BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 2 2012, pages 151–159
doi:10.1093/bioinformatics/btr644

Genome analysis

Advance Access publication November 21, 2011

FSR: feature set reduction for scalable and accurate multi-class cancer subtype classification based on copy number

Gerard Wong[1,2,*], Christopher Leckie[1,2] and Adam Kowalczyk[1,3]

[1]National ICT Australia, Victoria Research Laboratory, Parkville, [2]Department of Computer Science and Software Engineering, The University of Melbourne, Carlton and [3]Department of Electrical Engineering, The University of Melbourne, Parkville, Australia

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Feature selection is a key concept in machine learning for microarray datasets, where features represented by probesets are typically several orders of magnitude larger than the available sample size. Computational tractability is a key challenge for feature selection algorithms in handling very high-dimensional datasets beyond a hundred thousand features, such as in datasets produced on single nucleotide polymorphism microarrays. In this article, we present a novel feature set reduction approach that enables scalable feature selection on datasets with hundreds of thousands of features and beyond. Our approach enables more efficient handling of higher resolution datasets to achieve better disease subtype classification of samples for potentially more accurate diagnosis and prognosis, which allows clinicians to make more informed decisions in regards to patient treatment options.

Results: We applied our feature set reduction approach to several publicly available cancer single nucleotide polymorphism (SNP) array datasets and evaluated its performance in terms of its multiclass predictive classification accuracy over different cancer subtypes, its speedup in execution as well as its scalability with respect to sample size and array resolution. Feature Set Reduction (FSR) was able to reduce the dimensions of an SNP array dataset by more than two orders of magnitude while achieving at least equal, and in most cases superior predictive classification performance over that achieved on features selected by existing feature selection methods alone. An examination of the biological relevance of frequently selected features from *FSR*-reduced feature sets revealed strong enrichment in association with cancer.

Availability: *FSR* was implemented in MATLAB R2010b and is available at http://ww2.cs.mu.oz.au/~gwong/FSR

Contact: gwong@csse.unimelb.edu.au

Supplementary information: Supplementary data are available from *Bioinformatics* online.

Received on December 13, 2010; revised on October 17, 2011; accepted on November 16, 2011

1 INTRODUCTION

An open problem in the field of microarray analysis is the classification of samples in terms of their cancer subtype using copy number data derived from high-density single nucleotide polymorphism (SNP) arrays. Accurate classification of clinically important subtypes can help define subgroups of patients who would benefit from more individualized treatment regimes. While classification of microarray data has received considerable interest in the gene expression domain, challenges of a different nature arise when dealing with copy number in high-density SNP array datasets.

SNP microarray datasets have more features than gene expression microarrays. Typically, gene expression datasets can comprise tens of thousands of features (genes), while SNP array datasets can comprise hundreds of thousands of features (SNPs). This implies that in SNP array datasets, the number of features are typically several orders of magnitude larger than the available sample size. Consequently, feature selection approaches that work well on gene expression datasets may not necessarily scale well to the increased number of features in SNP array datasets. Moreover, SNP microarray datasets possess a lower signal to noise ratio in comparison to gene expression microarrays. The raw copy number measurements can thus be extremely noisy as compared with gene expression measurements on microarray datasets.

While there have been previous attempts at classifying datasets generated from SNP microarrays, these studies were largely based on the earliest generation of SNP microarrays (Hu *et al.*, 2005; Statnikov *et al.*, 2007; Wang *et al.*, 2006; Wang, 2006) with ~1000 features, which makes these datasets similar in dimensions to gene expression datasets. Therefore, the challenges in scalability are not as apparent in these datasets. Among these studies Statnikov *et al.* (2007) was based solely on genotype data from SNP microarray datasets and not copy number data, whereas Wang (2006) was based only on loss of heterozygosity (LOH) data that is derived from genotype calls.

Other earlier attempts at classifying samples based on copy number datasets were largely based on the lower resolution array Comparative Genomic Hybridization (aCGH) technology (Bastian *et al.*, 2003; O'Hagan *et al.*, 2003; Wang *et al.*, 2007). Again the issue of scalability was not explored since the number of features in aCGH is much smaller than those of the Affymetrix SNP array. The scalability of feature selection in SNP array datasets with large numbers of features for the purpose of sample classification is an open problem that needs to be addressed.

Our aim in this article is to propose a feature reduction approach that is computationally more efficient than existing feature selection approaches (Saeys *et al.*, 2007) while achieving at least comparable or superior classification accuracy, when applied to copy number

[*]To whom correspondence should be addressed.

data generated from SNP microarrays to obtain more accurate classification of samples by cancer subtype.

SNP microarrays have a significant amount of in-built redundancy in their probe design, and probe readings are constantly affected by experimental noise. This implies that the raw SNP microarray measurement values (log2 ratios) will contain many redundant as well as irrelevant features, which makes them suited to feature reduction. Our proposal is that by eliminating redundant and irrelevant features in a computationally efficient manner, we can effectively reduce the original feature set down to a compact set of interesting features where we can continue to apply existing feature selection approaches in a more efficient manner without sacrificing classification accuracy. The feature selection approaches that are suitable for use with our proposed Feature Set Reduction (*FSR*) method are largely filter-based feature selection approaches. Filter-based feature selection approaches tend to be more computationally efficient than other feature selections approaches, such as wrapper-based and embedded approaches, and also have the added advantage of having performance that is independent of the classifier used.

The challenges that we have identified and addressed in designing our proposed method *FSR* are (i) how to reduce feature set size to enable more efficient feature selection; (ii) how to achieve at least equivalent if not superior predictive classification performance on classifiers trained on selected features, as opposed to features selected by the original feature selection algorithm on the entire feature set; and (iii) how to reduce the overall feature selection runtimes and scalability with respect to sample size and sample resolution.

Accordingly, we demonstrate the effectiveness of *FSR* in terms of: (i) reduction in empirical execution time for selecting the top-ranked features; (ii) improvement in predictive classification accuracy when applied with existing feature selection approaches; and (iii) scalability with respect to sample size and sample resolution over stand-alone feature selection methods.

2 AIM

In this section, we present the aim of our approach, along with the required inputs and expected outputs, as well as the requirements of our proposed *FSR* algorithm.

The aim of our approach, *FSR*, is to improve the efficiency and accuracy of feature selection for the purpose of scalable classification of high-dimensional SNP microarray samples. In particular, we focus on the application of our proposed approach to the task of classifying the cancer subtype of samples based on SNP microarray datasets.

The role of *FSR* is to efficiently filter uninformative features prior to using a more computationally expensive feature selection algorithm. The inputs to *FSR* are a set of samples, each described by a vector of real-valued features giving a dataset $r\epsilon\mathbb{R}^{M\times N}$ where M is the number of features and N is the number of samples. In addition, a class label $l_n\epsilon\{1,...,C\}$ is available for each sample n where C is the number of classes in the dataset. The corresponding output is a reduced set of discrete-valued features, $\hat{r}\epsilon\mathbb{Z}^{M''\times N}$ where $M''<<M$.

We require that *FSR* be computationally efficient, so that it can be used as a pre-filter preceding a filter-based feature selection approach. Specifically, it must be scalable to high-resolution microarray datasets with >10^5 features, typical of the current generation of SNP microarrays by Affymetrix and Illumina. It must be able to deal with datasets whose number of features, M, is several orders of magnitude larger than the number of available samples N, i.e. $M>>N$. In such datasets, not all combination of values are present in the available samples, and our approach needs to be able to estimate the utility of features given such a limited number of examples.

3 *FSR* ALGORITHM

In this section, we describe our *FSR* algorithm. We introduce two measures that we propose to gauge the relevance and discrimination strength of features, which we refer to as the *distinct segment value score* and *modal entropy*. Using these measures to assess the utility of a feature, we subsequently select a compact set of features that strongly characterize samples that belong to the same class, and help to discriminate samples belonging to different classes. In the Supplementary Material, we provide a worked example to illustrate the main steps of our algorithm.

The underlying principle of the *FSR* approach is that a useful feature is one whose values are highly consistent among samples from the same class, while being sufficiently differentiated from the values that appear among samples from the other classes.

Note that raw copy number measurements, also known as $\log_2$ ratios, are continuous values that indicate the copy number status of a fragment of DNA. For the purposes of feature selection and classification, discretization of $\log_2$ ratios into labels of copy number *gain*, *loss* or *no change* often conveys sufficient information about the copy number state of the DNA fragment in question. Consequently, the use of precise continuous measurement values is not required for classification.

Since single copy changes are most common (e.g. the duplication of a single copy of a DNA segment) and changes in copy number of two or more are far less common, we choose to employ a three-level discretization to simplify the representation of copy number in SNP array datasets.

The thresholds we choose have been commonly employed in previous studies (Haverty *et al.*, 2009; Lesch *et al.*, 2010). Specifically, the selected thresholds for copy number gain and loss are $\tau_{\text{gain}}=0.3$ and $\tau_{\text{loss}}=-0.3$, i.e, $\log_2$ ratios beyond 0.3 are discretized to 1, values below -0.3 are discretized to -1 and all other values to 0. We denote the discretized matrix as $\tilde{r}\epsilon\{-1,0,+1\}^{M\times N}$.

3.1 Distinct segment value score

For each feature, there is a need to measure the consistency of feature values in each class among the samples of that class, as well as the variability of values in samples outside of this class, to help gauge the usefulness of the feature in discriminating classes. For this purpose, we have designed a measure referred to as the *distinct segment value score*. We define a segment to be a set of samples that belong to the same class. We also define a distinct feature value in our context as a discretized feature value that is frequent or consistent in one class but is infrequent or inconsistent in other classes. We begin by sorting the samples in the dataset in numeric class order so that samples of the same class are adjacent to each other. Each feature can assume one of three discretized values $\{-1, 0, +1\}$ corresponding to loss, no change and gain, respectively. In deriving the distinct segment value score, we must first define the notions of class modes and normalized counts.

3.1.1 Normalized count We define a normalized count to be the proportion of samples in a given class having a given feature value in that class. This provides an indication of how frequently a value occurs within samples of a specific class. The normalized count of feature m having value v in class segment c is defined as

$$\eta_{m,c,v}=\frac{f(m,c,v)}{f(c)} \qquad (1)$$

where the discretized features value $v\epsilon\{-1,0,+1\}$, and the count is defined as

$$f(c)=\sum_{n\epsilon 1,\ldots,N}\delta(l_n,c),\delta(a,b)=\begin{cases}1, & \text{if } a=b\\ 0, & \text{otherwise}\end{cases}.$$

More specifically, the count of samples with value v for feature m in class segment c is defined as

$$f(m,c,v)=\sum_{n\epsilon 1,\ldots,N \text{ where } l_n=c}\delta\left(\tilde{r}_{m,n},v\right)$$

3.1.2 Class mode A class mode is the most frequently occurring discretized feature value in samples belonging to a specific class. The class mode for feature m in class segment c is defined as

$$\pi_{m,c}=\underset{v\epsilon\{-1,0,+1\}}{\operatorname{argmax}}\ f(m,c,v) \tag{2}$$

3.1.3 Normalized count for class mode The normalized count for the class mode is defined intuitively as

$$\psi_{m,c}=\max_{v\epsilon\{-1,0,+1\}}\eta_{m,c,v} \tag{3}$$

3.1.4 Computation of score This score can be used to measure the consistency of the mode of a feature in the samples of a class segment. We propose the distinct segment value score as a method for rating the ability of two different values v and v' for a feature m to discriminate between samples that belong to a class c and samples that belong to the other classes $\{1,\ldots,C\}\setminus c$. For example, consider a feature m that corresponds to the first row in Figure 1a. In this idealized example, all samples in the same class have the same discretized feature values. In this example, the feature values $v=-1$ (loss) and $v'=0$ (no change) have the ability to discriminate samples that belong to class c_1 from samples that belong to the other classes c_2, c_3 and c_4. Our aim is to find a measure that can be used to identify features with these types of discriminative values. We quantify this intuition in our distinct segment value score as follows:

$$\vartheta_{m,c,v,v'}=\eta_{m,c,v}+\frac{1}{(C-1)}\sum_{c'\epsilon\{1,\ldots C\}\setminus c}\eta_{m,c',v'} \tag{4}$$

where $\eta_{m,c,v}$ corresponds to the normalized count of the value v for feature m in class segment c, while the second term is the average normalized count of value $v'\neq v$ for feature m in all other class segments $c'\neq c$. Note that if all samples in class c have the same value v for feature f, then $\eta_{m,c,v}=1$. Likewise, if all samples in all the other class segments $c'\neq c$ have the value $v'\neq v$ for feature f, then $\sum_c \eta_{m,c',v'}=C-1$. Thus $0\leq\vartheta_{m,c,v,v'}\leq 2$, where a high score corresponds to a desirable feature with values v and v' that differentiate samples of one class from samples of the other classes. For example, $\vartheta_{m,1,0,-1}$ refers to the distinct segment value score for feature m, when feature m assumes a value of 0 in class segment 1 and a value of -1 in the other class segments.

3.1.5 Feature pruning (Stage 1) In total, $M\times C\times D\times(D-1)$ combinations of the distinct segment value scores, $\vartheta_{m,c,v,v'}$ need to be computed. D is the number of discretization levels and for the *FSR* algorithm is fixed at a constant of 3. A high ϑ score indicates consistency and differentiability and thus is used as a measure to prune features. We define the empirical mean of all distinct segment value scores to be μ_ϑ, and the empirical SD of all distinct segment value scores to be σ_ϑ. For each feature, m, we compute $C\times D\times(D-1)$ distinct segment value scores. Thus the scores can be stored in an M by $(C\times D\times(D-1))$ matrix, where each column represents the scores for a particular (c,v,v'), which we refer to as a *class-versus-rest value combination*, i.e. the column of distinct segment value scores when the value v occurs in a class segment c and the value v' occurs in all other class segments. For an illustrative example, see Supplementary Figure C. The efficiency of the *FSR* algorithm can be further enhanced by parallelizing the computation of $\vartheta_{m,c,v,v'}$, since the distinct segment value score for each class-versus-rest value combination can be computed independently of each other. Parallelization here would speed up this stage of the algorithm by a factor of $C\times D\times(D-1)$ and is an area we have identified for future development.

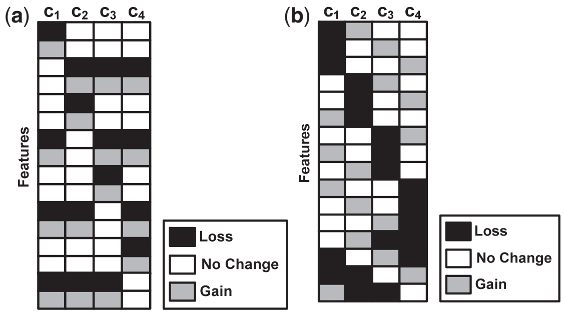

Fig. 1. Each column c_i corresponds to a class segment of samples that belong to class c_i. The figures show an idealized scenario where all samples within a class have the same feature value, which can be one of gain (grey), loss (black) or no change (white). The dataset on the left contains examples of features with distinct segment values, where one feature value can uniquely identify one of the class segments. The dataset on the right contains examples of 'two-unique' feature patterns where two of the class segments are each identified by a unique feature value. (**a**) Ideal distinct value patterns; (**b**) ideal 'two-unique' patterns.

For each class-versus-rest value combination, we consider only the indices of the top $F=C\times D$ scores as explained below. For pruning purposes, we retain the original index of the feature (in the top F) if its score exceeds the threshold of $\mu_\vartheta+\sigma_\vartheta$. This is employed to ensure that overfitting or over-representation of features for any particular class-versus-rest value combination does not occur. For three-level discretization, the value of D is constant, fixed at 3, so the limit F essentially depends on the number of classes present in the dataset. The more fragmented the dataset is in terms of classes, the greater the threshold F to ensure sufficient feature representation. After considering all class-versus-rest value combinations, duplicate feature indices that were retained are merged.

In the next subsection, we discuss the second feature-value pattern that is of interest to us. It occurs when a feature has unique values in exactly two classes such as the feature corresponding to the first row in Figure 1b. We term this type of feature-value pattern as 'two-unique', which suggests that there are exactly two class segments where the feature has a distinct unique value. Such a feature would be able to discriminate simultaneously between two classes that bear these unique values. To handle this scenario, we introduce next the concept of modal entropy.

3.2 Modal entropy

For certain features, two of the classes may each have their own unique class modes. Such features are useful in helping to discriminate between two classes simultaneously. To enable efficient identification of features that possess this property, we use the concept of *modal entropy*. This approach helps us to avoid exhaustive feature value comparisons, which could otherwise result in a combinatorial explosion. Modal entropy is the entropy of the class modes for each class segment of samples.

We define 'two-unique' as a feature that has distinct modes in exactly two class segments as shown in Figure 1b. Since it is possible to compute the modal entropy for a 'two-unique' feature pattern for datasets with any number of classes, such feature value patterns can be identified efficiently through the computation of modal entropy. The modal entropy for feature m is defined as:

$$\varphi_m=-\sum_{v\epsilon\{-1,0,+1\}}\frac{f(m,v)}{C}\log_2\frac{f(m,v)}{C} \tag{5}$$

Consider a four-class example with three-level feature value discretization. A 'two-unique' pattern for any feature occurs when we have either one of the following combinations of modes in each class segment: $\{1,0,-1,-1\}$, $\{-1,0,1,1\}$, $\{0,1,-1,-1\}$, $\{1,-1,0,0\}$, $\{-1,1,0,0\}$, $\{0,-1,1,1\}$. Note the combinations of modes listed are not exhaustive. The modal entropy of any combination shown or otherwise in this example is equal to $-0.25\times\log_2 0.25-0.25\times\log_2 0.25-0.5\times\log_2 0.5=1.5$. This value is referred to as φ_{two} in Algorithm 1, i.e. it is the modal entropy value which corresponds to a feature that possesses the 'two unique' pattern and can be pre-computed for any number of classes. After computing the modal entropy for all features, we can filter the feature indices based on the pre-computed threshold φ_{two}.

3.2.1 Feature pruning (Stage 2) After the set of features displaying 'two-unique' patterns has been identified, this set is further pruned to ensure that the modes in each class segment have a clear majority within that class segment.

For this purpose, we introduce a measure known as the *sum of normalized counts for class modes*. As defined previously, the normalized counts indicate the proportion of samples in each class segment that have a particular discretized feature value. When applied to class modes, the normalized counts for class modes simply refer to the proportion of samples in each class segment that possess the most frequently occurring discretized feature value. For example, if there is a normalized count of $\Psi_{1,1}=0.66$ for feature 1 in class segment 1 compared to a normalized count of $\Psi_{2,1}=0.5$ for feature 2 in class segment 1, then feature 1 has a mode that is more consistent in the class segment and is more likely to be better at discriminating for class segment 1 than feature 2.

The sum of normalized counts for class modes γ_m for a feature m is simply the aggregation of all normalized counts for class modes across all class segments for the 'two-unique' feature m.

$$\gamma_m=\sum_{c\epsilon 1,\dots,C}\psi_{m,c} \tag{6}$$

A higher sum indicates a more consistent 'two-unique' pattern for that feature. The maximum score attainable is equal to the number of classes in the dataset, C. A score equal to C implies that the feature has an ideal 'two-unique' pattern where all the samples in each class segment have the same value (i.e. the mode) and exactly two class segments have distinct modes. This ideal scenario is illustrated in Figure 1b. In reality, it is unlikely that this ideal scenario will occur. Therefore, we select a more realistic threshold of $\gamma_m>\frac{2C}{3}$, where C is the number of classes in the dataset, to retain the set of features that exhibit a sufficiently strong 'two-unique' pattern. This threshold ensures that the mode is sufficiently consistent within each class segment without being too stringent in pruning the features. Refer to the Supplementary Material for examples of the empirical distribution of γ_m in three different evaluation datasets.

3.2.2 Output feature set The output of the *FSR* algorithm is the union set of features selected by the two filtering stages, i.e. distinct value and 'two-unique'. This feature set can then be used as the input to a subsequent feature selection algorithm for ranking. The ordered features can then be used to train a classifier for the task of classifying samples into their cancer subtype. This *FSR* algorithm is summarized in Algorithm 1. Note that this algorithm can also be applied to binary class datasets ($C=2$) as discussed in Supplementary Section 2.

The *FSR* algorithm reduces the feature set by identifying features that consistently exhibit either of the two patterns of discrimination discussed. The *FSR* algorithm does not collectively rank the two reduced sets of features. We do not feel a compelling need to further implement a collective ranking / selection stage for *FSR* as after the dimensions of the feature sets have been reduced substantially by at least two orders of magnitude, existing feature selection algorithms can be employed rather efficiently for the purpose of feature selection or ranking.

Algorithm 1 *FSR*

```
Input: $r_{m,n}$ : raw copy number of feature $m$ in sample $n$
    $l_{1,\dots,N}\epsilon\{1,\dots,C\}$, $l_n$ : numeric class label for sample n,
    $\tau_{\text{loss}}=-0.3$, $\tau_{\text{gain}}=0.3$, $D=3$ (discretisation levels)
Output: reduced feature set, $\hat{r}_{1,\dots,M'',1,\dots,N}$ where $M''<<M$
/* Class Segment Partitioning */
    Sort columns of $r$ in ascending numeric class order, $\tilde{r}\leftarrow sortcolumns(r)$
Discretize $\tilde{r}$ based on thresholds of $\tau_{\text{loss}}$ and $\tau_{\text{gain}}$
Compute the normalized counts, $\eta_{m,c,v}$, for all values in each class segment
Determine the mode, $\pi_{m,c}$ for each class segment and its associated normalized count, $\psi_{m,c}$
Compute distinct segment value scores $\vartheta_{m,c,v,v'}$, for every pair of feature values $(v,v')$
/* Computation of Population Statistics for Distinct Segment Value Scores */
    Compute the overall empirical mean, $\mu_\vartheta$ and standard deviation, $\sigma_\vartheta$ of distinct segment value scores, $\vartheta$
/* Feature Pruning Stage 1 */
    For each class-versus-rest value combination $(c,v,v')$ retain the top $C\times D$ features for each if their scores $\vartheta_{m,c,v,v'}>(\mu_\vartheta+\sigma_\vartheta)$ avoiding duplicate features, giving the pruned feature set, $F'$.
Calculate class modal entropy, $\varphi_m$, for each feature
Calculate the sum of normalized counts of class modes $\gamma_m$ for each feature
Compute the entropy for unique modes in exactly two class segments ('two-unique' pattern), $\varphi_{\text{two}}$
/* Feature Pruning Stage 2 */
    Prune $\tilde{r}$ and retain features whose $\gamma_m>\frac{2C}{3}$ and $\varphi_m=\varphi_{\text{two}}$. Residual features form the second pruned feature set, $\tilde{r''}$
Remove duplicate features from the union of the two pruned feature sets, $\tilde{r'}$ and $\tilde{r''}$, the remaining features form the ultimate reduced feature set, $\hat{r}$
return $\hat{r}$
```

4 EVALUATION

In this section, we begin by describing the aims of our evaluation. Next, we discuss the choice of datasets and classifier used to evaluate the performance of *FSR* in our experiments, and the procedure for preprocessing the raw data files (CEL files) acquired to produce raw copy number data ($\log_2$ ratios). We then discuss the choice of feature selection approaches selected for use with *FSR*, and the choice of classifier used to evaluate the predictive classification accuracy of selected features from an *FSR*-reduced feature set. We subsequently present our results on predictive classification accuracy over the selected datasets with and without the application of *FSR*, as well as the speedup in execution times for feature selection with the use of *FSR*, and the scalability of *FSR* with respect to sample size, sample resolution as well as the number features selected. We also examined the biological relevance of the top features selected by *FSR*-paired approaches.

4.1 Aim of evaluation

The aims of our evaluation are to assess the performance of *FSR* in terms of predictive classification accuracy and change in execution times when paired with existing filter-based feature selection approaches, as well as its scalability with regards to sample size, sample resolution and the number of features selected.

4.2 Datasets

4.2.1 Choice of datasets The datasets (sarcoma, lymphoma and leukemia) and their respective CEL files used in our experiments were acquired from the Gene Expression Omnibus (GEO) high throughput public data repository, which is built and maintained by the National Center for Biotechnology Information (NCBI) (Barrett *et al.*, 2009). The datasets were chosen because they were publicly available, had relatively large sample sizes and provide some diversity in terms of the range of subtypes among the samples. The classification task applied to the selected datasets is that of cancer subtype discrimination, where subtype classes were previously determined in the original references as shown in Supplementary Table S1.

We also obtained three independent validation datasets based on mucosa-associated lymphatic tissue (Rinaldi *et al.*, 2011), mantle cell lymphoma (Kawamata *et al.*, 2009) and chronic lymphocytic leukemia (Gunnarsson *et al.*, 2011) to evaluate the robustness of our trained classification models to independent datasets.

4.2.2 Preprocessing The $\log_2$ ratios were processed from the SNP array CEL files using the publicly available CNAG software package (Nannya *et al.*, 2007). Tumor samples were normalized from a pool of 20 unmatched references randomly selected from the HapMap project (The International HapMap Consortium, 2003) for the specific array type. For gender-based diseases, the references selected were of the same sex. No subsequent segmentation or smoothing was applied to the generated raw copy number data ($\log_2$ ratios). Non-autosomal probesets are excluded from the analysis for non-gender specific diseases to avoid confounding copy number differences in the sex chromosomes. A summary of the size of the complete feature set (probesets) for all datasets used can be found in Supplementary Table S1.

4.3 Feature selection approaches for coupling

Filter-based approaches and wrapper-based approaches are the two main categories of feature selection approaches. Filter-based approaches evaluate the usefulness of a feature subset by examining the intrinsic characteristics of the data, based on the relation of each single feature with the class label according to a pre-defined metric. Features are ranked in terms of the metric based on some probabilistic or distance measure, with the top scoring features chosen to build the classifier. In essence, filter-based approaches perform feature selection independent of the classification method. Classifier independence suggests that it is possible that the selected features might not be optimal for the eventual classifier. However this also implies that overfitting is also less likely to occur with filter-based feature selection. Filter-based approaches are commonly used for gene selection in microarray sample classification tasks (Dudoit *et al.*, 2002; Kononenko, 1994; Ooi *et al.*, 2006; Peng *et al.*, 2005).

With wrapper-based feature selection approaches, the feature selection algorithm is coupled with the classification algorithm. The wrapper-based approach is not extensively used in microarray sample classification, primarily due to the large combination of feature subsets that needs to be explored. For this reason, we choose to focus on filter-based feature selection approaches.

We have selected three approaches to be coupled with *FSR* for the purpose of assessing the predictive classification accuracy of a classifier built on the selected features. The feature selection approaches (mRMR, reliefF and F-test) were selected because of their contrasting optimization functions, their relative computational efficiency and on the basis that they are commonly used in the domain of gene expression microarray analysis, which is closest to our application domain of SNP microarrays since no specific approach has been developed for feature selection on copy number data produced on SNP microarrays. The minimum Redundancy–Maximum Relevance (mRMR) (Peng *et al.*, 2005) feature selection approach frames feature selection as a dual optimization problem, maximizing relevance with class labels and minimizing redundancy among features. mRMR handles both continuous and discrete-valued features. The authors have proven that it is an optimal first-order incremental selection. ReliefF (Kononenko, 1994) is an approach designed to estimate the quality of attributes in problems with strong dependencies between attributes. The key idea of reliefF is to estimate the quality of attributes according to how well their values distinguish between instances that are similar to each other. The F-test is a statistical approach that performs ranking according to the F-statistic between features and class labels. The F-statistic test is a generalization of the t-statistic for two classes.

Our MATLAB implementation of mRMR, reliefF and F-test was based on the libraries of LIBGS (Zhang *et al.*, 2010).

4.4 Classifier selection

We used a multiclass SVM as our classifier, as it is stable in performance with highly competitive classification accuracy over other classification and clustering approaches. The multi-class SVM used in our experiments is based on the one-versus-rest strategy (Crammer and Singer, 2002) and was implemented in MATLAB as part of the LIBLINEAR package version 1.8 (Fan *et al.*, 2008). LIBLINEAR is a linear classifier package that supports data with millions of instances and features, making it highly suited to the dimensions of SNP array-based datasets.

4.5 Results

4.5.1 Predictive classification accuracy We evaluate the predictive classification accuracy of our feature prefiltering approach by employing 3-fold cross-validation on an SVM trained on the prefiltered feature set. The 3-fold cross-validation performed was stratified, i.e. we ensured that the class distribution of samples in each fold was approximately equal. *FSR* was only applied to the folds that were selected for training the classifier, and the samples in the test fold were completely concealed from the feature reduction stage. In Figure 2, we report the accuracy through 3-fold cross-validation achieved on each of the test datasets. We also report in Table 1 the corresponding execution times for the *FSR*-paired approaches in comparison to the stand-alone feature selection approaches.

In summary, the three multiclass datasets to which we applied our *FSR* approach achieved predictive accuracies that were significantly higher than the majority class percentages of each respective dataset, which is our performance baseline. This supports the effectiveness in the use of SNP microarray datasets in the task of disease subtype determination in the context of the various cancers that we have considered. Generally, for all three datasets

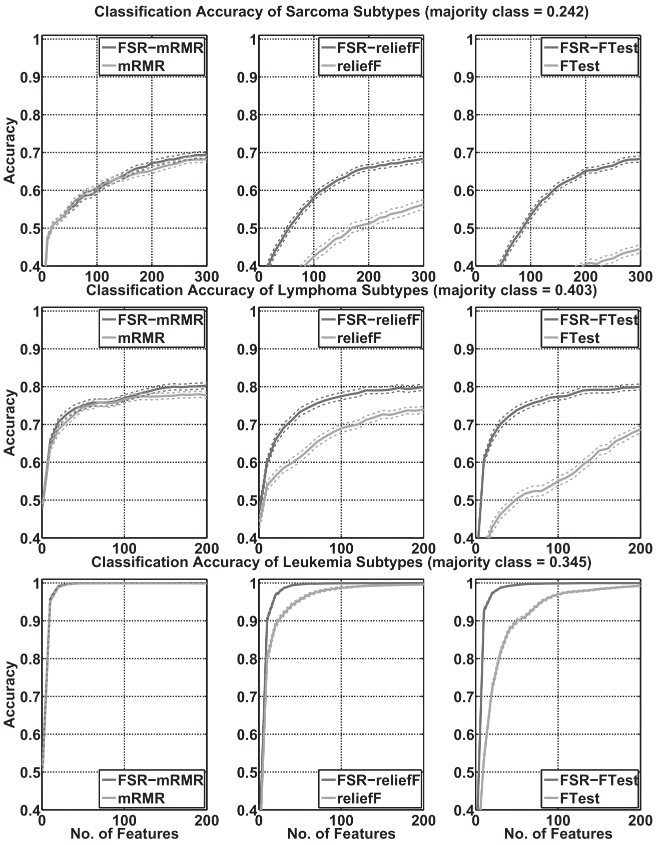

Fig. 2. The 3-fold cross-validation accuracy. The dotted lines represent the 95% confidence intervals for mean accuracy. Non-overlapping confidence intervals are deemed to be statistically significant. We observed that *FSR*-coupled approaches are generally at least of equal performance to the corresponding stand-alone feature selection approaches and in most cases exhibit a statistically significant improvement over the stand-alone feature selection approach. The classification accuracy achieved with *FSR* coupling was clearly superior to the majority class percentages, which is our baseline for each dataset.

the classification accuracy of the *FSR*-paired approaches exceeds the accuracy achieved by the respective stand-alone approaches. There were, however, three cases (Sarcoma: mRMR and reliefF, Lymphoma: reliefF) where for small numbers of features (<100), the stand-alone approaches achieved marginally better classification accuracy than the *FSR*-paired approach, but eventually leveled off to a significantly lower accuracy as the number of features increase. It is also worth noting that across the three *FSR*-paired scenarios, there was much less variation in the maximum accuracy achieved using the different hybrid schemes (*FSR*-mRMR, *FSR*-reliefF and *FSR*-FTest). For example, for the sarcoma dataset, mRMR achieved a mean maximum accuracy of ∼0.68 (@ 300 features) while FTest achieved an accuracy of only 0.45 (@ 300 features), whereas for the paired approaches of *FSR*-mRMR and *FSR*-FTest, the respective accuracies were ∼0.695 and 0.68. This suggests that our *FSR*-paired approach achieves not only superior classification accuracy, but also more stable classification performance.

We also assessed our trained classification models on three independent cancer datasets. Based on the known cancer type (sarcoma, lymphoma and leukemia) of each given sample, the appropriate classifier is used to determine the cancer subtype of the sample. Each of the three classifiers were trained with the top ω most frequently selected features from 3-fold cross-validation, where ω is determined as the number of features at which point classification accuracy starts to level-off (this was identified from the plots shown in Fig. 2). The number of selected features used for training the lymphoma cancer subtype classifier was 190 while for the leukemia subtype classifier this was 150.

After training, we classified all the samples from the three independent validation datasets with the relevant classifier (as determined by its cancer type). The accuracies obtained for each of the three independent test datasets is summarized in Supplementary Table S4. For each dataset, the accuracy achieved with the *FSR*-coupled approach was superior to the classification achieved with the corresponding stand-alone approach. The classification accuracies obtained for each subtype dataset was comparable to the accuracy obtained previously via 3-fold cross-validation.

In Section 6 in Supplementary Material, we compare the classification accuracy and execution time of using an SVM classifier based on all features (i.e. without feature selection) compared with an SVM classifier on the *FSR*-reduced datasets. We show that by using *FSR* + SVM, we can achieve a statistically significant improvement in accuracy compared with SVM alone, with an order of magnitude reduction in execution time.

The impact of different discretization levels and segmentation on predictive classification accuracy was also assessed and documented in detail in Sections 7 and 8 of the Supplementary Material. In summary, our selected discretization threshold of 0.3 was robust and performed better than other thresholds, and the application of segmentation on copy number did not improve predictive classification accuracy, presumably since narrow informative regions of copy number change are lost as a result. Furthermore, in Section 10 of the Supplementary Material we show that feature reduction using other traditional approaches such as principle component analysis are less effective than *FSR* at achieving good classification performance on SNP datasets.

4.5.2 Execution time for feature selection We implemented *FSR* on MATLAB R2010b. In Table 1, we compare the empirical runtimes for selecting the top ranked features for each dataset. All experiments were executed on a PC equipped with an Intel Core-i7 920 processor with 12 GB DDR3 RAM. Our results show that the use of *FSR* reduces the mean empirical runtimes for feature selection by at least an order of magnitude and extends to beyond two orders of magnitude for reliefF.

4.5.3 Feature set reduction In Table 2, we compare the size of the original feature set to the size of the *FSR*-reduced feature set. Essentially, a dataset of hundreds of thousands of features is typically reduced to hundreds of features. As demonstrated by the classification accuracy achieved on the pruned datasets in Figure 2, we can conclude that the pruned feature set provides a compact and informative representation of the entire feature set as the features selected from the pruned feature set have yielded at least equal and, more frequently, a statistically significant

Table 1. A comparison of empirical runtimes (in seconds) for selection of the top ϕ ranked features in 3-fold cross validation

	ϕ	mRMR	*FSR*-mRMR	Speedup	reliefF	*FSR*-reliefF	Speedup	F-test	*FSR*-F-test	Speedup
Sarcoma	300	263.3	11.7	22.5×	189.4	1.89	100.1×	36.3	1.42	25.6×
Lymphoma	200	126.0	2.52	49.9×	133.5	1.16	115.5×	54.4	1.06	51.2×
Leukemia	200	98.8	3.59	27.2×	270.5	1.91	141.8×	53.9	1.74	30.9×
Mean				33.2×			119.1×			35.9×

Overall, feature selection with our hybrid approach was at least an order of magnitude faster than stand-alone feature selection using mRMR, reliefF or F-test alone. The empirical runtimes recorded here for the *FSR*-paired approaches consist of the time required to reduce the original feature set with *FSR* as well as the time to rank the reduced feature set with each of the three selected feature selection approaches. The maximum memory used by *FSR* for each dataset is ≃190 MB for the sarcoma dataset, ≃122 MB for the lymphoma dataset and ≃332 MB for the leukemia dataset.

Table 2. For the tested SNP array cancer datasets, *FSR* was able to reduce each feature set by more than two orders of magnitude and hence enhance the efficiency of the paired filter-based feature selection approach in ranking the top features from the *FSR*-pruned feature space

	No. of original features	Size of reduced feature set	As a percentage of original feature set (%)	Factor of reduction
Sarcoma	238 300	789	0.33	302×
Lymphoma	255 866	238	0.09	1075×
Leukemia	255 866	326	0.13	785×

The size of the reduced feature set depends on the level of redundancy of the features in the dataset, the variance of the values in the dataset as well as the number of classes.

improvement in classification accuracy compared with their stand-alone counterparts.

4.5.4 Scalability Scalability is an important consideration for feature selection algorithms as the cost of higher resolution arrays falls and the use of higher resolution arrays becomes commonplace. It is also important that feature selection algorithms scale to large sample sizes as studies with several thousand samples are gradually becoming more common.

Number of features selected: for certain feature selection algorithms, selecting a different number of top ranked features does not impact runtimes since the entire set of features are already ranked. For approaches such as mRMR, this is not the case. Other approaches such as DDP (Ooi *et al.*, 2006) scale even more poorly with respect to the number of top ranked features selected. However, with *FSR*-mRMR, the complexity with respect to the number of features selected is reduced from quadratic to linear as shown in Figure 3A.

Sample resolution: sample resolution refers to the resolution of the microarray used in the experiment. For Affymetrix SNP microarrays, it refers to the number of probesets present on the chip. The latest generation of Affymetrix microarrays has ∼1.8 million probesets, which are a combination of polymorphic probesets covering SNPs and non-polymorphic probesets interrogating copy number exclusively. In Figure 3B, we assess the runtimes of each stand-alone algorithm and also each *FSR*-paired approach in selecting the top 100 features from a simulated dataset of 4 classes and 200 samples. The log2 ratios were generated from a $\mathbf{N}(0,0.35^2)$ distribution, which is reasonably close to the distribution parameters of the natural experimental datasets for cancer subtypes explored in this article.

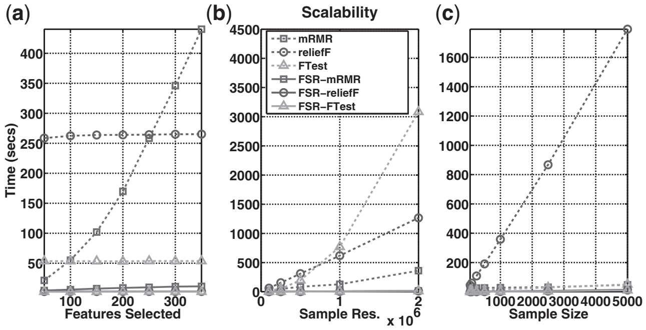

Fig. 3. Scalability of *FSR*-paired approaches. (**a**) mRMR has quadratic complexity to the number of features selected. *FSR*-mRMR improves on this with a linear complexity to the number of features selected with a very gentle gradient. This is largely due to the greatly reduced search space with *FSR* pairing. Our results show that *FSR* reduces the feature space by more than two orders of magnitude. (**b**) reliefF and mRMR appear to scale linearly to sample resolution while the F-test appears to scale quadratically. All three *FSR*-paired combinations execute in close to constant time with respect to sample resolution (i.e. number of features). (**c**) reliefF has clear linear complexity to sample size (albeit a rather steep one) while mRMR and F-test are less evidently linear as well. The *FSR*-pairings show near constant complexity to sample size.

Sample size: for the purpose of classification performance, learning from more samples will help improve classification accuracy if the samples possess a good-quality signal. Certain feature selection algorithms are not suited to large sample sizes and hence are unable to exploit the availability of a larger training set. In Figure 3C, we see that the time complexity of the *FSR*-paired approaches are almost constant with respect to the sample size.

5 DISCUSSION

5.1 Biological relevance of features selected

A further aim was to determine if the associated genes of the most frequently selected SNPs using the *FSR*-coupled feature selection approaches were biologically relevant. Detailed results of our evaluation of biological relevance are included in the Supplementary Material. In summary, for the lymphoma dataset on which we based our analysis, we found strong enrichment of genes associated with cancer and more specifically with many subtypes of lymphoma. In contrast, for each set of genes associated with the most frequently selected SNPs in the stand-alone approaches, the association with cancer and lymphoma was less frequent. This provides evidence that the most frequently selected SNPs by *FSR*-coupled approaches are not spurious but highly involved in cancer progression. Biological relevance in the most frequently selected features could partly explain better generalization achieved with *FSR*-coupled approaches leading to improved classification accuracy.

5.2 Comment on utilizing *FSR* for the discovery of biologically relevant regions

FSR selects SNPs that are representative of copy number changes either in focal regions or longer contiguous regions without a bias for the selection of either. Selected SNPs are essentially markers for the regions. In some cases, more than one SNP will be selected from a particular region depending on the *FSR* computed scores. However, it is unlikely that all SNPs will be selected for any particular region. A potential advantage is that the selection of features (SNPs) distributed over multiple genomic loci is likely to be more robust in classification performance compared with an approach that utilizes all SNPs within a few regions. Many of the SNPs within each region are highly correlated and therefore will result in a highly redundant feature set with many features providing equivalent information. For a fixed number of features, this could adversely impact the generalization ability and ultimately the classification accuracy resulting from an over-representation of SNPs in particular regions and insufficient representation in other regions. Having multiple markers over multiple regions is likely to be more robust for classification purposes.

FSR is intended for usage as a tool for ‘compressing’ the representation of a SNP dataset with minimal loss of informative SNPs, and therefore minimizing redundancy is an important consideration. From the identified markers, the user can be directed to regions of the genome where aberrations are consistent across genomes. The approach, however, does not provide the actual genomic boundaries of the aberration. For this purpose, we would suggest the use of a segmentation tool such as HaarSeg (Ben-Yaacov and Eldar, 2008). However with segmentation, there is a bias for long contiguous regions of copy number change and narrow regions are often ‘lost’ through the application of a segmentation algorithm on copy number. These ‘smoothed’ narrow regions of copy number change could potentially be informative for subtype discrimination and therefore their loss could result in poorer classification performance.

5.3 Comment on use of prior biological knowledge

Certain feature selection approaches (Guan *et al.*, 2009; Zhao *et al.*, 2009), most commonly in the gene expression domain, incorporate the use of prior biological knowledge in terms of known genes associated with disease to prune the dimensions of the dataset and thus enhance feature selection efficiency and performance. Reliance on prior biological knowledge, however, would limit the discovery of new biomarkers that can potentially discriminate well between disease subtypes. Thus for this reason, we have chosen not to incorporate the use of prior biological knowledge in our *FSR* algorithm to ensure the discovery of all relevant markers in the dataset for cancer subtype discrimination.

5.4 Comment on copy number versus gene expression for cancer subtype profiling

In recent years, gene expression analysis has been conducted in tandem with genotyping experiments on SNP arrays to provide an integrative understanding of the underlying mechanisms of disease progression and to rationalize changes observed in gene expression. The objective of our proposed approach *FSR* parallels this. We are proposing the use of copy number profiles for classification of cancer subtypes to enhance the robustness of subtype determination in various cancers over the exclusive use of gene expression data. Copy number, an absolute measure of physical change in the genome, has been shown to be a stable and reliable measure for the discrimination of several cancer subtypes.

In ovarian cancer (Ramakrishna *et al.*, 2010), gene expression analysis was applied to discover relevant pathways and subgroups of clinical importance. However, variability in percentage cancer epithelium and the use of diverse control tissues were cited for the lack of consensus between related ovarian cancer studies based on gene expression alone. More importantly, the authors also state that genomic alterations may provide a more robust and reliable basis for predicting the location of driver genes over the use of gene expression. Gorringe *et al.* (2007) demonstrated that genomic aberrations that characterize most ovarian cancers were highly comparable at a global level i.e. very similar regions of frequent copy number aberrations were identified in related studies.

In the context of breast cancer (Bergamaschi *et al.*, 2006), distinct spectra of copy number aberrations were identified underlying the different subtypes of breast cancer, suggesting that breast cancer subtypes progress along distinct genetic pathways. The distinct copy number profiles observed in these subtypes provide further impetus to pursue the potential utility of DNA copy number for accurate cancer subtype profiling and classification.

In view of these examples, we believe that copy number can serve as an accurate alternative basis for the classification and profiling of various cancer subtypes.

5.5 Application of *FSR* to next-generation high-throughput sequencing data

Data from sequencing can be used to determine copy number variation over very narrow regions. Data of this nature is well-suited for *FSR* reduction. The sequencing data first needs to be aligned with a reference genome. Then copy number measurements are taken as the ratio of counts between each sample and a reference. These ratios need to be computed for each unique region. These copy number measurements can then take the place of the copy number measurements produced on SNP microarrays. The nature of sequencing data facilitates the detection of more fine-scaled markers of copy number change to help discriminate between cancer subtypes.

6 CONCLUSION

In this article, we have presented a novel feature set reduction method, *FSR*. We have demonstrated its effectiveness in pruning complete feature sets involving a variety of cancers down to compact sets of potentially interesting features, which ultimately yielded better predictive classification accuracy for an SVM trained on these features. We have shown that *FSR* is computationally efficient. The speedup in empirical runtimes achieved with an *FSR*-hybrid feature selection approach was at least an order of magnitude faster as compared with feature selection using the respective stand-alone approach. *FSR*-paired approaches also demonstrated significantly better classification accuracy over their stand-alone counterparts in many of the test scenarios.

The performance of *FSR* has implications and applications in the field of cancer genomics. Cancer is often regarded as a single disease with many variations. Each subtype of cancer has a different set of risk factors, different rates of progression, different treatment options and a different prognosis. Being able to classify cancer subtypes more accurately from selected biomarkers can be useful in achieving a more accurate prognosis of the disease as well as in making more informed choices about treatment options.

From a computational perspective, the ability to prune feature sets with minimal loss of information facilitates the subsequent application of a wider selection of feature selection approaches, some of which may have been intractable on the original feature set due to its size. *FSR* is suited for application on high-density microarray datasets to analyze diseases such as cancer that typically have a small to moderate number of classes reflecting subtypes, grades or stages of the disease. As microarray resolution grows to improve coverage of the human genome, *FSR* can be a useful tool in enhancing the computational efficiency and predictive accuracy of disease subtype classification tasks performed on the datasets produced on these microarrays for potentially more accurate disease prognosis and personalized treatment.

ACKNOWLEDGEMENTS

We would like to thank our anonymous reviewers for their valuable feedback which has lead to various improvements in this manuscript and Justin Bedo for helpful discussions.

Funding: NICTA. NICTA is funded by the Australian Government through the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

Conflict of Interest: none declared.

REFERENCES

Barrett,T. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885.

Bastian,B. *et al.* (2003) Classifying melanocytic tumors based on DNA copy number changes. *Am. J. Pathol.*, **163**, 1765.

Ben-Yaacov,E. and Eldar,Y. (2008) A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*, **24**, i139–i145.

Bergamaschi,A. *et al.* (2006) Distinct patterns of dna copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer*, **45**, 1033–1040.

Crammer,K. and Singer,Y. (2002) On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, **2**, 265–292.

Dudoit,S. *et al.* (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Fan,R. *et al.* (2008) LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, **9**, 1871–1874.

Gorringe,K. *et al.* (2007) High-resolution single nucleotide polymorphism array analysis of epithelial ovarian cancer reveals numerous microdeletions and amplifications. *Clin. Cancer Res.*, **13**, 4731–4739.

Guan,P. *et al.* (2009) Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method. *J. Exp. Clin. Cancer Res.*, **28**, 1–7.

Gunnarsson,R. *et al.* (2011) Array-based genomic screening at diagnosis and follow-up in chronic lymphocytic leukemia. *Haematologica*, **96**, 1161–1169.

Haverty,P. *et al.* (2009) High-resolution analysis of copy number alterations and associated expression changes in ovarian tumors. *BMC Med. Genomics*, **2**, 21.

Hu,N. *et al.* (2005) Genome-wide association study in esophageal cancer using GeneChip mapping 10K array. *Cancer Res.*, **65**, 2542–2546.

Kawamata,N. *et al.* (2009) Identified hidden genomic changes in mantle cell lymphoma using high-resolution single nucleotide polymorphism genomic array. *Exp. Hematol.*, **37**, 937–946.

Kononenko,I. (1994) Estimating attributes: Analysis and extensions of relief. In Bergadano,F. and Raedt,L. D. (eds) *European Conference on Machine Learning*. Springer, New York, pp. 171–182.

Lesch,K.P. *et al.* (2010) Genome-wide copy number variation analysis in attention-deficit / hyperactivity disorder: association with neuropeptide Y gene dosage in an extended pedigree. *Mol. Psychiatry*, **1**, 13.

Nannya,Y. *et al.* (2007) Evaluation of genome-wide power of genetic association studies based on empirical data from the HapMap project. *Hum. Mol. Genet.*, **16**, 2494–2505.

O'Hagan,R. *et al.* (2003) Array comparative genome hybridization for tumor classification and gene discovery in mouse models of malignant melanoma. *Cancer Res.*, **63**, 5352.

Ooi,C. *et al.* (2006) Differential prioritization between relevance and redundancy in correlation-based feature selection techniques for multiclass gene expression data. *BMC Bioinformatics*, **7**, 320.

Peng,H. *et al.* (2005) Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 1226–1238.

Ramakrishna,M. *et al.* (2010) Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis. *PLoS One*, **5**, e9983.

Rinaldi,A. *et al.* (2011) Genome-wide DNA profiling of marginal zone lymphomas identifies subtype-specific lesions with an impact on the clinical outcome. *Blood*, **117**, 1595.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507.

Statnikov,A. *et al.* (2007) Effects of environment, genetics and data analysis pitfalls in an esophageal cancer genome-wide association study. *PloS One*, **2**, 958.

The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Wang,S. *et al.* (2007) An interval tree based feature reduction method for cancer classification using high-throughput DNA copy number data. In *International Conference on Bioinformatics and Computational Biology, BIOCOMP*, vol. 1. CSREA Press, pp. 248–55.

Wang,Y. *et al.* (2006) Tumor classification based on DNA copy number aberrations determined using SNP arrays. *Oncol. Rep.*, **15**, 1057–1060.

Wang,Y. (2006) Cancer classification using loss of heterozygosity data derived from single-nucleotide polymorphism genotyping arrays. In *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2006. EMBS'06*. IEEE Service Center, Piscataway, NJ, pp. 5864–5867.

Zhang,Q. *et al.* (2010). CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics*, **26**, 464.

Zhao,C. *et al.* (2009) Noninvasive detection of candidate molecular biomarkers in subjects with a history of insulin resistance and colorectal adenomas. *Cancer Prev. Res.*, **2**, 590.