BMC Bioinformatics

BioMed Central

Research article

Open Access

Integrated analysis of DNA copy number and gene expression microarray data using gene sets

Renée X Menezes*[1,2,3,4], Marten Boetzer[1], Melle Sieswerda[1], Gert-Jan B van Ommen[1,4] and Judith M Boer*[1,3,4]

Address: [1]Center for Human and Clinical Genetics, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands, [2]Pediatric Oncology Laboratory, Erasmus Medical Center, Rotterdam, The Netherlands, [3]BioRange, Netherlands Bioinformatics Centre, Nijmegen, The Netherlands and [4]Center for Medical Systems Biology, Leiden, The Netherlands

Email: Renée X Menezes* - r.x.menezes@lumc.nl; Marten Boetzer - mboetzer86@hotmail.com; Melle Sieswerda - melle.sieswerda@gmail.com; Gert-Jan B van Ommen - gjvo@lumc.nl; Judith M Boer* - j.m.boer@lumc.nl

* Corresponding authors

Published: 29 June 2009

BMC Bioinformatics 2009, **10**:203 doi:10.1186/1471-2105-10-203

Received: 16 December 2008
Accepted: 29 June 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/203

Abstract

Background: Genes that play an important role in tumorigenesis are expected to show association between DNA copy number and RNA expression. Optimal power to find such associations can only be achieved if analysing copy number and gene expression jointly. Furthermore, some copy number changes extend over larger chromosomal regions affecting the expression levels of multiple resident genes.

Results: We propose to analyse copy number and expression array data using gene sets, rather than individual genes. The proposed model is robust and sensitive. We re-analysed two publicly available datasets as illustration. These two independent breast cancer datasets yielded similar patterns of association between gene dosage and gene expression levels, in spite of different platforms having been used. Our comparisons show a clear advantage to using sets of genes' expressions to detect associations with long-spanning, low-amplitude copy number aberrations. In addition, our model allows for using additional explanatory variables and does not require mapping between copy number and expression probes.

Conclusion: We developed a general and flexible tool for integration of multiple microarray data sets, and showed how the identification of genes whose expression is affected by copy number aberrations provides a powerful approach to prioritize putative targets for functional validation.

Background

Tumor cells accumulate genetic damage, including changes in DNA copy number, sequence and methylation, resulting in the dysfunctioning of key regulators [1]. The advent of microarray technology has allowed genome-wide monitoring of these molecular changes at the DNA and RNA level. Gene expression profiling has facilitated classification of cancers into biologically and clinically distinct categories [2-7]. High-resolution array-based comparative genomic hybridization (array-CGH) has allowed the delineation of recurrent DNA copy number alterations in tumors [8-10]. Gene dosage changes play an important role in tumor development; oncogenes may be enhanced by DNA amplification and tumor suppressor genes may be inactivated by a physical deletion. Therefore, integrated analysis of both copy number and gene expression microarray data could give additional information about the role of copy number alterations in the development of cancer.

Combined analysis of DNA copy number and gene expression microarrays of the same or similar tumor samples has revealed a major and direct effect of allelic imbalance on gene expression in a variety of cancer types, including breast [11,12], pancreatic [13], colorectal [14], skin [15], head and neck [16,17], prostate [18], multiple myeloma [19], and lung [20] cancer. On a global level, 40–60% of the genes in higher level amplifications showed elevated expression, while circa 10% of highly overexpressed genes were amplified [11,12]. In low-level copy number aberrations, only about 10% of the genes have been reported to show concordant changes in gene expression [11,12,21].

Several approaches have been described to identify those genes whose expression levels are most significantly associated with copy number changes of the corresponding genomic region.

In the context of natural copy number variation in human populations, Stranger and co-authors [22] used a linear regression model to study associations between gene expression and copy number within a 2 Mb window. For the analysis of tumor microarray data, some authors performed a simultaneous exploratory analysis of the different microarray datasets, ordered along the genome, to search for regions where both copy number and gene expression are affected [12,14,23,24], or gene expression and DNA methylation [25]. While this can be clarifying if an effect is found, due to the small effect sizes and the often low signal-to-noise ratio in array data this approach tends to be inefficient. For example, a two-fold change in DNA copy number was observed to be accompanied on average by 1.5-fold changes in mRNA levels in breast tumors [12].

Other cancer studies classified samples according to the presence of chromosomal abnormalities, and subsequently tested for differences in gene expression between altered and unaltered samples. Some studies use a genewise test statistic similar to the Student's t-statistic [11,13,16,19] or a one-sided Wilcoxon rank-sums test [25-27]. Garraway and co-authors [15] used supervised analysis looking for gene expression differences between cell lines with and without 3p amplification. Adler and co-authors [28] used a classification approach as the first step in their stepwise linkage analysis of microarray signatures, where they test for differences in copy number between groups of breast cancer samples with and without the wound expression signature. While known and novel tumor-related genes were identified, these approaches may be unable to detect associations between low-level copy number changes and expression variation due to the categorization.

Low-level gains and losses, representing the most common types of genetic alterations in most cancers, were shown to have a significant influence on expression levels of genes in the regions affected, but these effects were more subtle on a gene-by-gene basis [11,21]. However, the impact of low-level gains on the dysregulation of gene expression patterns in cancer may be equally important if not more important than that of high-level amplifications [11-13]. Therefore, the search for DNA regions that might be involved in the initiation and progression of cancer must be powerful enough to detect subtle gene-specific effects that are possibly consistent across many genes. Moreover, the analysis method must take into account the high-dimensionality of the problem, and provide careful control of the error.

We propose to look for associations between copy number and expression not only using individual genes, but also using gene sets. Such a model can improve the power to detect associations, as neighbouring genes may also display association. If different microarray platforms are used to measure copy number and expression, it involves less arbitrariness because no mapping between copy number and expression probes is necessary. To illustrate these points, we first run a simulation study and then apply our model to two publicly-available breast cancer datasets. We will show that the use of gene sets is relevant not only when studying the impact of large-amplitude copy number changes (more than one copy gained or lost), but also in case of more subtle changes, either of low amplitude or spanning a small (<1 Mb) genomic region. The discussion that follows includes other possible applications and useful extensions.

Results

We wish to find which individual copy number changes affect gene expression levels within the same chromosomal region. For this, we propose to model copy number as a function of the expression levels of many genes at the same time. Statistically significant associations are indicated by a significant p-value for the copy number probe, and the genes with expression levels the most associated with this outcome are prioritized in a heatmap. We evaluate the power of this model in particular experimental setups via a simulation study below. After this, we apply both the gene-set model (2) and the gene-to-gene model (1) to experimental datasets.

Simulation study

In order to illustrate the power of the model to identify association patterns, we run a simulation study. We assume for simplicity that both copy number and expression measurements are obtained using the same arrays, so that there is a one-to-one correspondence between them. The study is designed to represent various situations commonly encountered in practice, where typically 10–50% of the samples display mild copy number effects spanning a sizable genomic region (here 26% of the total probes), such as part of a chromosome arm, or strong copy number effects spanning a small region (here between 2 and 8% of the total probes), typical of amplifications. Each probe's expression level is assumed to be a function of its own copy number, with various degrees of association. Key parameters were estimated from publicly available datasets, amongst them the amount of variability of copy number and expression measurements, as well as the distribution of the associations between copy number and expression. Sample sizes of 25, 50 and 100 are considered. We evaluate the results by producing receiver-operating characteristic (ROC) curves of the regional model for each case, and consider that an effect is detectable if there is power of at least 60% to detect it using an FDR of 10%. For more details about the study setup, see Appendix. The results are reassuring, as the ROC curves in figure 1 illustrate. The association most reliably detected is the amplification (log-ratio 2) spanning 20 probes, detectable in all situations. But as the total region size increases the proportion of affected probes decreases, with the effect becoming diluted, in particular if the sample size is small (25).

The effect least reliably detected is the one-copy gain due to the small amplitude (log-ratio 0.5): it is only detectable when 50% of 100 samples are affected and the region spans at least 500 probes. The one-copy loss is detectable in all situations with sample size at least 50. Note that, because the mild effects involve a fixed proportion of 26% of the probes, the power to detect them increases with the region length.

In practice, the amplitudes of recurrent copy number changes vary more, therefore associations between continuous copy number measurements and gene expression levels can also be detected for smaller sample sizes and less-frequent aberrations.

Breast cancer I: Pollack

Pollack and co-authors [12] produced and were the first to analyse this dataset, consisting of copy number and expression array data for 37 breast tumors and 4 breast-tumor cell lines, produced on the same cDNA microarrays. The datasets pre-processed by the authors were downloaded, consisting of log-ratios per gene. The curated dataset involved 4696 genes with both copy number and expression log-ratios available. Here we wish to investigate if there are copy number changes that affect expression levels within the same chromosome arm. We report results controlling the FDR at 10%. Considering the copy number data on a continuous scale, we fit the gene-set model explaining measurements for each copy number probe by the expression levels of all probes on the same chromosome arm. This model found evidence of association between copy number and expression levels, with in total 343 probes significant out of 4696 (figure 2). The gene-to-gene model can be applied directly as the same microarray was used both for copy number and for expression. This model selected 272 clones, 114 of which also selected by the gene-set model (see figures 1 and 2A in Additional Files 1 and 2). This shows that the gene-to-gene model finds some of the same effects as the gene-set model, but each one finds unique effects: 158 by the gene-to-gene model, 229 by the gene-set model.

Some of the effects found by both models identify copy-number aberrant regions such as those on 8p, 8q, 17q and 20q (figure 2). In particular, 17q includes genomic regions with high-amplitude copy number effects (amplifications) highly associated with resident-genes' expression levels (figure 3A). One of these regions contains the ERBB2 gene, another one contains the TRAF4 gene. These genes were also found by Pollack [12], and are known to be involved in breast cancer development. Other candidate oncogenes were identified in gained regions 8p11-12 (including LSM1, BAG4 and FGFR1) and on 20q (including NCOA3). In other instances the models yield different results. In general, the gene-set model finds regions of association, i.e. it tends to find associations involving neighbouring copy number probes. In contrast, the gene-to-gene model focuses on effects on individual probes, so it often finds single probes with association with no other effects on neighbouring probes. On chromosome arms 3p, 3q, 14q and 18q (figure 2D), both models pick up at least one copy number probe as having association with expression within those arms, but the effects span genomic regions under the gene-set model, whereas they are restricted to individual probes under the gene-to-gene model. Looking in more detail at 18q, many samples have a mild copy loss of up to -0.4 on the median smooth scale, with a handful of samples displaying slightly larger losses (see figure 4A). A couple of samples display mild copy gain in the same region. It turns out that expression is affected, as indicated by the many statistically significant associations found by the gene-set model. The gene-to-gene model, however, only finds three of those associations statistically significant.

In general, the gene-set model particularly benefits from (mild) associations spanning multiple genes. If expression levels of resident genes are affected, many copy number probes mapping the aberrant region will show significant associations, thereby highlighting the region. On the other hand, copy number changes that affect only one or a few genes are picked up by the gene-to-gene model, but may become diluted when the entire chromosome arm is analysed by the gene-set model. For example,

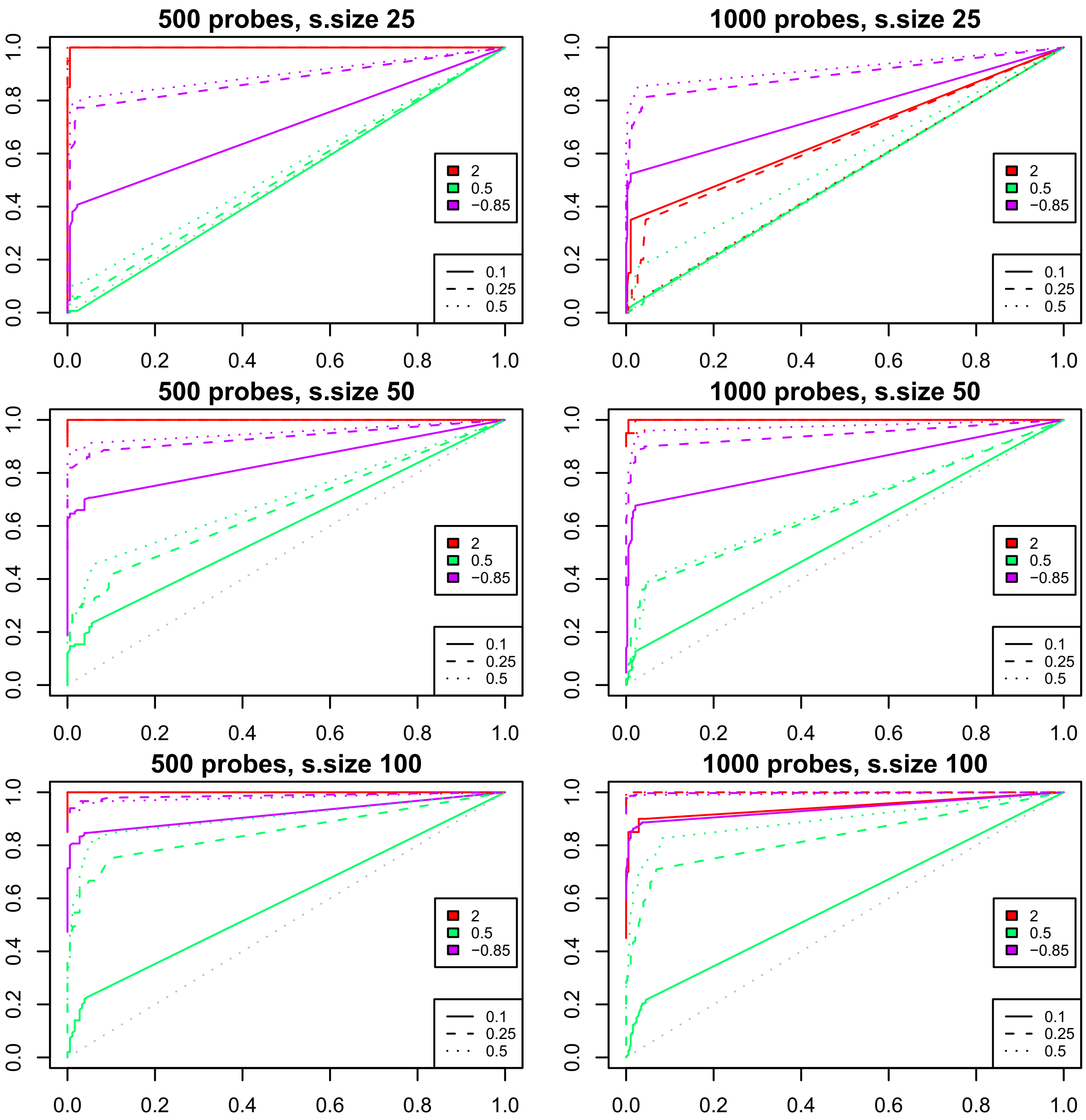

Figure 1

Receiver-operator characteristic curves for the simulation study. Conditions shown are: sample size 25, 50 and 100 (rows from top to bottom); length of the studied region 500 and 1000 probes (columns from left to right); proportion of samples with each effect 10, 25 and 50 percent (solid, dashed and dotted lines); and copy number effect sizes 2, 0.5 and -0.85 (line colours red, green and purple). For details about the study setup, see appendix.

associations on 9q and 11q are only detected with the gene-to-gene model (Figure 2C).

Breast cancer II: Chin

Let us now consider a set of 89 samples of breast tumor tissue, profiled both on a 2.5 K BAC array CGH and on an A3ymetrix U133A array from Chin and co-authors [29]. In this case there is no correspondence between copy number and expression probes, and the genomic coverage is rather different from the one yielded by Pollack's cDNA arrays, with the copy number arrays having about half the number of clones as in Pollack's data, and the expression arrays having over four times as many probes as Pollack's. Only probes with genomic annotation were used, totalling 2083 BACs and 21339 expression probe sets. Here we used again in the gene-set model all genes on a chromosome arm as a gene set. For all models applied to this dataset, associations found are those considered statistically

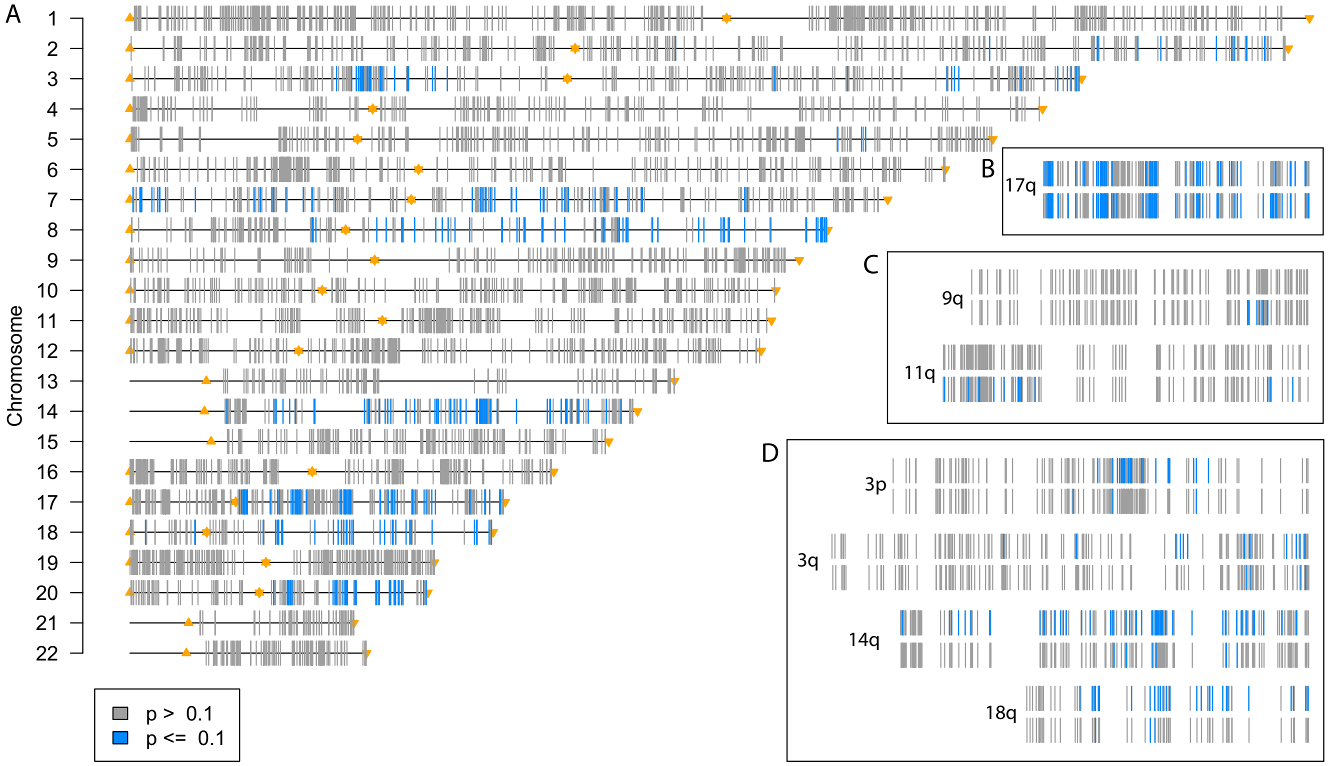

Figure 2

Overview of associations found with Pollack's breast cancer data and the gene-set model. Chromosomes are represented by horizontal bars, and the arms are the gene sets used, with gene set edges marked by triangles and stars. Each vertical bar represents one copy number probe. The colour of the bar indicates the test result: blue, significant(FDR ≤ 0.10); grey, not significant (FDR > 0.10). Results on the left-hand side (A) refer to all tests done on the Pollack data set using the gene-set model. Insets on the right-hand side display gene-set (top) and gene-to-gene (bottom) p-values for a selection of chromosome arms (B: 17q, C: 9q and 11q; D: 3p, 3q, 14q and 18q).

significant with FDR control at 1%. This more strict threshold is used to make results more comparable with those from the previous example, where the number of samples (41) was less than half of the number of samples here.

First we fit overlap and window gene-to-gene models. The overlap gene-to-gene model, measuring association between copy number on BAC clones and expression probe sets included in it, yields 991 comparisons representing less than half of the copy number probes observed (2083). In contrast, the 2 Mb-window gene-to-gene model involves 2030 comparisons. So by considering only expression probes located within BACs more than half of the BACs is neglected. In addition, the associations considered by the overlap model are also considered by the window model by definition, unless the window used in the latter is smaller than the BACs, which is not the case here. So we expect to identify the same associations with both models, which indeed happens: of the 239 statistically significant associations identified by the overlap model, only 13 were not identified by the window model too. This is merely because, by involving a much smaller number of tests, the overlap model results involve a less severe multiple-testing correction. From this viewpoint the overlap model yields the same patterns as the window model, but the latter makes better use of the observed data. For this reason, we will focus hereafter on comparisons between the window gene-to-gene and the gene-set models. The genomewide associations found between copy number and gene expression with these two models can be seen in Additional File 3.

There are many associations found both by the gene-set and by the gene-to-gene model, but also associations found by only one (see figure 1B in Additional File 1). The many effects found by both models refer to those involving large enough copy number changes and/or expression changes. For example, we found a pattern of association on 17q with both models, very similar to what had also been found in the previous example (figure 3B). This may sound obvious, but it is less so considering the widely dif-

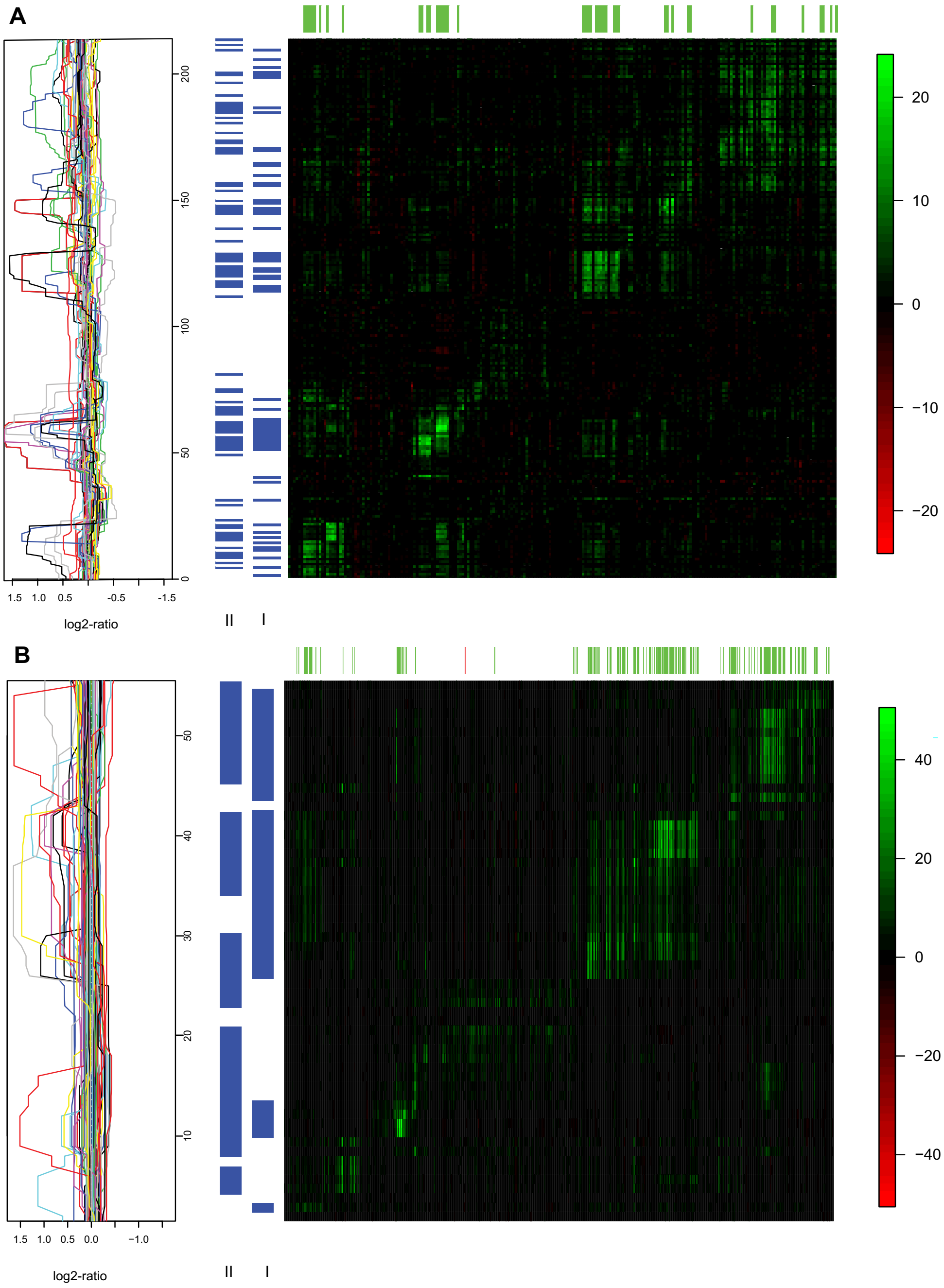

Figure 3

Association patterns between copy number and gene expression found on 17q for two independent breast cancer studies. Heatmap of association structure (green, positive; red, negative; black, no association) for the datasets of Pollack (A) and Chin (B), with rows representing copy number probes from centromere (bottom) to telomere (top), and columns representing gene expression probes from centromere (left) to telomere (right). The vertical bars on the left-hand side represent the p-values for the copy number probes, as calculated by the gene-set model (bar I) and by the window gene-to-gene model (bar II), with blue indicating the significant ones (for Pollack FDR ≤ 0.10, for Chin FDR ≤ 0.01). The top horizontal bar indicates expression probes with strong positive (green) or negative (red) association with the significant results from the gene-set model (mean z-score across significant tests ≥ 3). In the left panel, the log-ratio copy number values for all samples are represented by their smoothed medians, on the same probe spacing (equal space between each pair of consecutive probes) as used in the heatmap for comparability.

ferent microarray platforms used in the two studies, with markedly different genomic coverages, and the fact that independent samples are involved. Reflecting this, the number of probe sets mapping 17q differs markedly between the two studies: Pollack has 215 clones measuring both copy number and expression levels, whilst Chin has only 59 measuring copy number and as many as 913 measuring expression.

For some regions, the gene-set and gene-to-gene models yield different results. As in Pollack's data, for 18q the gene-set model finds clear association between copy number and expression for all but three clones in the region, in contrast with a weaker association detected by the gene-to-gene model (figure 4B). Indeed, copy number changes in this region are mild, with most changes being a loss of no more than 0.4 on the smoothed median scale. But many samples do display this loss, and its impact on expression drives the association. This mild effect is harder to be picked up by the gene-to-gene model.

An intermediate model between the gene-set on the chromosome arm and the gene-to-gene (on a 2 Mb window) models would be one that considers a smaller gene set, helping focus the search for associations in the region around the copy number probe. Such a gene set may be defined in various ways. Here we consider as gene set all gene expression measurements within a 2 Mb region centered around the copy number probe under study. As expected, this model finds many associations also identified by the other models, but also finds some more (see Additional File 4). Of the 2030 associations tested by all three models, 1344 (66%) were found to be statistically significant by at least one model and, of those, 519 (39%) were found by all three models. The largest overlap was found between the gene-to-gene and the gene-set on the 2 Mb window, with 775 associations found in common, which is reasonable. However, there were also associations found by each model individually, with the different models specializing in different effect types. This is further illustrated by re-examining significant associations found on chromosome 17 (Additional File 5).

Population data: HapMap

Gene dosage, as copy number variation is often referred to in a biological context, plays a role in regulating gene expression in normal individuals, as was shown by the analysis of copy number and expression data from the HapMap2 samples [22]. Such effects tend to be milder than those found in cancer data, since copy number changes will be typically smaller in size here. Here we illustrate that the gene-set model also has the power to find associations in this context, where mild effects are measured by higher-density arrays. We have re-analysed the data using the gene-set model over each chromosome arm. For comparability with Stranger's results, our model explains the expression levels of each individual probe by the copy number values of all BAC clones in the same chromosome arm, in contrast with the first two examples. This means that around 18 K gene expression probes were tested for association with copy number. Our results yielded similar numbers of gene expression probes selected at the same p-value threshold, but the overlap with the probes selected by Stranger was relatively small (see supplementary table 1 in Additional File 6). A thoroughcomparison between our results and Stranger's is not our objectivehere. Rather we wish to show the added-value of the gene-set model compared to gene-to-gene models, such as the one used by Stranger. The gene-set model had more power to detect subtle associations that span at least a few probes. Indeed, we have identified a region on 6p where gene expression displays association with some of the copy number probes, within each of the populations. The region goes from 32.593 to 32.817 Mb and includes four expression probes. Using a gene-to-gene model, Stranger only picked up two of these four gene expression probes as being associated with copy number, and only for one of the four populations. The gain in power was thus significant by including many BACs in the model, in spite of the fact that the relevant association only involves a small number of BACs (four), compared to the total included in the model (607) on 6p. With the largest Pearson correlation between BAC clones and expression probes in this region being less than 80%, the effect seems not to be strong enough to be picked up using individual BACs as Stranger did. Note that the four expression probes selected map major histocompatibility complex class II genes, known to harbour polymorphisms that are commonly genotyped prior to organ transplants. Thus these known polymorphic areas have copy number-regulated gene expression, but that was only detected by the gene-set model.

Discussion

We propose to jointly analyse DNA copy number and mRNA expression array data by modelling one (copy number, say) as a function of the values of the other (expression) for all genes in the same chromosomal arm or an independently defined region. This yields a gain in power to detect associations, as genome-based regulatory mechanisms tend to affect neighbouring genes. Considering the coordinate behaviour of groups of genes instead of individual genes was shown to be a useful strategy to improve robustness in gene expression analysis [30-32], but has not been previously used in the context of integrating expression data to another data type. Because the global test, the basis of our approach, has optimal power to detect subtle but consistent association between phenotype and expression signature [33], it enables us to detect associations between expression levels and low-

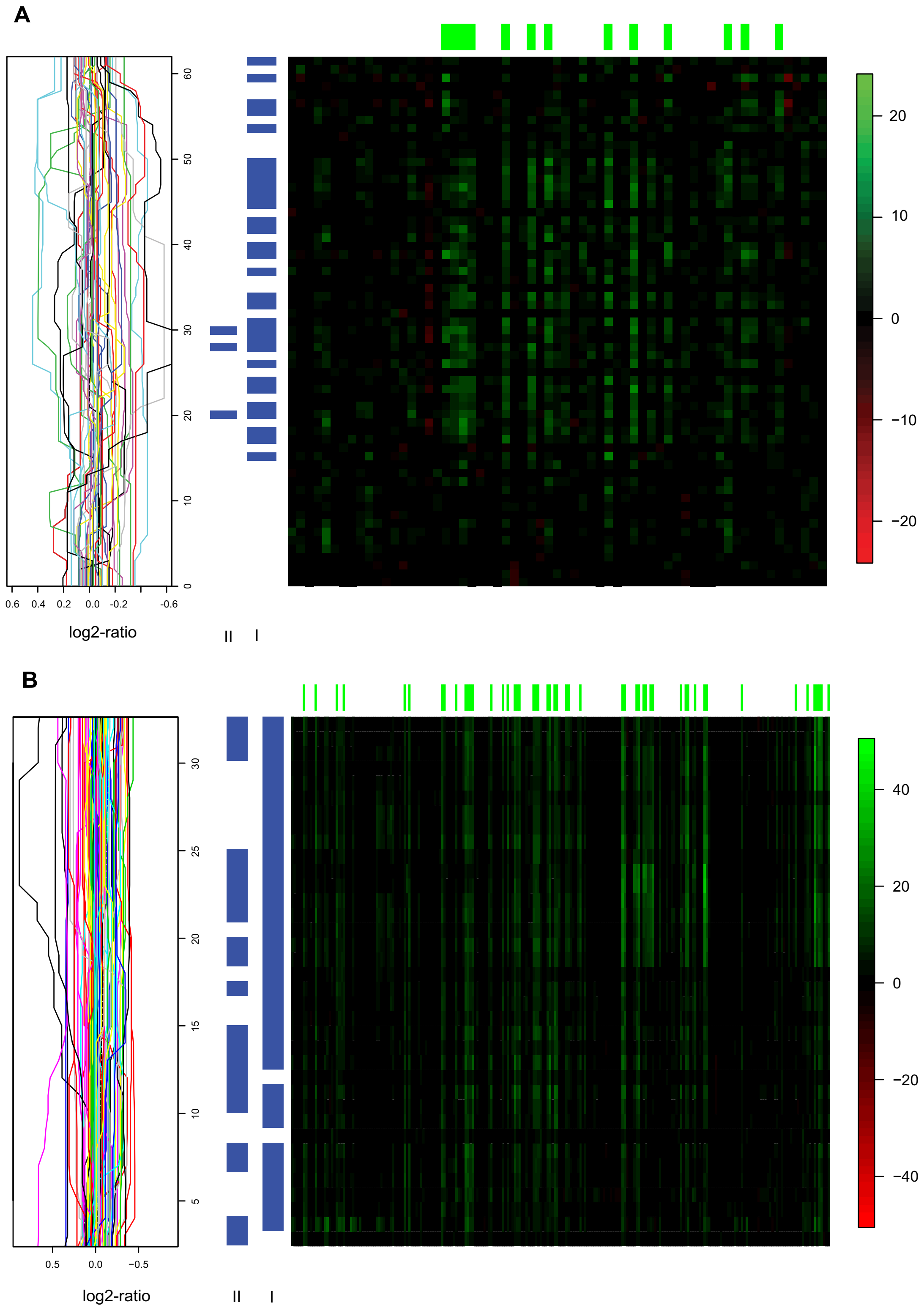

Figure 4

Association patterns between copy number and gene expression found on 18q for two independent breast cancer studies. Heatmap of associations for 18q. See figure 3 for a detailed description.

level gains, in contrast with previous papers which had only been able to detect associations involving high-level gains [11,13,15,23,28].

Our approach is unique in several ways. Firstly, by considering the association between each probe in the dependent data and a gene set in the independent data, rather than a single gene, it stands a better chance of detecting subtle but consistent effects across many genes. Indeed, we have shown in a simulation study that subtle effects can indeed be found if at least 50 samples are studied. As key parameters of the study were estimated from tumor datasets, such as copy number and expression variability as well as their association, results can be extended directly to other studies. As further confirmation, associations between large- and small-amplitude copy number changes and gene expression levels were also found in the breast cancer datasets studied. The use of gene sets keeps focus on consistent changes that are unlikely to be data-dependent, as we showed by obtaining similar patterns of association for two independent datasets, in spite of widely different microarray platforms having been used.

Secondly, the use of a regression framework means that our model enables control of confounder effects. For the samples studied by Chin estrogen-receptor status was known. We applied the gene-set model using this variable as a confounder. Associations found were similar to those without considering the confounder for most chromosomes, except for five of them: for 1p, 5q, 6p and 12q, no features were selected with ER-status adjustment whilst the unadjusted model selected between 20% and 60% of the probes, and for 19q, no features were found with the unadjusted model, but about 40% of the BACs were selected with ER-status adjustment. More importantly, in each of these chromosome arms a handful of BACs was assigned an FDR-corrected p-value in one analysis below 0.01, whilst in the other the p-value was larger than 0.20. These results suggest that copy number-based mechanisms of gene expression regulation differ according to estrogen-receptor status in breast cancer.

Thirdly, our model can be used with continuous copy number, as log-ratios like in our examples or on the original copy number scale, as well as segmented or discretized copy number data. Here we point out that there is no consensus as to whether or not association testing would benefit from segmentation of DNA copy number data [22,24,34]. However, we recommend using continuous data at least in cancer studies, because a non-integer number of copies may represent the average number of copies found on the sample of cells collected for that particular tumor, in which case a sharp cut-off is likely to introduce a bias.

Finally, our approach avoids introducing bias via matching between copy number and expression probes on the genome, as it rightly focuses on finding relevant associations regardless of their genomic location. We compared the performance of the gene-set model to that of the gene-to-gene model. While the former has more power to identify regions with coordinated association, the latter yields individual associations unrelated to possible effects in its neighbourhood, as expected. Associations involving a large proportion of samples displaying a large-amplitude copy number change are typically picked up by both models, as is the case on 17q, known to harbour regions of large amplifications that play a causal role in breast cancer. Mild associations, because either the copy number change has small amplitude, or the effect on gene expression is limited, or even the proportion of samples involved is small, are less likely to be identified by the gene-to-gene model than by the gene-set model. A clear example is that of 18q, where in both examples the gene-set model identifies the effect spanning a large region, but the gene-to-gene model just selects a handful of probes. Note, however, that the overlap between results of the two models for the Chin dataset is positively associated with window size and, as the 2 Mb window size used included most of the associations available, considerable overlap was the result.

From the biological viewpoint, associations found for many copy number probes within the same region are reassuring, as they are less likely to be driven by pure noise. Because it tends to find regions of association, the gene-set model is more robust to noise than the gene-to-gene model.

These two models represent two extremes. An intermediate model may be used to diminish the dilution, whilst still being robust to noise. Such a model could be a gene-set model over a window centered around the copy number probe, as used in the Breast Cancer II example. This sort of model is particularly useful when interest lies in mild associations, either spanning small regions, involving low-amplitude copy number changes or having a limited impact on expression levels. It might be particularly useful when high-density arrays are involved. Nevertheless, the gene-set model remains the least arbitrary and, while dilution might be a concern, individual effects may still be identified by visual inspection of heatmaps representing association patterns found. In such cases, focus shifts from finding statistically significant associations to finding consistent association patterns between copy number and expression. The nature of the problem turns then from that of hypothesis testing to an exploratory one.

We considered two ways in which the gene-to-gene model (1) can be used in a study where different microarray platforms were used to measure copy number and expression. The first one involved considering only copy number and expression probes that overlap, so that only measurements for the same locus are considered. The second way was to calculate associations between each copy number probe and expression probes located within a certain distance from it. Another possible way is to interpolate the copy number measurements, say using quantile smoothing as suggested by Eilers and Menezes [35], and thus obtain copy number estimates corresponding to all loci for which expression was measured. This would avoid the problem from the first approach that non-overlapping probes are neglected, and the arbitrariness of defining a distance on the second approach, so making use of all observed measures. However, it relies on good approximations via interpolation. If the density of the copy number probes is high with reasonably small intervals between probes, interpolated values tend to estimate well the true copy number. On the other hand, with large between-probe distances such as 1 Mb, this is less likely to be the case. In all cases some arbitrariness is involved, which the gene-set model avoids.

The gene-set model can be formulated in alternative ways to answer different questions. Perhaps the most intuitive formulation is to use expression as outcome and copy number as explanatory variable, best suited to find genes which expression is regulated by copy number changes in the region around it. However, if the objective is to find DNA-based markers that regulate gene expression on the same region, then the best formulation uses copy number as outcome and expression as explanatory variable, as we did in the analysis shown here. By considering the expression values of many genes simultaneously, this formulation is also able to capture coordinated variability in expression levels across genes, such as co-regulation, which would not be possible otherwise due to noise. This is relatively less important in copy number data, which typically displays relatively less noise compared to the signal.

It is straightforward to extend the model to analyse other types of high-dimensional data. For example, another type of expression regulation mechanism is DNA methylation, which can be measured via CpG-island arrays. In a similar way to simultaneous analysis of copy number and expression array data, there could be interest in analysing DNA-methylation and expression. The use of gene sets are still likely to improve power to detect associations, as DNA-methylation may affect the expression levels of multiple genes, like copy number.

A second interesting extension is to consider more than two types of array data in model (2). For example, gene expression can be regulated by different mechanisms in addition to copy number, including transcription factor levels, sequence changes, DNA methylation, loss of heterozygosity, and chromatin structure. Our method can be generalized to analyze the association between gene expression and other types of genomic information simultaneously. This extension is beyond the scope of this paper and will appear elsewhere. We hope that such a model taking into account multiple data sources simultaneously will shed light on the influence of different genetic and epigenetic mechanisms on gene regulation.

Finally, while the gene-set model serves as a starting point to identify copy number changes that are associated with expression patterns, additional experiments are needed to validate the possible role of these changes in the causation or maintenance of the phenotype under study.

Conclusion

We have proposed, and given proof of principle for, a new approach to identify association between high-throughput genomic copy number and gene expression profiling data, which can be used to identify putative candidate genes involved in tumorigenesis. By considering the expression levels of many genes simultaneously in the model, our approach identifies regions of association even if low-amplitude copy number changes are involved. The regression is able to control for confounder effects. Finally, it requires neither matching between copy number and expression probes on the genome, nor categorization of copy number, both of which are possible sources of bias.

Methods

We assume that each sample is profiled both on a copy number and on an expression array, that the copy number and the expression array data were separately pre-processed, adequately normalized and that probe annotation including identifier, chromosome number and location in base pairs is available.

We shall focus on answering the following question: which copy number changes affect gene expression within the same chromosomal region? This question typically arises when searching for DNA-based markers that regulate expression via copy number change.

The gene-to-gene model

Since our main interest is to find DNA-based markers that are associated with expression changes, it makes sense to consider copy number as the dependent variable, so expression is handled as the independent variable. The simplest model to consider is

$$E(Y_{ni}) = \alpha + \beta_i X_{ni}, \quad n = 1, \ldots, N, \quad i = 1, \ldots, I, \qquad (1)$$

where Y_{ni} represents the copy number measured for sample n and array-CGH copy number probe i ($i = 1,...,I$) and X_{ni} represents the expression level for sample n and expression probe i. This model is written assuming that there is a one-to-one correspondence between copy number and expression probes, for simplicity, which holds for example if the same array is used to measure both copy number and expression levels. Such a model is especially useful if, in addition to using the same array, the study goal is to find associations between copy number change and expression variation at the same locus. We shall refer to (1) as the *gene-to-gene* model.

In many cases interest lies in studying effects of copy number change on expression of resident genes, i.e., genes within the same region where copy number was measured. But it can be the case that not the same array was used to measure copy number and expression. Then model (1) can still be used if the data and/or the model are adapted. The first thing that can be done is to consider only expression probes that fall within copy number probes, so as to ensure that measurements relate to the same locus. We refer to this as the *overlap* approach. This is rather strict, and may result in many expression probes not being considered, even if they are close to the copy number probe. To relax this, we can consider expression probes that are within a certain distance from the middle of the copy number probe in each direction, and then fit model (1) to each one separately. We refer to this as the *window* approach, where the window size is the length of the entire interval considered. This may yield more than one test for some copy number probes, whilst for others with no expression probes near it no tests are done. So some arbitrariness is involved in defining the distance, which directly affects which tests are considered. Here we shall use a window of size 2 Mb (similar to Stranger *et al.* [22]), centered around the start of the copy number probe.

The gene-set model

In practice, it is not commonly the case that the same array platform is used for copy number and expression. Moreover, the arbitrariness of the window definition is undesirable, and considering only overlapping probes leads potentially to loss of valuable information. An ideal way to avoid these problems is to include in the model all expression probes within the same large region. This leads us to the model, for each copy number probe i,

$$E(Y_{ni}) = \alpha + \sum_{j=1}^{J} \beta_j X_{nj}, \quad n = 1,\dots,N. \qquad (2)$$

By considering the expression levels of many genes simultaneously, this model suits well most situations where copy number changes produce an effect spanning many expression probes, in a possibly subtle but consistent way. Because (2) makes use of a set of genes as independent variables, we shall refer to it as the *gene-set* model.

Note that the gene-set model (2) is not estimable if $J > n$, in a classic linear regression context. Since our main objective is to test whether copy number change is associated in general with expression levels $\{X_{nj}, j = 1,...,J\}$, it is natural to study the distribution of $\beta \equiv (\beta_1,...,\beta_J)^t$, a vector of independent random variables. We assume that each β_j has a certain distribution and, under the null hypothesis of no association between X and Y, has mean 0 and variance $\tau^2 \equiv 0$. The assumption $\beta \sim (0, \tau^2 I_J)$ means that model (2) is a random-effects model, and a natural distribution to assign to the vector β is the multivariate normal with a covariance matrix $\tau^2 I_J$, where I_J represents the identity matrix with J rows. The random-effects model framework arises thus naturally from the question under study and the biological context. Moreover, it guarantees that model (2) is identifiable, which would not be the case in the classic linear regression model framework if $J >> N$, which is often the case. Thus, model (2) can be fitted using methods for random-effects models.

Under the alternative hypothesis of association, the mean of each β_j may still be zero, but their variance should be strictly positive ($\tau^2 > 0$), suggesting that a non-empty subset of the $\{X_{nj}, j = 1,..., J\}$ is associated with copy number measurements for probe i. Therefore, we shall focus on testing $H_0 : \tau^2 = 0$ against $H_a : \tau^2 > 0$. A test to compare such null and alternative hypotheses was proposed by [30] for testing association between expression levels of a set of genes, e.g. those belonging to a biological pathway, with a clinical outcome. This approach has been shown to have more power to detect subtle associations than by performing separate tests and correcting the resulting p-values for multiple testing [33]. We shall make use of this global test as the basis for our approach in this new context, as well as consider extensions of interest.

By modelling the copy number at each locus by expressions within a large gene set, the gene-set model (2) takes advantage of the typically larger signal-to-noise ratio in copy number compared with gene expression microarray data. By considering the expression levels of many genes jointly, coordinated expression changes, such as co-regulation, are more likely to be detected than if gene expression levels were considered separately.

Considering covariates

Because models (2,1) are constructed within a regression framework, other explanatory variables which can act as confounders can be included. This is very important in more complex designs, such as when more than one sample is collected per patient, when patients are related, or when clinical variables are to be taken into account, for example tumor location and age.

Note that, as in multivariate regression analysis, the inclusion of a confounder can weaken or even eliminate an effect, if the association is limited to one confounder-defined subgroup. On the other hand, it may explain part of the copy number variation bringing out new associations.

Multiple testing correction

Both models (1) and (2) yield one *p*-value per copy-number probe tested. Copy number levels of neighbouring probes are likely to be associated, but genomic breaks such as centromeres and telomeres may break this association. Therefore, it seems reasonable to treat chromosomal arms independently in the analysis, including in what concerns multiple testing correction. The correction must allow for dependency between the tests, and currently the most adequate method available has been suggested by [36].

Note that the overlap and window approaches used with the gene-to-gene model (1) involve different numbers of tests, implying different multiple testing corrections, unless of course the same microarray is used to measure both copy number and expression. Indeed, the overlap approach typically will involve a smaller number of tests than there are copy number probes, in contrast with the window approach which may generates a larger number of tests than there are copy number probes.

Definition of genomic region

The length of the genomic region under study may affect the results of the model fit via the amount of multiple-testing needed and, to a lesser extent, via the number of explanatory variables, i.e. gene expressions, in the model. To avoid introducing bias in this way, we suggest that the genomic regions to be studied be determined *a priori*. In our experience chromosome arms are sensible such regions, as are minimal common regions of recurrent copy number aberrations.

Visualization and prioritization of genes

For each copy number probe, the test statistic can be decomposed into the individual contributions of the genes' expression levels [30]. After standardization per probe, we display the separate contributions of the gene expressions to copy number variability by means of a heatmap, where rows and columns represent copy number and expression probes respectively, both kept in their genomic order. If a copy number change spanning roughly the same genomic region across a subset of samples is positively associated with gene expression on the same region, it will be represented by a green rectangle on the diagonal. For each copy number probe, discretised *p*-values computed with the test can be displayed as a vertical bar next to the heatmap, so that significantly associated genomic regions are highlighted.

It is often of interest to identify the gene expression probes with the largest contributions to the test results, in some sense. We choose to compute the mean standardized contribution over all significant tests, per expression probe, then rank them to generate candidate genes for future investigation. This yields candidate genes whose expression levels are highly associated with copy number. By considering only the significant tests, we want to avoid diluting a possible association spanning a relatively small area, compared to the entire area under study.

Software used

We have used R version 2.5.1 [37] for all our analyses. In addition, we used the following R packages: globaltest, marray, multtest and quantsmooth. An R package called SIM implementing this approach has been made available via BioConductor.

Authors' contributions

RXM developed the concept, was involved in the statistical analyses, and wrote the manuscript. JMB helped develop the concept, was involved in interpretation of the results, and revised the manuscript. MS and MB performed part of the analyses and participated in developing the BioConductor package SIM. GJVO participated in the study design and coordination. All authors read and approved the final manuscript.

Appendix

Simulation study setup

Assumptions

In order to evaluate how the regional integration model works, we run a simulation study. For this, we assume for simplicity that both copy number and expression measurements are obtained using the same arrays, so that there is a one-to-one correspondence between them. We also assume that the copy number is measured in terms of log-ratios, that the expression is measured in terms of intensities, and that the datasets were normalized separately as adequate. All these assumptions are made for the sake of simplicity, being unimportant for the qualitative results of this study.

For each probe $i = 1,...,I$, the data consists of copy number measurements $Y_{i1},...,Y_{iN}$ and expression measurements $X_{i1},...,X_{iN}$ for samples $1,...,N$. We assume that samples are independent and that data distribution is independent of sample, so we ignore the sample index from now on for simplicity. The true log-ratio copy number of gene i is represented by ξ_i. We represent by Z_i the binary indicator variable that copy number of gene i regulates its expression, so that the vector $Z = (Z_1,...,Z_I)^t$ represents expression-regulating copy number regions. We assume that there are three such regions: region I, spanning 20 probes which all have true copy number 5, so a gain of 3 copies; region II, spanning around 26% of the probes which all have true copy number 3, so a gain of one copy; and region III, spanning also 26% of the probes which all have true copy number 1, so a loss of one copy.

Copy number measurements $Y_1,...,Y_I$ are assumed to be independent, each Y_i with distribution $\mathbf{N}(\xi_i, \sigma_i^2)$. Note that if gene i is located in any of the three regions of expression-regulating copy number, $Z_i = 1$ and $\xi_i \neq 0$ and, for all genes outside these regions, $Z_i = \xi_i = 0$.

Expression measurements $X_1,...,X_I$ are assumed to be conditionally independent, given $Y_1,...,Y_I$, with $(X_i|Y_i) \sim \mathbf{N}(\mu_i\{1+\alpha_i\xi_i\}, \tau_i^2)$, where μ_i represents gene i's baseline expression level, and α_i is a variable representing the extent to which copy number of gene i regulates its expression.

Parameter values

The mean log-ratios of copy numbers $\xi_i|Z_i = 1$ were estimated from [38], which made use of commercially available DNA samples with known copy number. These data yielded $\xi_i = 2$, $\xi_i = 0.5$ and $\xi_i = -0.85$ for genes in regions I, II and III respectively. Note that there is signal compression of the expected amplitude. Log-ratios of copy numbers $\{Y_i, i = 1,...,I\}$ were drawn independently from a $N(\xi_i, \sigma_i^2)$ where its dispersion $1/\sigma_i^2$ follows a $\Gamma(2, 2)$, so the mean dispersion is 4. Given Y_i, μ_i is drawn from a $N(9.5, 2.3^2)$, independently for each i. The dispersion of $X_i|Y_i$, $1/\tau_i^2$, is in its turn drawn from a $\Gamma(2, 0.5)$, implying that the mean dispersion is 1.

Copy-number impact on expression per probe i, α_i, was estimated using three subregions of the breast cancer data from [12], as in this case the same array was used to produce both copy number and expression. These three regions were chosen so as to contain copy number changes, some of them having been found to be associated with expressions, and were located in chromosomes 1q, 8p and 17q. First of all, the relationship between copy number and expression for the 303 selected probes can be described reasonably well by a linear function (data not shown). Moreover, the normal distribution seems to yield a reasonable approximation to the empirical distribution of α_i, with mean and variance estimated as -4.6 and 8.3, respectively (data not shown). So, in our simulation we draw α_i from this distribution. Other parameters that need to be fixed and values used are given in supplementary table 2 (see Additional File 6).

Note that the direction of the effect (positive or negative) is unimportant for its detectability by the model, only the log-ratio is important.

Additional material

Additional file 1

Overview of associations found with Pollack's breast cancer data and the gene-to-gene model*. Results obtained with the gene-to-gene model, where association is measured between each pair of copy number and gene expression measures obtained with the same cDNA clone. Chromosomes are represented by horizontal bars, with telomeres and centromeres marked by triangles and stars respectively. Each vertical bar represents one copy number probe. The colour of the bar indicates the test result: blue, significant(FDR ≤ 0.10); grey, not significant (FDR > 0.10).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-203-S1.pdf]

Additional file 2

Venn diagram of associations found by two models in two independent breast cancer studies*. Overlap of associations between copy number and expression found significant by the gene-set (right, in blue) and gene-to-gene (left, in red) models. For the gene-set model, the chromosome arm was used as gene set. A – Pollack's data (FDR ≤ 0.10); B – Chin's data (FDR ≤ 0.01), for which the gene-to-gene model was applied on a 2 Mb window.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-203-S2.pdf]

Additional file 3

Overview of associations found with Chin's breast cancer data*. Chromosomes are represented by horizontal bars, with centromerers and telomeres marked by triangles and stars. Each vertical bar represents one copy number probe. The colour of the bar indicates the test result: blue, significant (FDR ≤ 0.01); grey, not significant (FDR > 0.01). A: gene-set model; B: gene-to-gene model.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-203-S3.pdf]

Additional file 4

Venn diagram of associations found by three models with Chin's breast cancer data*. Overlap of associations between copy number and expression found significant by the gene-set model using chromosome arm (right), the gene-set model using only gene expression probes on a 2 Mb window around the copy number probe (bottom) and the gene-to-gene model applied to the same 2 Mb window. Significance threshold was taken as FDR ≤ 0.01.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-203-S4.pdf]

Additional file 5

Detail of associations found on 17q by three models with Chin's data. *Each vertical bar represents one copy number probe, and each horizontal bar one model: gene-set model using chromosome arm (top), gene-set model using only gene expression probes on a 2 Mb window around the copy number probe (middle) and gene-to-gene model applied to the same 2 Mb window (bottom). A: region where associations are found only with the models considering the 2 Mb window; B: region where associations are found only with the model considering the entire chromosome arm as gene set.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-203-S5.pdf]

Additional file 6

Supplementary tables: 1. Number of gene expression probes associated with copy number in HapMap data; 2. Parameter values used in the simulation study. *In table 1, the significance threshold used for uncorrected p-values was 0.001, for comparability with Stranger et al (2007).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-203-S6.pdf]

Acknowledgements

We thank P.A.C. 't Hoen, J.T. den Dunnen, R. Fodde, and J. J. Goeman for fruitful discussions. This study was supported by a grant from the Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO). This work is part of the BioRange bioinformatics research program supported by the Netherlands Bioinformatics Centre (NBIC). This research was partially funded through the European Community's Seventh Framework Programme CFP7 (2007–2013), ENGAGE project, grant agreement HEALTH-F4-2007-201413.

References

1. Vogelstein B, Kinzler K: **Cancer genes and the pathways they control.** *Nature Medicine* 2004, **10:**789-799.
2. Alizadeh A, Eisen M, Davis R, Ma C, Lossos I, Rosenwald A, Boldrick J, Sabet H, Tran T, Yu X, Powell J, Yang L, Marti G, Moore T, Hudson J Jr, Lu L, Lewis D, Tibshirani R, Sherlock G, Chan W, Greiner T, Weisenburger D, Armitage J, Warnke R, Levy R, Wilson W, Grever M, Byrd J, Botstein D, Brown P, Staudt L: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511.
3. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
4. Perou C, Sorlie T, Eisen M, Rijn M van de, Jeffrey S, Rees C, Pollack J, Ross D, Johnsen H, Akslen L, Fluge O, Pergamenschikov A, Williams C, Zhu S, Lonning P, Borresen-Dale A, Brown P, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
5. Roepman P, Wessels L, Kettelarij N, Kemmeren P, Miles A, Lijnzaad P, Tilanus M, Koole R, Hordijk G, Vliet P van der, Reinders M, Slootweg P, Holstege F: **An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas.** *Nature Genetics* 2005, **37:**182-186.
6. Sorlie T, Perou C, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen M, Rijn M Van de, Jeffrey S, Thorsen T, Quist H, Matese J, Brown P, Botstein D, Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98:**10869-10874.
7. van't Veer L, Dai H, Vijver M van de, He Y, Hart A, Mao M, Peterse H, Kooy K van der, Marton M, Witteveen A, Schreiber G, Kerkhoven R, Roberts C, Linsley P, Bernards R, Friend S: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530-536.
8. Albertson D, Collins C, McCormick F, Gray J: **Chromosome aberrations in solid tumors.** *Nature Genetics* 2003, **34:**369-376.
9. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo W, Chen C, Zhai Y, Dairkee S, Ljung BM, Gray J, Albertson D: **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.** *Nature Genetics* 1998, **20:**207-211.
10. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P: **Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances.** *Genes Chromosomes Cancer* 1997, **20:**399-407.
11. Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, Rozenblum E, Ringner M, Sauter G, Monni O, Elkahloun A, Kallioniemi OP, Kallioniemi A: **Impact of DNA amplification on gene expression patterns in breast cancer.** *Cancer Research* 2002, **62:**6240-6245.
12. Pollack J, Sorlie T, Perou C, Rees C, Jeffrey S, Lonning P, Tibshirani R, Botstein D, Borresen-Dale A, Brown P: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99:**12963-12968.
13. Aguirre A, Brennan C, Bailey G, Sinha R, Feng B, Leo C, Zhang Y, Zhang J, Gans J, Bardeesy N, Cauwels C, Cordon-Cardo C, Redston M, DePinho R, Chin L: **High-resolution characterization of the pancreatic adenocarcinoma genome.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101:**9067-9072.
14. Tsafrir D, Bacolod M, Selvanayagam Z, Tsafrir I, Shia J, Zeng Z, Liu H, Krier C, Stengel R, Barany F, Gerald W, Paty P, Domany E, Notterman D: **Relationship of gene expression and chromosomal abnormalities in colorectal cancer.** *Cancer Research* 2006, **66:**2129-2137.
15. Garraway L, Widlund H, Rubin M, Getz G, Berger A, Ramaswamy S, Beroukhim R, Milner D, Granter S, Du J, Lee C, Wagner S, Li C, Golub T, Rimm D, Meyerson M, Fisher D, Sellers W: **Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma.** *Nature* 2005, **436:**117-122.
16. Jarvinen A, Autio R, Haapa-Paananen S, Wolf M, Saarela M, Grenman R, Leivo I, Kallioniemi O, Makitie A, Monni O: **Identification of target genes in laryngeal squamous cell carcinoma by high-resolution copy number and gene expression microarray analyses.** *Oncogene* 2006, **25:**6997-7008.
17. Masayesva B, Ha P, Garrett-Mayer E, Pilkington T, Mao R, Pevsner J, Speed T, Benoit N, Moon C, Sidransky D, Westra W, Califano J: **Gene expression alterations over large chromosomal regions in cancers include multiple genes unrelated to malignant progression.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101:**8715-8720.
18. Phillips J, Hayward S, Wang Y, Vasselli J, Pavlovich C, Padilla-Nash H, Pezullo J, Ghadimi B, Grossfeld G, Rivera A, Linehan W, Cunha G, Ried T: **The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis.** *Cancer Research* 2001, **61:**8143-8149.
19. Carrasco D, Tonon G, Huang Y, Zhang Y, Sinha R, Feng B, Stewart J, Zhan F, Khatry D, Protopopova M, Protopopov A, Sukhdeo K, Hanamura I, Stephens O, Barlogie B, Anderson K, Chin L, Shaughnessy J, Brennan C, DePinho R: **High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients.** *Cancer Cell* 2006, **9:**313-325.
20. Tonon G, Wong K, Maulik G, Brennan C, Feng B, Zhang Y, Khatry D, Protopopov A, You M, Aguirre A, Martin E, Yang Z, Ji H, Chin L, DePinho R: **High-resolution genomic profiles of human lung cancer.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102:**9625-9630.
21. Mao R, Wang X, Spitznagel E, Frelin L, Ting J, Ding H, Kim J, Ruczinski I, Downey T, Pevsner J: **Primary and secondary transcriptional effects in the developing human Down syndrome brain and heart.** *Genome Biology* 2005, **6:**R107.

22. Stranger B, Forrest M, Dunning M, Ingle C, Beazley C, Thorne N, Redon R, Bird C, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer S, Tavaré S, Deloukas P, Hurles M, Dermitzakis E: **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science* 2007, **315:**848-853.
23. Lui F, Park F, Lai W, Maher E, Chakravarti A, Durso L, Jiang X, Yu Y, Brosius A, Thomas M, Chin L, Brennan C, DePinho R, Kohane I, Carroll R, Black P, Johnson M: **A genome-wide screen reveals functional gene clusters in the cancer genome and identifies EphA2 as a mitogen in glioblastoma.** *Cancer Research* 2006, **66:**10815-10823.
24. Lee H, Kong S, Park P: **Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes.** *Bioinformatics* 2008, **24:**889-896.
25. Chari R, Coe B, Wedseltoft C, Benetti M, Wilson I, Vucic E, MacAulay C, Ng R, Lam W: **SIGMA2: A system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes, and transcriptomes.** *BMC Bioinformatics* 2008, **9:**422.
26. Chin S, Teschendorff A, Marioni J, Wang Y, Barbosa-Morais N, Thorne N, Costa J, Pinder S, Wiel M van de, Green A, Ellis I, Porter P, Tavaré S, Brenton J, Ylstra B, Caldas C: **High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer.** *Genome Biology* 2007, **8:**R215.
27. van Wieringen W, Belien J, Vosse S, Achame E, Ylstra B: **ACE-it: a tool for genome-wide integration of gene dosage and RNA expression data.** *Bioinformatics* 2006, **22:**1919-1920.
28. Adler A, Lin M, Horlings H, Nuyten D, Vijver M van de, Chang H: **Genetic regulators of large-scale transcriptional signatures in cancer.** *Nature Genetics* 2006, **38:**421-430.
29. Chin K, DeVries S, Fridlyand J, Spellman P, Roydasgupta R, Kuo WL, Lapuk A, Neve R, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung B, Esserman L, Albertson D, Waldman F, Gray J: **Genomic and transcriptional aberrations linked to breast cancer pathophysiologies.** *Cancer Cell* 2006, **10:**529-541.
30. Goeman J, Geer S van de, de Kort F, van Houwelingen H: **A global test for groups of genes: testing association with a clinical outcome.** *Bioinformatics* 2004, **20:**93-99.
31. Mootha V, Lindgren C, Eriksson K, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Nicholas Houstis N, Daly M, Patterson N, Mesirov J, Golub T, Tamayo P, Spiegelman B, Lander E, Hirschhorn J, Altshuler D, Groop L: **PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nature Genetics* 2003, **34:**267-273.
32. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer.** *Nature Genetics* 2004, **36:**1090-1098.
33. Goeman J, Geer S van de, van Houwelingen H: **Testing against a high dimensional alternative.** *Journal of the Royal Statistical Society Series B* 2006, **68:**477-493.
34. Willenbrock H, Fridlyand J: **A comparison study: applying segmentation to array CGH data for downstream analyses.** *Bioinformatics* 2005, **21:**4084-4091.
35. Eilers P, de Menezes R: **Quantile smoothing of array CGH data.** *Bioinformatics* 2005, **21:**1146-1153.
36. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Annals of Statistics* 2001, **29:**1165-1188.
37. R Development Core Team: *R: A Language and Environment for Statistical Computing* 2007 [http://www.R-project.org]. R Foundation for Statistical Computing, Vienna, Austria [ISBN" 3-900051-07-0]
38. Cardoso J, Molenaar L, de Menezes R, Rosenberg C, Morreau H, Moslein G, Fodde R, Boer J: **Genomic profiling by DNA amplification of laser capture microdissected tissues and array CGH.** *Nucleic Acids Research* 2004, **32:**e146.