_{i=1}^{N} \left(\mathrm{E}_{z_i|\boldsymbol{x}_i}\left(z_i\right)\right)^2 \\ &+\; \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{z_i|\boldsymbol{x}_i}\left(\left(z_i \;-\; \mu_{z_i|\boldsymbol{x}_i}\right)^2\right) \\ &=\; \frac{1}{N}\sum_{i=1}^{N} \mu_{z_i|\boldsymbol{x}_i}^2 \;+\; \sigma_{z|\boldsymbol{x}}^2 \;, \end{aligned} \tag{4}$$

where the noise part is

$$\sigma_{z|\boldsymbol{x}}^2 \;=\; \mathrm{var}(z \mid \boldsymbol{x}) \tag{5}$$

and the signal part

$$\frac{1}{N}\sum_{i=1}^{N} \mu_{z_i|\boldsymbol{x}_i}^2 \;=\; 1 \;-\; \sigma_{z|\boldsymbol{x}}^2 = \; 1 \;-\; \mathrm{var}(z \mid \boldsymbol{x}) \;. \tag{6}$$

According to eq. (3) $\sigma_{z|\boldsymbol{x}}^2$ is independent of $\boldsymbol{x}_i$ and serves as I/NI call in FARMS [12]:

$$\begin{aligned} \mathrm{I/NI} \;&=\; \mathrm{var}(z \mid \boldsymbol{x}) \\ &=\; \sigma_{z|\boldsymbol{x}}^2 \;=\; \left(1 \;+\; \boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda}\right)^{-1} \;. \end{aligned} \tag{7}$$

Low I/NI calls values identify probe sets that contain a high signal and low noise. Therefore, detections for which the I/NI call is beyond a threshold are filtered out.

The I/NI calls has been applied to 30 real-life transcriptomics data and excluded 70 to 99% (in mean 84($\pm$1.5)%) of all probe sets (genes) while never excluding a gene that was known to be biologically meaningful [12].

Properties of the I/NI Call

As mentioned in the introduction an appropriate filter must fulfill three conditions: (a) it should enrich the remaining hypotheses for with p-values, (b) it must not introduce dependencies between hypotheses, and (c) it must be independent of the subsequent test statistic in order to control the type I error rate [7].

Item (a) ensures that the power of study increases because the filter keeps most hypotheses with low p-values while removing many with high p-values. Condition (a) is fulfilled by the I/NI call because false

detections that are filtered out are unlikely to obtain low p-values in a test. We will verify this important property of the I/NI call filter in the experiments (see section "CNV Detection on HapMap").

Item (b), the independence of p-values, follows from the locality of the latent variable model: a separate model is constructed for each hypothesis given by a probe set.

Item (c) ensures that an uniform p-value distribution for the null hypotheses is kept after filtering. This condition prohibits the latent variable model from changing assumptions used in the subsequent test. For example, the t-test assumes that null hypotheses stem from the same Gaussian distribution. That the I/NI call filter complies with condition (c) is shown in Theorem 1 for permutation invariant test statistics and for the t-test statistic T. The theorem guarantees type I error rate control if first hypotheses are filtered by the I/NI call, then are tested, and finally the test result is corrected for multiple testing.

Theorem 1. *For permutation invariant test statistics like the Wilcoxon rank sum statistic and for t-test statistic, the I/NI call filter applied to null hypotheses is independent of the statistic.*

Proof. See Appendix. □

Note, that for equal noise on each probe set, the I/NI call is equivalent to variance filtering. Also for a low noise level relative to the signal, I/NI call is very similar to variance filtering.

Results

We verify that the I/NI call filter reduces the false detection rate (FDR) and quantify the efficiency of the I/NI call by estimating the FDR on HapMap data from the "The International HapMap Project" and on cancer genome data sets.

CNV Detection on HapMap

In this subsection we verify and quantify the FDR reduction by I/NI call filtering. The goal is to identify true CNV regions in Affymetrix SNP 6.0 array data from the "The International HapMap Project" phase 2.

We define as "true CNV regions" those regions which were multiple detected by different platforms in Conrad et al. [21]. In Conrad et al. [21], 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 56bp. From the identified CNVs, random control samples were selected and successfully verified by quantitative PCR. The CNV regions extracted with NimbleGen tiling arrays served to design CNV-typing Agilent CGH arrays comprising 105,000 long oligonucleotide probes. With these Agilent arrays, 4,978 CNVs were detected

on 450 HapMap phase 3 samples and then completed by 59 CNV regions from McCarroll et al. [22]. The third platform, Illumina Infinium genotyping (Human660W), detected CNVs of which 87% were already known from Agilent CGH arrays. Almost all CNVs from Conrad et al. [21] were confirmed by at least two different platforms (NimbleGen tiling arrays, Agilent CGH, or Illumina Human660W). Of these 5,037 CNV regions, we only selected CNV regions from the 60 CEU HapMap phase 2 samples (CEU trios without children). Finally, we obtained 2,515 true CNV regions as reference for our experiment.

For detecting CNV regions we applied the I/NI call filter and a variance-based filter on probe sets summarized by by CRMA_v2 [23] and dChip [24].

Using the true CNVs, we can assess the false detection rate (FDR). To compute the FDR we would need both for the I/NI call filter and the variance-based filter a threshold which trades off the FDR against the true positive rate. To allow a fair comparison of the filtering approaches, we present the CNV detection and filtering 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 1 shows the PRC plots where I/NI call filter has indeed lower FDRs compared to the variance-based filtering methods. The corresponding areas under the precision-recall curves are listed in Table 1. A larger value means that the filter leads to lower FDR averaged over different given recall values. We observed that the FDR was significantly lower with cn.FARMS' I/NI call filter than with variance-based filtering with CRMA_v2 and dChip.

Figure 2 shows I/NI call plots across chromosome 4 for 3-loci and 5-loci regions. The y-axis gives the I/NI call and for both CRMA_v2 and dChip the variance across samples. Filter values are scaled such that the maximum is one. Local calling densities are encoded by blue color shades. True CNVs (reported in Conrad et al. [21]) are marked as light-rose bars and calls at these loci by red circles. A perfect filter would call all true CNVs (red circles at 1) and would not call others (dark-blue background at 0). The I/NI call filter 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 85mb in the upper I/NI call filter panel, indicate verified CNVs which were detected by one method, in this case I/NI call filter, but not by variance-based filters. The I/NI call filter reduces the FDR more efficiently than variance-based filtering methods.

CNV detection at HapMap data involves highly unbalanced data sets because only few individuals show copy numbers different from 2 [25]. Only few samples contribute to the signal, therefore variance-based

filtering methods struggle at distinguishing locations with a signal from locations without a signal. However, the I/NI call filter relies on the signal variance and, therefore, outperforms the variance-based filters on unbalanced data sets.

CNAs in Tumor Genomes

In contrast to heritable copy number variation (CNVs), are copy number aberrations (CNAs) the result of genomic instability in somatic tumor tissue [26]. In this subsection we assess the efficiency of the I/NI call in reducing the false detection rate (FDR) on a tumor genome association study from Chiang et al. [27]. Following cell lines from the American type culture collection were included into this study: HCC1143 (breast ductal carcinoma) with matched normal HCC1143BL and HCC1954 (breast ductal carcinoma) with matched normal HCC1954BL. Affymetrix Genome-Wide Human SNP Arrays 6.0 were used to measure 21 replicates of HCC1143, 21 replicates of HCC1143BL, 13 replicates of HCC1954, and 11 replicates of HCC1954BL.

To reduce the false positives at CNA detection with for Affymetrix SNP 6 arrays, 8 consecutive probe sets were merged to a segment in Chiang et al. [27]. Thereafter the segmentation software `DNAcopy` which implements the Circular Binary Segmentation algorithm [28] found 454 segments in HCC1954 and 300 segments in HCC1143. From these candidate segments, 153 CNA were detected in the HCC1954 cell line and 93 in the HCC1143 cell line. The number of false detections were estimated on matched normals to be 22 for HCC1954 and 16 for HCC1143.

We also used `DNAcopy` for segmentation but increased the resolution by joining 2 consecutive probe sets instead of 8. The I/NI call filter was based on a 3 probe set FARMS model which was selected without any information from the segmentation step. The I/NI call filter removed segments having a median I/NI call above a threshold.

Segments for which the I/NI call was larger than 0.01 are filtered out, where the threshold of 0.01 was adjusted on the matched normal samples to obtain about 10 false positives. Table 2 shows the results without filtering as reported in Chiang et al. [27] and those obtained by I/NI call filtering. The number of detections with I/NI call filtering increased 5 fold because of the higher resolution during segmentation. Despite this increase of detections, I/NI call filtering reduced the FDR 18 to 22 fold compared to the original results without filtering.

Conclusion

We showed that the I/NI call filter is an appropriate filter. First, we have proven that after I/NI call filtering type I error control by correction for multiple testing is still valid if the test is permutation invariant or t-test related. In experiments the I/NI call filter was found to enrich hypotheses that pass the filter with true detections. We found that the I/NI call filter outperformed variance-based filtering methods on data with rare events. The experiments further showed that the false detection rate reduces up 18 to 22 fold compared to the detections without filtering.

The theoretical properties of the I/NI call filter together with its experimental verified efficiency to reduce the FDR suggest it as an ideal filtering tool to increase the power of bio-medical studies.

Author's contributions

All authors contributed to analysis of the data. SH proved the independence of the I/NI filter. DAC and SH drafted the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that there are no competing interests.

Appendix: Proof of Theorem 1

We proof Theorem 1, where we assume that probe sets are summarized by Robust Multi-array Average (RMA) and single probes are Gaussian distributed. First we need some results on summarization with RMA for Gaussian noise and for probe sets containing an additional signal. These results are given in the following lemmas.

RMA Summarization of Gaussian Probes

Robust Multi-array Average (RMA) [29, 30] summarizes a probe set by median polish. Median polish as used in RMA first removes the median of each probe and thereby levels the probe distributions by centering the probes. Then RMA basically computes the median of the probe set.

We assume a probe set with $(2n+1)$ probes. According to Chu [31], for $(2n+1)$ samples drawn from a normal distribution with density $f(x) \sim \mathcal{N}(\xi, \sigma)$ and cumulative distribution function $F(x)$, the distribution

of samples' median is

$$p(x) \;=\; \frac{(2n+1)!}{n!\;n!}\;\left(F(x)\;(1-F(x))^n\;\right.\;f(x)\;. \tag{8}$$

According to Chu [31], $p(x)$ is asymptotically normal which is is formulated in following lemma.

Lemma 1. *For $2n+1$ samples randomly drawn according to a normal distribution $f(x) \sim \mathcal{N}(\xi, \sigma)$, the sample median is asymptotically normal distributed with mean ξ and variance*

$$\sigma_n^2 \;=\; \frac{1}{4\;f^2(\xi)\;(2n+1)}\;. \tag{9}$$

Proof. This lemma is shown in Chu [31]. □

Chu [31] states that the distribution of the median "tends 'rapidly' to normality." Using the bounds in Chu [31], for a probe set of 16 probes (a standard Affymetrix probe set), the factor deviating from a normal distribution is between 0.986 and 1.023.

RMA Summarization of Probe Sets with a Signal

Now we consider summarization for a probe set with correlated probes where the correlation is caused by varying concentrations of the mRNA they target. We assume a signal ξ_k for sample k, where ξ_k is the intensity of the probes caused by a specific mRNA concentration in sample k. The probe intensities also contain white Gaussian measurement noise $\mathcal{N}(0, \sigma)$, therefore the median of the probes follows for fixed ξ_k the Gaussian distribution $\mathcal{N}(\xi_k, \sigma_n)$. The signal ξ_k is drawn from a Gaussian signal distribution $\mathcal{N}(\mu_s, \sigma_s)$, where (μ_s, σ_s) determine the signal strength. This setting correlates the probes via a mRNA signal, where (μ_s, σ_s) determines the strength of correlation.

Another way to introduce signal into the probes of a probe set would have been to scale the signal for each sample. However, this approach is equivalent to above approach. Assume that a multiplicative factor ρ_k, which scales the reference signal μ, follows a Gaussian $\mathcal{N}(\mu_r, \sigma_r)$. Then the new mean values ξ_k follow a Gaussian $\mathcal{N}(\mu\;\mu_r, \mu^2\;\sigma_r)$. This reveals the equivalence to above approach by setting $\mu_s = \mu\;\mu_r$ and $\sigma_s = \mu^2\;\sigma_r$. Introducing correlations in other ways would not change the results but the convolution for non-Gaussian signal distributions might be more complicated.

Because the signal distribution determines the mean of the median distribution, the distribution of the median is the convolution of two Gaussian distributions $\mathcal{N}(\mu_s, \sigma_s)$ and $\mathcal{N}(0, \sigma_n)$. The result of this convolution is presented in the next lemma.

Lemma 2. *If the signal of probes of a probe set is drawn from a Gaussian distribution $\mathcal{N}(\mu_s, \sigma_s)$ and the noise of the probes is $\mathcal{N}(0, \sigma_n)$, then the median distribution is*

$$\mathcal{N}(\mu_s, \sigma_x) \ , \tag{10}$$

where

$$\begin{aligned} \sigma_x^2 &= \sigma_s^2 \ + \ \sigma_n^2 \ = \ \sigma_s^2 \ + \ \frac{1}{4 \ f^2_{\mathcal{N}(0,\sigma)}(\xi) \ (2n+1)} \\ &= \ \sigma_s^2 \ + \ \frac{\pi \ \sigma^2}{2 \ (2n+1)} \ , \end{aligned} \tag{11}$$

Proof. The lemma follows from Lemma 1 which says that the distribution of the median is $\mathcal{N}(\xi_k, \sigma_n)$ for fixed ξ_k. If ξ_k is drawn according to $\mathcal{N}(\mu_s, \sigma_s)$ then the median distribution is obtained by the convolution of $\mathcal{N}(0, \sigma_n)$ and $\mathcal{N}(\mu_s, \sigma_s)$. The distribution given in the lemma is the result of this convolution of two Gaussians. □

Other Summarization Methods

That the summarization value is a Gaussian if signal and probes are normally distributed also holds for other summarization methods like mean, MAS5, FARMS, dChip (model-based expression indexes — MBEI). The summarized value for a null hypothesis is a Gaussian composed of a signal and a noise part.

Of course, summarization by computing the mean of the probes results in a Gaussian that comprises signal and noise. Also weighted, robust averages have similar behavior as the mean or the median like Tukey biweight which is used by MAS5 summarization.

Our FARMS summarization method (see main text) supplies the signal part as the summarized value. The signal is estimated by a latent variable model where both the noise and the latent variable are assumed to be normally distributed. For the null hypotheses the true signal is normally distributed and therefore FARMS will recover this signal as a normally distributed latent variable $(z \mid \boldsymbol{x})$. If a minimal signal variance is assured for FARMS summarization, then it can be used together with the I/NI call filter.

Similar statements hold for least square estimates as used by model-based expression indexes (MBEI of dChip), where a normally distributed signal is also recovered from null hypotheses.

Independence of I/NI Filter and Test Statistic for Null Hypotheses

The Informative/Non-Informative (I/NI, [12]) call tries to access the noise part σ^2 of the overall variance by $\text{var}(z \mid \boldsymbol{x})$. Thus, the amount of signal σ_s in the probe set is estimated.

More specifically, the I/NI call is

$$\text{var}(z \mid \boldsymbol{x}) \;= \left(\frac{(2n+1)\;\sigma_s^2}{\sigma^2} \;+\; 1 \right)^{-1} \;<\; 0.5\;, \tag{12}$$

if all probes have the same noise and signal. Here $\frac{\sigma_s^2}{\sigma^2}$ is the squared signal-to-noise ratio for a single probe.

Probe sets are normally distributed regardless of signals. However, probes sets containing a signal result in larger variance because the signal variance σ_s is added to the variance of the median according to Lemma 2.

We define permutation invariant test statistic for m samples where we use the notation in Bourgon et al. [7].

Definition 1. *A test statistic U^{II} is* permutation invariant *if for fixed $\boldsymbol{Y}_i \in \mathbb{R}^m$, $i \in \mathcal{H}_0$, and Π drawn uniformly from S_m (the set of all permutations on m elements), the distribution of the test statistic $U^{II}(\boldsymbol{Y}_i)$ is equal to the distribution of $U^{II}(\Pi(\boldsymbol{Y}_i))$.*

Here $\mathcal{H}_0$ is the set of null hypotheses.

Now we can formulate our main theorem that for permutation invariant test statistics and for the t-test statistic T, the I/NI call filter applied to null hypotheses is independent of the statistic. The theorem guarantees type I error rate control if applying correction for multiple testing.

Theorem 1. *For permutation invariant test statistics like the Wilcoxon rank sum statistic and for t-test statistic T, the I/NI call filter applied to null hypotheses is independent of the statistic.*

Proof. First we note that the I/NI call for one probe set does not dependent on another probe set as the models are independently selected for each probe set.

A) *Permutation invariant test statistics*:

For permutation invariant test statistics the statement follows directly from the permutation invariance of the I/NI call filter. The I/NI call is permutation invariant because the I/NI call model selection objective, the *a posteriori* of the parameters, is independent of the permutation of the samples. Further, the implementation of the algorithm uses only the data covariance matrix [16] which is independent of permutations of the samples.

All assumptions on the filter of the the proposition "Marginal Independence: Permutation Invariance" in Bourgon et al. [7] are fulfilled. The independence between the I/NI call filter and permutation invariant test statistics is shown.

B) *t-test statistic* T:

As pointed out by Bourgon et al. [7] in their supplementary, the test statistics T for the t-test is invariant to scaling and shifting of the mean. If the noise level σ is equal for each probe set, then I/NI call is equivalent to variance filtering because only the signal variance σ_s determines the overall variance. The more interesting case is where signal and noise differ at each probe set, thus variance filtering and I/NI calls yield different results.

For probe set i the signal is drawn from a Gaussian distribution $\mathcal{N}(\mu_{si}, \sigma_{si})$. According to Lemma 2 the RMA summarized data follows the Gaussian $\mathcal{N}(\mu_{si}, \sigma_{xi})$, where $\sigma_{xi} = \sqrt{\sigma_{si}^2 + \frac{\pi\ \sigma_i^2}{2\ (2n+1)}}$). The signal strength and the noise level $(\mu_{si}, \sigma_{si}, \sigma_i)$ are assumed to be drawn from some distribution $P_{(\mu_{si},\sigma_{si},\sigma_i)}$.

The data $\boldsymbol{Y}_i$ can be generated by first drawing $(2n+1)$ samples from a standard normal distribution giving $\boldsymbol{X}_i \in \mathbb{R}^{2n+1}$, where $P_{\boldsymbol{X}_i} \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{2n+1})$ with $\boldsymbol{0}$ as the $(2n+1)$-dimensional zero vector and $\boldsymbol{I}_{2n+1}$ as the $(2n+1)$-dimensional identity matrix. Then $\boldsymbol{X}_i$ is scaled by $\sigma_{xi} = \sqrt{\sigma_{si}^2 + \frac{\pi\ \sigma_i^2}{2\ (2n+1)}}$ and shifted component-wise by μ_{si}. The shifting and scaling values are drawn from $P_{(\mu_{si},\sigma_{si},\sigma_i)}$ which is independent from $P_{\boldsymbol{X}_i}$.

For a null hypothesis $i \in \mathcal{H}_0$, we assume that both distributions $P_{\boldsymbol{X}_i}$ and $P_{(\mu_{si},\sigma_{si},\sigma_i)}$ are independent of the conditions $\boldsymbol{C}$.

For showing the independence of filtering U^I and test statistic U^{II}, we are interested in the probability of the event $\{U_i^I \in \mathcal{A}, U_i^{II} \in \mathcal{B}\}$. Here we define $U_i^I(\boldsymbol{Y}) = \mathrm{var}(z \mid \boldsymbol{x})(\boldsymbol{Y})$ with $\mathcal{A} = \{u \mid u < 0.5\}$ and $U_i^{II}(\boldsymbol{Y}) = T(\boldsymbol{Y}, \boldsymbol{C})$ for t-test statistic T, conditions $\boldsymbol{C}$, and $\mathcal{B} = \{u \mid u > \theta\}$. Let $\delta_{\mathcal{A}}$ and $\delta_{\mathcal{B}}$ be indicator functions for $\mathcal{A}$ and $\mathcal{B}$.

We consider a probe set $\boldsymbol{Y}_i$ for which $i \in \mathcal{H}_0$ (a true null hypothesis).

$$
\begin{aligned}
& P\big(U_i^I \in \mathcal{A}, U_i^{II} \in \mathcal{B}\big) \qquad (13)\\
&= \int \delta_{\mathcal{A}}\big(U^I(\boldsymbol{Y}_i)\big)\ \delta_{\mathcal{B}}\big(U^{II}(\boldsymbol{Y}_i)\big)\ dP_{\boldsymbol{Y}_i}\\
&= \int\int \delta_{\mathcal{A}}\big(U^I(\mu_{si}\ \boldsymbol{1}\ +\ \boldsymbol{X}_i\ \sigma_{xi})\big)\\
&\quad\ \delta_{\mathcal{B}}\big(U^{II}(\mu_{si}\ \boldsymbol{1}\ +\ \boldsymbol{X}_i\ \sigma_{xi})\big)\ dP_{\boldsymbol{X}_i}\ dP_{(\mu_{si},\sigma_{si},\sigma_i)}\\
&= \int\int \delta_{\mathcal{A}}\big(U^I(\sigma_{si}, \sigma_i)\ \delta_{\mathcal{B}}\big(U^{II}(\boldsymbol{X}_i)\ dP_{\boldsymbol{X}_i}\ dP_{(\mu_{si},\sigma_{si},\sigma_i)}\\
&= \int \delta_{\mathcal{A}}\big(U^I(\sigma_{si}, \sigma_i)\ dP_{(\mu_{si},\sigma_{si},\sigma_i)}\ \int \delta_{\mathcal{B}}\big(U^{II}(\boldsymbol{X}_i)\big)\ dP_{\boldsymbol{X}_i}\\
&= P\big(U_i^I \in \mathcal{A}\big)\ P\big(U_i^{II} \in \mathcal{B}\big)\ ,
\end{aligned}
$$

where

$$\sigma_{xi} = \sqrt{\sigma_{si}^2 + \frac{\pi\ \sigma_i^2}{2\ (2n+1)}}) \tag{14}$$

and $\mathbf{1}$ is the vector of ones with length n. The equality of the 3rd/4th line to the 5th line is obtained by the shift and scale invariance of U^{II} and the fact that U^I depends only on σ_{si} and σ_i. □

References

1. Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, Heward JM, Gough SC, de Smith A, et al.: **FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity**. *Nature Genet.* 2007, **39**(6):721–723.
2. Wellcome-Trust-Case-Control-Consortium: **Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls**. *Nature* 2010, **464**(7289):713–720.
3. Frayling TM: **Genome-wide association studies provide new insights into type 2 diabetes aetiology**. *Nature Rev. Genet.* 2007, **8**(9):657–662.
4. Boutros P, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao MS, Penn LZ, Jurisica I: **Prognostic gene signatures for non-small-cell lung cancer**. *Proc. Natl. Acad. Sci. USA* 2009, **106**(8):2824–2828.
5. Estivill X, Armengol L: **Copy Number Variants and Common Disorders: Filling the Gaps and Exploring Complexity in Genome-Wide Association Studies**. *PLoS Genet.* 2007, **3**(10):e190.
6. Marioni J, Thorne N, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews T, Stranger B, Lynch A, et al.: **Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization**. *Genome Biol.* 2007, **8**(10):R228.
7. Bourgon R, Gentleman R, Huber W: **Independent filtering increases detection power for high-throughput experiment**. *Proc Natl Acad Sci USA* 2010, **107**(2):9546–9551.
8. Talloen W, Hochreiter S, Bijnens L, Kasim A, Shkedy Z, Amaratunga D, Göhlmann H: **Filtering data from high-throughput experiments based on measurement reliability**. *Proc. Natl. Acad. Sci. USA* 2010, **107**(46):173–174.
9. Liu W, Mei R, Di X, Ryder T, Hubbell E, Dee S, Webster T, Harrington C, Ho M, et al.: **Analysis of high density expression microarrays with signed-rank call algorithms**. *Bioinformatics* 2002, **18**(12):1593–1599.
10. McClintick J, Edenberg H: **Effects of filtering by Present call on analysis of microarray experiments**. *BMC Bioinformatics* 2006, **7**:49.
11. Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, Pawitan Y: **Filtering genes to improve sensitivity in oligonucleotide microarray data analysis**. *Nucleic Acids Res.* 2007, **35**(16):e102.
12. Talloen W, Clevert DA, Hochreiter S, Amaratunga D, Bijnens L, Kass S, Göhlmann HWH: **I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data**. *Bioinformatics* 2007, **23**(21):2897–2902.
13. Kasim A, Lin D, Sanden SV, Clevert DA, Bijnens L, Göhlmann H, Amaratunga D, Hochreiter S, Shkedy Z, Talloen W: **Informative or noninformative calls for gene expression: a latent variable approach**. *Stat. Appl. Genet. Molec. Biol.* 2010, **9**.
14. Fodor AA, Tickle TL, Richardson C: **Towards the uniform distribution of null P values on Affymetrix microarrays**. *Genome Biol.* 2007, **8**(5):R69.
15. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci USA* 2003, **100**(16):9440–9445.
16. Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for Affymetrix probe level data**. *Bioinformatics* 2006, **22**(8):943–949.
17. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant S, 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 Res.* 2007, **17**:1665–1674.
18. Clevert DA, Mitterecker A, Mayr A, Klambauer G, Tuefferd M, Bondt AD, Talloen W, Göhlmann H, Hochreiter S: **cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate**. *Nucleic Acids Res.* 2011, **39**(12):e79.
19. Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S: **cn.MOPS: mixture of Poissons for discovering copy number variations in next generation sequencing data with a low false discovery rate**. *Nucleic Acids Res.* 2012, **40**.

20. Dempster AP, Laird NM, Rubin DB: **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *J. Roy. Statistical Society B* 1977, **39**:1–22.

21. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, et al.: **Origins and functional impact of copy number variation in the human genome**. *Nature* 2010, **464**(7289):704–712.

22. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PIW, Maller JB, et al.: **Integrated detection and population-genetic analysis of SNPs and copy number variation**. *Nature Genet.* 2008, **40**(10):1166–1174.

23. Bengtsson H, Wirapati P, Speed TP: **A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6**. *Bioinformatics* 2009, **25**(17):2149–2156.

24. Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, Li C: **dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data**. *Bioinformatics* 2004, **20**(8):1233–1240.

25. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, Kasim A, Khamiakova T, Sanden SV, Lin D, et al.: **FABIA: factor analysis for bicluster acquisition**. *Bioinformatics* 2010, **26**(12):1520–1527.

26. Albertson D, Collins C, McCormick F, Gray J: **Chromosome aberrations in solid tumors**. *Nature Genet.* 2003, **34**:369–376.

27. Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES: **High-resolution mapping of copy-number alterations with massively parallel sequencing**. *Nat. Methods* 2009, **6**:99–103.

28. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data**. *Biostatistics* 2004, **5**(4):557–572.

29. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249–264.

30. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data**. *Nucleic Acids Res.* 2003, **31**(4):1–8.

31. Chu JT: **On the Distribution of the Sample Median**. *Ann. Math. Statist.* 1955, **26**:112–116.

32. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov J: **GenePattern 2.0**. *Nature Genet.* 2006, **38**(5):500–501.

Tables

Table 1: Area under the precision-recall curves on HapMap SNP 6.0 arrays for I/NI call filter and variance-based filters based on CRMA_v2 and dChip at detecting previously multiple confirmed CNVs reported in Conrad et al. [21]. A larger value means that the filter leads to a lower FDR averaged over different given recall values. "Area under the PRC for combined loci of" reports the area under the precision-recall curves for different number of combined loci. Note, that large windows can increase the FDR again because CNV regions are overestimated. The I/NI call filter clearly outperforms variance-based filters.

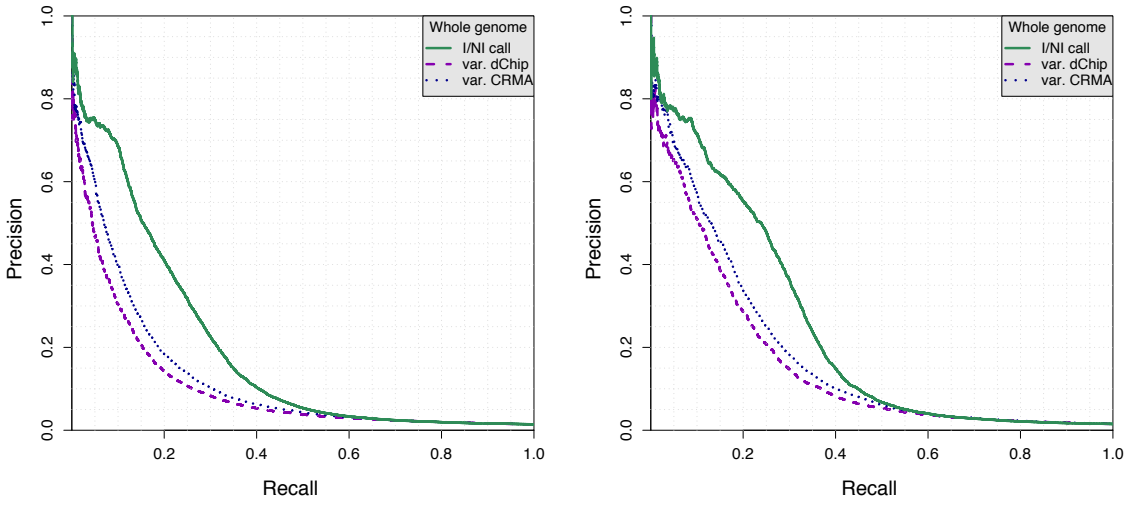
	Area under the PRC for combined loci of			
Method	3	4	5	7
I/NI call filter	0.20	0.22	0.24	0.26
variance-based filter CRMA_v2	0.13	0.16	0.18	0.21
variance-based filter dChip	0.11	0.14	0.16	0.19

Table 2: Results for CNA detection filtering approaches with Affymetrix SNP 6.0 arrays on the breast cancer data set from Chiang et al. [27]. The authors used the GenePattern software [32] and call segments by DNAcopy. The I/NI call filter calls segments found by DNAcopy (with relaxed parameters undo.SD=1 and 2 consecutive probe sets) as detections if the mean of I/NI calls within the segment is larger than 0.01. "called" is the number of called segments. "FP" and "FDR" is the number of falsely detected segments and the false detection rate on the normal cell lines, respectively. The I/NI call filter reduces the FDR 18 to 22 fold.

	I/NI call filter			no filtering			
	called	FP	FDR	called	FP	FDR	ratio
HCC1954	473	3	6.0e-03	153	22	1.4e-01	22
HCC1143	419	4	9.5e-03	93	16	1.7e-01	18

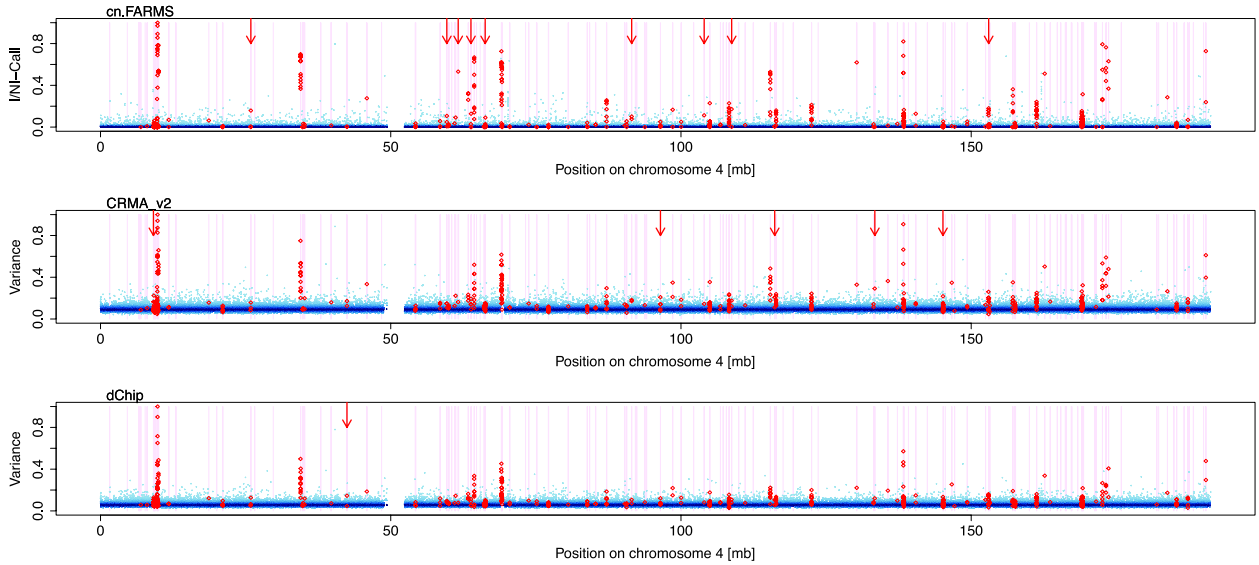
Figures

Figure 1: Precision-recall curves (PRCs) on HapMap SNP 6.0 arrays for I/NI call filter and variance-based filters based on CRMA_v2 and dChip at detecting previously multiple confirmed CNVs reported in Conrad et al. [21]. 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. Panel (A) and panel (B) gives the PRC for the whole genome for 3 loci and for 5 loci, respectively. I/NI call filter (solid green) has a clear advantage over variance-based filters using dChip (dashed purple) and CRMA_v2 (dotted blue). I/NI call filter has a considerable lower FDR compared to variance-based filtering.

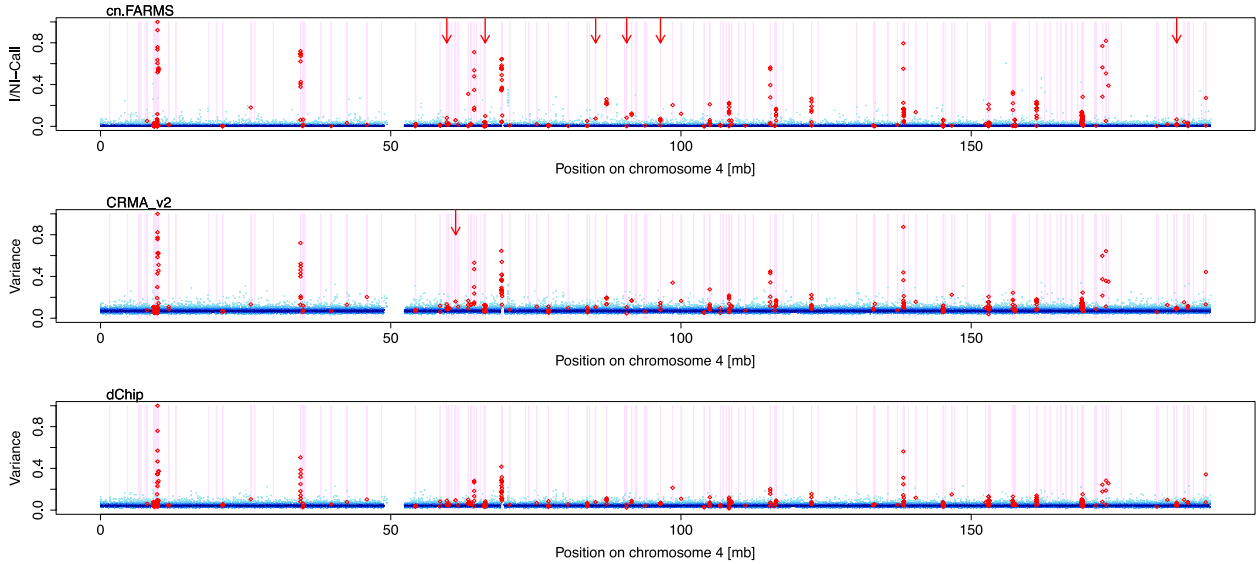


(a) (b)

Figure 2: **(a)**The y-axis gives the I/NI call estimated by cn.FARMS and for both CRMA_v2 and dChip it gives the value of the variance-based filter. Filter values are scaled such that the maximum is one. Local filter value densities are encode by blue color shades. True CNVs (reported in Conrad et al. [21]) are marked as light-rose bars and calls at these loci by red circles. A perfect filtering method would call all true CNVs (red circles at 1) and does not call others (dark-blue background at 0). The I/NI call filter 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 85mb in the upper I/NI call filter panel, indicate verified CNVs which were detected by one method, in this case I/NI call filter, but not by variance-based filters. The I/NI call filter reduces the FDR more efficiently than variance-based filtering methods. **(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, I/NI call filter leads to a lower FDR than variance-based filtering methods.



(a)



(b)