Bioinformatics, 31(24), 2015, 3961–3969
doi: 10.1093/bioinformatics/btv498
Advance Access Publication Date: 26 August 2015
Original Paper

Systems biology

CoD: inferring immune-cell quantities related to disease states

Amit Frishberg, Yael Steuerman and Irit Gat-Viks*

Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 3, 2015; revised on July 26, 2015; accepted on August 18, 2015

Abstract

Motivation: The immune system comprises a complex network of genes, cells and tissues, coordinated through signaling pathways and cell−cell communications. However, the orchestrated role of the multiple immunological components in disease is still poorly understood. Classifications based on gene-expression data have revealed immune-related signaling pathways in various diseases, but how such pathways describe the immune cellular physiology remains largely unknown.
Results: We identify alterations in cell quantities discriminating between disease states using ' Cell type of Disease' (CoD), a classification-based approach that relies on computational immune-cell decomposition in gene-expression datasets. CoD attains significantly higher accuracy than alternative state-of-the-art methods. Our approach is shown to recapitulate and extend previous knowledge acquired with experimental cell-quantification technologies.
Conclusions: The results suggest that CoD can reveal disease-relevant cell types in an unbiased manner, potentially heralding improved diagnostics and treatment.
Availability and implementation: The software described in this article is available at http://www.csgi.tau.ac.il/CoD/.
Contact: iritgv@post.tau.ac.il
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Immune-cell heterogeneity is central to systems-level functions in health and disease. It has therefore been the focus of a substantial research effort to identify immune-cell types that are relevant to a disease of interest. The main approach has been to quantify each individual immune-cell type using cell-measurement tools such as immunohistochemical staining or FACS analysis, and then utilize the resulting measurements to identify particular cell types that are relevant to patient prognosis (e.g. Ge *et al.*, 2013; Piersma *et al.*, 2007). However, as hundreds of immune-cell subsets work in coordination, manual measurement of relatively small numbers of subsets poses significant obstacles to successful identification of disease-relevant immune cells. One such impediment, attributable to the strict requirement for the use of antibodies of high quality, is the strong bias toward immune-cell types that are well studied. Another major hurdle has been the large amounts of material required for quantification, as well as the high costs and time-consuming nature of the process. As a result, most studies have been focused on preselected subpopulations of cells and have rarely investigated a broad range of immune-cell types.

One possible solution to these obstacles would be to combine disease classification based on transcription profiles with predefined groups of gene markers that relate to specific immune-cell types. Inferring the activity of relevant gene groups as a clue to disease classification is already an area of active research (Dinu *et al.*, 2007; Hanzelmann *et al.*, 2013; Lee *et al.*, 2008) but it has not, to our knowledge, been applied in the case of immune-cell types. Typically, these methods take as their starting point a list of signaling pathway-related gene groups from, for example, the Molecular Signatures Database (Liberzon *et al.*, 2011). The first phase is to determine the activity of each gene group in the transcription profiles of the various samples, and the second phase is to classify the samples based on the inferred activity levels of the pathways. A key question, however, is whether these group activity-based methods can be used to automatically infer the quantities of specific immune-cell subpopulations that are relevant to a disease state. To understand the challenges involved in this task, it is first necessary to realize that most genes are assigned a quantitative level in every cell type. This is fundamentally unlike the case of molecular pathways, in which the genes either participate in the pathway or do not.

Several computational approaches have been developed for identifying quantitative changes in particular immune-cell types within a complex tissue under study (e.g. Altboum *et al.,* 2014; Newman *et al.,* 2015; Qiao *et al.,* 2012). These methods integrate transcriptional profiling of a given complex tissue with prior genomic knowledge about immune cells into a mathematical framework enabling a robust decomposition of tissues into immune-cell quantities. For example, the DCQ algorithm (Altboum *et al.,* 2014) has been shown to identify over 200 immune-cell subpopulations simultaneously, opening the way to a global view of immune physiology through computational modeling. Therefore, computational immune-cell decomposition methods can be used to aid identification of immune-cell types related to specific diseases.

Here, we developed 'Cell type of Disease' (CoD), a methodology for identifying *all* disease-relevant immune-cell types. CoD was designed to reveal the entire set of relevant cell types, including redundant cell types that are completely obscured by other cell types, thus providing a comprehensive understanding of disease mechanisms. To that end, CoD utilizes computational immune-cell decomposition on gene-expression data to predict cell type quantities (Altboum *et al.,* 2014), and then exploits the cell quantities as features in a random forest classification scheme (Breiman, 2001). Finally, permutation tests are utilized to select the entire set of relevant features (rather than a 'minimal-optimal' set of non-redundant features).

Our synthetic data analysis showed that CoD achieves better accuracy in predicting immune-cell types of relevance for disease classification than existing feature selection and activity-based classification methods. We tested our approach by applying it in the case of whole-body ionizing radiation (Zheng *et al.,* 2014) as well as in both the primary tumor and distal organs in mouse models of breast cancer (Schoenherr *et al.,* 2011). The results indicated the ability of CoD to successfully recapitulate many of the known immune-cell changes in these systems, and showed that the predictions can provide novel leads even when the immune-cell composition has already been measured experimentally.

2 Methods

2.1 An overview of the CoD pipeline

We have developed CoD, a method for selecting all types of immune cells of relevance to a particular disease or treatment. CoD takes as input gene-expression data over the samples, where each sample is labeled by one of two classes (e.g. cases versus controls, good versus poor prognosis, with or without a certain treatment). The CoD pipeline proceeds in three stages (Fig. 1). First, it generates 'cell-type features' based on the abovementioned DCQ algorithm and using prior quantitative knowledge about a collection of immune-cell types. In this study, we used a pre-compiled 'immune-cell compendium' of 207 gene-expression profiles that were monitored in isolated murine immune-cell types consisting of both inflammatory and resident immune cells (the Immgen database; Heng and Painter, 2008) (Supplementary Table S1). Second, CoD exploits the 'random forest' classification method (Breiman, 2001) to extract all disease-relevant immune-cell types while also considering complex relationships between these cell types. Finally, in order to select all disease-relevant cell types, statistical significance of features is calculated based on permutation tests.

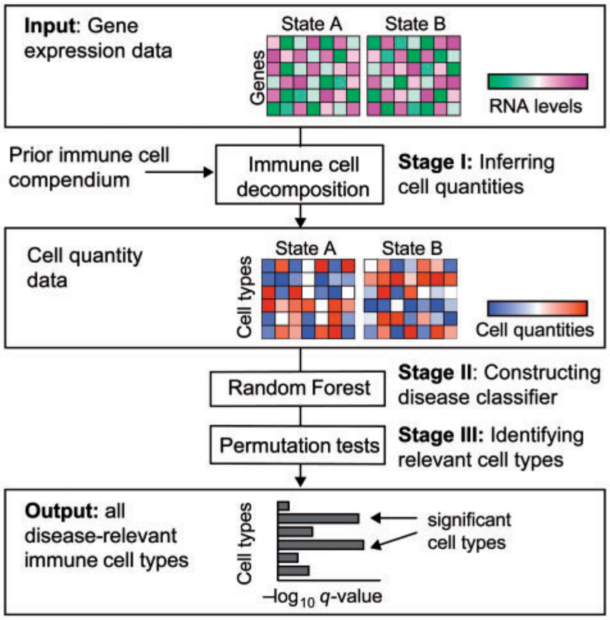
Input: Gene expression data
State A
State B
Genes
RNA levels
Prior immune cell compendium
Immune cell decomposition
Stage I: Inferring cell quantities
Cell quantity data
State A
State B
Cell types
Cell quantities
Random Forest
Stage II: Constructing disease classifier
Permutation tests
Stage III: Identifying relevant cell types
Output: all disease-relevant immune cell types
Cell types
significant cell types
−log10 q-value

Fig. 1. Overview of the CoD approach. The input data (top panel) include the expression of genes over the samples, where the samples are labeled by one of two disease states. CoD incorporates three stages. In stage 1, an immune-cell decomposition algorithm is applied in order to infer the quantity of each cell type in each of the samples. In stage 2, a random forest classification algorithm is applied on the cell-quantity dataset from stage 1 to identify disease-relevant immune-cell types. Stage 3 (bottom) calculates the statistical significance (*q*-values) of these predictions based on permutation tests

2.1.1 Stage 1: Inferring cell quantities

We start by normalizing the gene-expression values of each sample by subtracting from them the average value of all samples of the same gene. We refer to the normalized values as 'differential RNA levels'. Next, we infer the immune-cell quantities of each sample using the decomposition-based DCQ algorithm (Altboum *et al.,* 2014). In brief, the DCQ algorithm integrates two input types: (1) the immune-cell compendium together with a pre-defined list of the 61 cell-surface markers that were actually used to isolate (by FACS) the individual samples in the immune-cell compendium (taken from Altboum *et al.,* 2014); and (2) the differential RNA levels of a single sample. Once given a sample, and relying on the assumption that all RNAs are affected by the same changes in immune-cell quantities, DCQ models the differential RNA levels of each gene as the sum of changes in quantities of all 207 immune-cell types that are part of the immune-cell compendium. This is done by means of the regularized 'elastic net' regression technique (Zou and Hastie, 2005) while exploiting the quantitative signature of the 61 gene markers in each immune-cell type, even when the same markers are expressed in many of these cell subpopulations. The output of the DCQ algorithm reflects differential immune-cell quantity values (in accordance with its input differential RNA levels), but for simplicity of terminology we refer to these quantities throughout this study as 'cell quantities' or 'cell features'. Accordingly, the 'cell-quantity data' matrix consists of the inferred immune-cell-type features over the samples, where each sample is labeled by one of two classes (e.g., Fig. 1, middle panel).

2.1.2 Stage 2: Constructing an immune-cell-based classifier

CoD utilizes the data on immune-cell quantities from stage 1 in a disease-classification framework in order to identify all 'relevant cell types' that discriminate well between the two labeled classes. To that end we first apply the random forest ensemble learning method on the cell-quantity data matrix (Breiman, 2001). This provides a 'cell-based classifier', a collection of decision trees (here, 1000 trees) that use immune-cell types as features. Each tree is trained on the basis of a random sample of the learning data, and is capable of classifying a given sample into one of the two classes. The overall prediction for a given sample is calculated as the average of all predictions made by all the individual trees. The ability of the classifier to generalize for unseen data can be assessed by the 'generalization error' measure (Breiman, 2001; originally termed 'out-of-bag error'). This is calculated as the average error across all training samples, where the error values for each sample are calculated across the trees that did not use the same sample as part of their training data.

The random-forest algorithm natively produces a feature-importance score (here, 'cell-importance' score) reflecting the role of each cell feature in the complex interactions within the classifier model. This is done by calculating the contribution of each cell type to the learning accuracy of the classifier, as proposed by Breiman (2001).

2.2.3 Stage 3: Selection of all disease-relevant immune-cell types

In order to identify *all* the relevant features, we assign a *P*-value score to each cell type based on its calculated importance score. Most feature selection methods search for a minimal-optimal set of non-redundant features (Witold *et al.,* 2015). However, several recent studies have highlighted the importance of systematic identification of all relevant features—including both redundant and non-redundant features—in order to obtain a comprehensive view of disease mechanisms (denoted the 'all-relevant' feature selection problem; Witold *et al.* 2015). The problem of identifying all-relevant features was previously shown to be distinct from and more difficult than the traditional problem of minimal-optimal feature selection (Kursa, 2014; Kursa and Rudnicki, 2011; Nilsson *et al.,* 2007; Witold *et al.,* 2015). In particular, previous analyses have shown that given random correlations in data, it is difficult to discern whether a high importance score truly reflects a relevant feature. To tackle this problem, the addition of non-informative ('shadow') variables has been proposed to achieve an unbiased selection of features (reviewed in Witold *et al.,* 2015).

Here we devised a novel 'all-relevant' feature selection approach that relies on permutations and is implemented as a wrapper around a shadow-based random-forest algorithm. We use the shadow cell types only during calculation of the random-forest classifier and importance scores, and remove them before calculating the empirical *P*-values. The procedure is carried out as follows: we first extend the cell-quantity matrix by adding N_d shadows features, where the quantities of shadow cell types are generated by permuting the sample labels of the original cell types. Next, we create N_p permutations of this extended cell-quantity data matrix. A random-forest classifier and cell-importance scores are then calculated for each of the extended (original or permuted) matrices. The '*empirical P-value*' for a certain importance score *x* is defined as the fraction of original cell types in permuted datasets in which the importance score is higher than *x*. To compensate for cell–cell dependencies, empirical *P*-values are compared against the null model to determine the 'inflation coefficient' λ_{med}, which is the ratio of the median of the observed log empirical *P*-value distribution to the expected median. To calculate 'expected median', we start by applying CoD stage 1 and 2 on 100 permuted datasets, producing 100 importance scores vectors; the median values of these scores are then used as input to CoD stage 3 to calculate empirical *P*-values; the 'expected median' is defined as the median of these empirically derived null *P*-values. Next, the original empirical *P*-values are corrected for inflation ($\log P_{\text{corr}} = \log P/\lambda_{\text{med}}$) and then accounted for multiple testing using a standard false discovery rate (FDR) *q*-value procedure. 'Relevant cell types' are selected on the basis of an FDR *q*-value score that is lower than a predefined cutoff.

2.2 Synthetic data analysis

We carried out a simulation study using 'synthetic collections', where a single collection consisted of 100 samples (50 in each of the two sample classes A and B). We generated each sample by mimicking transcription profiles of immune cells in a complex tissue. Synthetic gene expression y_{jk} for gene *j* in sample *k* includes a mix of isolated cell types:

$$y_{jk} = \sum_{i=1\ldots L} f_i^k \cdot b_{ij} + \varepsilon_{jk} \tag{1}$$

where *L* is the total number of cell types in the immune-cell compendium (here, the ImmGen dataset—Heng and Painter, 2008—with $L = 207$); b_{ij} denotes the (log scaled) gene-expression value of gene *j* in cell type *i* (b_{ij} values were taken from the ImmGen dataset); and $\varepsilon_{jk} \sim N(0, \sigma_j{}^2)$ where $\sigma_j{}^2 = \gamma_g \cdot \Sigma_{i=1\ldots L} b_{ij}/L$ is based on a noise factor γ_g. The fraction of cell type *i* in sample *k* is denoted by $f_i{}^k$. For a sample in class A, the samples were generated by mixing all cell types with equal fractions. For a sample in class B, the fractions of certain cell types were altered by either increasing or decreasing the cell quantities. Formally,

$$f_i^k = \frac{c_i^k}{\sum\limits_{i=1,..,L} c_i^k} + \varepsilon_{ik} \tag{2}$$

$$c_i^k = \begin{cases} 1/L & k \in A \\ 1/L & k \in B, \quad i \notin \{U, D\} \\ (1+s)/L & k \in B, \quad i \in U \\ (1-s)/L & k \in B, \quad i \in D \end{cases} \tag{3}$$

where $\varepsilon_{ik} \sim N(0, \gamma_c \cdot c_{ik}/\Sigma_{i=1\ldots L} c_{ik})$ is based on the noise factor γ_c; the two cell-type lists *U* and *D* were preselected as the cell types whose fractions are increasing or decreasing, respectively; and *s* is the amount of alteration in cell quantity in these two subsets, denoted the *effect size.*

We selected the altered cell-type lists *U* and *D* while considering a common situation in real biological tissues where closely related cell subpopulations undergo an orchestrated change in quantities. Accordingly, the cell-type selection was based on a previously published partitioning into 58 groups of closely related cell types (Jojic *et al.*, 2013; Supplementary Table S1). Here, for each synthetic data collection we arbitrarily chose the *m*-increasing and *m*-decreasing groups out of the 58 groups, thereby generating the *U* and *D* lists of cell types, respectively. Hereafter, we refer to the value of 2 *m* as the '*number of altered cell-type groups*'. Overall, unless stated otherwise, a single '*synthetic dataset*' consists of 100 collections, each carrying 100 samples that were generated using the same effect size and number of altered cell-type groups. We consider datasets with various combinations of nine different effect sizes (s ranging from 0 to 0.9), five different noise factor levels ($\gamma_c = 10^{-6}$ to 10^{-2}; $\gamma_g = 10^{-6}$ to 10^{-2}) and seven different numbers of altered cell-type groups ($2m = 2$, 4, 6, 8, 10, 14, 20; this range is in agreement with our observations in real datasets; see Supplementary Table S2). This basic simulation setting was altered in several specific tests as detailed in Supplementary Information Sections 1–3. Unless stated otherwise, we use the default parameters $s = 0.01$ and $\gamma_g = \gamma_c = 10^{-4}$, whose level of complexity is similar to the situation in real biological datasets (Supplementary Fig. S1).

To assess the ability of the method to correctly detect the relevant cell types that differ between the two sample classes, we rely on a standard Receiver Operating Characteristic (ROC) analysis. The area under this curve captures the performance over a range of P-value cutoffs, and is referred to as the '*cell-type accuracy*' metric. Unless stated otherwise, CoD was applied on synthetic data using $N_{\mathrm{p}} = 100$ and $N_{\mathrm{d}} = 9 \cdot L$ (Supplementary Fig. S2).

2.3 Comparison with alternative methods

We compared our CoD pipeline with the 'all-relevant' Boruta method (Kursa and Rudnicki, 2011). In brief, Boruta constructs a random forest based on a selected subset of significant features—together with their shadow variables—which are then evaluated using a statistical t-test.

CoD is also compared with 'group activity'-based methods, including GSVA (Hanzelmann *et al.*, 2013), SAM-GS (Dinu *et al.*, 2007), PAC-all and PAC-best (Lee *et al.*, 2008), which aim to rank gene lists by their ability to discriminate between the two sample classes. To apply these methods, for each of the cell types we generated a separate gene list consisting of all gene markers that are commonly used to isolate that particular cell type (Supplementary Table S1). All compared methods are detailed in Supplementary Information Section 4.

2.4 Real data analysis

To test CoD we applied it on two different datasets. The first dataset consisted of blood-expression profiles of mice exposed to high or low levels of ionizing radiation at both early and late time points (at least 20 individuals in each of the sample classes; Zheng *et al.*, 2014). The second was a dataset from 25 mice with breast tumors and 25 control mice. Five tissues were profiled from each individual mouse (Schoenherr *et al.*, 2011). The CoD pipeline was applied on each of these tissues separately, using $N_p = 10\ 000$ (similar results were obtained with a larger number of permutations; data not shown) and $N_d = 9 \cdot L$. For each dataset we normalized the ($\log_2$-transformed) value of each gene in each sample relatively to its average value in all samples of the same dataset (and the same tissue).

3 Results

3.1 CoD accurately assesses alterations in the quantities of immune-cell types

To test the performance of the CoD algorithm we conducted a simulation study using synthetic 'datasets', each consisting of 100 collections that carry 50 samples in each of two labeled classes. For each collection we randomly selected particular groups of closely related cell types and then altered the quantity of cells within each of these groups between the two sample classes. The amount of alteration is referred to as the 'effect size' (see Section 2). We tested a total of 315 synthetic datasets that were generated using particular combinations of altered cell-type group numbers (2, 4, 6, 8, 10, 14 or 20), effect-size values (0–0.9) and noise factors ($\gamma_c = 10^{-6}$ to 10^{-2}; $\gamma_g = 10^{-6}$ to 10^{-2}). We observed that intermediate values of these parameters ($\gamma_c = \gamma_g = 10^{-4}$ and $s = 0.01$) coincide with the complexity that arises naturally in real biological datasets (Supplementary Fig. S1). The performance is evaluated based on the area under a standard 'ROC' curve, denoted 'cell-type accuracy': the higher the cell-type accuracy, the better the prediction of alteration in cell types between the two sample classes (see Section 2). Notably, the synthetic data analysis confirmed the utility of shadow features (Supplementary Fig. S2) and the selection of our particular gene-expression normalization scheme (Supplementary Fig. S3).

We first tested the algorithm's performance across different numbers of altered cell-type groups and across varying levels of noise and effect sizes. As shown in Figure 2a, cell-type accuracy is reduced with increasing numbers of cell-type groups, consistently with the increasing complexity of the problem due to larger numbers of altered cell types. The accuracy is reduced with higher noise and increased with increasing effect sizes, as expected (Fig. 2a). Qualitatively similar results were obtained when other noise factor values were used (e.g. Supplementary Fig. S4) and when the cell quantities of rare or abundant cell types were altered (Supplementary Information Section 1 and Supplementary Fig. S5). Notably, the accuracy decreases slightly with very high effect sizes (Fig. 2a, left panel). These results suggest, as might be expected, that the CoD algorithm has limited applicability for exceptionally high or low effect sizes: whereas extremely low effect sizes are challenging due to their weak signals, the existence of very high effect sizes may blur the signal, probably because of interactions between high effects of different cell types.

We hypothesized that the information about the performance of a classifier in classifying new samples (that is, the 'generalization error'; see Section 2) can be utilized to enhance the identification of true alterations in cell-type quantities. We found, as expected, that lower generalization-error values coincided with an increase in the information used to classify the samples (here, a higher effect size or a larger number of altered immune-cell-type groups; Supplementary Fig. S6a). Comparison of generalization error and cell-type accuracy across synthetic datasets indicated that high cell-type accuracy values are less likely to be attained in cases of high generalization error ($r^2 = 0.64$; Supplementary Fig. S6b, left panel). For example, cell-type accuracy values that were attained using datasets with a generalization error ≤ 0.2 obtained an average cell-type accuracy of 0.776, whereas the average cell-type accuracy with a generalization error > 0.2 was only 0.538. Taken together, by restricting our attention to datasets of low generalization error we can attain an improved cell-type accuracy.

We then compared the utility of the CoD algorithm with four previous group activity-based approaches: SAM-GS (Dinu *et al.*, 2007), GSVA (Hanzelmann *et al.*, 2013), PAC-all and PAC-best (Lee *et al.*, 2008). In our implementation, GSVA and PAC provide alternatives to utilization of the DCQ algorithm; SAM-GS, in contrast, is inherently different from the entire pipeline of CoD (see Supplementary Information Section 4). The cell-type accuracy attained with the CoD was significantly higher than that of the compared methods over a wide range of synthetic data parameters (Fig. 2b and Supplementary Fig. S7a–c). For example, for effect size = 0.01 and two cell groups, t-test $P < 10^{-14}$ in all cases (Supplementary Fig. S7d). Similar results are obtained when unequal sizes of the positive-effect and negative-effect cell-type groups are

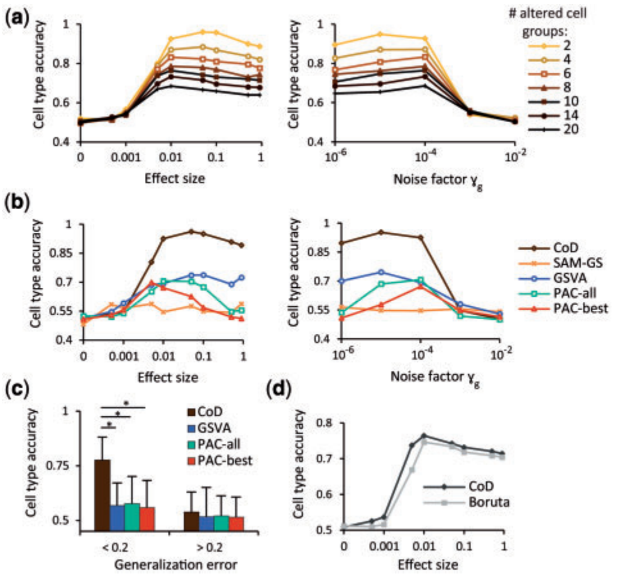

Fig. 2. Performance analysis in synthetic datasets. (**a**) Performance of CoD. Shown is the cell-type accuracy attained by CoD (*y*-axis) across datasets of different effect sizes (left) or different noise factors (right; *x*-axis) and different numbers of altered cell groups (color coded). (**b**) Performance of group activity-based methods. Shown is the cell-type accuracy (*y*-axis) of different methods (color coded) across different effect sizes (left) or different noise factors (right; *x*-axis), when the number of altered cell groups is 2. (**c**) Cell-type accuracy (*y*-axis) across various group activity-based methods (color-coded as in **b**) using synthetic data collections whose generalization errors are lower or higher than 0.2 (*x*-axis). The results demonstrate the substantially better performance of CoD than that of existing methods in the relevant case of generalization error <0.2 (**t*-test $P << 10^{-150}$). (**d**) Comparison with a state-of-the-art 'all-relevant' feature selection method. Shown is the cell-type accuracy metric (*y*-axis) over effect sizes (*x*-axis) for the CoD (black) and an alternative 'all-relevant' feature selection method ('Boruta'; gray) using synthetic data of 10 altered cell-type groups. The results suggest the superiority of the CoD over existing algorithms. Unless stated otherwise, effect size = 0.01 and $\gamma_g = 10^{-4}$

used (Supplementary Fig. S7e and Supplementary Information Section 2). Notably, compared with the alternative methods, CoD attained better correlation between the generalization error and the cell-type accuracy across datasets (Supplementary Fig. S6b), although its generalization error is not necessarily the lowest (Supplementary Figs S6c and S8a). In agreement, Figure 2c indicates the low accuracy of all methods in the case of high generalization error, and the superiority of CoD over existing methods in the relevant range of generalization errors <0.2 (*t*-test $P < 10^{-150}$ for CoD compared with each of the methods).

We next aimed to characterize CoD as a feature selection approach, with the goal of identifying all the features of relevance to a particular disease. For this purpose we compared the performance of CoD with that of the Boruta algorithm (Kursa and Rudnicki, 2011), recently reported to be the best-performing method for selection of 'all relevant features' (Kursa, 2014; Witold *et al.*, 2015). In particular, in this comparison we used stages 1 and 2 of CoD and replaced only stage 3 with the Boruta algorithm (Supplementary Information Section 4). As shown in Figure 2d, both CoD and Boruta appear to perform best with high effect sizes, but Boruta typically attains slightly lower accuracy scores (especially when the number of altered cell types is large; Supplementary Fig. S9a). For example, using 10 cell types and effect size = 0.005, $P < 10^{-11}$ and $P < 10^{-10}$ (*t*-test) when using all datasets or only datasets that attained generalization errors of <0.2, respectively (Supplementary Fig. S9b and c). All in all we conclude that CoD performs well on a broad range of data parameters, outperforming existing state-of-the-art methodologies.

Notably, the classification challenge may become substantially more difficult in the case of variable expressivity, i.e. when a certain disease gives rise to several different physiological states (or clinical conditions). To establish whether the CoD algorithm is robust in this biomedical scenario, we generated synthetic 'variable expressivity datasets'. In these datasets, class B consists of two different groups of individuals, each of whom is distinct from all those in class A (Supplementary Information Section 3). As shown in Supplementary Figure S10a and b, CoD is only slightly reduced when applied on datasets carrying variable expressivity, probably because it considers combinations of features. As expected, a naive approach that relies on assessment of discriminative features by testing each cell type independently is likely to fail in this case (e.g. utilizing a two-sample *t*-test or its variants—Golub *et al.*, 1999, Hedenfalk *et al.*, 2001; Supplementary Fig. S10a, right panel).

3.2 CoD successfully identifies immune-cell alterations during radiation treatment

To demonstrate the ability of the CoD pipeline to assess changes in cell quantities, we examined its performance in the case of ionizing radiation treatment in murine blood, focusing on possible changes in immune-cell quantities between high (8–10.5 Gy) and low (1–2 Gy) whole-body ionizing radiation. Since immune responses are dynamic and may change over time, we tested differences in cell-type quantities between low and high radiation over two time periods: 6–24 and 120–168 h post-radiation treatment (denoted 'early' and 'late', respectively). The input transcription profiles were obtained from a published dataset (Zheng *et al.*, 2014) that includes samples from C57BL/6 mouse strains, with at least 20 individuals in each sample class and variable expressivity at early time points (Section 2 and Supplementary Fig. S10c).

Our observations pointed to granulocytes as the key cell type revealing differential quantities between high and low radiation at early time points (five of seven top-ranked cell types, *q*-value <0.05, generalization error = 0.27; Fig. 3a) and to macrophages as the main cell type exhibiting changes in cell quantities at late time points (eight of 11 top-ranked cell types, *q*-value <0.005, generalization error = 0.024; Fig. 3b and Supplementary Table S2). Several lines of evidence support these predictions, although the levels of these cell types have not been previously reported for the blood or for these radiation levels. For example, accumulation of neutrophil granulocytes was monitored in the spleen and thymus of C57BL/6 mice over 6–24 h following 4 Gy ionizing radiation (Lorimore *et al.*, 2001; Uchimura *et al.*, 2000), and alterations in the quantities of macrophages in the spleen and bone marrow were observed mostly at late stages after 4 Gy radiation (Coates *et al.*, 2008). Altogether, five of seven cell types at early time points and 11 of 29 cell types at late time points (all with *q*-values <0.05) are in agreement with the literature (see details in Supplementary Table S2). The radiation example, therefore, illustrates the ability of the CoD pipeline to correctly identify differential cell quantities between two sample classes.

The ionizing radiation dataset provides a clear example of the advantages and limitations of CoD in comparison with a standard gene-level classification analysis. We obtained gene-based classification results by using the random forest algorithm applied on the gene-expression dataset at late periods (Supplementary Fig. S8b). We further annotated each gene with its known role in immune

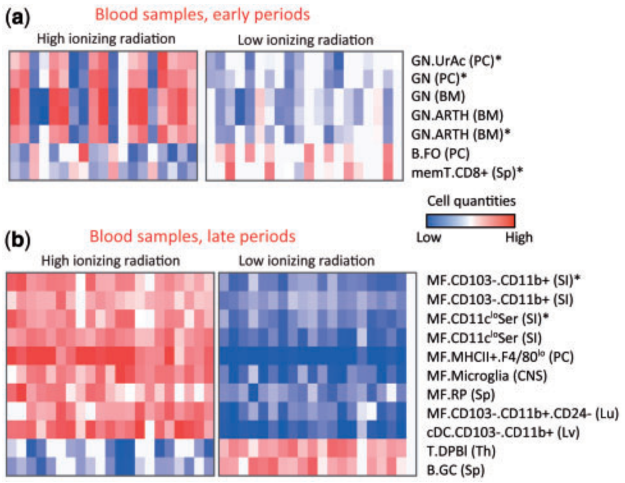

Fig. 3. Immune-cell responses to low (1–2 Gy) and high (8–10.5 Gy) whole-body ionizing radiation. Differential cell quantities are shown at (**a**) early (6–24 h) and (**b**) late (120–168 h) periods after radiation treatment. Presented are differential cell quantities, high (red) and low (blue) across individual samples (columns: high radiation, left; low radiation, right) and immune-cell types (rows) that attained q-values lower than 0.05 (**a**) or 0.005 (b). BM, bone marrow; SI, small intestine; PC, peritoneal cavity; Sp, spleen; Lu, lung; Lv, liver; Th, thymus; *Stimulated cell type

physiology based on the Ingenuity Knowledge Base database (Ingenuity, Redwood City, CA). Out of 30 top-scored genes, only five have a known role in immune physiology, and only one of these relates to regulation of macrophage activation (the *Hamp* gene, ranked 15th; Li *et al.*, 2012). When we examined the generalization error, we found that the gene-level approach outperformed the cell-type-level approach (Supplementary Fig. S8b). This result suggests that compared with gene-level classification, CoD attains better immunological interpretability but with a higher generalization error.

3.3 CoD builds an immune-cell quantity model for breast cancer

We next investigated the ability of CoD to capture the variety of immune-cell types in the surrounding environment of tumors and in distant organs. To that end we examined data from *ErbB2*-driven transgenic mice, a spontaneous estrogen-independent mammary cancer model (the 'MMTV-Neu' model; Muller *et al.*, 1988). Overexpression of the human *ERBB2* (*HER2*) oncogene is associated with about 20% of human breast cancers (Mitri *et al.*, 2012). Accordingly, induced expression of *ErbB2* in the mammary epithelial cells of MMTV-Neu mice leads to estrogen-independent mammary tumorigenesis with 100% penetrance and lung metastases at ~15 weeks post-pregnancy (Muller *et al.*, 1988). The dataset (downloaded from Schoenherr *et al.*, 2011) consists of gene-expression values from the primary breast tissue, the peripheral circulation (blood), and three additional distant organs (liver, spleen and thymus), all isolated from tumor-bearing and from healthy control mice (25 individuals in each sample class; see Section 2). The CoD algorithm was applied (separately) on each of the five tissues and attained low classification errors in breast (0.06), spleen (0.06), blood (0.08) and liver (0.06) (Fig. 4a and Supplementary Table S2). In view of the low error achieved by our algorithm for four of the tissues, we next interpreted the cell types that emerged from the CoD algorithm in light of previous findings.

Based on CoD predictions, the peripheral circulation and spleen contained increased subpopulations of Cd11b + Gr1 + (granulocytes; q-value<0.05, <0.005, respectively; Fig. 4b and c), and the tumor site (breast) primarily contained induced dendritic cells, macrophages, $\gamma\delta$T and pre-T cells, as well as repressed CD8+T cells (q-value < 0.0005; Fig. 4d). These predictions are indeed consistent with the fact that a specialized population of Cd11b + Gr1 + granulocytes, called 'myeloid-derived suppressor cells' (MDSCs), are known to be expanded in the bone marrow, blood and spleen of tumor-bearing individuals (Gabrilovich and Nagaraj, 2009; Morales *et al.*, 2010). For example, the abundance of MDSCs is expanded from 2.5% of the total splenocyte population in healthy individuals to >40% in cancer patients (Youn *et al.*, 2008). Following this expansion, MDSCs migrate to the tumor site and differentiate into dendritic cells and macrophages (Franklin *et al.*, 2014; Gabrilovich and Nagaraj, 2009; Norian *et al.*, 2009), consistently with the predictions of CoD. Recent reports document two broad categories of solid (e.g. mammary) tumors, the first category containing infiltrated T cells and the second category lacking infiltrated T cells in the tumor site (Gajewski et al., 2013). The *ErbB2*-based mouse model was indeed reported to have a non-CD8+T-cell-infiltrating tumor (Norian *et al.*, 2009) (that is, a lack of infiltrated CD8+ T cells in the tumor site), consistently with the predictions of CoD. Notably, although tumor studies have focused mainly on CD8+ T cells, tumor infiltration of $\gamma\delta$T regulatory cells and CD20+ B cells has also been documented (Nelson, 2010; Wesch *et al.*, 2014), as predicted by CoD in the primary breast tissue (Fig. 4d). The predicted decrease in B cells in the spleen (Fig. 4c) has not been previously reported, suggesting a possible new cell-type hypothesis for further research. Details about how the prediction of each cell type is supported by previous experiments are given in Supplementary Table S2.

The analysis of the liver tissue is particularly interesting. Previous reports have described the formation of a hospitable microenvironment in secondary organs ('pre-metastatic niche') that precedes the initiation of metastases and is an essential prerequisite for the proliferation of tumor cells in a secondary site (Psaila and Lyden, 2009). Whereas breast cancer metastases in the MMTV-Neu model arise primarily in the lung and lymph nodes (Fantozzi and Christofori, 2006), it was not known whether such a pre-metastatic niche also occurs in the liver. By applying CoD we found that all known components of a pre-metastatic niche were indeed predicted to be induced in the liver (Fig. 4e). This includes a predicted increase in hematopoietic progenitor cells (CLP and CDP; q-value < 0.03, <0.015, respectively) that is in agreement with previous evidence (Psaila and Lyden, 2009); an inferred elevation of monocytes, macrophages and progenitor B cells (q-value <0.015), consistently with previous reports (Gil-Bernabe *et al.*, 2012); and a predicted decrease in stem cells (q-value < 0.015), which was previously reported in a pre-metastatic niche (Freitas *et al.*, 2003) (see details in Supplementary Table S2). The concordance of predictions and previous literature suggests that a pre-metastatic niche has already evolved in the liver.

The pre-metastatic niche hypothesis is consistent with reported findings of the secretion of a variety of factors from the pre-metastatic niche, including the inflammatory S100a8 and S100a9 chemokines, serum amyloid Saa3, Vegfa and Mmp9 (Psaila and Lyden, 2009). Our liver dataset indeed confirms the alterations in levels in the majority of these factors (six of eight factors with $P < 10^{-3}$; t-test, Bonferroni-corrected; Fig. 4f and Supplementary Table S3). In humans, 50% of breast cancer patients show metastases in bone, lung and liver, unlike in MMTV-Neu mice, which mainly form metastases in the lung and lymph nodes but rarely in bone or liver. Taken together, our results point to the presence of a pre-metastatic niche in the liver of MMTV-Neu mice, although such niches rarely develop into actual metastases (Fig. 4g).

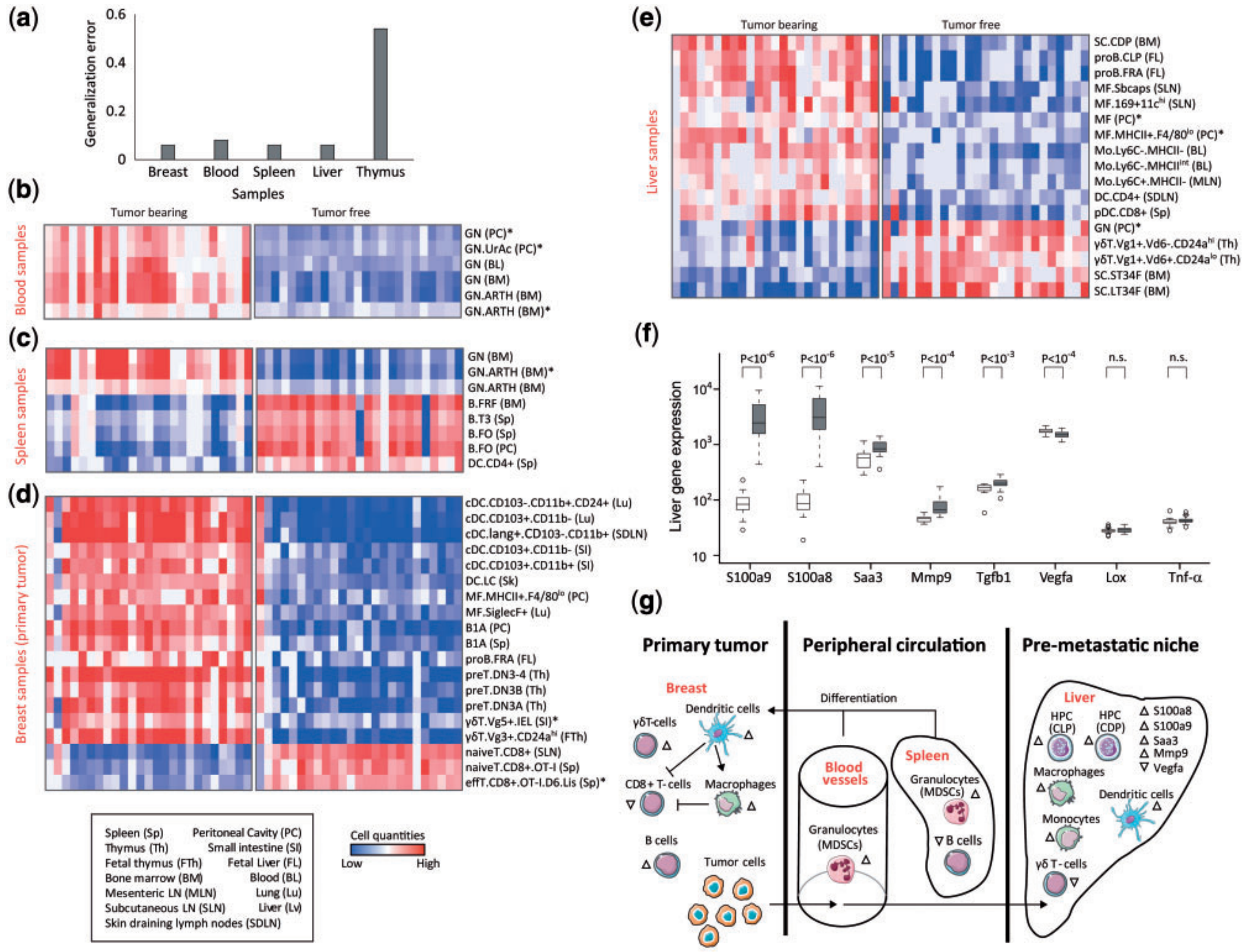

Fig. 4. An immune-cell subpopulation model for mammary cancer in *ErbB2*-driven transgenic mice. (**a**) Generalization error. Shown is the generalization error attained by CoD (*y*-axis) across datasets from primary breast tissue and distant organs (*x*-axis) isolated from tumor-bearing and healthy control mice. Four tissues attained low errors and are therefore subjected to further analysis. (**b–e**) Differential cell quantities, high (red) and low (blue), are shown in blood (**b**), spleen (**c**), breast (**d**) and liver (**e**) tissues across individual samples (columns: tumor-bearing, left; tumor-free, right) and immune-cell types (rows). Presented are all cell types that attained *q*-values < 0.005 (c), < 0.0005 (d) or < 0.05 (b, e). *Stimulated cell type. (**f**) Pre-metastatic factors in the liver. Shown are liver gene-expression box-plots (*y*-axis, log-scaled) of known factors typically secreted from a pre-metastatic niche in both human and mouse (*x*-axis) for all samples isolated from tumor-bearing (gray) and healthy control (white) mice. Observed significant alterations in secreted pre-metastatic factors are consistent with the pre-metastatic composition of immune-cell types in (**e**). (**g**) A model for mammary cancer in primary breast tissue (left), circulation and spleen (middle) and the suggested pre-metastatic niche in the liver (right). Shown are predicted changes in cell quantities and pre-metastatic factors that are in agreement with literature reports (increasing/decreasing changes are marked by upward/downward arrows)

4 Discussion

In this article, we present CoD, a novel framework for the identification of immune-cell subsets whose quantities discriminate between different types of disease. Based on a computation decomposition algorithm, CoD generates cell-type-based features and exploits the random forest classification method to extract the cell types that discriminate between disease states (Fig. 1). A key advantage of CoD is its ability, unlike existing group activity-based approaches, to integrate prior quantitative information about each of the relevant immune-cell types. In simulations CoD shows high accuracy, outperforming existing state-of-the-art group activity-based classification methods (Fig. 2a and b) as well as 'all-relevant' feature selection methods (Fig. 2d), predicting disease-relevant cell subsets more accurately. This relatively high accuracy can be improved even further by focusing only on those cases in which the generalization error is low (Fig. 2c and Supplementary Fig. S9b). Thus, although all classification methods (ours as well as others) are under-constrained and prone to overfitting (∼200 cell types and ≤50 individuals in each sample class), the CoD algorithm seems more accurate. In the tumor-bearing dataset, for example, although our analysis was restricted to 50 individuals, a large number of candidates were identified as relevant cell types (Fig. 4). Given larger datasets, the CoD algorithm can potentially capture additional relevant cell types.

Our results in mice point to the utility of the CoD pipeline for real data analysis. First, whereas standard gene-level analyses cannot attain interpretable immunophysiological results, CoD successfully identifies many of the previously reported alterations in cell types, albeit with a higher generalization error (Supplementary Fig. S8 and Supplementary Table S2). Second, the identification of 'all-relevant' features, including redundant features (CoD stage 3), is crucial for identifying the entire set of disease-related cell types. In agreement, many of the identified cell types are tightly correlated with each other, while others may be quite distinct (Supplementary Fig. S11). This study presents two alternative all-relevant methods (CoD stage

3 and Boruta—Kursa and Rudnicki, 2011) that obtain qualitatively similar results in both real and synthetic datasets (CoD is only slightly better; Fig. 2d and Supplementary Table S2), thus providing additional evidence for the reliability of a cell-level classification approach. Finally, some of the reconstructed cell types lead to novel hypotheses. For example, predictions with respect to liver metastases matched prior knowledge about the pre-metastatic microenvironment in distal organs (Fig. 4e). This prediction was further confirmed by the observed increase in a variety of pre-metastatic niche-specific factors, such as Mmp9, S100a8, S100a9 and Saa3 (Fig. 4f).

Clearly, the next step is to adjust and apply the CoD algorithm to humans in order to attain a comprehensive view of immune-cell physiology in health and disease. While our method has been applied only in mice, we believe that it is potentially applicable to humans. An important first step will be the construction of a list of cell-surface markers for human compendium datasets (e.g. Abbas *et al.*, 2005; Newman *et al.*, 2015), required as input for the DCQ algorithm. Second, alternative decomposition methods (tailored for human data) should be tested within the CoD pipeline (Newman *et al.*, 2015; Qiao *et al.*, 2012; Zhong *et al.*, 2013). Third, human single-cell RNA-seq data may open the way to an extremely large collection of cell-type signatures within the prior compendium, providing a unique opportunity to uncover changes in the quantity of cell types that have not yet been characterized or cannot be isolated by a FACS sorter.

CoD can be viewed as a pipeline for improving medical decisions. For example, CoD successfully recognizes the lack of T cells in the mammary tumor microenvironment (Fig. 4d and g), thus allowing the identification of tailored therapeutic targets and clinical interventions that fit such a non-T-cell-infiltrated tumor. Because the immune-cell state commonly correlates with patient prognosis (e.g. Ge *et al.*, 2013; Piersma *et al.*, 2007), CoD can be viewed as a promising tool for the clinical improvement of diagnostics and therapeutics.

Acknowledgement

We thank Shirley Smith for scientific editing.

Funding

This work was supported by the Israeli Centers of Research Excellence (I-CORE): Center No. 41/11 (A.F., Y.S.) and by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University (A.F., Y.S.). Research in the I.G-V. lab was supported by the European Research Council (637885), the Israeli Science Foundation (1643/13) and the Broad-ISF program (1168/14). I.G-V. is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University and an Alon Fellow.

Conflict of Interest: none declared.

References

Abbas,A.R. *et al.* (2005) Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun.*, **6**, 319–331.

Altboum,Z. *et al.* (2014) Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.*, **10**, 720.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Coates,P.J. *et al.* (2008) Ongoing activation of p53 pathway responses is a long-term consequence of radiation exposure in vivo and associates with altered macrophage activities. *J. Pathol.*, **214**, 610–616.

Dinu,I. *et al.* (2007) Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8**, 242.

Fantozzi,A. and Christofori,G. (2006) Mouse models of breast cancer metastasis. *Breast Cancer Res.*, **8**, 212.

Franklin,R.A. *et al.* (2014) The cellular and molecular origin of tumor-associated macrophages. *Science*, **344**, 921–925.

Freitas,I. *et al.* (2003) Stem cell recruitment and liver de-differentiation in MMTV-neu (ErbB-2) transgenic mice. *Anticancer Res.*, **23**, 3783–3794.

Gabrilovich,D.I. and Nagaraj,S. (2009) Myeloid-derived suppressor cells as regulators of the immune system. *Nat. Rev. Immunol.*, **9**, 162–174.

Gajewski,T.F. *et al.* (2013) Innate and adaptive immune cells in the tumor microenvironment. *Nat. Immunol.*, **14**, 1014–1022.

Ge,Y. *et al.* (2013) Circulating CD31+ leukocyte frequency is associated with cardiovascular risk factors. *Atherosclerosis*, **229**, 228–233.

Gil-Bernabe,A.M. *et al.* (2012) Recruitment of monocytes/macrophages by tissue factor-mediated coagulation is essential for metastatic cell survival and premetastatic niche establishment in mice. *Blood*, **119**, 3164–3175.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hanzelmann,S. *et al.* (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.

Hedenfalk,I. *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *N Engl J Medi*, **344**, 539–548.

Heng,T.S. and Painter,M.W. (2008) The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.*, **9**, 1091–1094.

Jojic,V. *et al.* (2013) Identification of transcriptional regulators in the mouse immune system. *Nat. Immunol.*, **14**, 633–643.

Kursa,M.B. (2014) Robustness of random forest-based gene selection methods. *BMC Bioinformatics*, **15**, 8.

Kursa,M.B. and Rudnicki,W.R. (2011) Feature selection with the Boruta package. *J. Stat. Softw.*, **36**, 11.

Lee,E. *et al.* (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, **4**, e1000217.

Li,J.J. *et al.* (2012) Hepcidin destabilizes atherosclerotic plaque via overactivating macrophages after erythrophagocytosis. *Arterioscler. Thromb. Vasc. Biol.*, **32**, 1158–1166.

Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

Lorimore,S.A. *et al.* (2001) Inflammatory-type responses after exposure to ionizing radiation in vivo: a mechanism for radiation-induced bystander effects? *Oncogene*, **20**, 7085–7095.

Mitri,Z. *et al.* (2012) The HER2 receptor in breast cancer: pathophysiology, clinical use, and new advances in therapy. *Chemother. Res. Pract.*, **2012**, 743193.

Morales,J.K. *et al.* (2010) GM-CSF is one of the main breast tumor-derived soluble factors involved in the differentiation of CD11b-Gr1- bone marrow progenitor cells into myeloid-derived suppressor cells. *Breast Cancer Res. Treat.*, **123**, 39–49.

Muller,W.J. *et al.* (1988) Single-step induction of mammary adenocarcinoma in transgenic mice bearing the activated c-neu oncogene. *Cell*, **54**, 105–115.

Nelson,B.H. (2010) CD20+ B cells: the other tumor-infiltrating lymphocytes. *J. Immunol.*, **185**, 4977–4982.

Newman,A.M. *et al.* (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, **12**, 453–457.

Nilsson,R. *et al.* (2007) Consistent feature selection for pattern recognition in polynomical time. *J. Mach. Learn. Res.*, **8**, 589–612.

Norian,L.A. *et al.* (2009) Tumor-infiltrating regulatory dendritic cells inhibit CD8+ T cell function via L-arginine metabolism. *Cancer Res.*, **69**, 3086–3094.

Piersma,S.J. *et al.* (2007) High number of intraepithelial CD8+ tumor-infiltrating lymphocytes is associated with the absence of lymph node metastases in patients with large early-stage cervical cancer. *Cancer Res.*, **67**, 354–361.

Psaila,B. and Lyden,D. (2009) The metastatic niche: adapting the foreign soil. *Nat. Rev. Cancer*, **9**, 285–293.

Qiao,W. *et al.* (2012) PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol.*, **8**, e1002838.

Schoenherr,R.M. *et al.* (2011) Proteome and transcriptome profiles of a Her2/Neu-driven mouse model of breast cancer. *Proteomics Clin. Appl.*, **5**, 179–188.

Uchimura,E. *et al*. (2000) Transient infiltration of neutrophils into the thymus in association with apoptosis induced by whole-body X-irradiation. *J. Leukoc. Biol.*, **67**, 780–784.

Wesch,D. *et al*. (2014) Human gamma delta T regulatory cells in cancer: fact or fiction? *Front. Immunol.*, **5**, 598.

Witold,R. *et al*. (2015) All relevant feature selection methods and applications. In: Stanczyk,U. and Jain,L.C. (eds), *Feature Selection for Data and Pattern Recognition*. Springer, Berlin.

Youn,J.I. *et al*. (2008) Subsets of myeloid-derived suppressor cells in tumor-bearing mice. *J. Immunol.*, **181**, 5791–5802.

Zheng,L. *et al*. (2014) Biological pathway selection through Bayesian integrative modeling. *Stat. Appl. Genet. Mol. Biol.*, **13**, 435–457.

Zhong,Y. *et al*. (2013) Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, **14**, 89.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, **67**, 301–320.