Genome Medicine

This Provisional PDF corresponds to the article as it appeared upon acceptance. Copyedited and fully formatted PDF and full text (HTML) versions will be made available soon.

A kernel-based integration of genome-wide data for clinical decision support

Genome Medicine 2009, **1**:39 doi:10.1186/gm39

Anneleen Daemen (anneleen.daemen@esat.kuleuven.be)
Olivier Gevaert (olivier.gevaert@esat.kuleuven.be)
Fabian Ojeda (fabian.ojeda@esat.kuleuven.be)
Annelies Debucquoy (annelies.debucquoy@med.kuleuven.be)
Johan AK Suykens (johan.suykens@esat.kuleuven.be)
Christine Sempoux (christine.sempoux@clin.ucl.ac.be)
Jean-Pascal Machiels (jean-pascal.machiels@uclouvain.be)
Karin Haustermans (karin.haustermans@uz.kuleuven.ac.be)
Bart De Moor (bart.demoor@esat.kuleuven.be)

ISSN 1756-994X

Article type Research

Submission date 4 November 2008

Acceptance date 3 April 2009

Publication date 3 April 2009

Article URL http://www.genomemedicine.com/content/1/4/39

This peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in *Genome Medicine* are listed in PubMed and archived at PubMed Central.

For information about publishing your research in *Genome Medicine* go to

http://www.genomemedicine.com/info/instructions/

© 2009 Daemen *et al.* , licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A kernel-based integration of genome-wide data for clinical decision support

Anneleen Daemen*[1], Olivier Gevaert[1], Fabian Ojeda[1], Annelies Debucquoy[2], Johan A K Suykens[1], Christine Sempoux[3], Jean-Pascal Machiels[4], Karin Haustermans[2]and Bart De Moor[1]

[1]Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, Leuven, Belgium
[2]Department of Experimental Radiotherapy, Katholieke Universiteit Leuven, UZ Herestraat 49, Leuven, Belgium
[3]Department of Pathology, Université Catholique de Louvain, St Luc University Hospital, 10 Avenue Hippocrate, Brussels, Belgium
[4]Department of Medical Oncology, Université Catholique de Louvain, St Luc University Hospital, 10 Avenue Hippocrate, Brussels, Belgium

Email: Anneleen Daemen - anneleen.daemen@esat.kuleuven.be;

*Corresponding author

Abstract

Background: Although microarray technology allows the investigation of the transcriptomic make-up of a tumour in one experiment, the transcriptome does not completely reflect the underlying biology due to alternative splicing, post-translational modifications, as well as the influence of pathological conditions (e.g., cancer) on transcription and translation. This increases the importance of fusing more than one source of genome-wide data such as the genome, transcriptome, proteome, and epigenome. The current increase in the amount of available omics data emphasizes the need for a methodological integration framework.

Methods: We propose a kernel-based approach for clinical decision support in which many genome-wide data sources are combined. Integration occurs within the patient domain at the level of kernel matrices before building the classifier. As supervised classification algorithm, a weighted Least Squares Support Vector Machine is used. We apply this framework on two cancer cases, namely, a rectal cancer data set containing microarray and proteomics data and a prostate cancer data set containing microarray and genomics data. For both cases, multiple outcomes are predicted.

Results: For the rectal cancer outcomes, the highest leave-one-out (LOO) areas under the receiver operating characteristic curves (AUC) were obtained when combining microarray and proteomics data gathered during therapy and ranged from 0.927 to 0.987. For prostate cancer, all four outcomes had a better LOO AUC when

combining microarray and genomics data, ranging from 0.786 for recurrence to 0.987 for metastasis.

Conclusions: For both cancer sites the prediction of all outcomes improved when more than one genome-wide data set was considered. This suggests that integrating multiple genome-wide data sources increases the predictive performance of clinical decision support models. This emphasizes the need for comprehensive multi-modal data. We acknowledge that in a first phase this will substantially increases costs; however, this is a necessary investment to ultimately obtain cost-efficient models usable in patient tailored therapy.

Background

Kernel Methods are a powerful class of methods for pattern analysis. In recent years, they have become a standard tool in data analysis, computational statistics, and machine learning applications [1]. Based on a strong theoretical framework, their rapid uptake in applications such as bioinformatics [2], chemoinformatics, and even computational linguistics is due to their reliability, accuracy, and computational efficiency. In addition, they have the capability to handle a very wide range of data types (e.g. kernel methods have been used to analyze sequences, vectors, networks, phylogenetic trees, etc). The ability of kernel methods to deal with complex structured data makes them ideally positioned for heterogeneous data integration. More specifically in this contribution, we used a weighted Least Squares Support Vector Machine (LS-SVM), an extension of the Support Vector Machine (SVM) for supervised classification [3], [4], [5]. The LS-SVM is, compared to the SVM, easier and faster for high dimensional data because the quadratic programming problem is converted into a linear problem. To account for the unbalancedness in many two-class problems, this linear problem is extended with weights, different for the positive and negative class.

The growing amount of data combined with factors such as time, cost, and personalized treatment is complicating clinical decision making. Using advanced mathematical models such as the above mentioned LS-SVM can aid clinical decision support because information arising from clinical risk factors (e.g. tumour size, number of lymph nodes) is not accurate enough to reliably predict patient prognoses. Patients with the same clinical and pathological characteristics but different clinical outcomes can potentially be

discerned with microarray technology. This technology investigates the transcriptomic make-up of a tumour in one experiment. A decade ago, it was first used in cancer studies to classify tissues as cancerous or non-cancerous [6], [7]. Within the domain of cancer, microarray technology has earned a prominent place for its capacity to characterize the underlying tumour behaviour in detail. Although the first gene expression profile signature is being validated in clinical trials [8], [9], [10], the microarray technology can not measure the complete transcription due to the limited number of probes per gene on a chip, nor does the transcriptome completely reflect the biology underlying a disease. Besides transcription, pathological conditions such as cancer also influence alternative splicing, chromosomal aberrations, and methylation [11], [12].

For example, chromosomal aberrations have been found in the general population as well as in all major tumour types [13], [14]. These regions of increased or decreased deoxyribonucleic acid (DNA) copy number can be detected using e.g. array comparative genomic hybridization (CGH) technology. This technique measures copy number variations (CNV) within the entire genome of a disease sample compared to a normal sample [11]. Many small aberrations have emerged as prognostic and predictive markers. Numerous aberrations, however, also affect large genomic regions, encompassing multiple genes or whole chromosome arms.

Secondly, due to differential splicing or post-translational modifications such as phosphorylation or acetylation, the proteome is many orders of magnitude bigger than the transcriptome. This makes the proteome, which reflects the functional state of the cell, a potential, richer source for unraveling diseases [15]. It can be measured using mass spectrometry [16], or protein or antibody microarrays [17]. Additionally, other available omics data such as epigenomics, namely, the study of epigenetic changes such as DNA methylation and histone modifications [12], and single nucleotide polymorphisms genotyping [18] should be considered as they promise refinements for the unraveling of cancer mechanisms and their molecular descriptions. Although the technologies are available, joint analysis of multiple hierarchical layers of biological regulation is at a preliminary stage.

In this contribution, we will investigate whether the integration of information from multiple layers of biological regulation improves the prediction of cancer outcome.

Related work

Other research groups have already proposed the idea of data integration, but most groups only investigated the integration of clinical and microarray data. Tibshirani and colleagues proposed such a framework by reducing the microarray data to one variable, addable to models based on clinical characteristics such as age, grade, and size of the tumour [19]. Nevins and colleagues combined clinical risk factors with metagenes (i.e. the weighted average expression of a group of genes) in a tree-based classification system [20]. Wang *et al.* combined microarray data with knowledge on two clinicopathological variables by defining a gene signature only for the subset of patients for who the clinicopathological variables were not sufficient to predict outcome [21].

A further evolution can be seen in studies in which two omics data sources are simultaneously considered, in most cases microarray data combined with proteomics or array CGH data. Much literature on such studies involving data integration already exists. However, the current definition of the integration of high-throughput data sources as it is used in literature differs from our point of view.
In a first group of integration studies, heterogeneous data from different sources are analyzed sequentially (i.e. one data source is analyzed while keeping the second data source as confirmation of the found results or for further deepening the understanding of the results) [22]. Such approaches are used for biological discovery and a better understanding of the development of a disease, but not for predictive purposes. For example, Fridlyand and colleagues found three breast tumour subtypes with a distinct copy number variation pattern based on array CGH data. Microarray data were subsequently analyzed to identify the functional categories that characterized these subtypes [23]. Tomioka *et al.* analyzed microarray and array CGH data of patients with neuroblastoma in a similar way. Genomic signatures resulted from the array CGH data, while molecular signatures were found after the microarray analysis. The authors suggested that a combination of these independent prognostic indicators would be clinically useful [24].
The term data integration has also been used as a synonym for data merging in which different data sets are concatenated at the database level by cross-referencing the sequence identifiers, requiring semantic compatibility among data sets [25], [26]. Data merging is a complex task due to a.o. the use of different identifiers, the absent of a 'one gene - one protein' relationship, alternative splicing, and measurement of multiple signals for one gene. In most studies, the concordance between the merged data sets and their interpretation in the context of biological pathways and regulatory mechanisms are investigated. Analyses on the merged data set by clustering or correlating the protein and microarray data can help identify

candidate targets when changes in expression occur at both the gene and protein level. However, there has been only modest success from correlation studies of gene and protein expression. Bitton *et al.* combined proteomics data with exon array data which allowed a much more fine grained analysis by assigning peptides to their originating exons instead of mapping transcripts and proteins based on their id [27].

Our definition for the combination of heterogeneous biological data is different. We integrate multiple layers of experimental data into one mathematical model for the development of more homogeneous classifiers in clinical decision support. For this purpose, we present a kernel-based integration framework. Integration occurs within the patient domain at another level as described so far in literature. Instead of merging data sets or analyzing them subsequently, the variables from different omics data are treated equally. This leads to the selection of the most relevant features from all available data sources which are combined in a machine learning-based model.
We were inspired by the idea of Lanckriet and colleagues [28]. They presented an integration framework in which each data set is transformed into a kernel matrix. Integration occurs on this kernel level without referring back to the data. They applied their framework on amino-acid sequence information, expression data, protein-protein interaction data, and other types of genomic information to solve a single classification problem: the classification of transmembrane versus non transmembrane proteins.
In the setting of Lanckriet, all considered data sets were publicly available. This requires a computationally intensive framework for determining the relevance of each data set by solving an optimization problem. Within our set-up however, all data sources are derived from the patients themselves. This makes the gathering of these data sets highly costly and limits the number of data sets, but guarantees more relevance for the problem at hand.

We previously investigated whether the prediction of distant metastasis in breast cancer patients could be improved when considering microarray data besides clinical data [29]. In this manuscript, not only microarray data but high-throughput data from multiple biological levels are considered. Three different strategies for clinical decision support are proposed: the use of individual data sets (referred to as step A), an integration of each data type over time by manually calculating the change in expression (step B), and an approach in which data sets are integrated over multiple layers in the genome (and over time) by treating variables from the different data sets equally (step C).

We apply our framework on two cases, summarized in Table 1. In the first case on rectal cancer, tumour

regression grade, lymph node status, and circumferential margin involvement are predicted for 36 patients based on microarray and proteomics data, gathered at two time points during therapy. The second case on prostate cancer involves microarray and copy number variation data of 55 patients. Tumour grade, stage, metastasis, and occurrence of recurrence were available for prediction [30], [31].

Methods

Data set I on rectal cancer

Patients and treatment

Fourty patients with rectal cancer (T3-T4 and/or N+) from seven Belgian centres were enrolled in a phase I/II study investigating the combination of cetuximab, capecitabine, and external beam radiotherapy in the preoperative treatment of patients with rectal cancer [32]. These patients received preoperative radiotherapy (1.8 Gy, 5 days/week for 5 weeks) in combination with cetuximab (initial dose 400 mg/m^2 intravenous given one week before the beginning of radiation followed by 250 mg/m^2/week for 5 weeks) and capecitabine for the duration of radiotherapy (650 mg/m^2 orally twice-daily, first dose level; 825 mg/m^2 twice-daily, second dose level, including weekends). Details of the eligibility criteria, pretreatment evaluation, radiotherapy, chemotherapy and cetuximab administration, surgery, follow-up, and histopathological assessment of response to chemoradiation are published in [32].

Data preprocessing

Tissue and plasma samples were gathered at three time points: before treatment (T_0), after the first loading dose of cetuximab but before the start of radiotherapy with capecitabine (T_1), and at moment of surgery (T_2). All experimental procedures were done following standard laboratory procedures, or following the manufacturers' instructions. Because of the exclusion of some patients due to missing outcome value, death before surgery, or lacking surgery, ultimately the data set contained 36 patients.

The frozen tissue samples were hybridized to Affymetrix human U133 2.0 plus gene chip arrays. The resulting data were first preprocessed for each time point separately using robust multichip analysis (RMA) [33]. Secondly, the number of features was reduced from 54613 probe sets to 27650 genes by taking

the median of all probe sets that matched on the same gene. Probe sets that matched on multiple genes were excluded because of danger for cross-hybridization. Taking into account the low signal-to-noise ratio of microarray data, we finally filtered out genes with low variation across all samples. Only retaining the genes with a variance in the top 25% reduces the number of features to 6913 genes.

Ninety six proteins known to be involved in cancer were measured in the plasma samples using a Luminex 100 instrument. Proteins that had absolute values above the detection limit in less than 20% of the samples were excluded for each time point separately. This resulted in the exclusion of six proteins at T_0, four at T_1, and six at T_2. The proteomics expression values of transforming growth factor alpha (TGFα), which had too many values below the detection limit, were replaced by the results of ELISA tests performed at the Department of Experimental Oncology in Leuven, Belgium. For the remaining proteins the missing values were replaced by half of the minimum detected for each protein over all samples, and values exceeding the upper limit were replaced by the upper limit value. Because most of the proteins had a positively skewed distribution, a log transformation (base 2) was performed.

In this paper, only the data sets at T_0 and T_1 were used because our goal is to predict the four different outcomes before therapy or early in therapy.

Response classification

A semiquantitative classification system has been described by Wheeler *et al.* [34] for determining the histopathological tumour regression (i.e. the therapy response). There are also two prognostic factors important in rectal cancer, i.e. pathologic lymph node involvement and circumferential margin involvement [35]. Because the completeness of tumour resection relies on the assessment of resection margins by the pathologist, knowledge of the circumferential margin involvement before therapy provides important prognostic information for local recurrence and for development of distant metastasis and survival [36].

These three outcomes were registered for 36 patients at moment of surgery. For all these outcomes, "responders" are distinguished from "nonresponders". The grading of regression established by Wheeler and colleagues [34] (from now on referred to as WHEELER) is a modified pathological staging system for irradiated rectal cancer. It includes a measurement of tumour response after preoperative therapy: grade 1,

good responsiveness (tumour is sterilized or only microscopic foci of adenocarcinoma remain); grade 2, moderate responsiveness (marked fibrosis but with still a macroscopic tumour); grade 3, poor responsiveness (little or no fibrosis with abundant macroscopic tumour). Tumours are classified as "responder" when assigned to grade 1 (26 patients) and "nonresponder" when assigned to grade 2 or 3 (10 patients). Response can also be evaluated with the pathologic lymph node stage at surgery (pN-STAGE). The "responder"-class contains 22 patients with no lymph nodes found at surgery while the "nonresponder"-class contains 14 patients with at least 1 regional lymph node. The circumferential margin involvement (CRM) was measured according to the guidelines of Quirke *et al.* [37]. The CRM was considered positive when the distance between the tumour and the mesorectal fascia was less than or equal to 2mm. Tumours with a negative CRM are classified as "responder" (27 patients), while tumours with a positive CRM belong to the "nonresponder"-class (9 patients). Thirteen patients belong to the "responder"-class for all three outcomes, while there is an overlap of two patients between the "nonresponder"-classes.

Data set II on prostate cancer

Patients and treatment

We also applied our method on a publicly available data set of prostate cancer. Lapointe and colleagues first profiled gene expression in 71 prostate tumours of which 62 primary and 9 lymph node metastases. All tumours were removed by radical prostatectomy (i.e. the surgical removal of the prostate gland). A complementary DNA (cDNA) microarray was used, containing 39711 human cDNAs representing 26260 mapped genes [30]. Additionally, DNA CNVs were profiled on cDNA microarrays for CGH, for 64 prostate tumours among which 55 primary tumours and 9 pelvic lymph node metastases. The arrays were obtained from the Stanford Functional Genomics Facility and included 39632 human cDNAs corresponding to 22279 genes [31]. Among the primary tumours, fifty-five were in common for which gene expression and genomics data are available.

Data preprocessing

The median fluorescence ratios were calculated for genes represented by multiple arrayed cDNAs. Missing gene expression values were imputed unsupervised using the k-nearest neighbours method of Troyanskaya *et al.* [38]. The parameter k was set to 15 such that a missing value for a spot S in a sample is estimated as the weighted average of the 15 spots that are most similar to spot S in the remaining samples. The same unsupervised prefiltering as applied on the rectal cancer data set was used for both the microarray and genomics data set. Features with a variance in the top 50% were retained, reducing the data sets to 6974 genes and 7305 CNVs, respectively.

Response classification

Two pathological variables stage and grade, metastasis of the tumour, as well as the outcome after prostatectomy defined as recurrence are considered. For grade (from now on referred to as GRADE), the Gleason Grading system was used which is based on the most common and second most common architectural patterns of the glands of the tumour [39]. Two groups could be distinguished based on the architecture of the most common pattern: 36 tumours were well differentiated (i.e. low-grade), 19 were poorly differentiated (i.e. high-grade). According to the extent of the primary tumour (STAGE), 25 samples were of stage T2 (i.e. the cancer is confined within one lobe of the prostate gland), while 25 samples were of advanced stage T3 (i.e. the tumour has extended through the fibrous tissue surrounding the prostate gland but no other organs are affected). The stage of the remaining 5 patients was not known. The cancer had metastasized to distant lymph nodes in 12 tumours, while the cancer had not spread beyond the regional lymph nodes in 38 of the tumours (METASTASIS). Tumour recurrence was defined as a rise in prostate-specific antigen (PSA) of at least 0.07 ng/ml or as occurrence of clinical metastasis (RECURRENCE). Seven tumours recurred while 22 tumours did not. The recurrence status of the remaining 26 patients was not available.

Kernel methods and weighted Least Squares Support Vector Machines

Kernel methods are a group of algorithms that can handle a very wide range of data types, such as vectors, sequences, networks, etc. They map the data x from the original input space to a high dimensional feature

space with the mapping function $\Phi(x)$. This embedding into the feature space is performed by a mathematical object $K(x_k, x_l)$, called a 'kernel function'. This function efficiently computes the inner product $\langle \Phi(x_k), \Phi(x_l) \rangle$ between all pairs of data items x_k and x_l in the feature space, resulting in the kernel matrix. The size of this matrix is determined only by the number of data items, whatever the nature or the complexity of these items. For example, a set of 100 patients each characterized by 6913 gene expression values is still represented by a 100×100 kernel matrix [40]. The representation of all data sets by this real-valued square matrix, independent of the nature or complexity of the data to be analyzed, makes kernel methods ideally positioned for heterogeneous data integration.
Any symmetric, positive semidefinite function is a valid kernel function, resulting in many possible kernels, e.g. linear, polynomial, and diffusion kernels. They all correspond to a different transformation of the data, meaning that they extract a specific type of information from the data set. In this paper, the normalized linear kernel function

$$\tilde{K}(x_k, x_l) = K(x_k, x_l) / \sqrt{K(x_k, x_k) K(x_l, x_l)} \tag{1}$$

with $K(x_k, x) = x_k^T x$ is used instead of the linear kernel function $K(x_k, x_l) = x_k^T x_l$. With the normalized version, the values in the kernel matrix will be bounded because the data points are projected onto the unit sphere while these elements can take very large values without normalization. Normalizing is thus required when combining multiple data sources to guarantee the same order of magnitude for the kernel matrices of the data sets.

A kernel algorithm for supervised classification is the Support Vector Machine (SVM) developed by Vapnik [41] and others. Contrary to most other classification methods and due to the way data is represented through kernels, SVMs can tackle high dimensional data (e.g. microarray data). Given a training set $\{x_k, y_k\}_{k=1}^N$ of N samples with feature vectors $x_k \in R^n$ and output labels $y_k \in \{-1, +1\}$, the SVM forms a linear discriminant boundary $y(x) = \text{sign}[w^T \Phi(x) + b]$ in the feature space with maximum distance between samples of the two considered classes, with w representing the weights for the data items in the feature space and b the bias term. This corresponds to a non-linear discriminant function in the original input space. A modified version of SVM, the Least Squares Support Vector Machine (LS-SVM), was developed by Suykens *et al.* [3], [4]. On high dimensional data sets, this modified version is much faster for classification because a linear system instead of a quadratic programming problem needs to be solved.

The constrained optimization problem for an LS-SVM has the following form:

$$\min_{w,b,e} \left(\frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 \right)$$

$$\text{subject to } y_k [w^T \Phi(x_k) + b] = 1 - e_k, \ k = 1, \ldots, N$$

with e_k the error variables, tolerating misclassifications in case of overlapping distributions, and γ the regularization parameter which allows tackling the problem of overfitting. It has been shown that regularization seems to be very important when applying classification methods on high dimensional data [42].

In many two-class problems, data sets are skewed in favour of one class such that the contribution of false negative and false positive errors to the performance assessment criterion are not balanced. We therefore used a weighted LS-SVM in which a different weight ζ_k is given to positive and negative samples in order to account for the unbalancedness in the data set [5]. The objective function changes into

$$\min_{w,b,e} \left(\frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^{N} \zeta_k e_k^2 \right)$$

with

$$\zeta_k = \begin{cases} \frac{N}{2N_P} & \text{if } y_k = +1 \\ \frac{N}{2N_N} & \text{if } y_k = -1 \end{cases}$$

and N_P and N_N representing the number of positive and negative samples, respectively.

Feature selection

Univariate feature selection techniques are computationally simple but do not incorporate feature-feature interactions. However, due to small sample size limitations, multivariate approaches are often not appropriate to discover the underlying complex, multivariate correlations. Because it has been shown that univariate gene selection methods lead to good and stable performances across many cancer types and yield in many cases consistently better results than multivariate approaches [43], we used the method DEDS (Differential Expression via Distance Synthesis) [44]. This technique is based on the integration of different univariate test statistics via a distance synthesis scheme because features highly ranked simultaneously by multiple statistics are more likely to be differential expressed than features highly ranked by a single test statistic. The statistical tests combined are ordinary fold changes, ordinary

t-statistics, SAM-statistics (significance analysis for microarrays) and moderated t-statistics. DEDS is available as a BioConductor package in R.

We applied DEDS on the microarray data sets as well as on the genomics data set. From our experience, DEDS is less appropriate for data with a limited set of features (data not shown). Since the proteomics data on rectal cancer only contain 90-92 cancer-related proteins, one test statistic suffices for which we chose the Wilcoxon rank sum test.

Model building

To determine the optimal number of features, we use a leave-one-out cross-validation (LOO-CV) approach in which we increase the number of included features iteratively according to the obtained feature ranking but in which we do not include more features than the number of samples in the data set on which the optimal number of features is determined, as discussed by Li *et al.* [45]. Besides the number of features, also the parameters of the kernel method (parameter γ for LS-SVM with normalized linear kernel) need to be selected. This selection occurs on a k-dimensional grid with k-1 the number of data sets included. We considered 40 possible values for γ, ranging from 10^{-4} to 10^{6} on a logarithmic scale. In each LOO iteration, a sample is left out, feature selection is performed on the remaining n-1 samples, and models are built for all possible combinations of parameters on this grid. Each model with the instantiated parameters is evaluated on the left out sample. This whole procedure is repeated for all samples. The model parameters are chosen corresponding to the model with the highest LOO AUC (area under the receiver operating characteristic curve). If multiple models with equal AUC, the model with the lowest balanced error rate and an as high as possible sum of sensitivity and specificity is chosen. For each considered outcome, the AUC of the best performing model is compared with the AUC of the other models using the method of Hanley and McNeil [46]. The final features are chosen as the ones that occurred most often in the top rankings determined in each LOO iteration.

Three kinds of model building strategies are proposed, different in the degree of integration. Figure 1 shows these strategies more in detail. The data sets are represented as matrices with rows corresponding to patients and columns corresponding to genes, proteins, or CNVs. The matrices representing microarray or genomics data are larger than those for the proteomics data to emphasize the difference in dimensionality.

All three types of strategies were applied on the microarray and proteomics data sets of rectal cancer. For the prostate cancer data set however, only two strategies were applicable due to lack of measurements repeated over time. For all models the parameters were trained according to the same approach which makes the corresponding LOO results comparable for each outcome separately.

Step A models: single data set

In a first step, LS-SVM models are built on each data set separately, mimicking the results that would have been obtained when only static data from one platform was available. For rectal cancer, the single data sets are microarray at T_0, microarray at T_1, proteomics at T_0, and proteomics at T_1 for the prediction of a regression grading system and two prognostic factors (see Figure 1A). For prostate cancer, LS-SVM models are built on the microarray and genomics data separately for the prediction of grade, stage, metastasis, and recurrence. Because of only one set of features, a 2-dimensional grid is used for the optimization of the regularization parameter and the number of features.

Step B models: manual integration of data over time

When measurements are repeated at multiple time points, knowledge over time can be exploited. For rectal cancer, data were available before and early in therapy and therefore can be combined in the models. This is done for each data type separately by manually calculating the change in gene expression or protein abundance between the first two time points ($T_0 - T_1$). These changes over time are used as features for the models as shown in Figure 1B. Also for these models, a 2-dimensional grid suffices for the optimization of the regularization parameter and the number of features.

Step C models: multiple omics integration approach

The previous two types of models (step A and B) are considered to verify whether complex integration of data over multiple layers of biological regulation is crucial. The ability of kernel methods to deal with complexly structured data makes them ideally positioned for a more advanced integration of heterogeneous data sources. We will use the intermediate integration method proposed in [47] in which a kernel matrix is

computed for each data source separately. Subsequently, these data sources can be integrated in a straightforward way by summing the multiple kernel matrices. Positive semidefiniteness of the linear combination of kernel matrices is guaranteed by constraining the weights of the kernels to be non-negative. A weighted LS-SVM is trained on the explicitly heterogeneous kernel matrix. The choice of the weights to give to each data set is important. A kernel framework for optimizing weights is proposed in [48]. This optimization is important when dealing with many data sets of which only several are relevant. However, when the number of data sets is limited and most of them are reliable and relevant to the problem at hand, a trade-off needs to be made between performance and computational burden (e.g., extra required cross-validation loops). Due to the rather small sample size in both case studies, weights were chosen equally. Moreover, our aim is to emphasize that classification becomes more accurate when data from multiple layers in the genome are available and to offer a machine learning-based method for integrating these data sources, rather than to improve an algorithm for the optimization of weights (e.g. [48]). A 3-dimensional grid is used for the optimization of the parameters, i.e. the regularization parameter, the number of genes selected from the microarray data sets, and the number of proteins or CNVs obtained from the proteomics data sets or the genomics data set, respectively. For the data on rectal cancer, the number of genes/proteins selected at T_0 and T_1 were taken equally when data from both time points were considered. Figure 1C gives an overview of the strategy.

Results

Study I on rectal cancer

Using the methodologies shown in Figure 1, models were built on microarray and proteomics data of 36 rectal cancer patients at two time points during therapy for the prediction of three outcomes registered at moment of surgery: a tumour regression grading system (WHEELER) and two prognostic factors, pathologic N stage at surgery (pN-STAGE) and the circumferential resection margin (CRM). The models with the highest AUC, lowest balanced error rate and an as high as possible sum of sensitivity and specificity are shown in Table 2. The step A models are MT_0 (model based on microarray data at T_0), MT_1 (model based on microarray data at T_1), PT_0 (model based on proteomics data at T_0), and PT_1 (model based on proteomics data at T_1). The step B models consist of $MT_0 - T_1$ (model based on change in gene expression between T_0 and T_1) and $PT_0 - T_1$ (model based on change in protein abundances

between T_0 and T_1). Finally, the step C models comprise MT_{01} (model based on microarray data at both time points), PT_{01} (model based on proteomics data at both time points), MPT_0 (model based on microarray and proteomics data at T_0), MPT_1 (model based on microarray and proteomics data at T_1), all possible combinations of three data sets (using the same name convention), and MPT_{01} (model based on all data (microarray and proteomics data at both time points)). The number of genes and proteins were chosen to optimize the leave-one-out (LOO) performance of the LS-SVM models. The features selected most often in the 36 LOO iterations are listed and discussed. For each outcome, the receiver operating characteristic (ROC) curve of the best model was compared with the ROC curves of all other models [46]. The p-values of these significance tests are reported as well.

Table 2 shows the LS-SVM models for the considered combinations of data sets to predict WHEELER, pN-STAGE, and CRM with the optimal number of genes and proteins selected with DEDS and the Wilcoxon rank sum test, respectively. The corresponding ROC curves are shown in Additional file 1. The performance of the models based on three data sets is given in Additional file 2. Due to the slightly, not significantly better performance for each outcome of one model based on three data sets compared to models based on two data sets, we report the results for the best model combining two data sets. Such models would only require a sample to be taken at one time point (MPT_0, MPT_1) or one technology to be applied on two time points (MT_{01}, PT_{01}). For the prediction of WHEELER, the expression of 25 genes and 12 proteins at T_1 was best, although not significantly, with an AUC of 0.92692. Also for pN-STAGE, combining both data sets at T_1 using the expression of 21 genes and 14 proteins resulted into the best LOO AUC of 0.98701. This performance is significantly better than all step A and B models as well as PT_{01}. Finally, the inclusion of 7 genes and 33 proteins at T_1 lead to an AUC of 0.96296 for the prediction of CRM. Four models based on only one data type perform significantly worse compared to MPT_1. For all outcomes, none of the selected proteins are a product of the selected genes.

The contribution of the genes and/or proteins in rectal or colorectal cancer that were selected most often in the LOO iterations of MPT_1 and predicted most accurately WHEELER, pN-STAGE, or CRM are shown in Table 3. A protein important for CRM for example is the epidermal growth factor receptor (EGF-R), involved in signaling pathways affecting cellular growth, differentiation, and proliferation. This protein represents one of the most promising targets allowing progress in colorectal cancer treatment. It has been suggested that EGF-R polymorphisms as well as polymorphisms of other genes active in the EGF-R

pathway may be potential indicators of radiosensitivity in patients with rectal cancer treated with chemoradiation [49]. In colorectal cancer, proinflammatory cytokines such as interleukin 1 beta (IL-1B) and IL-6 may be accountable for the overexpression of Cox-2, important in the early stage and for progression [50]. TGFα, down-regulated in our patients with a good responsiveness to preoperative therapy, is implicated in metastatic spread of colon cancer cells [51]. The expression of IL-8 is associated with induction and progression of colorectal carcinoma and the development of colorectal liver metastases [52]. In our data set, it is down-regulated in the group of patients with no lymph nodes found at surgery. Finally, elevated carcinoembryonic antigen (CEA) and cancer antigen 19-9 are related to poor outcome in colorectal cancer [53]. Their levels are low in patients with no lymph nodes, while CEA is also less expressed in patients with a negative CRM, i.e. belonging to the class of "responders".
A complete list of the genes and proteins chosen by the models MPT_1 are shown, for each outcome separately, in Additional file 3. The predictions seem to depend on mainly different subsets of features. The gene PAI-2 is important for both WHEELER and CRM, while the proteins important for 2 of the 3 outcomes are interleukin-4, ferritin, apolipoprotein H, EGF, MMP-2, and lymphotactin. Notably, these genes and proteins were also selected by the other models based on microarray and/or proteomics data at T_1, although the specific feature ranking depends on the number of features included. Some of these genes and proteins were included as well in the models based on data at T_0.

Study II on prostate cancer

The same methodology was applied on microarray and genomics data of 55 patients with prostate cancer. Table 4 shows the results for the prediction of the grade and stage of the tumour (GRADE and STAGE), as well as the tumours that metastasized to distant lymph nodes (METASTASIS) or that recurred (RECURRENCE). Because the data were gathered at 1 time point, only step A and C models are applicable. The step A models are represented as M (model based on microarray data) and G (model based on genomics data), the step C model based on both microarray and genomics data as MG. Also here, after having optimized the essential number of features to be included using a LOO cross-validation, the final genes and CNVs were selected based on their number of occurrences in the top of the 55 LOO rankings.

We obtained similar results as for rectal cancer. Combining gene expression with measurements at the DNA level (MG) led, for all four outcomes, to an improvement in classification accuracy and was

significant in some cases (see Table 4). For the prediction of GRADE, 6 genes and 8 CNVs selected with DEDS resulted in an AUC of 0.9006. For STAGE, 42 genes and 22 CNVs were needed for a performance of 0.8528. The model *MG* for the prediction of METASTASIS had an AUC of 0.9868 when fusing the expression of 18 genes with 3 CNVs. Finally, the prediction of RECURRENCE was most difficult with an AUC of 0.7857 when combining 32 genes and 2 CNVs. Additional file 1 shows the ROC curves of the models listed in Table 4.

Several genes and CNVs have been selected by *MG* and are known to be involved in and important for prostate cancer, listed in Table 5. The gene ALOX15B is a suppressor of prostate tumour development [54] and is in this data set down-regulated in tumours of high-grade and in tumours that recurred. Both SFRP4 and CXCL14 on the other hand are inhibitors of prostate tumour growth [55], [56]. SFRP4 is up-regulated in tumours of high-grade, CXCL14 in tumours of advanced stage. A small deletion involving chromosomal band 21q22.3 fuses all coding exons of ERG to androgen-related sequences in the promoter of the prostate-specific TMPRSS2 gene. This chromosomal rearrangement is a highly prevalent oncogenic alteration in prostate tumour cells and leads to an aberrant expression of the ERG proto-oncogene, important for early prostate carcinogenesis [57]. In this data set, ERG is overexpressed in tumours in which the cancer metastasized to distant lymph nodes. It has been shown that this genetic biomarker is a strong prognostic factor for disease recurrence, and can be used for early detection and outcome prediction in prostate cancer [58]. VAV3, an oncogene involved in development and progression of prostate cancer, is up-regulated in tumours that metastasized [59]. It has previously been shown that a strong overexpression of TIAM1 is significantly associated with disease recurrence and a decreased disease-free survival [60]. Also JAG1 is significantly associated with recurrence [61] and plays a role in cell growth, progression, and metastasis. In this data set, both genes are up-regulated in the group of tumours that recurred. Finally, several germline mutations or variants in RNASEL have been observed among hereditary prostate cancer cases, indicating that polymorphic changes within the RNASEL gene may be associated with increased risk of familial but not sporadic prostate cancer [62].

A list of all the genes and CNVs selected by the models *MG* are shown in Additional file 3. As for rectal cancer, the outcomes for prostate cancer seem to be characterized by mainly different sets of features. Five genes overlap between at least 2 outcomes (ERG, AHSG, SEMA4G, F5, and ALOX15B), while the same holds for four CNVs at the genes GPD1L, KCTD12, SMYD5, and TRO.

Comparison with an ensemble approach

To assess the benefit of our kernel-based integration approach over standard data fusion techniques, we implemented an ensemble approach in which each data set gives rise to a separate LS-SVM classifier. These individual LS-SVM models were built similarly as the step A models, with the same number of genes, proteins or CNVs selected as included in the best models MPT_1 and MG. Subsequently as a late integration step, the continuous outputs of these models were added.

For the study on rectal cancer, the AUC values of the ensemble models integrating the microarray and proteomics data set gathered at T_1 and the corresponding AUC values of the best model obtained with our strategy (MPT_1) are shown in Table 6. The p-values of the significance tests comparing the ROC curves are reported as well [46]. For CRM, our strategy was significantly better than the ensemble approach at a significance level of 0.05. For WHEELER and pN-STAGE, the AUC values did not differ significantly. Similar for the study on prostate cancer, the AUC values of MG were compared with the AUC values of the ensemble models combining microarray and genomics data, shown in Table 6. For all four outcomes, the AUC of MG was better than the AUC of the ensemble models, although being significantly better for RECURRENCE only.

Correlation analysis

We additionally verified whether, in both cases, data from multiple layers of molecular biology were complementary. After mapping the entities of the data sets based on their entrez gene IDs, we investigated the correlation between the microarray and proteomics data of rectal cancer on the one hand, and between the microarray and genomics data of prostate cancer on the other hand. Using the Spearman correlation coefficient, there was no significant correlation for rectal cancer between the abundances of the 90/92 proteins and their corresponding transcripts at a significance level of 0.05. The microarray and genomics data set for prostate cancer were slightly more correlated. While for GRADE the six genes selected by the model MG did not correlate with their DNA expression, two of the 42 selected genes for STAGE were significantly correlated ($p<0.05$). For METASTASIS and RECURRENCE, there was a significant correlation for one and three genes, respectively. The regions, with involved CNVs selected from the genomics data, were also compared with the regions in which the selected genes from the microarray data

were located. For the majority of regions, there was no overlap. For the other regions with the same rough chromosomal location, the genes selected by both data sets were different.

Discussion

The proposed integration approach has been applied on two patient data sets, each with two high-throughput data sources. Microarray and proteomics were gathered from 36 patients with rectal cancer at two time points during preoperative treatment, while microarray and genomics were gathered from 55 patients with prostate cancer. To verify the merit of our integration approach over the use of a single omics data source, models were built for classifying cancer patients according to therapy response, prognostic factors, metastasis, or recurrence. In many studies, only single data sources are explored for the development of such profiles. However in our opinion, a single layer of molecular information is inadequate to explain the complete network of molecules underlying a disease. In this study shown in Figure 1, LS-SVMs were first built on all data sets individually. Next, we manually integrated data measured at multiple time points by building LS-SVMs using the change in expression between two time points. Because the integration of data may be more complex than the change in expression over time, we subsequently applied an intermediate integration approach in which multiple omics data were combined at the kernel level within the patient domain.

For the data on rectal cancer, all three outcomes, namely, a tumour regression grading system and two prognostic factors, could be predicted most accurately and most cost-efficiently with an AUC ranging from 0.92692 to 0.98701 when fusing microarray and proteomics data gathered during therapy (MPT_1) (see Table 2). For WHEELER for example, MPT_0 performance is better than each of the models based on data from an individual technology (MT_0 and PT_0), as is the case for MPT_{01} compared to MT_{01} and PT_{01}. This trend of increased performance when combining data from two different technologies was further confirmed by our second data set for prostate cancer patients. Best results for the prediction of grade, stage, metastasis, and recurrence were obtained when integrating microarray and genomics data (MG). The corresponding AUC values were 0.9006, 0.8528, 0.9868, and 0.7857, respectively (see Table 4). For many of the genes, proteins, and CNVs included in these models, involvement in rectal or prostate cancer has been defined, indicating the reliability of the selected features (Tables 3 and 5). These models were compared with models obtained with an ensemble approach in which classifiers are combined instead of data sets at

the kernel level. Globally, our approach performed better, although not always significantly (see Table 6).

By looking at the correlation between two data sets gathered from the same set of patients, we showed that data from different layers are mainly complementary. For rectal cancer, there was a lack of correlation between the selected genes and their corresponding proteins. Also the selected proteins did not significantly correlate with their transcript level, suggesting alternative splicing and post-translational modification. With newer technologies such as mass spectrometry, the whole proteome will become measurable. For prostate cancer, up to 3 genes included in the model MG were significantly correlated with their corresponding CNV.

More specific for the study on rectal cancer, we can conclude from Table 2 that data gathered after an initial dose of cetuximab are more informative for prediction of therapy response than data gathered before the start of the therapy. Neither microarray nor proteomics data can predict the outcomes more accurately at T_0 than T_1, except for the proteomics data at T_0 being more informative for the prediction of CRM. Moreover, when combining both data types at one time point (MPT_0 and MPT_1), the models applicable after the initial dose of cetuximab outperform those at T_0.

We acknowledge that the models proposed in this manuscript are quite expensive. Applying a model for rectal cancer would require microarray and/or proteomics data, gathered at one or two time points during therapy. However, we did an attempt to keep the cost to a minimum. The performance difference between models combining two data sets, only requiring a sample to be taken at one time point or one technology to be applied on two time points, and models requiring a sample to be taken at both time points and both technologies to be performed was minimal and not statistically significant. We therefore chose the best model among the models based on two data sets. We admit that there may exist other, less expensive data sources that can contain complementary information as well. Firstly, clinical information is routinely gathered during therapy, such as tumor size, tumor location and number of positive lymph nodes. However, we only had access to the clinical parameter age, for which we performed an additional analysis to verify whether this parameter could be of use. A univariate analysis based on the Wilcoxon rank sum test showed no significant difference in age between the two classes of samples according to the considered outcomes. In a multivariate logistic regression model, the parameter age was not significant as well. Secondly, there is an increasing need for multi-modal studies in which, among others, clinical, genomic and genetic data are collected. Also imaging, such as computed tomography (CT) and magnetic resonance imaging (MRI) can

be a potential predictor to use in combination with high-throughput data sources. Such studies are required to determine which data sets are most relevant for the problem at hand and which data sets should be combined to become good performing, affordable models that are clinically applicable.

Conclusions

The results suggest that the use of our integration approach on experimental data from multiple levels in the genome, can improve the performance of decision support in cancer. For both data sets studied in this manuscript, combining high-throughput data sets (transcriptomics with proteomics, or genomics with transcriptomics) outperformed the models based on data from a single layer of biological information, independent of the outcome considered for prediction. These results emphasize the need for comprehensive multi-modal data gathered with high-throughput technologies as well as imaging, because it is unknown which technologies and thus which levels of molecular biology are the most relevant for prognostic prediction. We acknowledge that this will substantially increase costs in a first exploratory phase. However, this is a necessary investment to ultimately obtain cost-efficient models usable in patient tailored therapy.

In the near future, we will compare our kernel-based integration method with a Bayesian network integration framework. These frameworks are complementary. We also plan to apply an ensemble approach for integrating these two frameworks because more accurate classifiers are not only obtained by combining different data types but also by combining individual decisions of multiple classifiers. In this way, the advantages of both methods can be exploited.

Abbreviations

AUC, area under the ROC curve; cDNA, complementary DNA; CGH, comparative genomics hybridization; CNV, copy number variation; CRM, circumferential margin involvement; CT, computed tomography; CV, cross-validation; DEDS, differential expression via distance synthesis; DNA, deoxyribonucleic acid; FS, feature selection; G, model based on genomics data; LOO, leave-one-out; LS-SVM, Least Squares Support Vector Machine; M, model based on microarray data; MG, model based on both microarray and genomics data; MPT_0, model based on microarray and proteomics data at T_0; MPT_1, model based on microarray and proteomics data at T_1; MPT_{01}, model based on all data (microarray and proteomics data at both

timepoints); MRI, magnetic resonance imaging; MT_0, model based on microarray data at T_0; MT_1, model based on microarray data at T_1; MT_{01}, model based on microarray data at both time points; $MT_0 - T_1$, model based on change in gene expression between T_0 and T_1; NF, number of features; PSA, prostate-specific antigen; PT_0, model based on proteomics data at T_0; PT_1, model based on proteomics data at T_1; PT_{01}, model based on proteomics data at both time points; $PT_0 - T_1$, model based on change in protein abundances between T_0 and T_1; RMA, robust multichip analysis; ROC, receiver operating characteristic; SAM, significance analysis for microarray; SVM, Support Vector Machine; T_0, time point before treatment; T_1, time point after the first loading dose of cetuximab but before the start of radiotherapy with capecitabine; T_2, time point at moment of surgery.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ADa performed the kernel-based integration modeling and drafted the manuscript. OG, FO, and JS participated in the design and implementation of the framework. ADa and OG performed pre-processing of the data. OG, JS, and BDM helped drafting the manuscript. ADe, JPM, and KH provided clinical input, looked up patient records in the database, performed sample annotation, and gathered follow-up of patients. All authors read and approved the final manuscript.

Acknowledgements

AD is research assistant of the Fund for Scientific Research - Flanders (FWO-Vlaanderen). BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. The authors are grateful to Anja von Heydebreck, Detlef Guessow and Christopher Stroh for their contribution at Merck Serono. This work is partially supported by: **1.** Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA, several PhD/postdoc & fellow grants. **2.** Flemish Government: **a.** FWO: PhD/postdoc grants, projects G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0318.05 (subfunctionalization),

G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); **b.** IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM-Endometriosis. **3.** Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011). **4.** EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

References

1. Shawe-Taylor J, Cristianini N: *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press 2004.
2. Bhaskar H, Hoyle DC, Singh S: **Machine learning in bioinformatics: A brief survey and recommendations for practitioners**. *Comp Biol Med* 2006, **36**:1104–1125.
3. Suykens JAK, Vandewalle J: **Least Squares Support Vector Machine classifiers**. *Neural Processing Letters* 1999, **9**:293–300.
4. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J: *Least Squares Support Vector Machines*. Singapore: World Scientific 2002.
5. Cawley GC: **Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs**. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)* 2006:1661–1668.
6. Alon A, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays**. *Proc Natl Acad Sci U S A* 1999, **96**:6745–6750.
7. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**:531–537.
8. Cardoso F, van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ: **Clinical application of the 70-gene profile: the MINDACT trial**. *J Clin Oncol* 2008, **26**:729–735.
9. Sparano JA: **TAILORx: trial assigning individualized options for treatment (Rx)**. *Clin Breast Cancer* 2006, **7**:347–350.
10. Sparano JA, Paik S: **Development of the 21-gene assay and its application in clinical practice and clinical trials**. *J Clin Oncol* 2008, **26**:721–728.
11. Pinkel D, Albertson DG: **Array comparative genomic hybridization and its applications in cancer**. *Nature Genetics* 2005, **37**:S11–S17.
12. Esteller M: **Epigenetics in cancer**. *N Engl J Med* 2008, **358**:1148–1159.
13. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews D, Fiegler H, Shapero MH, Carson W A R Chen, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME: **Global variation in copy number in the human genome**. *Nature* 2006, **444**:444–454.
14. Fröhling S, Döhner H: **Chromosomal abnormalities in cancer**. *N Engl J Med* 2008, **359**:722–734.
15. Kolch W, Mischak H, Pitt AR: **The molecular make-up of a tumour: proteomics in cancer research**. *Clinical Science* 2005, **108**:369–383.

16. Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**:198–207.

17. MacBeatch G, Schreiber SL: **Printing proteins as microarrays for high-throughput function determination**. *Science* 2000, **289**:1760–1763.

18. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA: **Systematic assessment of copy number variant detection via genome-wide SNP genotyping**. *Nat Genet* 2008, **40**:1199–1203.

19. Tibshirani RJ, Efron B: **Pre-validation and inference in microarrays**. *Statistical Applications in Genetics and Molecular Biology* 2002, **1**:Article 1.

20. Nevins JR, Huang ES, Dressman H, Pittman J, Huang AT, West M: **Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction**. *Hum Mol Genet* 2003, **12**:153–157.

21. Wang SM, Ooi LL, Hui KM: **Identification and validation of a novel gene signature associated with the recurrence of human hepatocellular carcinoma**. *Clin Cancer Res* 2007, **13**:6275–6283.

22. Mathew JP, Taylor BS, Bader GD, Pyarajan S, Antoniotti M, Chinnaiyan AM, Sander C, Burakoff SJ, Mishra B: **From bytes to bedside: data integration and computational biology for translational cancer research**. *PLoS Computational Biology* 2007, **3**:e12.

23. Fridlyand J, Snijders AM, Ylstra B, Li H, Olshen A, Segraves R, Dairkee S, Tokuyasu T, Ljung BM, Jain AN, McLennan J, Ziegler J, Chin K, Devries S, Feiler H, Gray JW, Waldman F, Pinkel D, Albertson DG: **Breast tumor copy number aberration phenotypes and genomic instability**. *BMC Cancer* 2006, **6**:96.

24. Tomioka N, Oba S, Ohira M, Misra A, Fridlyand J, Ishii S, Nakamura Y, Isogai E, Hirata T, Yoshida Y, Todo S, Kanedo Y, Albertson DG, Pinkel D, Feuerstein BG, Nakagawara A: **Novel risk stratification of patients with neuroblastoma by genomic signature, which is independent of molecular signature**. *Oncogene* 2008, **27**:441–449.

25. Waters KM, Pounds JG, Thrall BD: **Data merging for integrated microarray and proteomic analysis**. *Brief Funct Genomic Proteomic* 2006, **5**:261–272.

26. Goble C, Stevens R: **State of the nation in data integration for bioinformatics**. *J Biom Inf* 2008, **41**:687–693.

27. Bitton DA, Okoniewski MJ, Connolly Y, Miller CJ: **Exon level integration of proteomics and microarray data**. *BMC Bioinformatics* 2008, **9**:118.

28. Lanckriet GRG, De Bie T, Cristianini N, Jordan MI, Noble WS: **A statistical framework for genomic data fusion**. *Bioinformatics* 2004, **20**:2626–2635.

29. Daemen A, Gevaert O, Moor BD: **Integration of clinical and microarray data with kernel methods**. In *Conference Proceedings of the IEEE Engineering in Medicine and Biology* 2007:5411–5415.

30. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR: **Gene expression profiling identifies clinically relevant subtypes of prostate cancer**. *Proc Natl Acad Sci U S A* 2004, **101**:811–816.

31. Lapointe J, Li C, Giacomini CP, Salari K, Huang S, Wang P, Ferrari M, Hernandez-Boussard T, Brooks JD, Pollack JR: **Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis**. *Cancer Res* 2007, **67**:8504–8510.

32. Machiels JP, Sempoux C, Scalliet P, Coche JC, Humblet Y, Van Cutsem E, Kerger J, Canon JL, Peeters M, Aydin S, Laurent S, Kartheuser A, Coster B, Roels S, Daisne JF, Honhon B, Duck L, Kirkove C, Bonny MA, Haustermans K: **Phase I/II study of preoperative cetuximab, capecitabine, and external beam radiotherapy in patients with rectal cancer**. *Ann Oncol* 2007, **18**:738–744.

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

34. Wheeler JMD, Warren BF, Mortensen NJ, Ekanyaka N, Kulacoglu H, Jones AC, George BD, Kettlewell MGW: **Quantification of histologic regression of rectal cancer after irradiation**. *Dis Colon Rectum* 2002, **45**:1051–1056.

35. Machiels JP, Aydin S, Bonny MA, Hammouch F, Sempoux C: **What is the best to predict disease-free survival after preoperative radiochemotherapy for rectal cancer patients: tumor regression grading, nodal status or circumferential resection margin invasion?** *J Clin Oncol* 2006, **24**:1319–1321.
36. Adam IJ, Mohamdee MO, Martin IG, Scott N, Finan PJ, Johnston D, Dixon MF, Quirke P: **Role of circumferential margin involvement in the local recurrence of rectal cancer**. *Lancet* 1994, **344**:707–711.
37. Quirke P, Durdey P, Dixon MF, Williams NS: **Local recurrence of rectal adenocarcinoma due to inadequate surgical resection: histopathological study of lateral tumour spread and surgical excision**. *Lancet* 1986, **2**:996–999.
38. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays**. *Bioinformatics* 2001, **17**:520–525.
39. Gleason DF: **Classification of prostatic carcinomas**. *Cancer Chemother Rep* 1966, **50**:125–128.
40. Schölkopf B, Tsuda K, Vert JP: *Kernel methods in computational biology*. United States: MIT Press 2004.
41. Vapnik V: *Statistical Learning Theory*. New York: Wiley 1998.
42. Pochet N, De Smet F, Suykens J, Moor BD: **Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction**. *Bioinformatics* 2004, **20**:3185–3195.
43. Lai C, Reinders MJT, van't Veer LJ, Wessels LFA: **A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets**. *Bioinformatics* 2006, **7**:235–244.
44. Yang YH, Xiao Y, Segal MR: **Identifying differentially expressed genes from microarray experiments via statistic synthesis**. *Bioinformatics* 2005, **21**:1084–1093.
45. Li W, Yang Y: **How many genes are needed for a discriminant microarray data analysis**. In *Methods of Microarray Data Analysis*. Edited by Lin SM, Johnson KF, Kluwer Academic 2002:137–150.
46. Hanley JA, McNeil BJ: **A method of comparing the areas under receiver operating characteristics curves derived from the same cases**. *Radiology* 1983, **148**:839–843.
47. Pavlidis P, Weston J, Cai J, Grundy WN: **Gene functional classification from heterogeneous data**. In *Conference Proceedings of Computational Molecular Biology* 2001:242–252.
48. Zien A, Ong CS: **Multiclass multiple kernel learning**. In *Conference Proceedings of Machine Learning* 2007:1191–1198.
49. Zhang W, Park DJ, Lu B, Yang DY, Gordon M, Groshen S, Yun J, Press OA, Vallböhmer D, Rhodes K, Lenz HJ: **Epidermal growth factor receptor gene polymorphisms predict pelvic recurrence in patients with rectal cancer treated with chemoradiation**. *Clin Cancer Res* 2005, **11**:600–605.
50. Maihöfner C, Charalambous MP, Bhambra U, Lightfoot T, Geisslinger G, Gooderham NJ: **The Colorectal Cancer Group (2003) Expression of cyclooxygenase-2 parallels expression of interleukin-1beta, interleukin-6 and NF-kappaB in human colorectal cancer**. *Carcinogenesis* 2003, **24**:665–671.
51. Sawhney RS, Sharma B, Humphrey LE, Brattain MG: **Integrin α_2 and extracellular signal-regulated kinase are functionally linked in highly malignant autocrine transforming growth factor-α-driven colon cancer cells**. *J Biol Chem* 2003, **278**:19861–19869.
52. Rubie C, Frick VO, Pfeil S, Wagner M, Kollmar O, Kopp B, Gräber S, Rau BM, Schilling MK: **Correlation of IL-8 with induction, progression and metastatic potential of colorectal cancer**. *World J Gastroenterol* 2007, **13**:4996–5002.
53. Louhimo J, Carpelan-Holmström M, Alfthan H, Stenman UH, Järvinen HJ, Haglund C: **Serum HCG beta, CA 72-4 and CEA are independent prognostic factors in colorectal cancer**. *Int J Cancer* 2002, **101**:545–548.
54. Bhatia B, Maldonado CJ, Tang S, Chandra RD D Klein, Chopra D, Shappell SB, Yang RA P Newman, Tang DG: **Subcellular localization and tumor-suppressive functions of 15-lipoxygenase 2 (15-LOX2) and its splice variants**. *J Biol Chem* 2003, **278**:25091–25100.
55. Horvath LG, Lelliott JE, Kench JG, Lee CS, Williams ED, Saunders DN, Grvgiel JJ, Sutherland RL, Henshall SM: **Secreted frizzled-related protein 4 inhibits proliferation and metastatic potential in prostate cancer**. *Prostate* 2007, **67**:1081–1090.

56. Schwarze SR, Luo J, Isaacs WB, Jarrard DF: **Modulation of CXCL14 (BRAK) expression in prostate cancer**. *Prostate* 2005, **64**:67–74.

57. Furusato B, Gao CL, Ravindranath L, Chen Y, Cullen J, McLeod DG, Dobi A, Srivastava S, Petrovics G, Sesterhenn IA: **Mapping of TMPRSS2-ERG fusions in the context of multi-focal prostate cancer**. *Mod Pathol* 2008, **21**:67–75.

58. Nam RK, Sugar L, Yang W, Srivastava S, Klotz LH, Yang LY, Stanimirovic A, Encioiu E, Neill M, Loblaw DA, Trachtenberg J, Narod SA, Seth A: **Expression of the TMPRSS2:ERG fusion gene predicts cancer recurrence after surgery for localised prostate cancer**. *Br J Cancer* 2007, **97**:1690–1695.

59. Dong Z, Liu Y, Lu S, Wang A, Lee K, Wang LH, Revelo M, Lu S: **Vav3 oncogene is overexpressed and regulates cell growth and androgen receptor activity in human prostate cancer**. *Mol Endocrinol* 2006, **20**:2315–2325.

60. Engers R, Mueller M, Walter A, Collard JG, Willers R, Gabbert HE: **Prognostic relevance of Tiam1 protein expression in prostate carcinomas**. *Br J Cancer* 2006, **95**:1081–1086.

61. Santagata S, Demichelis F, Riva A, Varambally S, Hofer MD, Kutok JL, Kim R, Tang J, Montie JE, Chinnaiyan AM, Rubin MA, Aster JC: **JAGGED1 expression is associated with prostate cancer metastasis and recurrence**. *Cancer Res* 2004, **64**:6854–6857.

62. Silverman RH: **Implications for RNase L in prostate cancer biology**. *Biochemistry* 2003, **42**:1805–1812.

63. Raje D, Mukhtar H, Oshowo A, Clark CI: **What proportion of patients referred to secondary care with iron deficiency anemia have colon cancer?** *Dis Colon Rectum* 2007, **50**:1–4.

64. Ciardiello F, Tortora G: **Epidermal growth factor receptor (EGFR) as a target in cancer therapy: understanding the role of receptor expression and other molecular determinants that could influence the response to anti-EGFR drugs**. *Eur J Cancer* 2003, **39**:1348–1354.

65. Kim TD, Song KS, Li G, Choi H, Park HD, Lim K, Hwang BD, Yoon WH: **Activity and expression of urokinase-type plasminogen activator and matrix metalloproteinases in human colorectal cancer**. *BMC Cancer* 2006, **6**:211.

66. Uner A, Akcali Z, Unsal D: **Serum levels of soluble E-selectin in colorectal cancer**. *Neoplasma* 2004, **51**:269–274.

67. Eksioglu EA, Mahmood SS, Chang M, Reddy V: **GM-CSF promotes differentiation of human dendritic cells and T lymphocytes toward a predominantly type 1 proinflammatory response**. *Exp Hematol* 2007, **35**:1163–1171.

68. Zinzindohoué F, Lecomte T, Ferraz JM, Houllier AM, Cugnenc PH, Berger A, Blons H, Laurent-Puig P: **Prognostic significance of MMP-1 and MMP-3 functional promoter polymorphisms in colorectal cancer**. *Clin Cancer Res* 2005, **11**:594–599.

69. Zhang Y, Lai M, Lv B, Gu X, Wang H, Zhu Y, Zhu Y, Shao L, Wang G: **Overexpression of Reg IV in colorectal adenoma**. *Cancer Lett* 2003, **200**:69–76.

70. Ahn DH, Crawley SC, Hokari R, Kato S, Yang SC, Li JD, Kim YS: **TNF-alpha activates MUC2 transcription via NF-kappaB but inhibits via JNK activation**. *Cell Physiol Biochem* 2005, **15**:29–40.

71. Kummola L, Hämäläinen JM, Kivelä J, Kivelä AJ, Saarnio J, Karttunen T, Parkkila S: **Expression of a novel carbonic anhydrase, CA XIII, in normal and neoplastic colorectal mucosa**. *BMC Cancer* 2005, **5**:41.

72. Gröne J, Weber B, Staub E, Heinze M, Klaman I, Pilarsky C, Hermann K, Castanos-Velez E, Röpcke S, Mann B, Rosenthal A, Buhr HJ: **Differential expression of genes encoding tight junction proteins in colorectal cancer: frequent dysregulation of claudin-1, -8 and -12**. *Int J Colorectal Dis* 2007, **22**:651–659.

73. Viet HT, Wå gsäter D, Hugander A, Dimberg J: **Interleukin-1 receptor antagonist gene polymorphism in human colorectal cancer**. *Oncol Rep* 2005, **14**:915–918.

74. Kloor M, Michel S, Buckowitz B, Rüschoff J, Büttner R, Holinski-Feder E, Dippold W, Wagner R, Tariverdian M, Benner A, Schwitalle Y, Kuchenbuch B, von Knebel Doeberitz M: **Beta2-microglobulin mutations in microsatellite unstable colorectal tumors**. *Int J Cancer* 2007, **121**:454–458.

75. Youssef EM, Chen Xq, Higuchi E, Kondo Y, Garcia-Manero G, Lotan R, Issa JPJ: **Hypermethylation and silencing of the putative tumor suppressor** *Tazarotene-induced gene 1* **in human cancers**. *Cancer Res* 2004, **64**:2411–2417.

76. Muc-Wierzgon M, Nowakowska-Zajdel E, Kokot T, Kozowicz A, Zubelewicz B, Klakla K, Mazurek U, Cholewa K, Wilczok T, Wierzgon J, Sosada K: **Genetic disregulation of gene coding tumor necrosis factor alpha receptors (TNFalpha Rs) in colorectal cancer cells**. *J Exp Clin Cancer Res* 2004, **23**:651–660.

77. Maeda K, Kang SM, Sawada T, Nishiguchi Y, Yashiro M, Ogawa Y, Ohira M, Ishikawa T, Hirakawa YS, Chung K: **Expression of intercellular adhesion molecule-1 and prognosis in colorectal cancer**. *Oncol Rep* 2002, **9**:511–514.

78. Ferroni P, Palmirotta R, Spila A, Martini F, Raparelli V, Fossile E, Mariotti S, Del Monte G, Buonomo O, Roselli M, Guadagni F: **Prognostic significance of adiponectin levels in non-metastatic colorectal cancer**. *Anticancer Res* 2007, **27**:483–489.

79. Miyanaga K, Kato Y, Nakamura T, Matsumura M, Amaya H, Horiuchi T, Chiba Y, Tanaka K: **Expression and role of thrombospondin-1 in colorectal cancer**. *Anticancer Res* 2002, **22**:3941–3948.

80. Wan Y, Wu N, Wang Z, Ju X, Zhu J, Liu Y, Tang J, Huang Y: **Relationship between tissue factor expression and hepatic metastasis and prognosis in rectal cancer**. *Zhonghua Zhong Liu Za Zhi (Ch J Oncol)* 2002, **24**:378–380.

81. Bethke L, Webb E, Sellick G, Rudd M, Penegar S, Withey L, Qureshi M, Houlston R: **Polymorphisms in the cytochrome P450 genes CYP1A2, CYP1B1, CYP3A4, CYP3A5, CYP11A1, CYP17A1, CYP19A1 and colorectal cancer risk**. *BMC Cancer* 2007, **7**:123.

82. Cross NA, Chandrasekharan S, Jokonya N, Fowles A, Hamdy FC, Buttle DJ, Eaton CL: **The expression and regulation of ADAMTS-1, -4, -5, -9, and -15, and TIMP-3 by TGFbeta1 in prostate cells: relevance to the accumulation of versican**. *Prostate* 2005, **63**:269–275.

83. Hudolin T, Juretic A, Spagnoli GC, Pasini J, Bandic D, Heberer M, Kosicek M, Cacic M: **Immunohistochemical expression of tumor antigens MAGE-A1, MAGE-A3/4, and NY-ESO-1 in cancerous and benign prostatic tissue**. *Prostate* 2006, **66**:13–18.

84. Ishii K, Usui S, Sugimura Y, Yoshida S, Hioki T, Tatematsu M, Yamamoto H, Hirano K: **Aminopeptidase N regulated by zinc in human prostate participates in tumor cell invasion**. *Int J Cancer* 2001, **92**:49–54.

85. Diss JK, Faulkes DJ, Walker MM, Patel A, Foster CS, Budhram-Mahadeo V, Djamgoz MB, Latchman DS: **Brn-3a neuronal transcription factor functional expression in human prostate cancer**. *Prostate Cancer Prostatic Dis* 2006, **9**:83–91.

86. Cross DS, Burmester JK: **Functional characterization of the GDEP promoter and three enhancer elements in retinoblastoma and prostate cell lines**. *Med Oncol* 2008, **25**:40–49.

87. Wolfgang CD, Essand M, Lee B, Pastan I: **T-cell receptor gamma chain alternate reading frame protein (TARP) expression in prostate cancer cells leads to an increased growth rate and induction of caveolins and amphiregulin**. *Cancer Res* 2001, **61**:8122–8126.

88. Descazeaud A, de la Taille A, Allory Y, Faucon H, Salomon L, Bismar T, Kim R, Hofer MD, Chopin D, Abbou CC, Rubin MA: **Characterization of ZAG protein expression in prostate cancer using a semi-automated microscope system**. *Prostate* 2006, **66**:1037–1043.

89. Sahni A, Simpson-Haidaris PJ, Sahni SK, Vaday GG, Francis CW: **Fibrinogen synthesized by cancer cells augments the proliferative effect of fibroblast growth factor-2 (FGF-2)**. *J Thromb Haemost* 2008, **6**:176–183.

90. Bandyopadhyay S, Wang Y, Zhan R, Pai SK, Watabe M, Iiizumi M, Furuta E, Mohinta S, Liu W, Hirota S, Hosobe S, Tsukada T, Miura K, Takano Y, Saito K, Commes T, Piquemal D, Hai T, Watabe K: **The tumor metastasis suppressor gene Drg-1 down-regulates the expression of activating transcription factor 3 in prostate cancer**. *Cancer Res* 2006, **66**:11983–11990.

Figure legends

Figure 1 - Overview methodology

Overview of the three applied model building strategies: A - use of a single data set; B - manual integration of data over time; C - a genome-wide integration approach. The data sets are represented as matrices with rows corresponding to patients and columns corresponding to genes, proteins, or CNVs. In step A, LS-SVM models are built on each data set separately. A 2-dimensional grid is used for the optimization of the regularization parameter and the number of features. For step B, data sets over time are combined. By using the changes in expression or abundance as features, a 2-dimensional grid is sufficient. In step C, an intermediate integration method is used for the integration of all available data sets. A k-dimensional grid is required for optimizing the regularization parameter and the number of features selected from the (k-1) integrated data sets. (LOO = leave-one-out; FS = feature selection; NF = number of features; T = time point; M_i = model for parameter combination i)

Tables

Table 1 - Overview of the two case studies on rectal and prostate cancer

	Data set I on rectal cancer		*Data set II on prostate cancer*
nb of samples	36		55
data sources	microarray proteomics		microarray genomics
nb of features (after preprocessing)	T_0 6913 genes 90 proteins	T_1 6913 genes 92 proteins	 6974 genes 7305 CNVs
outcomes	WHEELER pN-STAGE CRM		GRADE STAGE METASTASIS RECURRENCE

Table 2 - LS-SVM models for the prediction of WHEELER, pN-STAGE and CRM in rectal cancer

outcome		model	NG$^{\mu}$	NP$^{\varsigma}$	AUC (SE)$^{\varphi}$	p-value$^{\varrho}$
WHEELER	A	MT_0	4		0.75385 (0.1085)	0.0987
		MT_1	29		0.90385 (0.0502)	0.6861
		PT_0		35	0.74231 (0.0867)	0.0540
		PT_1		11	0.90385 (0.0575)	0.7273
	B	$MT_0 - T_1$	32		0.68462 (0.1215)	0.0598
		$PT_0 - T_1$		5	0.86538 (0.0621)	0.4135
	C	MT_{01}	3^{π}		0.78077 (0.0985)	0.1320
		PT_{01}		21^{π}	0.76923 (0.0831)	0.0831
		MPT_0	3	35	0.84615 (0.0718)	0.2760
		MPT_1	**25**	**12**	**0.92692 (0.0425)**	
		MPT_{01}	2^{π}	31^{π}	0.88462 (0.0558)	0.4858
		MT_0PT_1	2	4	0.93846 (0.0444)	0.8101^{ϕ}
pN-STAGE	A	MT_0	25		0.64935 (0.0914)	2.315e-4
		MT_1	22		0.85065 (0.0665)	0.0362
		PT_0		2	0.67532 (0.0906)	6.659e-4
		PT_1		12	0.84091 (0.0652)	0.0238
	B	$MT_0 - T_1$	4		0.60714 (0.0986)	1.359e-4
		$PT_0 - T_1$		9	0.76623 (0.0900)	0.0153
	C	MT_{01}	24^{π}		0.92857 (0.0450)	0.1998
		PT_{01}		34^{π}	0.81818 (0.0695)	0.0145
		MPT_0	27	27	0.91883 (0.0469)	0.1591
		MPT_1	**21**	**14**	**0.98701 (0.0135)**	
		MPT_{01}	23^{π}	16^{π}	0.96104 (0.0280)	0.3421
		MT_0PT_{01}	26	20^{π}	1 (0)	0.3347^{ϕ}
CRM	A	MT_0	33		0.67901 (0.1016)	0.0072
		MT_1	9		0.92593 (0.0472)	0.4955
		PT_0		34	0.85185 (0.0624)	0.0935
		PT_1		34	0.76543 (0.0831)	0.0281
	B	$MT_0 - T_1$	6		0.91358 (0.0480)	0.4030
		$PT_0 - T_1$		2	0.82716 (0.0709)	0.0849
	C	MT_{01}	16^{π}		0.80658 (0.0846)	0.0468
		PT_{01}		3^{π}	0.75309 (0.0865)	0.0227
		MPT_0	7	27	0.84774 (0.0688)	0.1340
		MPT_1	**7**	**33**	**0.96296 (0.0344)**	
		MPT_{01}	2^{π}	3^{π}	0.82305 (0.0771)	0.0973
		MT_1PT_0	16	14	0.96296 (0.0376)	1
		$MT_{01}PT_1$	9^{π}	29	0.98765 (0.0146)	0.4924^{ϕ}

$^{\mu}$ number of genes selected in each LOO iteration
$^{\varsigma}$ number of proteins selected in each LOO iteration
$^{\pi}$ number of features used at both time points
$^{\varphi}$ area under the ROC curve (standard error) obtained with leave-one-out
$^{\varrho}$ comparison of AUC between each model and the best model in bold [46]
$^{\phi}$ this model is better than the model in bold we compare with

Table 3 - Features for (colo)rectal cancer selected by MPT_1 and known to be involved in this type of cancer

outcome$^{\mu}$	gene/protein	hits$^{\varphi}$	region	function	up/down$^{\varsigma}$	reference
W	Cox-2	36	1q25.2-q25.3	progression	up	[50]
W	IL-1B	36	2q14	inflammatory response	up	[50]
W	ferritin	36	11q13; 19q13.3-q13.4	iron storage	down	[63]
W	EGF	36	4q25	cell growth/proliferation/ differentiation	up	[64]
W	MMP-2	36	16q13-q21	invasion/metastasis	up	[65]
W	TGFα	36	2p13	angiogenesis/cell proliferation	down	[51]
W	SELE	25	1q22-q25	progression/metastasis	up	[66]
W	GM-CSF	24	5q31.1	maintenance of granulocytes/macrophages	up	[67]
W	MMP-1	15	11q22.3	tumour invasion/metastasis/ poor prognosis	up	[68]
N	Reg4	36	1p13.1-p12	early carcinogenesis	down	[69]
N	MUC2	36	11p15.5	deregulated by TNFα	down	[70]
N	CA1	36	8q13-q22.1	carbonate dehydratase activity	down	[71]
N	CA2	36	8q22	carbonate dehydratase activity	down	[71]
N	CLDN8	36	21q22.11	tumorigenesis	down	[72]
N	CEA	36	19q13.1-q13.2	cell adhesion; tumour marker for recurrence	down	[53]
N	IL-1ra	36	2q14.2	carcinogenesis	up	[73]
N	CA19-9	36		tumour marker for recurrence	down	[53]
N	ferritin	36	11q13; 19q13.3-q13.4	iron storage	down	[63]
N	IL-1beta	36	2q14	inflammatory response	down	[50]
N	beta2-microglobulin	36	15q21-q22.2	metastasis	up	[74]
N	RARRES1	31	3q25.32-q25.33	cell proliferation	down	[75]
N	IL-8	28	4q13-q21	progression/metastasis	down	[52]
N	TNFRII	24	1p36.3-p36.2	apoptosis	up	[76]
C	ICAM-1	36	19p13.3-p13.2	metastasis	down	[77]
C	CEA	36	19q13.1-q13.2	cell adhesion; tumour marker for recurrence	down	[53]
C	MMP-2	36	16q13-q21	invasion/metastasis	up	[65]
C	adiponectin	36	3q27	metabolic/hormonal processes	down	[78]
C	thrombos-pondin-1	36	15q15	angiogenesis/tumour growth	up	[79]
C	EGF-R	36	7p12	cell growth/differentiation/ proliferation	up	[49]
C	tissue factor	35	1p22-p21	angiogenesis/metastasis	up	[80]
C	CYP1B1	35	2p21	drug metabolism	down	[81]
C	EGF	32	4q25	cell growth/proliferation/ differentiation	up	[64]

CRC = (Colo)rectal Cancer

$^{\mu}$ W=WHEELER; N=pN-STAGE; C=CRM

$^{\varphi}$ number of occurrences of the gene/protein in the 36 LOO iterations

$^{\varsigma}$ up/down-regulation in the good responders w.r.t. moderate or poor responders; no lymph nodes w.r.t. at least one regional lymph node; negative CRM w.r.t. positive CRM

Table 4 - LS-SVM models for the prediction of GRADE, STAGE, METASTASIS and RECURRENCE in prostate cancer

outcome		model	NG$^{\mu}$	NC$^{\varsigma}$	AUC (SE)$^{\varphi}$	p-value$^{\varrho}$
GRADE	A	M	24		0.8304 (0.0623)	0.2727
		G		8	0.7822 (0.0632)	0.0503
	C	MG	**6**	**8**	**0.9006 (0.0413)**	
STAGE	A	M	18		0.6576 (0.0778)	0.0191
		G		32	0.7936 (0.0631)	0.3466
	C	MG	**42**	**22**	**0.8528 (0.0550)**	
METASTASIS	A	M	18		0.9759 (0.0178)	0.4392
		G		12	0.8114 (0.0755)	0.0166
	C	MG	**18**	**3**	**0.9868 (0.0121)**	
RECURRENCE	A	M	24		0.7208 (0.0936)	0.5392
		G		26	0.4481 (0.1433)	0.0354
	C	MG	**32**	**2**	**0.7857 (0.0934)**	

$^{\mu}$ number of genes selected in each LOO iteration

$^{\varsigma}$ number of copy number variations selected in each LOO iteration

$^{\varphi}$ area under the ROC curve (standard error) obtained with leave-one-out

$^{\varrho}$ comparison of AUC between each model and the best model in bold [46]

Table 5 - Features for prostate cancer selected by MG and known to be involved in this type of cancer

outcome$^{\mu}$	gene/CNV	hits$^{\varphi}$	region	function	up/down$^{\varsigma}$	reference
G	SFRP4	55	7p14.1	inhibitor of PT growth/invasion	up	[55]
G	VCAN	55	5q14.3	contributor to PC pathology	up	[82]
G	ALOX15B	36	17p13.1	suppressor of PT development	down	[54]
S	MAGEA4	50	Xq28	only expressed in PC (diagnosis & therapy)	down	[83]
S	ANPEP	50	15q25-q26	PT cell invasion	down	[84]
S	POU4F1	50	13q31.1	PC cell growth	down	[85]
S	CXCL14	48	5q31	inhibitor of PT growth	up	[56]
S	RNASEL	48	1q25	polymorphic changes as tumour suppressor in hereditary PC	up	[62]
S	GDEP	41	4q21.1	prostate-specific gene	down	[86]
M	ERG	50	21q22.3	proto-oncogene; early prostate carcinogenesis	up	[57]
M	AREG	49	4q13-q21	PC progression/growth via TARP	down	[87]
M	VAV3	49	1p13.3	oncogene; PC development/progression	up	[59]
M	ADAMTS1	26	21q21.2	negatively effected by TGFbeta1, which increases VCAN-expression	down	[82]
R	AZGP1	29	7q22.1	inversely associated to tumour stage; predictor of biochemical recurrence	down	[88]
R	TIAM1	29	21q22.1-11	predictor of decreased disease-free survival/recurrence	up	[60]
R	FGG	28	4q28	PC cell growth	down	[89]
R	ATF3	26	1q32.3	inversely related to invasion/ angiogenesis; positively correlated to metastases	down	[90]
R	JAG1	26	20p12.1-11.23	cell growth/progression/metastasis	up	[61]
R	ERG	14	21q22.3	proto-oncogene; early prostate carcinogenesis	up	[57]
R	ALOX15B	14	17p13.1	suppressor of PT development	down	[54]

PT = Prostate Tumour; PC = Prostate Cancer

$^{\mu}$ G=GRADE; S=STAGE; M=METASTASIS; R=RECURRENCE

$^{\varphi}$ number of occurrences of the gene/CNV in all LOO iterations (nb of LOO iterations for G=55, S=50, M=50, R=29)

$^{\varsigma}$ up/down-regulation in high-grade w.r.t. low-grade; advanced stage w.r.t. early stage; metastasis w.r.t. no metastasis; recurrence w.r.t. no recurrence

Table 6 - Comparison of our kernel-based integration approach with the ensemble approach

outcome	AUC (SE)$^{\mu}$ MPT_1 / MG	AUC (SE)$^{\mu}$ ensemble approach	p-value$^{\varsigma}$
WHEELER	0.92692 (0.0425)	0.95 (0.0339)	0.6160
pN-STAGE	0.98701 (0.0135)	0.92532 (0.0432)	0.1422
CRM	0.96296 (0.0344)	0.78601 (0.0783)	**0.0384**
GRADE	0.9006 (0.0413)	0.85673 (0.0521)	0.3745
STAGE	0.8528 (0.0550)	0.8304 (0.0582)	0.6836
METASTASIS	0.9868 (0.0121)	0.94518 (0.0309)	0.1313
RECURRENCE	0.7857 (0.0934)	0.45455 (0.1352)	**0.0182**

$^{\mu}$ area under the ROC curve (standard error) obtained with leave-one-out

$^{\varsigma}$ comparison in AUC between the best models obtained with our strategy (MPT_1 for rectal cancer, MG for prostate cancer) and the corresponding ensemble models based on the same number of features [46]

Additional file legends

Additional file 1

File format: PDF

Title: ROC curves of the models shown in Tables 2 and 4.

Description: The ROC curves of the optimal LS-SVM models for all considered combinations of data sets shown in Tables 2 and 4 are shown in these additional figures. Additional figures 1 to 3 show the ROC curves for the prediction of WHEELER, pN-STAGE, and CRM in rectal cancer, respectively. For prostate cancer, the ROC curves for the prediction of GRADE, STAGE, METASTASIS, and RECURRENCE are shown in additional figures 4 to 7, respectively.

Additional file 2

File format: PDF

Title: Additional LS-SVM models for rectal cancer.

Description: This additional file shows the results for the prediction of WHEELER, pN-STAGE, and CRM in rectal cancer, using step C models for which a sample is required at both time points and for which both technologies need to be performed. The AUC value and the number of included features are shown for each model. Significance tests were performed to compare these models with the best model based on two data sets shown in bold in Table 2 from the main manuscript.

Additional file 3

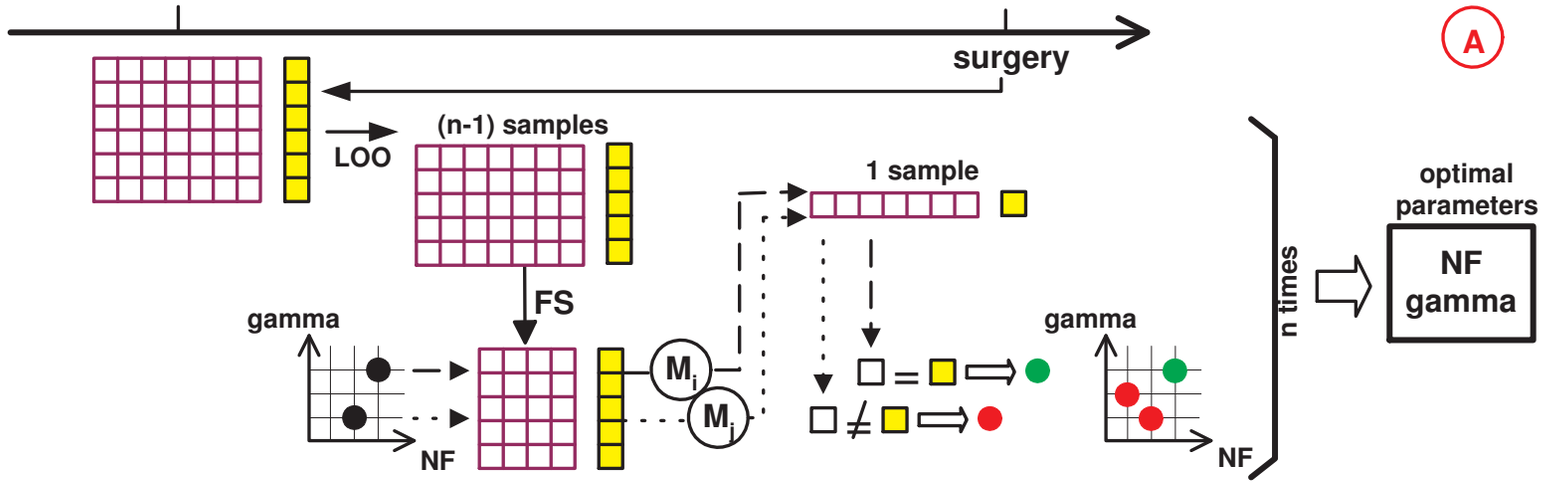
File format: PDF

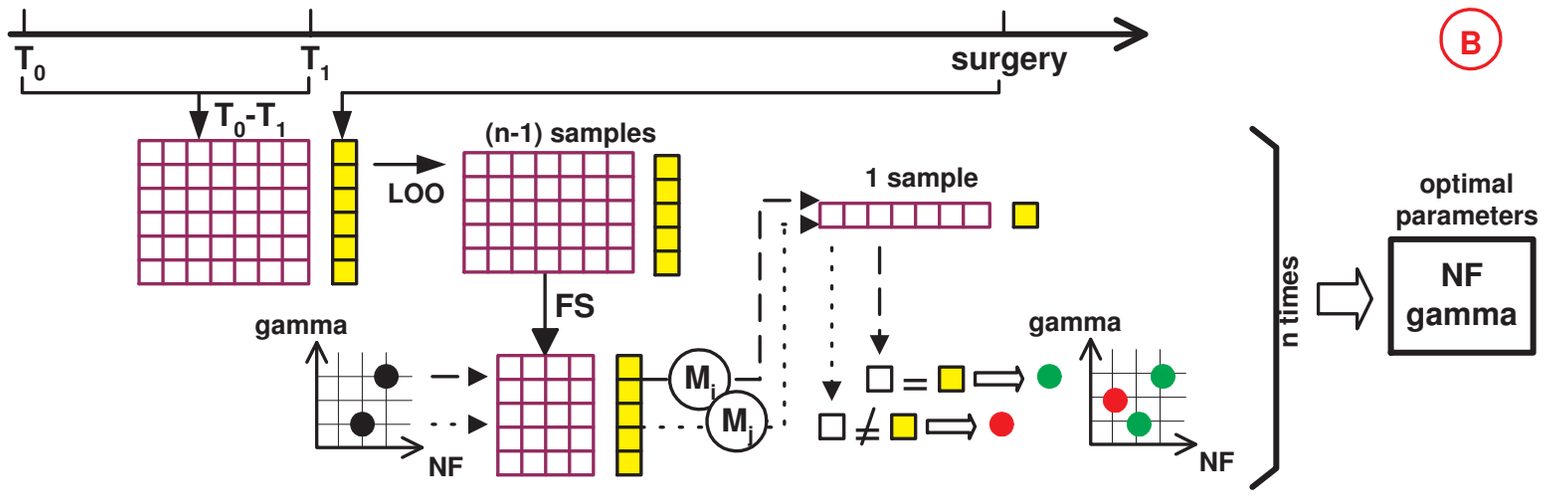
Title: Genes, proteins, and CNVs selected by the models MPT_1 and MG.

Description: Additional tables 1 to 3 show all genes and proteins selected by the best performing models MPT_1 for the prediction of WHEELER (25 genes, 12 proteins), pN-STAGE (21 genes, 14 proteins), and CRM (7 genes, 33 proteins) in rectal cancer. Additional tables 4 to 7 show, for prostate cancer the genes and CNVs selected by the best performing models MG for the prediction of GRADE (6 genes, 8 CNVs), STAGE (42 genes, 22 CNVs), METASTASIS (18 genes, 3 CNVs), and RECURRENCE (32 genes, 2

CNVs). All tables additionally show the number of LOO iterations in which each gene, protein, or CNV was selected, their chromosomal region, and whether it is up- or down-regulated.


features 1
features 2
outcome
n samples
data set 1
data set 2


A
surgery
LOO
(n-1) samples
1 sample
FS
gamma
NF
n times
optimal parameters
NF
gamma


B
T0
T1
surgery
T0-T1
LOO
(n-1) samples
1 sample
FS
gamma
NF
n times
optimal parameters
NF
gamma

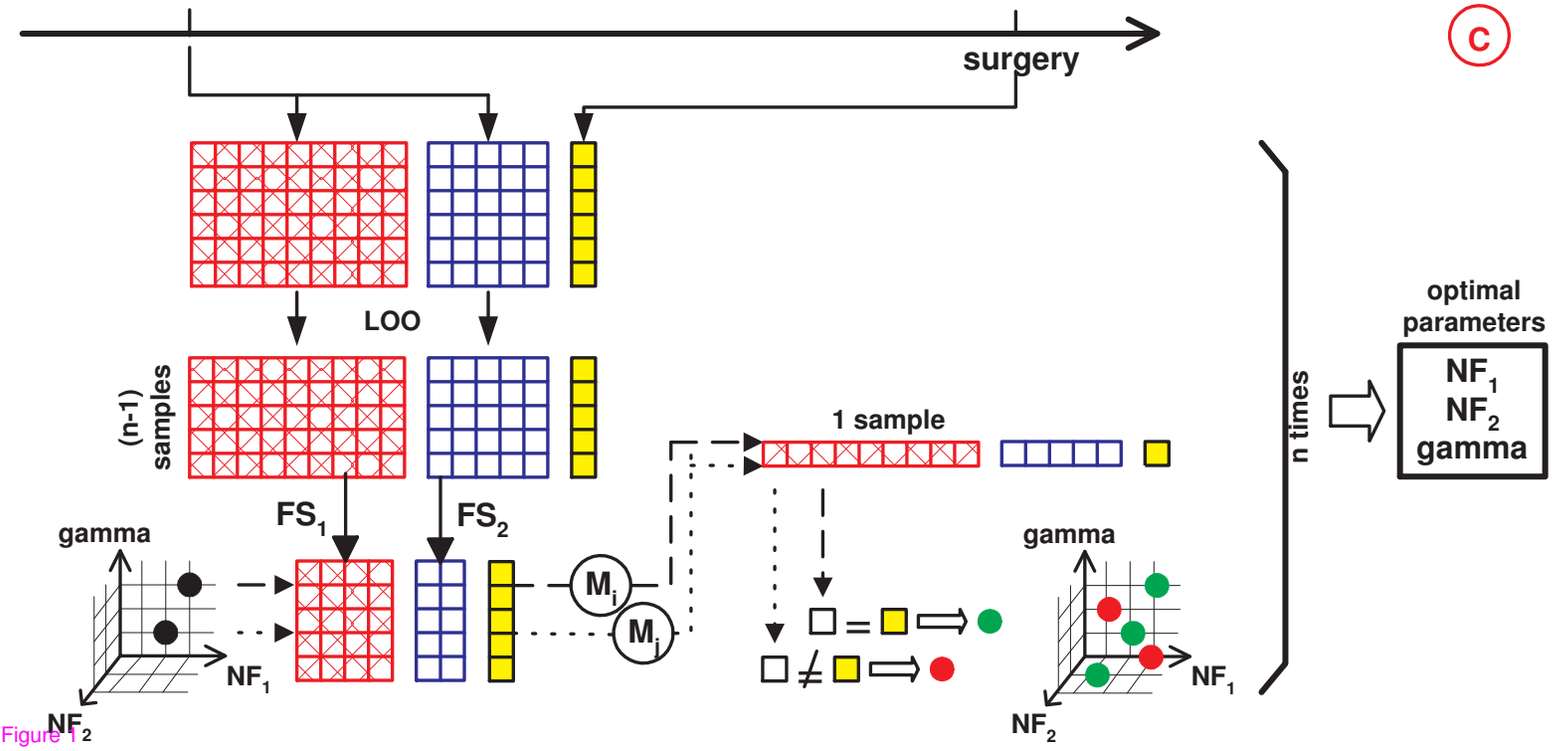

C
surgery
LOO
(n-1) samples
1 sample
FS1
FS2
gamma
NF1
NF2
n times
optimal parameters
NF1
NF2
gamma

Figure 1 2

Additional files provided with this submission:

Additional file 1: additionalfile1.pdf, 33K
http://www.genomemedicine.com/imedia/2043679482540134/supp1.pdf
Additional file 2: additionalfile2.pdf, 32K
http://www.genomemedicine.com/imedia/2025715217254013/supp2.pdf
Additional file 3: additionalfile3.pdf, 59K
http://www.genomemedicine.com/imedia/1535481098263075/supp3.pdf