

ViLoN—a multi-layer network approach to data integration demonstrated for patient stratification

Maciej M. Kańduła^{1,2}, Alexander D. Aldoshin¹, Swati Singh^{1,3}, Eric D. Kolaczyk^{4,*} and David P. Kreil^{1,*}

¹Institute of Molecular Biotechnology, Boku University Vienna, Austria, ²Janssen Pharmaceutica NV, Beerse, Belgium, ³Department of Biological Sciences and Bioengineering, Indian Institute of Technology Kanpur, Kanpur, India and ⁴Department of Mathematics and Statistics, McGill University, Canada

Received November 30, 2021; Revised October 11, 2022; Editorial Decision October 11, 2022; Accepted November 02, 2022

ABSTRACT

With more and more data being collected, modern network representations exploit the complementary nature of different data sources as well as similarities across patients. We here introduce the Variation of information fused Layers of Networks algorithm (ViLoN), a novel network-based approach for the integration of multiple molecular profiles. As a key innovation, it directly incorporates prior functional knowledge (KEGG, GO). In the constructed network of patients, patients are represented by networks of pathways, comprising genes that are linked by common functions and joint regulation in the disease. Patient stratification remains a key challenge both in the clinic and for research on disease mechanisms and treatments. We thus validated ViLoN for patient stratification on multiple data type combinations (gene expression, methylation, copy number), showing substantial improvements and consistently competitive performance for all. Notably, the incorporation of prior functional knowledge was critical for good results in the smaller cohorts (rectum adenocarcinoma: 90, esophageal carcinoma: 180), where alternative methods failed.

INTRODUCTION

With the general increase in human life expectancy, ever larger parts of the population need to be treated for complex diseases. A typical example is the treatment of cancers, where selecting an effective therapy for individual patients can be particularly challenging because of the heterogeneity of the disease. Disease progression and treatment response can vary widely across patients. In order to determine if a patient should receive a specific treatment, clinicians delineate subgroups of patients likely to react simi-

larly ('patient stratification'). Typical criteria include clinical records, such as the age at diagnosis, sex and comorbidities. It is by now widely recognized that the underlying mechanisms can vary widely across patients. Even for patients with the same specific clinical diagnosis, for instance, we find many subtypes of breast cancer (1) or adult acute myeloid leukaemia (AML) (2). Combining clinical records and histologic tests alone often cannot reliably identify the biological processes underlying a particular tumour type (3). Increasingly, therefore, molecular markers are now incorporated to improve the prediction of therapy response and prognosis (3–7). Common molecular markers include changes in gene activity, such as identifying characteristic gene sets or signatures (8,9), and genomic sequence variants, such as copy number changes from deletions and amplifications of certain genomic regions, like the amplification of the MYCN gene in Neuroblastoma patients (10), as well as smaller changes, such as single nucleotide polymorphisms (11).

Drawing clinically relevant insight has soon become the bottleneck that still remains rate-limiting for exploitation of the massive molecular profiles now collected from biomedical assays. A lot of hope is being placed in analysis approaches which combine measurements from different sources, be that horizontally, across patients (12,13) or vertically, across assay types (13–16). Vertical integration combines different assay types that may capture complementary aspects of information and, investigated together, could help shed new light on complex relationships of interest, such as the relationship between genotype and eventual consequences for tumour progression. Specifically, one would expect the impacts of gains and losses of gene copies to affect the expressions of *cis* and *trans* genes of relevance to cancer in a non-trivial way, and complementary measurements may facilitate accurate survival time prediction.

Employing both gene expression and copy number information was reported to improve clustering for subtype analysis using a probabilistic model of joint latent vari-

*To whom correspondence should be addressed. Tel: +43 1 47654 79009; Email: nar21@kreil.org
Correspondence may also be addressed to Eric D. Kolaczyk. Tel: +1 514 398 3805; Email: eric.kolaczyk@mcgill.ca

ables (17,18). Introducing dimensionality reduction by low-rank approximation, LRAcluster was reported to further improve the stratification of cancer patients (19), establishing a probabilistic model of different data types conditional on shared latent factors. Alternative approaches include network-based algorithms, which have seen a variety of successful applications, such as the identification of dysregulated pathways (20,21) or an optimization of biotechnological processes (22). In the Similarity Network Fusion (SNF) algorithm, network representations exploit not just the complementary nature of data sources but also similarities across patients (23), i.e. combining both vertical and horizontal data integration. First, patients are grouped into a network, with every node representing a patient. Distances in the network reflect the similarities of the patients with respect to a particular data type. These networks for the different data types are then merged through cross-diffusion by message passing across the networks. This effectively combines evidence across complementary data types and patients, with the hope of identifying relevant patterns of the underlying biology of the disease.

We here introduce ViLoN (Variation of Information fused Layers Of Networks), a novel network-based approach for the integration of multiple molecular data types. By design, our algorithm extracts information at multiple levels, starting with differential effect analysis of the molecular data. Going beyond earlier work, our approach then directly incorporates *functional knowledge* from expert curated sources, including the GeneOntology (GO) (24,25), which lists genes involved in known biological processes, as well as pathway annotations from the KEGG (26) database. Earlier work has demonstrated the value of incorporating functional knowledge directly into other algorithms (27,28). Finally, similarities across patients are exploited to build a multiple-layer network (29) that captures all the structured information discovered. This consolidated resource can then be exploited at various levels, be that the exploration of functional modules identified across patients, or the identification of clinically relevant groupings of patients in the network structure.

In the context of predicting patient survival, commonly a group of similar patients is identified that can be associated with an average risk change. Thus a grouping of patients ('patient stratification') is clinically relevant if one can find a sufficiently reduced or increased risk for a reasonable number of patients with significance. Notably, significance alone is insufficient for clinical relevance (30). Although there is no universally agreed threshold for the clinical relevance of hazard ratios, for death by cancer a risk change of 14% is typically considered to be small, changes of 47–90% can be considered moderate size effects, while risk changes of 90% or more are considered large (31). Large effects can often be identified more easily for more specific, smaller subsets of patients. For a meaningful comparison of different predictions we therefore need to consider the number of patients as well as the size of the risk change. To this end we introduce an *effective number of affected patients* and demonstrate its use as a balanced metric.

We explore the value of our approach on a range of cancers and data types using several complementary assays.

MATERIALS AND METHODS

Data

Neuroblastoma. Raw microarray gene expression and CGH data as well as RNA-Seq expression profiles were provided by the CAMDA organizers together with a sample description file for mapping patient samples between the different data types. Data were downloaded from the 'Challenges' section of the CAMDA website (www.camda.info) directly. The dataset used in this work consists of the 145 profiles matched between all three data types, i.e. all three data types are available for the same 145 patients (32–36). The clinical information provided includes survival times of the patients, the majority being right-censored (107 out of 145), i.e. a patient either dropped out of the study before it ended or at the end of the study was still alive. Additionally, a clinical classification into high/low risk patients is provided.

The microarray data are normalized with the quantile method (37). Copy number data are scale-normalized pre-analysis. For RNA-Seq data, in line with recommendations, we apply TMM normalization (38) and Voom transformation (39) on read counts.

We then employ standard linear models to compute empirical Bayes regularized *t*-statistics for each data type. Specifically, we employed the R package *limma* (40,41) and used the `lmFit()`, `contrasts.fit()` and `eBayes()` functions for data processing. From the obtained log-odds ratios (*B* values) for a differential signal between specific tumour and control samples (the remaining cohort of tumour samples) we then calculate the posterior probability of a differential effect occurring for a gene. These are further used as inputs to develop the patient-specific pathway networks. Here, normal samples could have been used as controls. It is, however, non-trivial to obtain matched normals in representative numbers for specific cancer datasets, with specific molecular profiles having no corresponding normals. Usually, only a few patients will have matched normal data. Thus, pooled normals would need to be used for the whole cohort. As we are interested in direct differences between the patients' diseases we instead here focus on the cancer data only, comparing each patient to all the others. In our experiments this approach has proven to work well.

TCGA cancers. Data for the TCGA cancers were acquired from the corresponding projects using the TCGA2STAT R package (42). We always used the largest matched sample size available for a pair of molecular profiles for which the corresponding clinical information was also available.

Specifically, for the cervical cancer dataset (CESC-TCGA) we used 279 matched tumour samples with RNA-Seq (RNAseq2 pipeline data) and CNV-SNP (Affymetrix Genome-Wide Human SNP Array 6.0) assays available. For colon and rectal cancer (COADREAD-TCGA) we used the same molecular profile types, with 369 matched tumour samples.

Thyroid carcinoma (THCA-TCGA), esophageal carcinoma (ESCA-TCGA) and rectum adenocarcinoma (READ-TCGA) were analysed using RNA-Seq (RNAseq2 pipeline data) and methylation (Illumina Infinium HumanMethylation450 BeadChip) assays, with 496, 183 and

90 sample available, respectively. Sex chromosomes were removed from the methylation data, medians of the beta values per gene were calculated and finally beta values were converted to M -values.

All molecular data were mapped to genes and processed similarly to neuroblastoma data, with methylation (unavailable in the neuroblastoma dataset) being scale-normalized.

Gene set data. For constructing the patient-specific pathway networks we use the GO biological process collection and the open-source version of KEGG pathways provided in the Molecular Signatures Database (43), MSigDB v6.0.

Algorithm implementation and validation

Introducing posterior probabilities of a differential effect. In the original publication by Pham *et al.* (20) the authors use t -scores to assess the differential effect for a gene. We here introduce the use of posterior probabilities of a differential effect occurring for a gene. The posterior probability is a natural choice for the edge weight because edge weights of the bipartite graph are multiplied when summarising information over the GO layer. These products of edge weights can then be interpreted as joint probabilities. The posterior probabilities provide the patient-specific input to our analysis pipeline, complemented by the functional knowledge represented by KEGG and GO. For the weighted stochastic block model, we use logit-transformed sums of posterior probabilities as weights, which are Gaussian for each cluster in the block model (data not shown).

Variation of Information calculations. In order to compare the clusterings of pathways between patients and construct the patient similarity graph we use the `compare()` function, available in the *igraph* R package (44) to calculate the variational information (VI). This is a distance reflecting the similarity between two nodes in the network, i.e. how closely related two patients are to each other. The smaller the VI and shorter the distance, the stronger the similarity between two nodes.

In order to construct a *distance*-based *integrated* patient similarity graph we combine patient similarity graphs from multiple molecular profile types, by averaging the VI values. After the VI distances are integrated, we normalize them as suggested by Meila (45). Specifically, we multiply each VI value by $1/\log(N)$, where N is the number of data points in the specific cancer dataset, i.e. the cohort size or number of patients. This bounds the range of values between 0 and 1. In order to obtain a weight corresponding to this distance we further subtract the normalized VI from 1. This results in a *weight*-based integrated patient similarity graph, where the higher the weight between two nodes the stronger the similarity between the two patients.

Constructing patient-specific pathway networks. We built a custom version of the LPIA algorithm (20) that outputs the bi-partite KEGG pathway networks per patient, using as input the KEGG and GO databases, together with differential effect data: the posterior probabilities for differential expression per gene.

Clustering of pathway networks by weighted stochastic block model. We apply a weighted stochastic block model (SBM) to network clustering (46,47), which gave better results than spectral clustering (48,49) in this context (data not shown). The established implementation in the *blockmodels* R package was used (version 1.1.1).

When clustering the patient-specific pathway networks the Integrated Classification Likelihood (50) was used for selecting the optimal number of clusters for each patient and data type.

Clustering of patient networks by weighted stochastic block model. For stratification from the patient graph we employ a weighted stochastic block model, where a variation of information metric averaged across data types is used to compute node weights as detailed above.

Considering only clinically relevant strata per model, we focus on 2–9 clusters for patient similarity graphs. While the Integrated Classification Likelihood (50) would support even more, smaller patient groups, a larger number of clusters would reduce generalization performance considering the available cohort sizes in the low hundreds. We show the number of clusters for which the best pairwise N_{eff} score was obtained. Notably, this was always for considerably fewer than 9 clusters.

Robustness of patient clustering. We test the robustness of the patient similarity graph on the neuroblastoma dataset with a leave-one-out approach. First we use the whole cohort of patients, i.e. all the matched patient samples, and run our method. Specifically, we construct the full network, and group patients into three subgroups using SBM. Next, for all patients, one patient at a time is removed from the data used for creating the network. Networks (without one patient each) are then generated and patients are grouped into three subgroups. Each of these new groupings of patients is compared to the grouping from the full cohort of patients. We calculate the accuracy and normalized mutual information (NMI) of the compared groupings.

Normalized mutual information (NMI) (51) is calculated with the function `calNMI()` from the *SNFtool* R package.

Survival analysis. All the survival analyses are performed with the `coxphf()` function from the *coxphf* R package, version 1.12. This package facilitates performing Cox regression with Firth correction for right-censored data, which is reliable in contrast to normal Cox regression when many data points are censored (52).

If a grouping of more than two groups is considered, i.e. a survival analysis with $N > 2$, the analysis naturally yields results for multiple groups, e.g. multiple pairwise hazard ratios (HR). For a 2-group analysis, as the survival model is always calculated relative to a reference group, there is only 1 result. For example, in an $N = 4$ analysis yields three HRs relative to a reference group. Notably, not all groups need to yield statistically significant results. HRs calculated and presented in this work (i.e. not HRs referenced from the literature), when a result of multiple cluster analysis, are always shown for the best statistically significant group for the specific model as assessed by the product metric. For example, if we show results for a survival analysis of four

clusters ($N = 4$) the HRs presented are from the result with the maximum product metric amongst all the resulting significant pairwise results for this model.

Moreover, hazard ratios are always obtained for a specific subset of patients compared to a distinct reference group of patients providing a baseline. It then is the size of the smaller of the two groups that determines how meaningful an observed high hazard ratio actually is. For a fair comparison of different methods and their results, for each patient clustering we make the conservative choice of selecting the largest group as the reference baseline, avoiding an artificial inflation of hazard ratios.

Kaplan–Meier curves are created using the `survfit` and `ggsurvplot` functions available in the `survminer` R package. The automatically generated p -values are based on a log-rank test, and approximated for display in the plots.

Comparison to other tools. We employ the `SNFtool` R package and perform the analysis exactly as described in the documentation. Final output of the analysis is a patient classification per each data type as well as for the integrative approach. We use this classification into groups directly in the survival analysis.

We apply `LRAcluster` (19) version 1.0, downloaded directly from the authors' webpage (<http://bioinfo.au.tsinghua.edu.cn/member/jgu/lracluster/>), following the instructions and the publication to guide the analyses. The tool yields a grouping based on the integrated molecular data, which we use for survival analysis.

In order to compare our performance with results obtained by state-of-the-art methods dedicated to analysis of this particular neuroblastoma dataset we also look at the survival data provided by other CAMDA 2017/2018 authors (available both in publications and presentations).

RESULTS

A novel algorithm for the construction of patient similarity graphs

We here introduce ViLoN, a new algorithm relating patients based on matched molecular profiles, such as gene expression, copy number information, or others. Rather than considering patient similarity based on the molecular profiles alone (23), we combine these with knowledge of biological processes from the GeneOntology, GO, (24,25) and different classes of KEGG pathways (26) into a multi-level patient/pathway network (Figure 1). Together, this allows for the first time a systematic exploitation of patient relationship graphs incorporating both functional knowledge and multiple molecular profiles.

First, pathway networks are built for each patient and type of molecular profile separately (Figure 1A), directly incorporating functional knowledge from GO and KEGG (20). In comparison to methods based only on molecular profiles, the newly introduced functional focus reduces dimensionality to extract higher-level actionable patterns, and attenuates noise. Specifically, at this point, each patient is represented by a network of KEGG pathways that are connected by weighted edges. Weights are computed from the probabilities of the pathway-associated genes (KEGG) showing profiling differences between the tested patient

and controls (i.e. differential effect), while considering co-occurrence of genes in the same molecular processes (GO). An integrated view of KEGG pathways can be obtained by summation over GO processes. A weighted stochastic block model (wSBM) (47,53) is then used to find robust (54) pathway clusters in that network. This set of pathway clusters can be used to characterize *the disease of a specific patient* as reflected in a particular molecular profile type and already incorporating functional knowledge. The shared information distance between the clusterings (55) for different patients is then a natural metric of patient similarities. The shared information distance of clusterings is also known as the Variation of Information (VI) metric, and has favourable robustness and locality properties (45). This corresponds to the similarity between two nodes in the top-level network, i.e. how closely related two patients are with each other.

The metric furthermore allows a meaningful direct comparison and, thus, combination of distances. We exploit the robustness property to construct a stable patient similarity graph (Figure 1B) for each molecular profile type (gene expression, copy number, etc.). In summary, we obtain vector-valued edge weights between patients with each vector coordinate corresponding to a particular molecular profile type. We then take advantage of the metric's locality to combine information across different molecular profile types, and average the VI values, yielding a single *integrated patient similarity graph* (Figure 1C). After the VI distances are integrated, we normalize them as suggested by Meila (45).

This integrated patient similarity graph now captures all the structured information discovered both from individual patients and across patient cohorts, as guided by the functional focus of incorporating knowledge about pathways and biological processes. The structure resulting from this *Variation of information fused Layers of Networks* (ViLoN) can then be exploited at various levels, be that in an exploration of functional modules identified across patients, the establishment of more powerful prognoses, or information rich patient stratification for precision medicine, seeking an effective assignment of patient specific treatments. We here demonstrate the effectiveness of clustering patients in the integrated patient similarity graph using a Stochastic Block Model (SBM) to successfully identify similar patients with shared risk profiles. Survival analysis confirms that we can find patient groups of high clinical relevance, where many patients are affected by high hazard ratios.

A metric for clinically relevant patient stratification

Patient stratification seeks to identify a group of similar patients that can be associated with an average risk change. Such a grouping of patients is clinically relevant if the reduced or increased risk is sufficiently large, affects a reasonable number of patients, and passes statistical significance tests. While significance alone is insufficient for clinical relevance (30), there is no universally agreed threshold for the clinical relevance of hazard ratios for death by cancer (31). Large effects can often be identified more easily for more specific, smaller subsets of patients (cf. Figure 2, right panel). For small groups, we could even observe hazard ratios above 10^{10} (data not shown). Conversely, small hazard

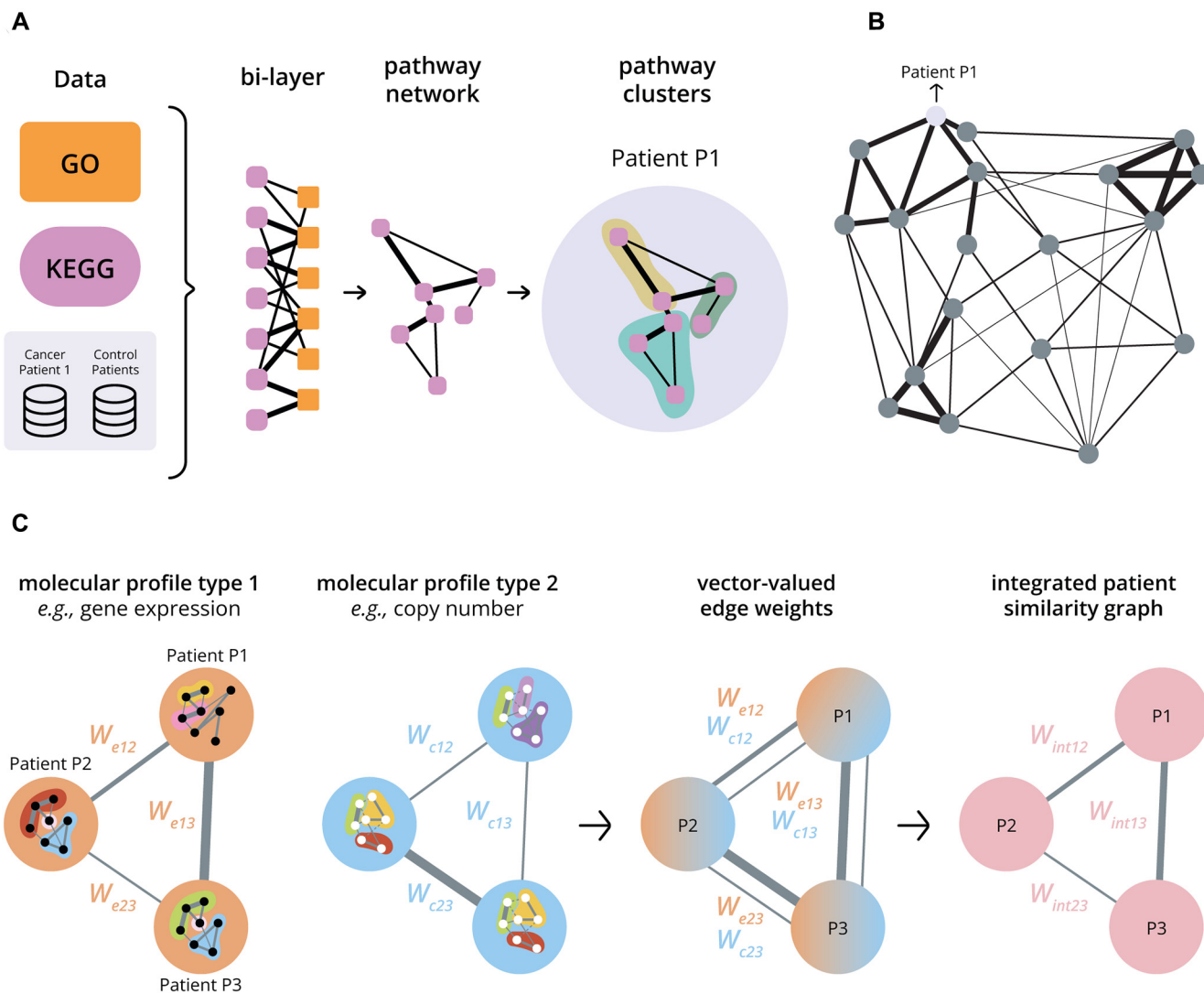


Figure 1. Overview of the multi-layer network structures exploited by the novel algorithm. **(A)** Building a patient-specific pathway network from a single molecular profile type that integrates functional knowledge. A patient-specific disease profile ('Cancer Patient 1') is compared to controls ('Control Patients'). The resulting differential effect (e.g. differential expression) is then assessed for genes of known biological processes (GO, orange) in a range of KEGG pathways (lavender). We next construct a bi-layer network with one layer of nodes representing KEGG pathways (lavender) and one layer of nodes representing biological processes (GO, orange). The overlap of genes between an orange and a lavender node (Jaccard Index) and the median posterior probability of a differential effect of these genes determine the edge weight (see *Methods*). This links pathways through their common functions in the cell in the context of the patient's disease. After integrating the bi-layer network by summation, each patient is thus represented by a network of KEGG pathways connected by weighted edges that reflect the similarities between pathways in the context of the patient's disease. The robust pathway clusters identified by a stochastic block model (shown in three colours for Patient P1) can be used as a characteristic fingerprint. Weights are illustrated by black lines with thickness representing weight strength (see Supplementary Figure S8 for details). **(B)** Building a patient similarity graph from a single molecular profile type using pathway clusters. The set of pathway clusters for a patient can be used to characterize the disease of that patient as reflected in a particular molecular profile type and already incorporating functional knowledge (A). The variation of information distance between the clusterings for different patients then forms a natural measure of patient differences, yielding a Variation of Information metric. Constructing a network of patients based on this metric thus allows an exploration of similarities between patients in the context of the disease. Nodes in the *patient similarity graph* represent patients and edge weights reflect pairwise similarities. The thicker a connecting edge, the stronger the similarity. (The figure omits some connections for visual clarity.) **(C)** Integrating multiple molecular profile types. A patient similarity graph is first constructed for each molecular profile type (cf. B). Nodes in the graph represent patients, and edge weights reflect the patient similarities for each molecular profile type. The graphs for different molecular profile types can be superimposed to give one graph with vector-valued weights for edges between patient nodes. An integrated patient similarity graph is then obtained by combining the information across molecular profile types for edges reflecting an average patient similarity (see Materials and Methods).

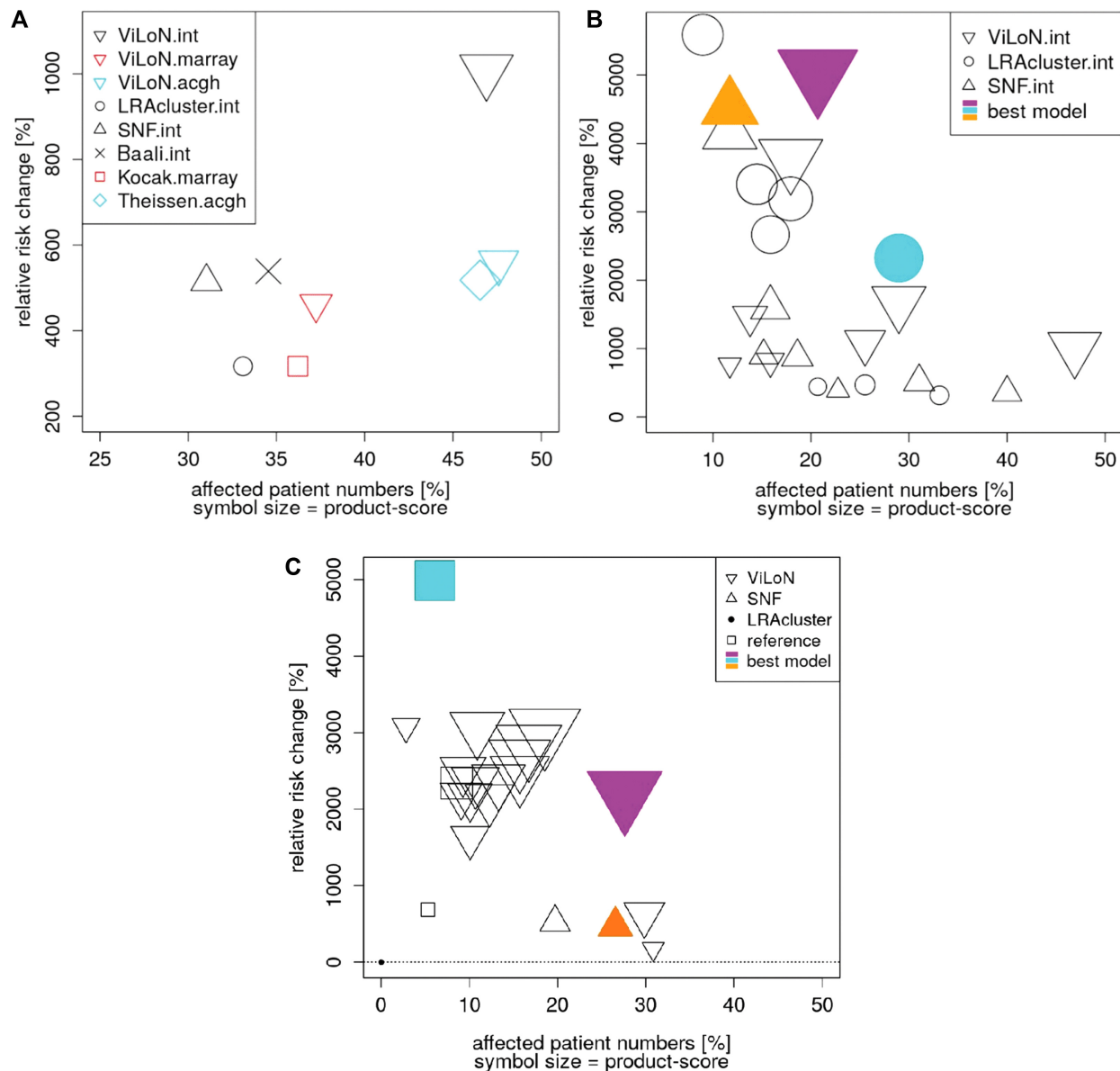


Figure 2. Effective numbers of affected patients. Model fit and other statistics are often not available for published patient splits. We can always, however, obtain the hazard ratio that characterizes the risk difference between patient groups as well as the number of patients affected. This can be used to construct an *effective number of affected patients* as a balanced measure N_{eff} . The clinically most relevant grouping will have a high hazard ratio and a large number of affected patients, as reflected in a large N_{eff} score. In this figure, the N_{eff} score is represented as symbol size. The x -axis indicates the actual number of patients affected, and the y -axis shows the relative risk change according to the hazard ratio. For N -group stratification (B, C), each symbol represents a distinct pairwise patient grouping, showing affected patient numbers (x -axis) and relative risk change vs the largest group as baseline reference (y -axis). Only groups with a significant risk change are shown. Results cover models for N groups of patients, for all N between 2 and 9. Colours highlight the best scores achieved by each method. (A) Two-group neuroblastoma stratification. We here show results for a widely studied large Neuroblastoma cohort for two-group stratification. ViLoN single profile results (coloured triangles point down) favourably compare with the best results reported earlier (square and diamond symbols). Assay type is shown by colour (microarrays: magenta, aCGH: cyan). The black symbols compare different approaches to vertical integration of assay types. The ViLoN model (black downwards triangle) clearly outperforms any single profile results. It also yielded much better results than alternative state-of-the-art integrative algorithms LRAcluster (black circle) and SNF (black triangle point up), as well as the best model identified in a recent CAMDA data analysis competition (black cross). (B) Integrative neuroblastoma analysis for N groups. We here show results for a widely studied large Neuroblastoma cohort for N -group stratification. We compare ViLoN (triangles point down) with alternative state-of-the-art integrative algorithms LRAcluster (circles) and SNF (triangles point up) for multi-group stratification. The clinically most relevant group was identified by ViLoN (large magenta triangle point down). (C) Integrative THCA analysis for N groups. We here show results for the TCGA thyroid carcinoma cohort (THCA-TCGA). We compare ViLoN (triangles point down) with alternative state-of-the-art integrative algorithms SNF (triangles point up) and LRAcluster (no significant models found) for multi-group stratification. The clinically most relevant group was identified by ViLoN (large magenta triangle point down).

ratios may be statistically significant but clinically not actionable. For a meaningful comparison of different splits of a patient cohort into groups we therefore need to consider the number of patients as well as the size of the risk change. To this end we introduce an *effective number of affected patients* as a balanced measure. Specifically, the product of the risk change and the number of affected patients gives a score reflecting an effective number of affected patients. To also allow a comparison of studies restricted to a subset of the full cohort (or other cohorts of different sizes), we standardize the score to a cohort of 1000 patients,

$$N_{\text{eff}} = 1000 \frac{N_g (2^{\text{abs}(\log_2 \text{HR})} - 1)}{N}, \quad (1)$$

for a hazard ratio HR and the affected group size N_g in a cohort size N .

Performance in the CAMDA cancer data integration challenge

The annual CAMDA data analysis challenge (www.camda.info) provides a well recognized forum for open-ended comparative exploration of novel algorithms (56–59). We here consider the neuroblastoma dataset (32–34,36). Neuroblastoma is the most common cancer in children (60) (www.cancer.gov/types/neuroblastoma/), causing 15% of cancer-related deaths at a young age (61). Despite modern therapies <40% of high-risk patients survive (www.cancer.gov/tcga). While some patients show spontaneous recovery it remains hard to predict who is at risk and assign appropriate therapy. A more precise prognosis and more effective treatment assignment is now expected to require an integration of molecular patient profiles (3–6).

In this cohort, three matched molecular profile types are available for 145 patients: RNA-Seq, microarray gene expression, and copy number data. Clinical records of both 97 low and 48 high risk patients include the survival times of the patients, with the majority of observations being right-censored, i.e. at the end of the study the patient was still alive or dropped off the study before it ended.

Groups of patients found by ViLoN were stable under removal of patients, i.e. the network was robust to sample changes. When we remove a patient from the dataset, the remaining patients mostly fall into the same clusters as obtained from the complete patient cohort. Leave-one-out robustness tests, removing each patient in turn, yielded an average accuracy of 98% (SE 11%) and an average normalized mutual information (51) of 97% (SE 6%). Values of 100% designate perfect agreement with the assignment from the original cluster.

ViLoN results for individual molecular profile types compare favourably with the most effective patient stratifications reported in the literature for gene expression profiles and copy number data (32–36). Several complementary statistics and metrics of model performance are displayed in Table 1A for gene expression profiles from microarrays ('marray') and copy number data from array comparative genomic hybridization ('acgh'). A small p value indicates statistical significance and the lower the Akaike Information Criterion, the better the model fit. The latter value is notably not often available for published patient splits. In

all cases, however, we can consider the hazard ratio (HR) that characterizes the risk difference between the patient groups, and the number of patients affected, which can be used to construct an *effective number of affected patients* as a balanced measure N_{eff} (see formula 1). The clinically most relevant grouping will have a high hazard ratio and a large number of affected patients, as reflected in a large N_{eff} score.

This N_{eff} score is represented as symbol size in Figure 2A, with the actual number of patients affected shown on the x -axis, and the relative risk change according to the two-group hazard ratio displayed on the y -axis. The ViLoN algorithm clearly improved on the best results reported in earlier work (Table 1, Figure 2A). ViLoN achieved an $N_{\text{eff}} = 1721$ versus 1247 for gene expression microarrays ('marray', magenta) and an $N_{\text{eff}} = 2660$ versus 1874 for array comparative genomic hybridization ('acgh', cyan). Vertically integrating patient networks across different molecular profile types considerably improved stratification performance further, achieving an N_{eff} of 4727.

We then compared the performance of ViLoN in two-group stratification with alternative state-of-the-art integrative algorithms: SNF (23) as an established network-based method, LRAcluster (19) implementing a modern probabilistic dimensionality reduction algorithm, and the best other model identified on this dataset in a recent CAMDA data analysis competition (62). In addition, we considered a method aiming to identify cancer driver genes in this competition (63). These driver genes may point to new therapy approaches but yielded lower stratification scores (data not shown), suggesting that drivers may not necessarily be the most prominent markers for patient stratification. ViLoN notably excelled already in two-group stratification (Table 1B, Figure 2A).

The grouping by ViLoN, interestingly, already covered most of the patients identified by SNF (96%) while expanding the group of high risk patients considerably (by 25, that is +58%) (Supplementary Figure S1). At the same time, the risk increase associated with group membership was not diluted and the group-associated risk was even considerably sharpened from 6.2 to 11.1 (Table 1B). Indeed, the ViLoN 2-cluster model (Table 1B) is already better than the 'high_risk' label by clinical experts (Table 1D, Supplementary Table S2). In a direct comparison, the high-risk labelled patients were strongly enriched in the group identified by ViLoN (Chi-square test p -value $< 10^{-14}$, Supplementary Figure S5). While the ViLoN group had a similar relative risk change (11.1 versus 11.7), the number of affected patients was considerably larger (+42%, 68 versus 48), allowing a clinically meaningful diagnosis for many more patients (Table 1B and D).

The standard benchmark Cox hazard model estimates the hazard ratio between two groups, allowing a direct comparison of a range of stratification methods, also relative to clinical reports in the literature. Several modern methods, however, can stratify into multiple groups, which can independently be assessed relative to a baseline, using the largest group as the reference. Indeed, considering stratification results from 2 to 9 clusters, the best pairwise performances for the compared integrative methods were found for a group of patients identified in stratifications by SNF, LRAcluster

Table 1. Comparisons of stratification performance: (A) state-of-the-art neuroblastoma stratifications by molecular profile data type, (B) comparison of integrative algorithms in two-group neuroblastoma stratifications, (C) the best pairwise splits in integrative multi-group neuroblastoma stratifications, (D) the clinical neuroblastoma ‘high_risk’ grouping, (E) THCA thyroid carcinoma stratifications, including state-of-the-art patient stratifications from the literature and the best results from established integrative algorithms. The table shows: Method, the name of the method for which results are listed in the corresponding table row; data type, the type of data used in the analysis (‘marray’, microarrays; ‘acgh’, copy number data; ‘int’, integrative model; ‘high_risk’, clinical label); #g, the number of strata that the model stratified patients into (between 2 and 9); p, the p -value indicating statistical significance of a non-zero log hazard ratio; adj.p.model (for #g > 2), the p -value of the whole model adjusted for multiple testing, actually yielding a q -value (75) indicating an upper bound for the false discovery rate (FDR); adj.p (for #g > 2), the p -value of the individual pairwise model with the largest N_{eff} for the whole stratification model adjusted for multiple testing, yielding a q -value (75), indicating an upper bound for the FDR; AIC, assessing the goodness of fit for the Cox regression by the Akaike Information Criterion, where lower values of the AIC are better; grSize, the size of the patient group with the largest hazard ratio in the model compared to the reference (largest) group; refSize, the size of the largest, i.e. reference, patient group in the model; fullSize, the size of the whole patient cohort analysed; absHR, the absolute hazard ratio; log₂HR, the corresponding log₂ of the hazard ratio – symmetrical for increased/decreased hazard ratio; N_{eff} , the effective number of affected patients score, indicating the clinical relevance of the model, incorporating both the hazard ratio via absHR, and the affected group size grSize, for a cohort of size fullSize. Note that the model with the largest N_{eff} and thus the largest clinical relevance is not necessarily the model with the best AIC (see text for Discussion). LRAcluster yielded no significant models for THCA thyroid carcinoma (adj.p for the corresponding largest yielded N_{eff} shaded in grey)

| A | Method | data type | #g | p | adj.p | AIC | grSize | refSize | fullSize | absHR | log ₂ HR | N_{eff} |
|---|-----------------------|-----------|----|--------------------|--------------------|-------|--------|---------|----------|-------|---------------------|------------------|
| | ViLoN | int | 2 | <10 ⁻⁹ | - | 317 | 68 | 77 | 145 | 11.1 | 3.5 | 4727 |
| | ViLoN | acgh | 2 | <10 ⁻⁷ | - | 329 | 69 | 76 | 145 | 6.6 | 2.7 | 2660 |
| | Theissen | acgh | 2 | - | - | - | 93 | 109 | 202 | 5.1 | 2.3 | 1874 |
| | ViLoN | marray | 2 | <10 ⁻⁶ | - | 329 | 54 | 91 | 145 | 5.6 | 2.5 | 1721 |
| | Kocak | marray | 2 | <10 ⁻³ | - | - | 208 | 366 | 574 | 4.4 | 2.2 | 1247 |
| B | Method | Data type | #g | p | adj.p | AIC | grSize | refSize | fullSize | absHR | log ₂ HR | N_{eff} |
| | ViLoN | int | 2 | <10 ⁻⁹ | - | 317 | 68 | 77 | 145 | 11.1 | 3.5 | 4727 |
| | Kazan | int | 2 | <10 ⁻⁹ | - | NA | 172 | 326 | 498 | 6.4 | 2.7 | 1863 |
| | SNF | int | 2 | <10 ⁻⁶ | - | 325 | 45 | 100 | 145 | 6.2 | 2.6 | 1598 |
| | LRAcluster | int | 2 | <10 ⁻⁴ | - | 336 | 48 | 97 | 145 | 4.2 | 2.1 | 1049 |
| C | Method | Data type | #g | adj.p.model | adj.p | AIC | grSize | refSize | fullSize | absHR | log ₂ HR | N_{eff} |
| | ViLoN | int | 5 | <10 ⁻⁵ | <10 ⁻⁷ | 308 | 30 | 34 | 145 | 51.3 | 5.7 | 10401 |
| | LRAcluster | int | 3 | <10 ⁻⁶ | <10 ⁻⁷ | 314 | 42 | 66 | 145 | 24.3 | 4.6 | 6734 |
| | SNF | int | 8 | <10 ⁻⁴ | <10 ⁻⁴ | 311 | 17 | 22 | 145 | 46.2 | -5.5 | 5300 |
| D | Method | Data type | #g | adj.p.model | adj.p | AIC | grSize | refSize | fullSize | absHR | log ₂ HR | N_{eff} |
| | clinical | high_risk | 2 | <10 ⁻¹¹ | <10 ⁻¹¹ | 305.7 | 48 | 97 | 145 | 11.72 | 3.6 | 3549 |
| E | Method | Data type | #g | adj.p.model | adj.p | AIC | grSize | refSize | fullSize | absHR | log ₂ HR | N_{eff} |
| | ViLoN | int | 4 | 0.002 | 0.005 | 150 | 137 | 206 | 496 | 23.1 | 4.5 | 6104 |
| | ViLoN | int | 5 | 0.002 | 0.005 | 147 | 83 | 171 | 496 | 29.5 | 4.9 | 4761 |
| | ViLoN | int | 6 | 0.002 | 0.008 | 148 | 78 | 152 | 496 | 25.6 | 4.7 | 3873 |
| | ViLoN | int | 7 | 0.002 | 0.006 | 144 | 54 | 117 | 496 | 31.6 | 5.0 | 3327 |
| | Agrawal <i>et al.</i> | clinical | 4 | <10 ⁻⁴ | <10 ⁻⁴ | - | 27 | 316 | 444 | 50.9 | 5.7 | 3034 |
| | ViLoN | int | 8 | 0.002 | 0.008 | 144 | 53 | 91 | 496 | 24.6 | 4.6 | 2521 |
| | ViLoN | int | 9 | 0.002 | 0.006 | 140 | 46 | 94 | 496 | 26.0 | 4.7 | 2318 |
| | ViLoN | int | 3 | 0.002 | 0.006 | 155 | 148 | 248 | 496 | 7.2 | 2.9 | 1853 |
| | SNF | int | 3 | 0.017 | 0.025 | 155 | 132 | 267 | 497 | 5.7 | 2.5 | 1259 |
| | LRAcluster | int | 9 | 0.043 | 0.668 | 157 | 52 | 85 | 497 | 6.2 | 2.6 | 539 |
| | ViLoN | int | 2 | 0.043 | 0.043 | 163 | 153 | 343 | 496 | 2.7 | 1.4 | 527 |
| | Agrawal <i>et al.</i> | clinical | 3 | 0.006 | 0.006 | - | 24 | 259 | 454 | 7.9 | 3.0 | 363 |

and ViLoN into 8, 3 and 5 clusters, respectively (Table 1C, Figure 2B).

The 5-cluster ViLoN stratification (Figure 3, right-hand side) yielded clearly distinct patient groups, with no deaths at all for the 34 patients of Group #1, and increasingly adverse outcomes for the other groups (Supplementary Figure S6). While the top group for SNF also had no deaths, it was only half the size (17 patients). The other groups for both SNF and LRAcluster were less consistently distinct and thus less easy to interpret.

In summary, already for two clusters our integrative ViLoN algorithm could stratify patients into clinically more relevant groups than the alternative integrative approaches (Table 1B), while our integrated model for five clusters identified an even more effective patient split, with the highest pairwise clinical relevance overall (Table 1C). The ViLoN clusters emerged directly from the molecular data alone, reflecting the underlying pathway networks in the context of

the patients’ disease. A preliminary investigation of differences in the pathway networks in our high-risk patient cluster highlighted the role of steroid sex hormone biosynthesis, identifying both genes with known roles in neuroblastoma as well as newly implicated genes (cf. Supplementary Figure S14 and accompanying discussion), corroborating the relevance of the emerged clusters.

Validation on other datasets

In order to further investigate the clinical relevance of our approach we now focus on a well established collection of cancer data: The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), collects data for over 30 types of cancer. It is the largest collection of molecular datasets of its kind, profiling patients with a range of assays, like RNA-Seq, methylation,

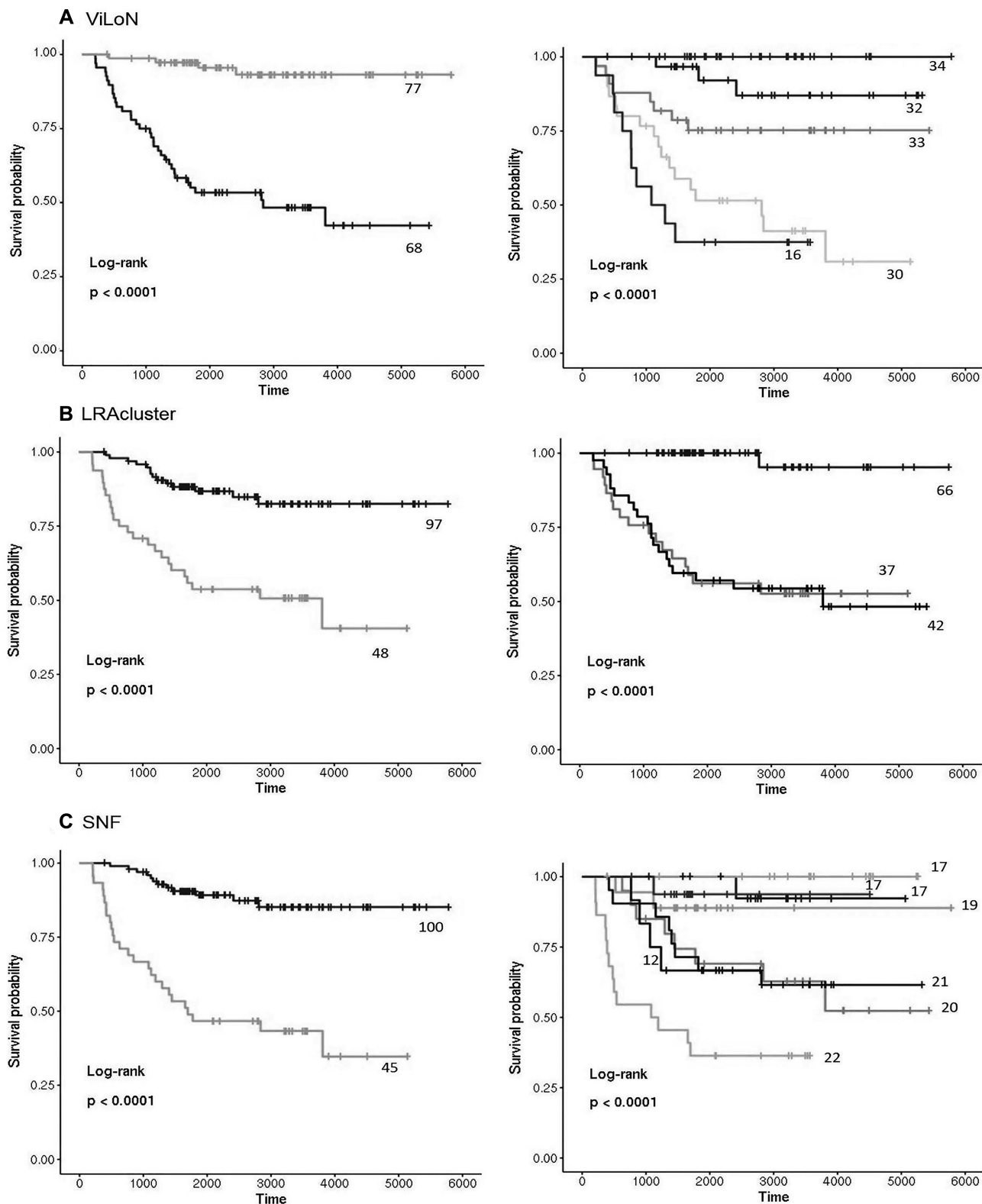


Figure 3. Kaplan–Meier curves showing the distinct survival profiles of groups found by the integrative algorithms for Neuroblastoma. Survival profiles of the patient groups found by (A) ViLoN, (B) LRAcluster, (C) SNF. For each algorithm, we show groups found by each algorithm in the two-group stratification (left-hand side) and the best reported patient stratification (right-hand side). The *p*-values are based on a log-rank test, and automatically approximated and plotted by the employed R function (from the *survminer* R package).

or copy number information (64). The TCGA datasets are publicly available and most of them were published in high-impact articles, where each cancer is examined by state-of-the-art methodology. This ensures a well-established baseline for any follow-up research for improving the prevention, diagnosis, and treatment of cancer. To emphasize the potential clinical impact of our novel method we focus on a few cancers where stratification for treatment is known to be difficult. Each of the cancer datasets contains matched profiles of two molecular profile types, specifically: RNA-Seq and either copy number or methylation data, together with corresponding clinical information. This allows us to build ViLoN networks for exploring integrated analysis. For each resulting integrated network, we stratify patients into subgroups for survival analysis. We then compare ViLoN stratification with models returned by two alternative state-of-the-art integrative methods—SNF and LRAcluster—as well as the original cancer-specific models developed for the seminal papers reporting the specific cancer datasets.

In order to highlight that the performance of ViLoN does not depend on the specific molecular profile types used in the neuroblastoma dataset, we next present a dataset that combines gene expression with DNA methylation data. In addition, we examined four different cancers profiled for gene expression as well as methylation or copy number. Analysis results are included in the *Supplement*, all of which corroborate the trends reported here.

Gene expression and DNA methylation profiles of TCGA thyroid carcinoma. The THCA-TCGA cohort comprises 500 thyroid carcinoma patients, all characterized by RNA-Seq and matched methylation profiles, making the repository one of the largest such multi-track resources. Thyroid carcinoma is the most common endocrine malignancy, yet identifying high-risk patients for adequate treatment and monitoring remains a major challenge (65). The clinical risk assessment for thyroid carcinoma in the seminal paper reporting the THCA-TCGA cohort (66) provides a state-of-the-art baseline reference. For a clinically meaningful comparison with modern computational stratification approaches we examined the effective numbers of affected patients given by the products of the pairwise hazard ratios and stratification group sizes (Table 1D, Figure 2C). Besides our ViLoN algorithm, we examined results by the leading established tools SNF (23) and LRAcluster (19) in the comparison.

LRAcluster yielding no significant models at all, however, reflects the difficulty of the stratification task. While the best stratification published to date (66) identified a group with an extremely high hazard ratio, the group diagnosis affected only 27 patients. Interestingly, ViLoN identified two much larger groups of high-risk patients that already recovered 20 of these 27 high-risk patients. The two groups of high-risk patients newly identified by ViLoN also featured very high hazard-ratios (23x and 31x), providing a clinically valuable diagnosis for 137 + 92 patients. In contrast, for the best SNF model, the highest risk group, while comprising 132 patients, achieved a lower group hazard ratio (<6x). While SNF recovered 5 of the 27 patients identified as highest risk in the literature (66) and shared 33 patients with the highest risk group identified by ViLoN, the

group diagnosis from SNF stratification was therefore clinically less actionable. This is also reflected in the Kaplan-Meier curves of the stratifications (Figure 3, Supplementary Figure S7).

The ViLoN clusters emerged directly from the molecular data alone, reflecting the underlying pathway networks in the context of the patients' disease. A preliminary investigation of differences in the pathway networks in our patient clusters highlighted the role of Notch signalling, suggesting a beneficial activation of Notch signalling in the lowest risk group of patients identified by ViLoN (cf. Supplementary Figure S15, Tables S8–S10, and accompanying discussion). This is interesting because of the complex and varying role of Notch signalling in different thyroid cancer subtypes (67). Notably, we also observed a significant enrichment of the follicular variant of papillary thyroid cancer in this patient group.

DISCUSSION

The success of the promise of integrated molecular profiling / patient profiling is highly dataset-specific

Recognizing differences in therapy response across individual patients, patient stratification is used to assign effective therapies by dividing patients into subgroups of individuals likely to react similarly to a treatment (1,2). A particular focus has been on cancer, a very heterogeneous disease, where patient stratification can remain challenging (3). For a detection of the wide variety of underlying mechanisms across patients, clinicians increasingly employ molecular markers in order to improve predictions of therapy response and prognosis (3–6). These include changes in gene activity (8,9), larger genomic sequence variants (10), as well as smaller changes, such as SNPs (11).

In fact, with the large amounts of molecular data now routinely collected in biomedical assays, the inference of clinically relevant insight remains the rate-limiting bottleneck. The very high dimensionality of genome scale profiles contributes to this challenge, (68,69) and much hope has been placed in an effective integration of heterogeneous data both across patients (12,13) and across assay types (13–15). For instance, combining data from gene expression and copy number information in a probabilistic framework was reported to improve clustering for subtype analysis (17,18) and the stratification of cancer patients (19). Network-based algorithms have successfully been applied to exploit not just the complementary nature of data sources but also similarities across patients, (23) both in basic research, such as identifying dysregulated pathways (20,21), and in applied research, such as the optimization of biotechnological processes (22). Surprisingly, despite these successes, integrated data analyses remain the exception rather than the rule and are, in fact, even falling in proportion of published studies (cf. Supplementary Figure S13). While this may partly be due to the higher complexity of such analyses, modern algorithms can be highly sensitive to dataset characteristics. Even well established methods may fail on new diseases or cohorts sufficiently different to those on which they were originally validated (cf. LRAcluster and SNF in Figure 2, and Supplementary Figures S9–S12), emphasising the need for comprehensive systematic benchmarks.

A novel clinically relevant metric allows a meaningful comparison of modern algorithms and classical literature

While joint benchmarks that rally the scientific community around common challenges and datasets have emerged as one successful approach to comparing the latest state of the art methods (cf. CAMDA, www.camda.info) (70–72), comparisons of patient stratification results have traditionally only considered statistical significance and hazard ratios. A meaningful comparison across different cohorts, however, needs to also consider the number of patients affected. The *effective number of affected patients* introduced here provides such a suitable, balanced measure. Importantly, it not only supports the comparison of the clinical relevance of patient groupings from different algorithms but also allows meaningful comparisons to the best patient stratifications from the clinical literature (also see Supplementary Table S7).

A combined network approach successfully exploits cohort structure, individual variations, and functional knowledge

In rigorous comparisons to established state-of-the-art algorithms and the clinical literature, we have developed and validated a novel framework for integrative data analysis, with a first application to cancer patient stratification into treatment groups. In ViLoN (*Variation of information fused Layers of Networks*), cancer patients are represented in a multi-layer network view, combining differential effect analysis of matched molecular profiles. In addition, we also incorporate external knowledge about biochemical pathways (KEGG) and molecular processes (GO terms). This builds on the LPIA approach to integrating functional information (20). Inspired by the patient networks of SNF (23), we use a multi-network approach to integrate both complementary molecular profile types and information across different patients. ViLoN goes beyond the SNF algorithm by both including information from differential effect analysis and taking advantage of a pathway level view. Together, we can compile a fully integrated characterization of the disease of a specific patient in relation to the rest of the cohort that summarizes complementary molecular data and exploits existing functional knowledge.

Good performance across different data types and cohort sizes

Indeed, the best performing models in the CAMDA cancer data integration benchmark challenge combined complementary information, integrating multiple molecular data types per patient with external domain knowledge about structure in the data. The performance improvement of our integrated analysis relative to models based on a single molecular data type was substantial, easily exceeding the state of the art in the literature, yielding both a higher pairwise hazard ratio as well as affecting a larger fraction of patients (cf. Table 1A).

We have successfully validated our novel algorithm in a comparison with state-of-the-art tools applied to a diverse set of cancer cohorts. ViLoN consistently outperformed both literature and alternative approaches substantially. In an integrated analysis of gene expression profiles and copy

number data of neuroblastoma patients we obtained the clinically most relevant pairwise split, with a $51.3\times$ hazard ratio for 30 patients. While the next best competing split was for a larger number of patients, it achieved a much lower hazard ratio ($24.3\times$, 42 patients), for a lower effective number of affected patients (cf. Table 1C).

Remarkably, despite improving patient stratification (Table 1D) our method builds solely on molecular information and external domain knowledge, without considering clinical data for its predictions. Still, patients clinically judged to be at high risk were significantly enriched in the identified high-risk patient groups. Our algorithm extended these to similar high risk patients that had not been identified by clinical parameters or established molecular markers alone (Supplementary Figures S2–S4, Table S1), while yielding more easily interpretable patient groups than other approaches (Figure 3).

Superior results were seen across different data types: We have established new state-of-the-art performance for the integrated analysis of gene expression profiles and copy number data in cohorts of neuroblastoma patients, cervical cancer patients (Supplementary Table S3, Figure S9), as well as colon and rectal cancer patients (Supplementary Table S4, Figure S10). Demonstrating that our approach generalizes to other data types, we have demonstrated similarly good results for integrated analysis of expression and methylation profiles: ViLoN substantially outperformed both alternative molecular approaches and the clinical reference stratification for thyroid cancer. We obtained the clinically most relevant pairwise split, achieving a $23.1\times$ hazard ratio for 137 patients. The next best competing split was actually from the clinical stratification (66). While it achieved a higher hazard ratio it was relevant for a much smaller number of patients ($50.9\times$, 27 patients), giving a much lower effective number of affected patients (cf. Table 1). Strikingly, all the other algorithms seriously struggled with this dataset. Similarly good performance of ViLoN integrating gene expression with methylation data could be observed for esophageal carcinoma (Supplementary Table S5, Figure S11) and rectal adenocarcinoma (Supplementary Table S6, Figure S12). Overall, we successfully benchmarked ViLoN on three data types and independent cohorts from six diseases.

Notably, and in contrast to other algorithms, our approach also performed robustly across different cohort sizes, ranging from only 90 patients for rectal adenocarcinoma to several hundred, e.g. for thyroid cancer (~ 500).

A powerfully generic extensible framework

In summary, we have introduced a novel powerfully generic framework for the integration of existing functional knowledge, complementary molecular profiles per patient, and similarities across patients for an integrated stratification of patient cohorts. We have comprehensively evaluated our approach discussing both established methods and through an innovative integrated metric for clinically relevant pairwise patient splits. The generated splits indeed capture clinically known risks while extending actionable patient groups considerably. ViLoN was successfully validated on three data types and independent cohorts from six diseases. Im-

proved results were particularly striking for the more difficult cohorts, where most or all other tested methods failed to yield significant splits: esophageal carcinoma and rectal adenocarcinoma. With the latter comprising just 90 patients, this improvement specifically for smaller cohorts may be promising for the relatively large number of rare diseases (73) considering that classical algorithms have traditionally struggled.

DATA AVAILABILITY

An implementation of the algorithm introduced in this manuscript (using Snakemake (74) workflows) together with documentation, tutorials, and pre-processed datasets for benchmarking and files supporting the analysis are available at <https://github.com/data-int/vilon/>.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

M.M.K. is a Marshall Plan Research Scholar and the research performed was in part supported by the Austrian Marshall Plan Foundation. The authors are grateful to Miriam Vázquez Carballido for the professional rendering of all schematic figures.

Author contributions: M.M.K. made key contributions to the conception and design of the study, was responsible for the computational analyses and the generation of figures, and contributed to the manuscript. E.K. and D.P.K. devised, designed, guided the study, and provided advice on the interpretation of results. D.P.K. wrote the manuscript. A.D.A. performed the molecular level pathway analyses and complementary comparisons to the state of the art, and contributed to the manuscript. SS preprocessed and prepared the neuroblastoma data for network analyses. All authors read and approved the final manuscript.

FUNDING

M.M.K. was partially supported by the Austrian Marshall Plan Foundation. Funding for open access charge: Janssen Pharmaceutica NV.

Conflict of interest statement. None declared.

REFERENCES

- Curtis,C., Shah,S.P., Chin,S.-F., Turashvili,G., Rueda,O.M., Dunning,M.J., Speed,D., Lynch,A.G., Samarajiwa,S., Yuan,Y. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.
- Papaemmanuil,E., Gerstung,M., Bullinger,L., Gaidzik,V.I., Paschka,P., Roberts,N.D., Potter,N.E., Heuser,M., Thol,F., Bolli,N. *et al.* (2016) Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.*, **374**, 2209–2221.
- Bagatell,R. and Cohn,S.L. (2016) Genetic discoveries and treatment advances in neuroblastoma. *Curr. Opin. Pediatr.*, **28**, 19–25.
- Oberthuer,A., Kaderali,L., Kahlert,Y., Hero,B., Westermann,F., Berthold,F., Brors,B., Eils,R. and Fischer,M. (2008) Subclassification and individual survival time prediction from gene expression data of neuroblastoma patients by using CASPAR. *Clin. Cancer Res.*, **14**, 6590–6601.
- Cohn,S.L., Pearson,A.D.J., London,W.B., Monclair,T., Ambros,P.F., Brodeur,G.M., Faldum,A., Hero,B., Iehara,T., Machin,D. *et al.* (2009) The international neuroblastoma risk group (INRG) classification system: an INRG task force report. *J. Clin. Oncol.*, **27**, 289–297.
- Fardin,P., Barla,A., Mosci,S., Rosasco,L., Verri,A., Versteeg,R., Caron,H.N., Molenaar,J.J., Ora,I., Eva,A. *et al.* (2010) A biology-driven approach identifies the hypoxia gene signature as a predictor of the outcome of neuroblastoma patients. *Mol. Cancer*, **9**, 185.
- Lindskrog,S.V., Prip,F., Lamy,P., Taber,A., Groeneveld,C.S., Birkenkamp-Demtröder,K., Jensen,J.B., Strandgaard,T., Nordentoft,I., Christensen,E. *et al.* (2021) An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nat. Commun.*, **12**, 2301.
- Milioli,H.H., Vimieiro,R., Tishchenko,I., Riveros,C., Berretta,R. and Moscato,P. (2016) Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset. *BioData Min.*, **9**, 2.
- Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 6567–6572.
- Mueller,S. and Matthay,K.K. (2009) Neuroblastoma: biology and staging. *Curr. Oncol. Rep.*, **11**, 431–438.
- Fagerholm,R., Schmidt,M.K., Khan,S., Rafiq,S., Tapper,W., Aittomäki,K., Greco,D., Heikkinen,T., Muranen,T.A., Fasching,P.A. *et al.* (2015) The SNP rs6500843 in 16p13.3 is associated with survival specifically among chemotherapy-treated breast cancer patients. *Oncotarget*, **6**, 7390–7407.
- Nguyen,T., Diaz,D., Tagett,R. and Draghici,S. (2016) Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. *Sci. Rep.*, **6**, 29251.
- Nguyen,T.C. (2017) In: *Horizontal and Vertical Integration Of Bio-Molecular Data*. Wayne State University Dissertations, Vol. **1728**.
- Searls,D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.*, **4**, 45–58.
- Hecker,M., Lambeck,S., Toepfer,S., van Someren,E. and Guthke,R. (2009) Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*, **96**, 86–103.
- El-Manzalawy,Y., Hsieh,T.-Y., Shivakumar,M., Kim,D. and Honavar,V. (2018) Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. *BMC Med. Genomics*, **11**, 71.
- Shen,R., Olshen,A.B. and Ladanyi,M. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.
- Shen,R., Mo,Q., Schultz,N., Seshan,V.E., Olshen,A.B., Huse,J., Ladanyi,M. and Sander,C. (2012) Integrative subtype discovery in glioblastoma using iCluster. *PLoS One*, **7**, e35236.
- Wu,D., Wang,D., Zhang,M.Q. and Gu,J. (2015) Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics*, **16**, 1022.
- Pham,L., Christadore,L., Schaus,S. and Kolaczyk,E.D. (2011) Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 13347–13352.
- Verbeke,L.P.C., Van den Eynden,J., Fierro,A.C., Demeester,P., Fostier,J. and Marchal,K. (2015) Pathway relevance ranking for tumor samples through network-based data integration. *PLoS One*, **10**, e0133503.
- Mizrachi,E., Verbeke,L., Christie,N., Fierro,A.C., Mansfield,S.D., Davis,M.F., Gjersing,E., Tuskan,G.A., Van Montagu,M., Van de Peer,Y. *et al.* (2017) Network-based integration of systems genetics data reveals pathways associated with lignocellulosic biomass accumulation and processing. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 1195–1200.
- Wang,B., Mezlini,A.M., Demir,F., Fiume,M., Tu,Z., Brudno,M., Haibe-Kains,B. and Goldenberg,A. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.
- Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.*

- (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
25. The Gene Ontology Consortium (2019) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
 26. Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
 27. Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. and Stuart, J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–i45.
 28. Elmarakeby, H.A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S.H., Salari, K., Kregel, S., Richter, C. et al. (2021) Biologically informed deep neural network for prostate cancer discovery. *Nature*, **598**, 348–352.
 29. Kivela, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y. and Porter, M.A. (2014) Multilayer networks. *J. Complex Netw.*, **2**, 203–271.
 30. Kühnberger, A., Fritz, A., Lermer, E. and Scherndl, T. (2015) The significance fallacy in inferential statistics. *BMC Res. Notes*, **8**, 84.
 31. Azuero, A. (2016) A note on the magnitude of hazard ratios. *Cancer*, **122**, 1298–1299.
 32. Stigliani, S., Coco, S., Moretti, S., Oberthuer, A., Fischer, M., Theissen, J., Gallo, F., Garavento, A., Berthold, F., Bonassi, S. et al. (2012) High genomic instability predicts survival in metastatic high-risk neuroblastoma. *Neoplasia*, **14**, 823–832.
 33. Coco, S., Theissen, J., Scaruffi, P., Stigliani, S., Moretti, S., Oberthuer, A., Valdora, F., Fischer, M., Gallo, F., Hero, B. et al. (2012) Age-dependent accumulation of genomic aberrations and deregulation of cell cycle and telomerase genes in metastatic neuroblastoma. *Int. J. Cancer*, **131**, 1591–1600.
 34. Kocak, H., Ackermann, S., Hero, B., Kahlert, Y., Oberthuer, A., Juraeva, D., Roels, F., Theissen, J., Westermann, F., Deubzer, H. et al. (2013) Hox-C9 activates the intrinsic pathway of apoptosis and is associated with spontaneous regression in neuroblastoma. *Cell Death. Dis.*, **4**, e586.
 35. Theissen, J., Oberthuer, A., Hombach, A., Volland, R., Hertwig, F., Fischer, M., Spitz, R., Zapatka, M., Brors, B., Ortmann, M. et al. (2014) Chromosome 17/17q gain and unaltered profiles in high resolution array-CGH are prognostically informative in neuroblastoma. *Genes Chromosomes Cancer*, **53**, 639–649.
 36. Zhang, W., Yu, Y., Hertwig, F., Thierry-Mieg, J., Zhang, W., Thierry-Mieg, D., Wang, J., Furlanello, C., Devanarayan, V., Cheng, J. et al. (2015) Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol.*, **16**, 133.
 37. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
 38. Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
 39. Law, C.W., Chen, Y., Shi, W. and Smyth, G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
 40. Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.
 41. Smyth, G.K. (2005) limma: linear models for microarray data. In: Gentleman, R., Carey, V., Huber, W., Irizarry, R.A. and Dudoit, S. (eds). *J. Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health*. Springer, NY, pp. 397–420.
 42. Wan, Y.-W., Allen, G.I. and Liu, Z. (2016) TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics*, **32**, 952–954.
 43. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P. and Tamayo, P. (2015) The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
 44. Csárdi, G. and Nepusz, T. (2006) The igraph software package for complex network research.
 45. Meilã, M. (2007) Comparing clusterings—an information based distance. *J. Multivar. Anal.*, **98**, 873–895.
 46. Mariadassou, M., Robin, S. and Vacher, C. (2010) Uncovering latent structure in valued graphs: a variational approach. *Ann. Appl. Stat.*, **4**, 715–742.
 47. Léger, J.-B. (2016) Blockmodels: a R-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates.
 48. Weiss, Y. (2001) On spectral clustering: analysis and an algorithm. *Adv. Neural Inform. Process. Syst.*, **14**, 849–856.
 49. Von Luxburg, U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.
 50. Biernacki, C., Celeux, G. and Govaert, G. (1998) Assessing a mixture model for clustering with the integrated classification likelihood.
 51. Pfitzner, D., Leibbrandt, R. and Powers, D.D. (2009) Characterization and evaluation of similarity measures for pairs of clusterings. *Knowl. Inf. Syst.*, **19**, 361–394.
 52. Heinze, G. and Dunkler, D. (2008) Avoiding infinite estimates of time-dependent effects in small-sample survival studies. *Stat. Med.*, **27**, 6455–6469.
 53. Côme, E. and Latouche, P. (2015) Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Stat. Model.*, **15**, 564–589.
 54. Biernacki, C., Celeux, G. and Govaert, G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**, 719–725.
 55. Meilã, M. (2003) Comparing clusterings by the variation of information. In: Schölkopf, B. and Warmuth, M.K. (eds). *Learning Theory and Kernel Machines, Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, Vol. **2777**, pp. 173–187.
 56. Johnson, K. and Lin, S. (2001) Call to work together on microarray data analysis. *Nature*, **411**, 885.
 57. Tilstone, C. (2003) DNA microarrays: vital statistics. *Nature*, **424**, 610–612.
 58. Nature Methods (2008) Going for algorithm gold. *Nat. Methods*, **5**, 659–659.
 59. Francescato, M., Chierici, M., Rezvan Dezfouli, S., Zandonà, A., Jurman, G. and Furlanello, C. (2018) Multi-omics integration for neuroblastoma clinical endpoint prediction. *Biol. Direct*, **13**, 5.
 60. Lam, C.G., Howard, S.C., Bouffett, E. and Pritchard-Jones, K. (2019) Science and health for all children with cancer. *Science*, **363**, 1182–1186.
 61. McGuire, S. (2016) World cancer report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015. *Adv. Nutr.*, **7**, 418–419.
 62. Baali, I., Acar, D.A.E., Aderinwale, T.W., Hafez Qorani, S. and Kazan, H. (2018) Predicting clinical outcomes in neuroblastoma with genomic data integration. *Biol. Direct*, **13**, 20.
 63. Suo, C., Deng, W., Vu, T.N., Li, M., Shi, L. and Pawitan, Y. (2018) Accumulation of potential driver genes with genomic alterations predicts survival of high-risk neuroblastoma patients. *Biol. Direct*, **13**, 14.
 64. Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J.M. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
 65. Raue, F. and Frank-Raue, K. (2016) Thyroid cancer: risk-stratified management and individualized therapy. *Clin. Cancer Res.*, **22**, 5012–5021.
 66. Cancer Genome Atlas Research Network (2014) Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, **159**, 676–690.
 67. Guenter, R., Patel, Z. and Chen, H. (2021) Notch signaling in thyroid cancer. *Adv. Exp. Med. Biol.*, **1287**, 155–168.
 68. Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003) Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, **19**, 563–570.
 69. Mak, M.-W. and Kung, S.-Y. (2006) A solution to the curse of dimensionality problem in pairwise scoring techniques. In: King, I., Wang, J., Chan, L.-W. and Wang, D. (eds). *Neural Information Processing, Lecture notes in computer science*. Springer Berlin Heidelberg, Berlin, Heidelberg, Vol. **4232**, pp. 314–323.
 70. Aniba, M.R., Poch, O. and Thompson, J.D. (2010) Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Res.*, **38**, 7353–7363.

71. Boutros,P.C., Margolin,A.A., Stuart,J.M., Califano,A. and Stolovitzky,G. (2014) Toward better benchmarking: challenge-based methods assessment in cancer genomics. *Genome Biol.*, **15**, 462.
72. Sczyrba,A., Hofmann,P., Belmann,P., Koslicki,D., Janssen,S., Dröge,J., Gregor,I., Majda,S., Fiedler,J., Dahms,E. *et al.* (2017) Critical assessment of metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063–1071.
73. Jia,J. and Shi,T. (2017) Towards efficiency in rare disease research: what is distinctive and important?*Sci. China Life Sci.*, **60**, 686–691.
74. Köster,J. and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
75. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B (Methodological)*, **57**, 289–300.