Briefings in Bioinformatics, 22(3), 2021, 1–12

https://doi.org/10.1093/bib/bbaa217
Problem Solving Protocol

Analysis, identification and visualization of subgroups in genomics

Gunnar Völkel†, Simon Laban†, Axel Fürstberger†, Silke D. Kühlwein†, Nensi Ikonomi†, Thomas K. Hoffmann, Cornelia Brunner, Donna S. Neuberg, Verena Gaidzik, Hartmut Döhner, Johann M. Kraus and Hans A. Kestler

Corresponding author: E-mail: hans.kestler@uni-ulm.de
†These authors contributed equally to this work.
Authors Kraus and Kestler are joint senior authors.

Abstract

Motivation: Cancer is a complex and heterogeneous disease involving multiple somatic mutations that accumulate during its progression. In the past years, the wide availability of genomic data from patients' samples opened new perspectives in the analysis of gene mutations and alterations. Hence, visualizing and further identifying genes mutated in massive sets of patients are nowadays a critical task that sheds light on more personalized intervention approaches. **Results:** Here, we extensively review existing tools for visualization and analysis of alteration data. We compare different approaches to study mutual exclusivity and sample coverage in large-scale omics data. We complement our review with the standalone software AVAtar ('analysis and visualization of alteration data') that integrates diverse aspects known from different tools into a comprehensive platform. AVAtar supplements customizable alteration plots by a multi-objective evolutionary algorithm for subset identification and provides an innovative and user-friendly interface for the evaluation of concurrent solutions. A use case from personalized medicine demonstrates its unique features showing an application on vaccination target selection. **Availability:** AVAtar is available at: https://github.com/sysbio-bioinf/avatar **Contact:** hans.kestler@uni-ulm.de, phone: +49 (0) 731 500 24 500, fax: +49 (0) 731 500 24 502

Key words: visualization; exploratory analysis; multi-objective optimization; vaccination targets

Gunnar Völkel is a postdoctoral researcher at the Institute of Medical Systems Biology (MSB), Ulm University, Ulm, Germany.
Simon Laban is a senior physician at the Department of Otorhinolaryngology, Head and Neck Surgery, Ulm University Medical Center, Germany.
Axel Fürstberger is a postdoctoral researcher at MSB and deputy of Core Unit Bioinformatics, Ulm University, Ulm, Germany.
Silke D. Kühlwein is a PhD student at MSB and the International Graduate School of Molecular Medicine, Ulm University, Germany.
Nensi Ikonomi is a PhD student at MSB and the International Graduate School of Molecular Medicine, Ulm University, Germany.
Thomas K. Hoffmann is Medical Director, Department of Otorhinolaryngology, Head and Neck Surgery, Ulm University Medical Center, Germany.
Cornelia Brunner is a research laboratory head at the Department of Otorhinolaryngology, Head and Neck Surgery, Ulm University Medical Center, Germany.
Donna S. Neuberg is a senior lecturer of Biostatistics at the Department of Biostatistics, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.
Verena Gaidzik is a senior physician at the Department of Internal Medicine III, Ulm University Medical Center, Germany.
Hartmut Döhner is Medical Director, Department of Internal Medicine III, Ulm University Medical Center, Germany.
Johann M. Kraus is a postdoctoral researcher at MSB, Ulm University, Germany.
Hans A. Kestler is Director, Institute of Medical Systems Biology, and head of Core Unit Bioinformatics, Ulm University, Ulm, Germany, and an associated group leader at the Leibniz Institute on Aging–Fritz Lipmann Institute Jena, Jena, Germany.
Submitted: 4 June 2020; **Received (in revised form):** 14 August 2020

Introduction

High-throughput biomolecular technologies make multi-modal data available for diverse biological and medical settings. Analysis and visualization of genomic and transcript-omic maps are important steps not only for illustration but also for exploratory analysis [35, 45, 46, 70, 86]. In this context, computer assistance is paramount in the objective analysis of data such as gene mutations or overexpression of genes [74]. Most importantly, intuitive visualization tools simultaneously integrating different data modalities are required [87]. These data modalities such as mutation, expression or methylation profiles can be depicted as alteration plots, comparing patients' coverage for the examined alteration to alteration exclusivity for the single patient's sample.

Commonly, alteration plots are created manually by researchers in time-consuming processes, although specialized analysis tools for alteration data may include basic visualization functionality, e.g. cBioPortal [14, 30], Gitools [79], UCSC Cancer Genomics Browser [85], Integrative Genomics Viewer [84], IntOGen [32], MAGI [57] and caOmicsV [117].

Visualization and annotation of genomic alteration are only the first steps in deepening the knowledge on disease development and progression. Therefore, a huge effort has also been made in the context of the analysis of genome alteration data. Detection of mutually exclusive alterations has been shown to provide crucial information in the context of cancer development and investigation of therapeutic approaches, also in light of personalized treatments [24]. Due to the extensive heterogeneity in cancer genomes, most patients possess only a single driver mutation [105]. Hence, groups of genes harboring driver mutations tend not to co-occur in the same sample, as also shown in different cancer cohorts [11, 15]. Many cancer-related genes are involved in the phenomenon of mutual exclusivity [28]. Exemplarily, BRAF and NRAS, both members of the MAPK pathway, are widely altered in patients with melanoma [9], thyroid carcinoma [25], myeloma [71] and colorectal cancer [64]. However, few patients harbor both alterations [9]. Strikingly, the forced expression of two mutually exclusive genes in lung adenocarcinoma [101], KRAS and EGFR, causes proliferation and survival disadvantage to cancer cells [41]. The biologically motivated hypothesis behind mutual exclusivity is based on either functional redundancy [15, 17, 21, 90, 109] of these genes or synthetic lethality [16, 27]. Hence, having efficient algorithms to subgroup genes based on their mutual exclusivity gives both insights on patients' specific sub-groupings and potential personalized therapeutic interventions.

As an illustrative example, the TCGA AML (acute myeloid leukemia) dataset [99], containing mutation data of 200 samples, is used. Figure 1A, shows the alteration plot for the top 10 genes with the highest sample coverage. In the figure, each patient sample is a column of the plot, whereas every row represents a gene that can be altered or not in each sample. Coverage is defined as the representativeness of alteration within all samples. The overlap is considered as the co-occurrence of multiple mutations within the same sample. In Figure 1 considerable overlap (less mutually exclusivity) between the genes can be observed. Here, to further optimize the gene set selections, different algorithms can be applied. The final aim is to find subsets of genes with high coverage together with high mutual exclusivity (Figure 1B). These approaches are based either on alteration data only (*de novo*) or integration with experimental knowledge or databases (knowledge-based approaches).

In the following, we will review available visualization and analysis tools for gene alterations. Besides one exception, none of the currently available visualization tools provides implementations for gene drivers selection. Hence, we introduce our software 'analysis and visualization of alteration data' (AVAtar) that incorporates both visualization and analysis approaches. AVAtar has unique features for import, filtering and a combination of private and database data together with numerous export functions. Different from all other analysis software, AVAtar provides a user-friendly interface suitable for life scientists without the need for programming skills. To explore the possibilities provided by AVAtar, we first compared its features to available tools for both visualization and analysis. Second, we summarized a use case from an already published study [29] regarding semi-personalized selection of candidate vaccination targets for head and neck squamous cell carcinomas (HNSCC).

Genomic analyses to semi personalized medicine

Since the first human genome has been made available [18], the investigation and interpretation of genomic data have been a focus of modern molecular biology [53]. Technical improvements and strong reduction of sequencing costs have permitted genomic information to be more and more included in medical practice [10]. Thanks to introduction of genomic analyses, cancers can now be defined by their molecular drivers. This is not only useful in cancer identification but also for treatment decisions [10]. In fact, some traditional treatments could be proved ineffective or have side effects in certain patient populations [10, 98]. For example, tamoxifen was long time used to treat breast cancer patients. However, nowadays it is known that patient-specific alterations affecting its active metabolite exist [37]. Hence, analyses of genomic alterations is an important step towards semi-personalized medicine. Moreover, minimal sets of intervention targets might be of interest.

Visualization tools

Tools for visualization of genomic data are relevant for understanding and describing connections between genomic alterations and cancer [76, 88, 110]. Different approaches for visualization have been suggested (Table 1), such as genomic coordinate views, heat maps, network views, aberration plots and transcript views. The advantages of genomic coordinates-based tools rely on the possibility of accessing detailed sequences and various types of alterations. However, they can display limited numbers of samples and genes simultaneously [117]. Examples of these tools are cBioPortal [14, 30], Integrated Genomics Viewer [84], MAGI [57] and USC Cancer Genome Browser [85]. Besides genomic coordinate view approaches, heat maps and network views allow visualization of multiple genomic alterations in broad groups of genes and samples [14, 108, 114]. Hence, for further analyses, these types of genomic visualization are preferred. In accordance, most of the available tools include heat map visualization. A distinguishing feature of these tools is the possibility of combining private and database data, as well as exporting both raw data and figures (Table 1). In this regard, AVAtar is the only tool that provides a complete set of features for both import and export.

Tools for gene selection of mutual exclusivity

Different computational approaches have been developed to investigate driver gene mutations. They are mainly divided into

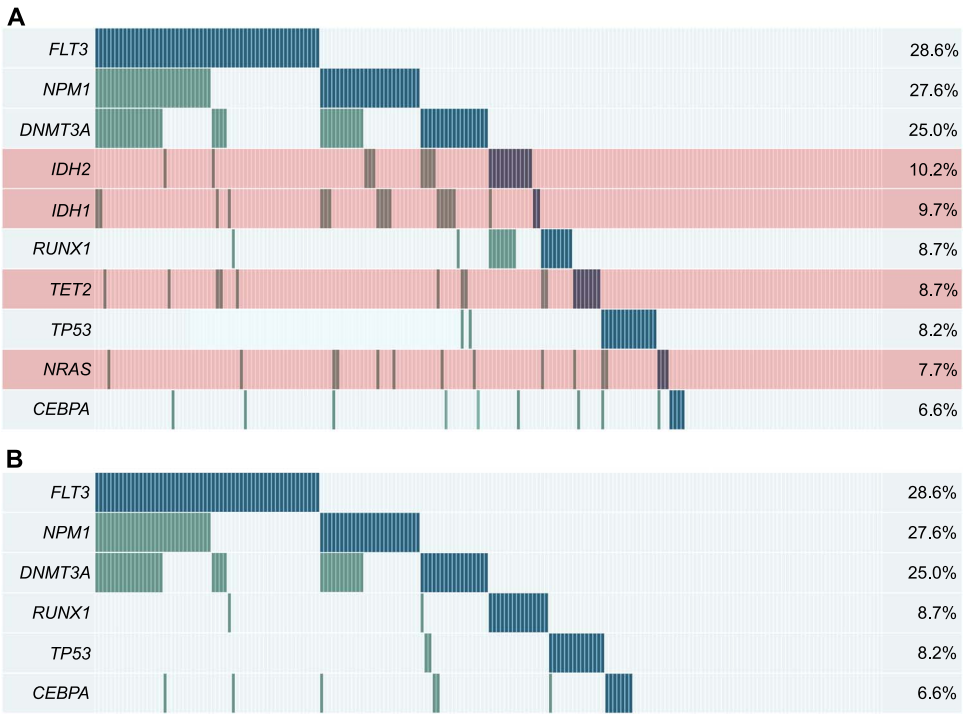

Figure 1. Alteration plots for gene selections of the TCGA AML dataset [99]. Every row represents a gene and every column a sample. If a sample has an alteration in a gene, the corresponding cell is marked (blue: first alteration in sample, green: overlapping alterations). (A) Top 10 genes with most frequent mutations sorted by sample coverage. (B) Example of optimization of the gene set selection with six genes (sorted by sample coverage). Genes highlighted in A (red) have been excluded by the optimization of the gene set selection.

de novo and knowledge-based approaches, where experimental information is integrated into the algorithms. Even if many tools for knowledge-based investigation are available [2, 5, 7, 8, 12, 13, 15, 31, 34, 38, 40, 47, 48, 59, 61, 65, 78, 80, 83, 92, 96, 97, 100, 113, 116, 118, 119], the fact that they require information on either pathways, interaction networks, or functional phenotypes data makes their broad application limited. Hence, *de novo* methods will be the focus of this review. In general, *de novo* methods are based only on alteration data. Two main strategies for the selection of mutually exclusive genes have been classically applied (Figure 2). One of the simplest approach to investigate mutual exclusivity is pairwise statistical analyses such as Fisher's exact test or likelihood ratio methods [24]. However, this approach has many limitations. First, it assumes that genetic alterations are evenly distributed across samples, which does not face the reality of alteration data [1, 3, 51, 54]. This problem was addressed by the WeSME approach that includes a weighted sampling proportional to the observed mutation frequency [49]. In addition, mutual exclusivity frequently does not involve only a few genes [115], making the pairwise test approach not suitable for investigating modules [67]. For this purpose, algorithms searching for modules of mutually exclusive mutations have been developed. Here, the new addressed task is to find sets of genes whose alterations cover high number of samples (coverage) together with low number of overlapping alterations in the set (high mutually exclusivity). This problem has been addressed from different perspectives. The combinatorial score approach, such as the Dendrix and its extension [56, 106], uses greedy algorithms to maximize the combination of these two objectives by expanding a seed set of genes. However, this method that tries to maximize both coverage and exclusivity can be biased towards high frequencies genes sets [105]. For this reason, other modules selection approaches have been implemented. Exemplarily, CoMet [60], MEGSA [77], and GAMToc [68] focus on selecting modules based on statical significance instead of maximization of scores (Figure 2). Another approach to overcome the problems behind score maximization has been implemented in AVAtar. Instead of defining weights a priori and searching for a single optimal gene selection, a set of gene selections consisting of the optimal trade-offs (defined as Pareto set) between the objectives can be identified (Figure 2). This allows the researcher to interactively explore the optimal trade-offs found and to choose gene sets based on task-specific background knowledge.

Another limitation of available tools to compute mutual exclusivity is their running environment (Table 2). Almost all tools require programming and bioinformatics skills of the user, thus excluding a wide range of life science researchers. On the other hand, AVAtar provides user-friendly graphical interface. In this context, also cBioPortal [14, 30] presents the possibility to perform mutual exclusivity analysis with a user-friendly graphics. However, the algorithm is based on pairwise tests that have already been shown to be quite limited [14, 24, 30]. Instead, AVAtar implements a modular search based on Pareto set optimization. Moreover, the user can preprocess data and combine sets of private and database searches in a unique form.

Multi-objective evolutionary algorithm implemented in AVAtar

Weighting the importance of contradicting aims is a common approach to resolve conflicts and allows the application of standard optimization algorithms. However, there is neither a generally applicable weighting for all datasets nor an obvious and objective weighting for a given dataset that guarantees an optimal selection. Here, multi-objective optimization [22] offers an alternative approach. Technically, the task consists in finding a

Table 1. Available visualization tools for genomic alterations. For each tool, its application type (R package, Web application with local installation or standalone software) is specified as well as analysis, visualization, import and export features. A ✓ indicates that the feature is available, '–' indicates it is not present

Name	Application type	Analysis/visualization				Import			Export	
		Alteration plots	Pathway subgrouping	Gene-set selection	Dataset preprocessing	Databases	Privatedata	Combined data	Raw data	Figures
caOmicsV [117]	R package	✓	–	–	–	–	✓	–	✓	✓
cBioPortal [14, 30]	Web/local	✓	✓	✓	–	✓	✓	–	✓	✓
Gitools [79]	Web/local	✓	–	–	–	✓	✓	–	✓	✓
Integrated Genomics Viewer [84]	Web/local	✓	–	–	✓	✓	✓	✓	–	–
IntOGen [32]	Web/local	✓	✓	✓	–	✓	✓	–	✓	–
MAGI [57]	Web/local	✓	✓	–	✓	✓	✓	✓	–	✓
UCSC Cancer Genome Browser [85]	Web/local	✓	✓	–	✓	✓	✓	✓	–	–
AVAtar	Standalone	✓	✓	✓	✓	✓	✓	✓	✓	✓

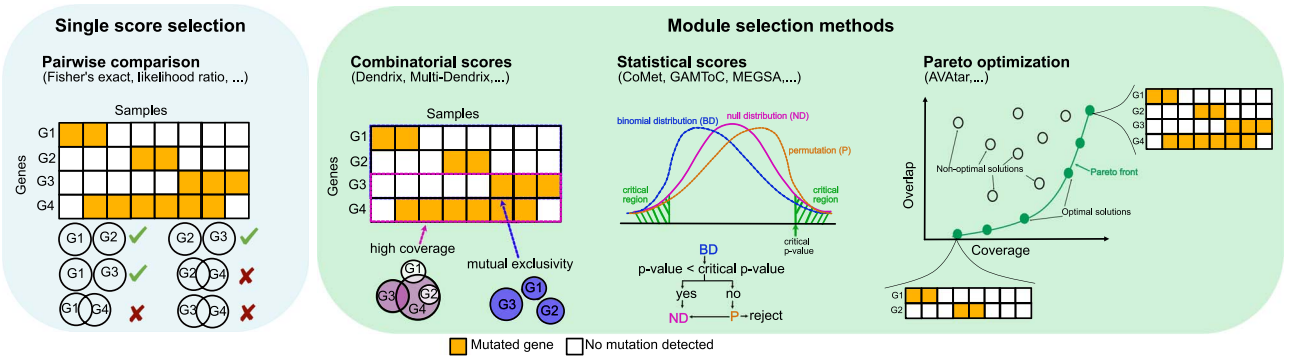

Figure 2. *De novo* approaches overview for investigation of mutual exclusivity. On the left blue box, single score selection is depicted. It is performed by applying pairwise test on alteration data. On the right green box, module selection methods are illustrated. Sets of mutually exclusive genes are selected by different approaches: combinatorial scores, statistical scores and Pareto optimization.

subset of genes G from a given set of genes $\mathcal{G}$ ($G \subseteq \mathcal{G}$) such that the number of covered samples $\gamma(G)$ is maximized and either the overlap $\omega(G)$ or the number of genes $|G|$ is minimized. This introduces a multi-objective optimization problem that results in finding a Pareto-optimal set $\mathcal{S}^* \subseteq \mathcal{S}$ of gene subsets within the set of all gene subsets $\mathcal{S} = G \subseteq \mathcal{G}$. To this purpose, we developed an evolutionary algorithm for the multi-objective gene selection task based on the Non-dominated Sorting Genetic Algorithm II [22, NSGA-II] (implemented by jMetal library v.5.3 [73]). This is a population-based metaheuristic that adapts concepts of the theory of evolution [81]. A set of solutions, called population, is evolved iteratively by applying recombination and mutation operators to the solutions. Finally, AVAtar facilitates the creation of objective and reproducible alteration plots by offering algorithmic sorting of genes and samples. The task of finding a gene order based on the additionally covered samples is formulated as a minimal set cover problem. A modified greedy algorithm [20] for set covering is applied [44]. Starting with an initial solution, the greedy algorithm incrementally adds the gene covering the most uncovered samples to its current partial solution. Further details of the algorithm are described in the Supplementary Information.

AVAtar

The multi-objective approach using a Pareto front was integrated in a readily usable standalone software AVAtar. There are no additional software requirements. After the extraction of the downloaded AVAtar archive, the user can immediately start to use it. The project repository of AVAtar available at https://github.com/sysbio-bioinf/avatar includes a detailed user manual with a stepwise walkthrough describing the application of AVAtar in the HNSCC analysis. The walkthrough already underlines our user-friendly interface, suitable for life scientists. AVAtar supports the import from different data sources: data files (Excel, Text), cBioPortal [14, 30] and Gene Expression Omnibus [6]. The import dialog with builtin search capabilities is shown in Figure 3. It is possible to combine data from multiple studies and to integrate different alteration types. Clinical attributes can be imported to define sample groups for

Table 2. Available software for identification of mutually exclusive genes in alteration data. For each of the tools, the method used to identify mutually exclusive alterations is stated. At last, running environment needed is reported. Tools marked by an asterisk (*) have both visualization and analysis features

Name	Method	Running environment
cBioPortal [14, 30]*	Pairwise test	Web tool
CoMET [60]	Statistical score	Python
Dendrix [106]	Combinatorial scores	Python
GAMToc [68]	Statistical score (entropy score)	Matlab
MEGSA [77]	Statistical score	R
Multi-Dendrix [56]	Combinatorial score	Python
MutExSL [94]	Pairwise test	Excel
RME [23]	Combinatorial score	Bash
TiMEX [19]	Statistical score	R
WeSME [49]	Stastical score	Python
WExT [58]	Statistical score	Python
AVAtar*	Pareto front	Standalone software

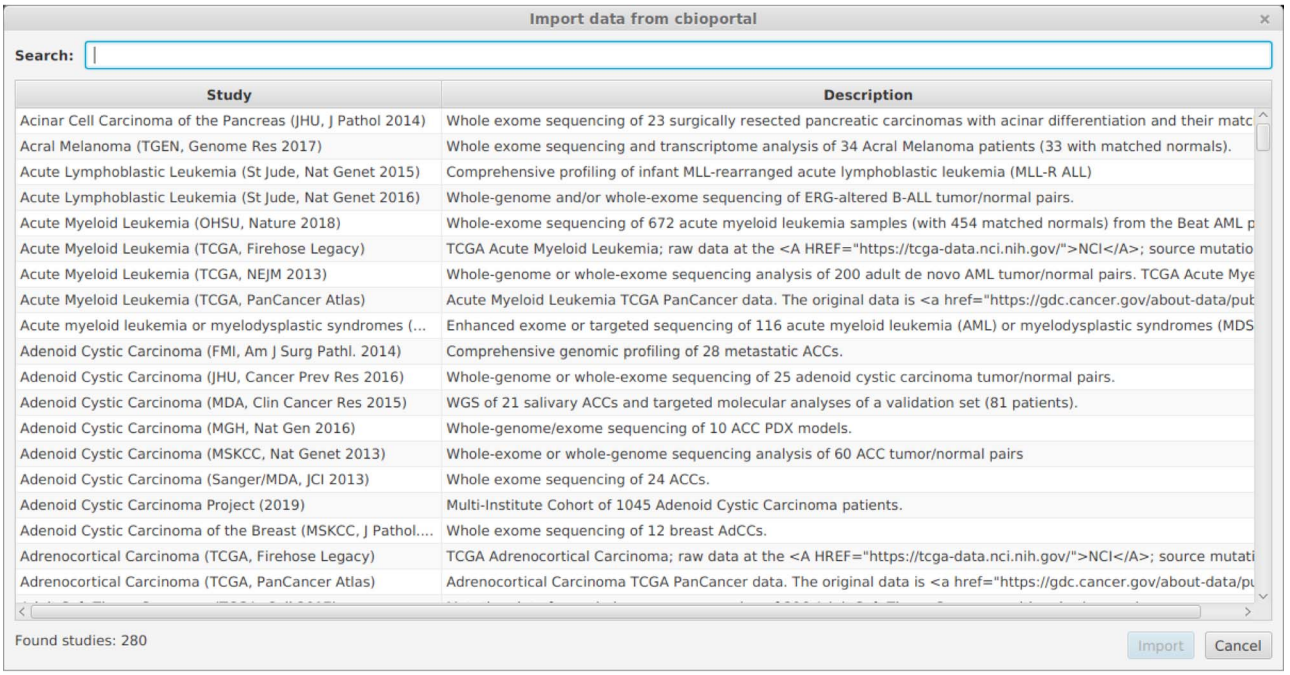

Figure 3. Dialog for selecting a study to import from cBioPortal. Studies can be searched by specifying terms that occur in their id, name or description.

analysis and visualization. Moreover, AVAtar's optimization setup is designed to offer default algorithm parameter values that are suitable for a broad range of applications. A batch mode to perform multiple optimizations in parallel on a compute server is available. It can be started as follows:

```
java11 -jar avatar.jar -f hnsc.avatar -o usecase.batch -t 6
Runs: 0/7 - Progress: 0.243% - Estimated Duration: 05:42:54
```

Resulting gene selection sets can be explored interactively as alteration plots and grouped by functional categories. Additionally, graphics contrasting alternative gene selections can also be created. Finally, all graphics obtained can be exported as publication ready vector graphics.

Comparison between AVAtar and the Multi-Dendrix approach

Deng *et al.* [24] already performed a benchmark comparison among a variety of available algorithms for mutual exclusivity investigation. From their analysis, the Multi-Dendrix algorithm [56] performed best. Hence, we compared the performance of AVAtar to this approach. We used the breast cancer dataset of the Multi-Dendrix publication [56]. The Multi-Dendrix algorithm uses a fixed a priori weight between coverage and overlap to find a specified small number of pairwise disjoint sets of genes with mostly mutually exclusive mutations within the sets. The genes of the found sets are considered as potential driver genes. The multi-objective evolutionary algorithm of AVAtar does not weight coverage and overlap of the gene sets but instead searches for the set of optimal tradeoffs between these two objectives. In Figure 4, the Pareto front and the corresponding set of selected genes resulting from our analysis are shown. Cells colored in blue are the genes also identified from the Multi-Dentrix. Genes with higher coverage are identified by both methods and are also represented in most of our solutions. In orange, we depicted new genes found by AVAtar. First, it can be observed that part of new hits is represented in solution with high coverage and overlap. However, we highlighted some interesting sets found by combinations of good coverage and low

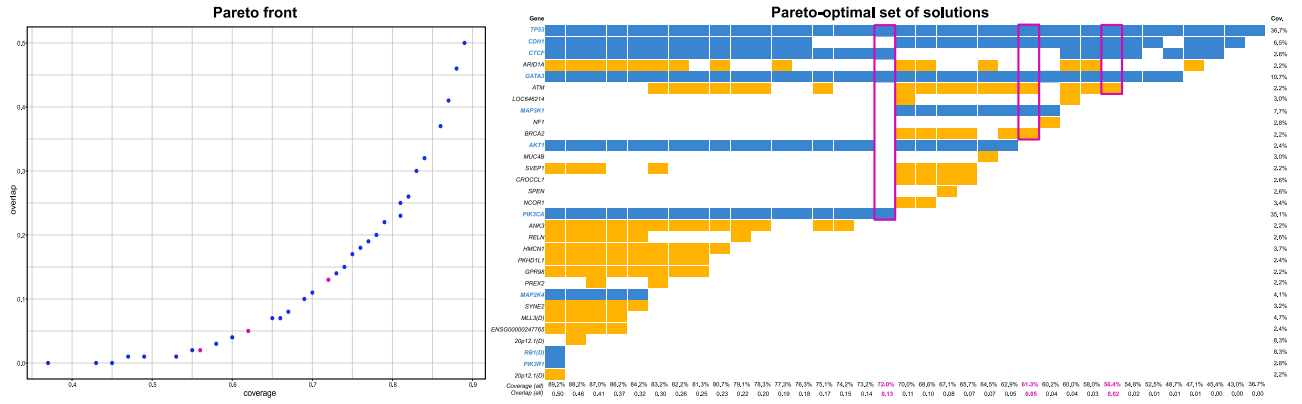

Figure 4. Visualization of the Pareto set resulting from a coverage maximization and overlap minimization on the breast cancer dataset used to evaluate the Multi-Dendrix algorithm [55]. On the left, the Pareto front is represented. Each point in the Pareto front represents a gene set solution for a certain coverage–overlap combination. Magenta dots represent the highlighted solution of the right figure. On the right, the corresponding Pareto set solutions for gene selection are depicted. Each column represents a solution of the Pareto set and each row contains a gene. The occurrence of a gene in a solution is marked by a colored rectangle. The solutions found also by the Multi-Dendrix algorithm are marked in blue, whereas new genes idenfied by AVAtar are marked in orange. The three selected solutions highlighted in magenta are shown in Figure 5.

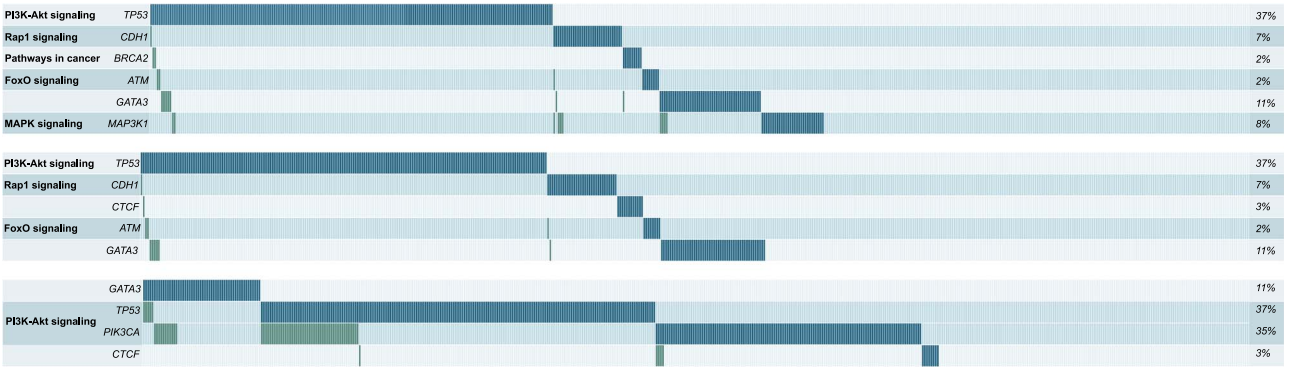

Figure 5. Mutation plot of the four gene solution of the Pareto set of Figure 4 with most promising gene set selections.

overlap (magenta rectangles). For these solutions, we provide the corresponding alteration plots in Figure 5. Moreover, similarly to the Multi-Dendrix approach, we provided also pathway subgrouping information. In these sets, AVAtar identified crucial genes for breast cancer as ATM [26, 33, 36, 66, 69, 112] and BRCA2 [42, 63, 72, 93, 102, 104] that were not found by the Multi-Dendrix approach. ATM is selected together with TP53 in two of our highlighted solutions. This supports our approach since ATM is widely reported to be mutually exclusive with TP53 in breast and also other types of cancers [39, 82, 91, 111]. Moreover, we could also select smaller sets with higher coverage on patients samples (72%). This set is better than the best set selected by the Multi-Dendrix approach in terms of coverage, overlap, and the number of genes selected. Thus, AVAtar empowers the user by providing all possible combination of the best trade-offs between coverage and mutual exclusivity. This possibility is of great relevance given that there is no commonly shared method to set a weight between these two conflicting objects.

Use case: vaccination targets for HNSCCs

In the following sections, we will further guide through the features of AVAtar by presenting the use case also available on the tutorial of the software. An in-depth medical description of the analysis performed on AVAtar can be found in [29]. The HNSCC analysis aims is to compare promising vaccination target sets for different subgroups of patients with HNSCC. Clinically, HNSCC shows distinct survival differences between the main three primary tumor sites: oral cavity (OC), oropharynx (OP) and larynx (L) [52]. Furthermore, HNSCC can be divided based on the main drivers of carcinogenesis: noxious agents (smoking, alcohol) or high-risk human papillomaviruses (HPV) [43, 107]. HPV-positive HNSCC is characterized by a much better prognosis compared to HPV-negative HNSCC, which has previously been shown for multiple treatment strategies [4, 62]. Shared cancer antigens, in particular cancer-testis antigens (CTA), could play an important role in future immunotherapy strategies for both HPV-negative and HPV-positive HNSCC [52, 95, 103].

Thus, a comprehensive analysis of the CTA repertoire as model antigens is needed to identify which antigens to target based on primary tumor site and HPV status. To rationalize vaccination efforts in clinical trials and to avoid the need for cumbersome individual testing of antigen expression, a semi-personalized off-the-shelve multi-antigen vaccine covering a high rate of the respective patient cohort is desired. In particular, for specific vaccination strategies, the number of target antigens that can reasonably be combined within one vaccine is limited. The task to find vaccination targets is formulated as multi-objective gene selection. Here, we applied coverage maximization and subset size minimization.

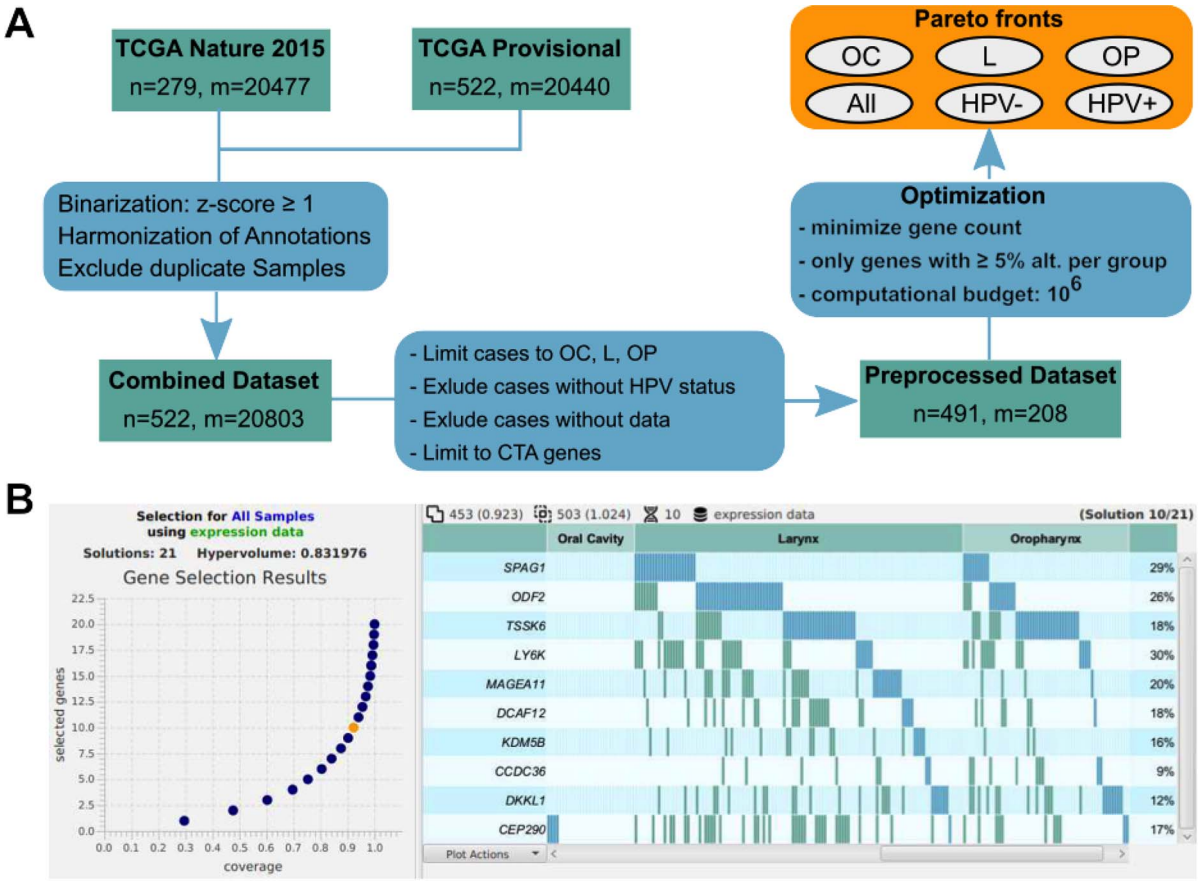

Figure 6. Preparation of the HNSCC analysis and AVAtar optimization result. (A) Analysis workflow for the HNSCC use case: data preparation steps and optimization settings. (B) Pareto set dialog of AVAtar. The found Pareto set (21 solutions) is shown on the left and a plot of the selected solution (orange circle) is shown on the right.

Dataset preparation and optimization setup for HNSCC analysis

AVAtar offers the possibility to pre-process data by deleting samples not relevant for the desired analysis, grouping samples according to clinical attributes, or filtering desired genes. More sophisticated preprocessing based on clinical data (e.g. harmonization of primary sites) can be accomplished by exporting the clinical data. Such editing and harmonization of clinical annotations are unique features of AVAtar.

In the specific use case, two publicly available large datasets of HNSCC patients (TCGA 2015 [75], TCGA provisional) have been combined for the analysis. The clinical annotation for HPV status is well defined in TCGA 2015 (> 1000 HPV E6/E7 RNA reads) but in TCGA provisional a surrogate marker for HPV-association was used (p16 immunohistochemistry). Therefore, the two datasets have been combined to obtain the well-defined HPV-status definition for the TCGA 2015 subset of patients. In our use case, the datasets have been downloaded on 25 October 2018. By accessing the preprocessing features of AVAtar cited above, the combined dataset has been prepared as follows (see also Figure 6A):

1. The expression data from the dataset TCGA 2015 have been inserted with alterations defined as overexpression using a threshold equal to the standard deviation.
2. The expression data from the dataset TCGA provisional (522 samples with expression profiles) have been imported with alterations defined as overexpression using a threshold equal to the standard deviation. Duplicate samples that are part of TCGA 2015 are excluded resulting in a total of 522 samples.
3. The primary tumor sites (clinical data) have been grouped and renamed for compliance with the other dataset. This has yielded the primary sites oral cavity, oropharynx, larynx and hypopharynx.
4. The HPV status (attribute HPV STATUS) for the samples from TCGA provisional has been determined based on the HPV-p16-status (attribute HPV STATUS P16) and primary site.
5. The samples have been grouped by the primary site. Samples with primary site lip ($m = 2$) and hypopharynx ($m = 10$) have been deleted.
6. The samples without an assigned HPV status ($m = 19$) have been deleted as well.
7. Genes without overexpression and non-CTA genes have been deleted.

The resulting dataset has 491 samples and 208 genes. Further, the user can proceed to perform a gene analysis to get mutually exclusive gene alterations within the selected and visualized subgroups. In our use case example, we ran the optimization with coverage maximization and gene count minimization on all samples and on the selected sample groups: oropharynx, larynx, oral cavity, HPV+ and HPV-. For the optimization of each sample group, we chose to consider only the genes with at least 5% alterations within that sample group. Optimization parameters for the HNSCC analysis are listed in Table 3. A fine-grained search with two swapped genes in expectation is used. This is compensated by a larger number of search steps (10^6). Due to

Figure 7. Optimization results for the HNSCC analysis. The objectives *maximal coverage* and *minimal number of genes* have been used in the optimization. (A) Alteration plot (overexpression) of the 10-gene solution resulting from the optimization for all samples. Clinical subgroups are shown. (B+C) The solution comparison tables show which genes are part (colored rectangles) of which solution (column). The solutions with at most 10 genes per primary site (B) and per HPV-status (C) are compared to the 10-gene solution for all samples.

the large computation budget of 10^6 iterations with 100 solutions resulting in 10^8 solution evaluations, the batch optimization mode of AVAtar has been used to perform the six optimization runs in parallel on a compute server.

Result visualization

An optimization using AVAtar has been performed for the whole cohort and the sample groups resulting from the clinical attributes of the primary tumor site and HPV status. For each Pareto-optimal set, the solution with the largest coverage and at most 10 genes have been selected. For the optimization within the whole cohort, a 10-gene solution with a coverage of 92.3% has been found (Figure 6B). The visualization of coverage and overlap by the primary site is displayed in Figure 6B and Figure 7A. However, the gene set optimized for coverage of all patients is dominated by OC patients, since these patients represent 64% of the cohort. This 10-gene set has coverage of 91.7% among OC, 96.5% among L and only 87.8% among OP indicating molecular differences among these primary sites. The optimizations for the different primary sites yield distinct gene sets (Figure 7B). Comparing these semi-personalized selections, it becomes evident that a selection of up to 10 genes optimized for the respective group of primary tumors results in an optimal coverage for the respective group with a suboptimal coverage in other primary sites. Since the OP cohort consisted primarily of HPV-positive patients (72.7%) in contrast to the other primary sites, this leads to the hypothesis that the main differences in the CTA repertoire may be due to HPV status. Optimizing for HPV status and comparing the gene selections of up to 10 genes for HPV-positive and HPV-negative patients, a distinct gene selection overlapping only in one gene can be observed (Figure 7C). The 10-gene selection found for the HPV-negative patients differs from the selection for all patients in five genes. The five-gene selection found for the HPV-positive patients contains completely distinct genes compared to all-patients selection and has only one common gene with the HPV-negative selection. An in-depth medical discussion of the obtained results can be found in [29].

Table 3. Optimization parameter setup of the use case. The parameters are given as the easier interpretable values from the 'simple setup' in AVAtar (second column) and the corresponding values from the algorithm description (third column)

Parameter	Value	Algorithm parameter value
Population size	100	$\mu = 100$
Solution combination count	5	$p_{cx} = 0.1$
Initially selected genes	20	$p_{sel} \approx 0.0962$
Swapped genes	2	$p_{mut} \approx 0.0096$
Selection pressure	10	$\tau = 10$
Search steps	10^6	$k = 10^6$

Conclusion

Nowadays, visualization and selection of gene alterations in large datasets are central issues in cancer research. In particular, in the context of providing intervention targets for personalized medicine approaches. Herewith, we revised available visualization and analysis tools and present AVAtar, our comprehensive software for gene visualization that also tackles the issue of target selection. An evolutionary algorithm is built into AVAtar to find trade-off solutions, which then can be explored interactively. Here, finding optimal gene subsets such as target selection for semi-personalized vaccination is formulated as a multi-objective optimization task. We further presented a walkthrough for the use of AVAtar by showing a real case scenario already applied in medical research [29]. The analysis of HNSCC expression data demonstrates the capabilities of AVAtar focusing on data import, visualization and optimization. In this context, we could show that subgroup-focused analysis can be performed with AVAtar by applying the optimization algorithm on different patient groups separately. If the optimization is executed separately for different clinical or molecular patient groups, distinct optimal gene selections become evident underlining the importance of subgroup-focused analyses in clinical trials. The diversity in the selected genes depending on the considered subgroups is in line with distinct molecular differences between HPV-positive and HPV-negative HNSCC [43, 50, 89]. Apart from the demonstrated use case, AVAtar can be used for explorative data analysis on binary or binarizable data, e.g. mutation, expression and methylation data. The interactive exploration of the trade-offs for the gene selection problem found by the optimization algorithm is a unique feature of AVAtar. To the best of our knowledge, AVAtar is the first software integrating visualization of alteration data and analysis via multi-objective optimization in an easily operable graphical user interface. This is complemented by the cBioPortal and Gene Expression Omnibus import functionality that provides access to a vast amount of published data. Finally, the gene selection optimization is a general method, which can be used for further research questions such as optimal gene selection for panel sequencing.

Key Points

- Overview of visualization and analysis tools for gene alterations in cancer with features and limitation. Potential improvement in the analysis and visualization of gene alterations.
- Introduced a comprehensive platform that integrates diverse aspects for gene alteration visualization and analysis. AVAtar provides a user-friendly interface that offers unique data-set processing features, gene selection analysis with an already implemented algorithm for multi-objective gene selection and exportable results.
- AVAtar was successfully applied to identify candidate vaccination targets for HNSCC in different sub-groups of patients.

Supplementary Material

Supplementary data are available at *Briefings in Bioinformatics* online.

Authors contribution

G.V., S.L., T.K.H., C.B. and V.G. conducted experiments and analyzed data. G.V. implemented the software. G.V., S.L., S.D.K., A.F., D.S.N., V.G., J.M.K., H.D. and N.I. analyzed data and prepared figures. G.V., S.L., N.I., J.M.K. and H.A.K. wrote the manuscript. H.A.K., N.I., S.L., H.D. and V.G. conceived the experiments and contributed to figure preparation.

Data availability

Datasets are provided on TCGA (https://portal.gdc.cancer.gov/projects/TCGA-LAML) and cBioPortal (http://www.cbioportal.org/study/summary?id=hnsc_tcga). AVAtar code and walkthrough is available at GitHub (https://github.com/sysbio-bioinf/avatar).

Funding

SFB 1074 (DFG) German Science Foundation (DFG, grant number 217328187); German Research Foundation GRK 2254 HEIST; Federal Ministry of Education and Research (BMBF, e:Med, CONFIRM, ID 01ZX1708C and TRANSCAN VI—PMTR-pNET, ID 01KT1901B).

References

1. Aguilera A, Gómez-González B. Genome instability: a mechanistic view of its causes and consequences. *Nat Rev Genet* 2008;**9**:204–17.
2. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics* 2004;**20**(4):578–80.
3. Alderton GK. Mutagenic clusters. *Nat Rev Cancer* 2012;**12**(7):452–3.
4. Ang KK, Harris J, Wheeler R, *et al.* Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med* 2010;**363**(1):24–35.
5. Babur Ö, Gönen M, Aksoy BA, *et al.* Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol* 2015;**16**(1):45.
6. Barrett T, Wilhite SE, Ledoux C, *et al.* NCBI GEO: archive for functional genomics data setsâŁupdate. *Nucleic Acids Res* 2013;**41**(Database issue):D991–5.

7. Bashashati A, Haffari G, Ding J, *et al*. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol* 2012;**13**(12): R124.
8. Beißarth T, Speed TP. GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* 2004;**20**(9):1464–5.
9. Brose MS, Volpe P, Feldman M, *et al*. BRAF and RAS mutations in human lung cancer and melanoma. *Cancer Res* 2002;**62**(23):6997–7000.
10. Brunicardi FC, Gibbs RA, Wheeler DA, *et al*. Overview of the development of personalized genomic medicine and surgery. *World J Surg* 2011;**35**(8):1693–9.
11. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;**455**(7216):1061.
12. Canisius S, Martens JWM, Wessels LFA. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol* 2016;**17**(1):261.
13. Cerami E, Demir E, Schultz N, *et al*. Automated network analysis identifies core pathways in glioblastoma. *PLoS One* 2010;**5**(2):e8918.
14. Cerami E, Gao J, Dogrusoz U, *et al*. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**(5):401–4.
15. Ciriello G, Cerami E, Sander C, *et al*. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* 2012;**22**(2):398–406.
16. Cisowski J, Sayin VI, Liu M, *et al*. Oncogene-induced senescence underlies the mutual exclusive nature of oncogenic KRAS and BRAF. *Oncogene* 2016;**35**(10):1328–33.
17. Cisowski J, Bergo MO. What makes oncogenes mutually exclusive? *Small GTPases* 2017;**8**(3):187–92.
18. Genome International Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**(6822):860–921.
19. Constantinescu S, Szczurek E, Mohammadi P, *et al*. TiMEx: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics* 2016;**32**(7):968–75.
20. Cormen TH, Leiserson CE, Rivest RL, and Stein C. *Introduction to Algorithms*, 3rd edn. Cambridge, MA, USA: The MIT Press, 2001.
21. Das K, Gunasegaran B, Tan IB, *et al*. Mutually exclusive FGFR2, HER2, and KRAS gene amplifications in gastric cancer revealed by multicolour FISH. *Cancer Lett* 2014;**353**(2):167–75.
22. Deb K. *Multi-Objective Optimization Using Evolutionary Algorithms*. NY, USA: John Wiley & Sons, Inc., 2001.
23. Deng Y, Luo S, Deng C, *et al*. Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. *Brief Bioinform* 2019;**20**(1):254–66.
24. Deng Y, Luo S, Deng C, *et al*. Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. *Brief Bioinform* 2019;**20**(1):254–66.
25. Di Cristofaro J, Marcy M, Vasko V, *et al*. Molecular genetic study comparing follicular variant versus classic papillary thyroid carcinomas: association of N-ras mutation in codon 61 with follicular variant. *Hum Pathol* 2006;**37**(7): 824–30.
26. Dörk T, Bendix R, Bremer M, *et al*. Spectrum of ATM gene mutations in a hospital-based series of unselected breast cancer patients. *Cancer Res* 2001;**61**(20):7608–15.
27. Etemadmoghadam D, Weir BA, Au-Yeung G, *et al*. Synthetic lethality between CCNE1 amplification and loss of BRCA1. *Proc Natl Acad Sci USA* 2013;**110**(48):19489–94.
28. Futreal AP, Coin L, Marshall M, *et al*. A census of human cancer genes. *Nat Rev Cancer* 2004;**4**(3):177–83.
29. Gangkofner DS, Holzinger D, Schroeder L, *et al*. Patterns of antibody responses to nonviral cancer antigens in head and neck squamous cell carcinoma patients differ by human papillomavirus status. *Int J Cancer* 2019;**145**(12): 3436–44.
30. Gao J, Aksoy BA, Dogrusoz U, *et al*. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**(269):pl1.
31. Grossmann S, Bauer S, Robinson PN, *et al*. Improved detection of overrepresentation of gene-ontology annotations with parent–child analysis. *Bioinformatics* 2007;**23**(22):3024–31.
32. Gundem G, Perez-Llamas C, Jene-Sanz A, *et al*. IntOGen: integration and data mining of multidimensional oncogenomic data. *Nat Methods* 2010;**7**(2):92–3.
33. Ho AY, Fan G, Atencio DP, *et al*. Possession of ATM sequence variants as predictor for late normal tissue responses in breast cancer patients treated with radiotherapy. *Int J Radiat Oncol Biol Phys* 2007;**69**(3):677–84.
34. Hou JP, Emad A, Puleo GJ, *et al*. A new correlation clustering method for cancer mutation analysis. *Bioinformatics* 2016;**32**(24):3717–28.
35. Hühne R, Kessler V, Fürstberger A, *et al*. 3D Network exploration and visualisation for lifespan data. *BMC Bioinform* 2018;**19**(390).
36. Iannuzzi CM, Atencio DP, Green S, *et al*. ATM mutations in female breast cancer patients predict for an increase in radiation-induced late effects. *Int J Radiat Oncol Biol Phys* 2002;**52**(3):606–13.
37. Ingle JN. Pharmacogenomics of endocrine therapy in breast cancer. *J Hum Genet* 2013;**58**(6):306–12.
38. Jerby-Arnon L, Pfetzer N, Waldman YY, *et al*. Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell* 2014;**158**(5):1199–209.
39. Jiang H, Reinhardt HC, Bartkova J, *et al*. The combined status of ATM and p53 link tumor development with therapeutic response. *Genes Dev* 2009;**23**(16):1895–909.
40. Junfei Z, Shihua Z, Wu L-Y, *et al*. Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* 2012;**28**(22):2940–7.
41. Kaelin WG. The concept of synthetic lethality in the context of anticancer therapy. *Nat Rev Cancer* 2005;**5**(9):689–98.
42. Karami F, Mehdipour P. A comprehensive focus on global spectrum of BRCA1 and BRCA2 mutations in breast cancer. *Biomed Res Int* 2013;**928562**:2013.
43. Keck MK, Zuo Z, Khattri A, *et al*. Integrative analysis of head and neck cancer identifies two biologically distinct HPV and three non-HPV subtypes. *Clin Cancer Res* 2015;**21**(4):870–81.
44. Hans A, Kestler LL, Lindner W, *et al*. On the fusion of threshold classifiers for categorization and dimensionality reduction. *Comput Stat* 2011;**26**:321–40.
45. Kestler HA, Müller A, Gress TM, *et al*. Generalized Venn diagrams: a new method of visualizing complex genetic set relations. *Bioinformatics* 2005;**21**(8):1592–5.

46. Kestler HA, Müller A, Kraus JM, *et al*. VennMaster: area-proportional Euler diagrams for functional GO analysis of microarrays. *BMC Bioinform* 2008;**9**(67).
47. Kim JW, Botvinnik OB, Abudayyeh O, *et al*. Characterizing genomic alterations in cancer by complementary functional associations. *Nat Biotechnol* 2016;**34**(5): 539–46.
48. Kim Y-A, Cho D-Y, Dao P, *et al*. MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics* 2015;**31**(12):i284–92.
49. Kim Y-A, Madan S, Przytycka TM. WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics* 2017;**33**(6):814–21.
50. Kostareli E, Holzinger D, Bogatyrova O, *et al*. HPV-related methylation signature predicts survival in oropharyngeal squamous cell carcinomas. *J Clin Invest* 2013;**123**(6):2488–501.
51. Kumar N, Rehrauer H, Cai H, *et al*. CDCOCA: a statistical method to define complexity dependence of co-occurring chromosomal aberrations. *BMC Med Genomics* 2011;**4**:21.
52. Laban S, Giebel G, Klümper N, *et al*. MAGE expression in head and neck squamous cell carcinoma primary tumors, lymph node metastases and respective recurrences-implications for immunotherapy. *Oncotarget* 2017;**8**(9):14719–35.
53. Lausser L, Schmid F, Platzer M, *et al*. Semantic multi-classifier systems for the analysis of gene expression profiles. *Arch Data Sci Ser A* 2016;**1**(1):157–76.
54. Lawrence MS, Stojanov P, Polak P, *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;**499**:214–8.
55. Mark DM, Leiserson DB, Sharan R, *et al*. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol* 2013;**9**(5):e1003054.
56. Leiserson MDM, Blokh D, Sharan R, *et al*. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol* 2013;**9**(5):e1003054.
57. Leiserson MDM, Gramazio CC, Hu J, *et al*. MAGI: visualization and collaborative annotation of genomic aberrations. *Nat Methods* 2015;**12**:483–4.
58. Leiserson MDM, Reyna MA, Raphael BJ. A weighted exact test for mutually exclusive mutations in cancer. *Bioinformatics* 2016;**32**(17):i736–45.
59. Leiserson MDM, Vandin F, Wu H-T, *et al*. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* 2015;**47**(2):106–14.
60. Leiserson MDM, Wu H-T, Vandin F, *et al*. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol* 2015;**16**(1):160.
61. Li H-T, Zhang Y-L, Zheng C-H, *et al*. Simulated annealing based algorithm for identifying mutated driver pathways in cancer. *BioMed Res Int* 2014;**375980**:2014.
62. Licitra L, Perrone F, Bossi P, *et al*. High-risk human papillomavirus affects prognosis in patients with surgically treated oropharyngeal squamous cell carcinoma. *J Clin Oncol* 2006;**24**(36):5630–6.
63. Liede A, Malik IA, Aziz Z, *et al*. Contribution of BRCA1 and BRCA2 mutations to breast and ovarian cancer in Pakistan. *Am J Hum Genet* 2002;**71**(3):595–606.
64. Loupakis F, Moretto R, Aprile G, *et al*. Clinico-pathological nomogram for predicting BRAF mutational status of metastatic colorectal cancer. *Br J Cancer* 2016;**114**(1):30–6.
65. Lu S, Lu KN, Cheng S-Y, *et al*. Identifying driver genomic alterations in cancers by searching minimum-weight, mutually exclusive sets. *PLoS Comput Biol* 2015;**11**(8):e1004257.
66. Maillet P, Bonnefoi H, Vaudan-Vutskits G, *et al*. Constitutional alterations of the ATM gene in early onset sporadic breast cancer. *J Med Genet* 2002;**39**(10):751–3.
67. Melamed RD, Wang J, Iavarone A, *et al*. An information theoretic method to identify combinations of genomic alterations that promote glioblastoma. *J Mol Cell Biol* 2015;**7**(3):203–13.
68. Melamed RD, Wang J, Iavarone A, *et al*. An information theoretic method to identify combinations of genomic alterations that promote glioblastoma. *J Mol Cell Biol* 2015;**7**(3):203–13.
69. Meyer A, John E, Dörk T, *et al*. Breast cancer in female carriers of ATM gene alterations: outcome of adjuvant radiotherapy. *Radiother Oncol* 2004;**72**(3):319–23.
70. Müller A, Holzmann K, Kestler HA. Visualization of genomic aberrations using Affymetrix SNP arrays. *Bioinformatics* 2006;**23**(4):496–7.
71. Mulligan G, Lichter DI, Di Bacco A, *et al*. Mutation of NRAS but not KRAS significantly reduces myeloma sensitivity to single-agent bortezomib therapy. *Blood* 2014;**123**(5):632–9.
72. Narod SA, Salmena L. BRCA1 and BRCA2 mutations and breast cancer. *Discov Med* 2011;**12**(66):445–53.
73. Nebro AJ, Durillo JJ, Vergne M. Redesigning the jMetal multi-objective optimization framework. In: *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation, GECCO Companion '15*. NY, USA: ACM, 2015, pp. 1093–100.
74. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet* 2012;**13**(9):667–72.
75. The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 2015;**517**:576–82.
76. Nielsen CB, Cantor M, Dubchak I, *et al*. Visualizing genomes: techniques and challenges. *Nat Methods* 2010;**7**(3): S5–S15.
77. Ordulu Z, Kammin T, Brand H, *et al*. Structural chromosomal rearrangements require nucleotide-level resolution: lessons from next-generation sequencing in prenatal diagnosis. *Am J Hum Genet* 2016;**99**(5):1015–33.
78. Paull EO, Carlin DE, Niepel M, *et al*. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* 2013;**29**(21):2757–64.
79. Perez-Llamas C, Lopez-Bigas N. Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PLoS One* 2011;**6**(5):e19541.
80. Pulido-Tamayo S, Weytjens B, De Maeyer D, *et al*. SSA-ME detection of cancer driver genes using mutual exclusivity by small subnetwork analysis. *Sci Rep* 2016;**6**: 36257.
81. Reeves CR. *Handbook of Metaheuristics, Chapter Genetic Algorithms*. New York, USA: Springer US, 2010, 109–39.
82. Reinhardt HC, Jiang H, Hemann MT, *et al*. Exploiting synthetic lethal interactions for targeted cancer therapy. *Cell Cycle* 2009;**8**(19):3112–9.
83. Remy E, Rebouissou S, Chaouiya C, *et al*. A modeling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis. *Cancer Res* 2015;**75**(19):4042–52.

84. Robinson JT, Thorvaldsdóttir H, Winckler W, *et al*. Integrative genomics viewer. *Nat Biotechnol* 2011;**29**(1):24–6.
85. Sanborn JZ, Benz SC, Craft B, *et al*. The UCSC cancer genomics browser: update 2011. *Nucleic Acids Res* 2011;**39**(Database-Issue):D951–9.
86. Schnattinger T, Schöning U, Marchfelder A, *et al*. RNA-Pareto: interactive analysis of Pareto-optimal RNA sequence-structure alignments. *Bioinformatics* 2013;**29**(23):3102–4.
87. Schroeder MP, Gonzalez-Perez A, Lopez-Bigas N. Visualizing multidimensional cancer genomics data. *Genome Med* 2013;**5**(9).
88. Schroeder MP, Gonzalez-Perez A, Lopez-Bigas N. Visualizing multidimensional cancer genomics data. *Genome Med* 2013;**5**(1):9.
89. Seiwert TY, Zuo Z, Keck MK, *et al*. Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. *Clin Cancer Res* 2015;**21**(3):632–41.
90. Seshagiri S, Stawiski EW, Durinck S, *et al*. Recurrent R-spondin fusions in colon cancer. *Nature* 2012;**488**(7413):660–4.
91. Shaheen M, Allen C, Nickoloff JA, *et al*. Synthetic lethality: exploiting the addiction of cancer to DNA repair. Blood. *J Am Soc Hematol* 2011;**117**(23):6074–82.
92. Sherman BT, Lempicki RA, Huang DW. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;**4**(1):44.
93. Shih HA, Nathanson KL, Seal S, *et al*. BRCA1 and BRCA2 mutations in breast cancer families with multiple primary cancers. *Clin Cancer Res* 2000;**6**(11):4259–64.
94. Srihari S, Singla J, Wong L, *et al*. Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. *Biol Direct* 2015;**10**(57).
95. Stevanović S, Pasetto A, Helman SR, *et al*. Landscape of immunogenic tumor antigens in successful immunotherapy of virally induced epithelial cancer. *Science* 2017;**356**(6334):200–5.
96. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Nat Acad Sci USA* 2005;**102**(43):15545–50.
97. Szczurek E, Beerenwinkel N. Modeling mutual exclusivity of cancer mutations. *PLoS Comput Biol* 2014;**10**(3):e1003503.
98. Taudien S, Lausser L, Giamarellos-Bourboulis EJ, *et al*. Genetic factors of the disease course after sepsis: rare deleterious variants are predictive. *EBioMedicine* 2016;**12**:227–38.
99. The Cancer Genome Atlas Research Network, Ley TJ, Miller C, *et al*. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* 2013;**368**(22):2059–74.
100. Torkamani A, Schork NJ. Identification of rare cancer driver mutations by network reconstruction. *Genome Res* 2009;**19**(9):1570–8.
101. Unni AM, Lockwood WW, Zejnullahu K, *et al*. Evidence that synthetic lethality underlies the mutual exclusivity of oncogenic KRAS and EGFR mutations in lung adenocarcinoma. *Elife* 2015;**4**:e06907.
102. Vahteristo P, Eerola H, Tamminen A, *et al*. A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families. *Br J Cancer* 2001;**84**(5):704–8.
103. van der Burg SH, Arens R, Ossendorp F, *et al*. Vaccines for established cancer: overcoming the challenges posed by immune evasion. *Nat Rev Cancer* 2016;**16**(4):219–33.
104. Van der Looij M, Szabo C, Besznyak I, *et al*. Prevalence of founder BRCA1 and BRCA2 mutations among breast and ovarian cancer patients in Hungary. *Int J Cancer* 2000;**86**(5):737–40.
105. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. *Genome Res* 2012;**22**(2):375–85.
106. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. *Genome Res* 2012;**22**(2):375–85.
107. Varier I, Keeley BR, Krupar R, *et al*. Clinical characteristics and outcomes of oropharyngeal carcinoma related to high-risk non-human papillomavirus16 viral subtypes. *Head Neck* 2016;**38**(9):1330–7.
108. Vaske CJ, Benz SC, Sanborn ZJ, *et al*. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 2010;**26**(12):i237–45.
109. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med* 2004;**10**(8):789–99.
110. Wang R, Perez-Riverol Y, Hermjakob H, *et al*. Open source libraries and frameworks for biological data visualisation: a guide for developers. *Proteomics* 2015;**15**(8):1356–74.
111. Wang X, Simon R. Identification of potential synthetic lethal genes to p53 using a computational biology approach. *BMC Med Genomics* 2013;**6**(1):30.
112. Weigelt B, Bi R, Kumar R, *et al*. The landscape of somatic genetic alterations in breast cancers from ATM germline mutation carriers. *J Natl Cancer Inst* 2018;**110**(9):1030–4.
113. Wendl MC, Wallis JW, Lin L, *et al*. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* 2011;**27**(12):1595–602.
114. Wong CK, Vaske CJ, Ng S, *et al*. The UCSC Interaction Browser: multidimensional data views in pathway context. *Nucleic Acids Res* 2013;**41**(W1):W218–24.
115. Yamamoto H, Shigematsu H, Nomura M, *et al*. PIK3CA mutations and copy number gains in human lung cancers. *Cancer Res* 2008;**68**(17):6913–21.
116. Yeang C-H, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J* 2008;**22**:2605–22.
117. Zhang H, Meltzer PS, Davis SR. caOmicsV: an R package for visualizing multidimensional cancer genomic data. *BMC Bioinform* 2016;**17**(141).
118. Zhang J, Wu L-Y, Zhang X-S, *et al*. Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics* 2014;**15**(1):271.
119. Zhang J, Zhang S, Wang Y, and Zhang X-S. Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. *BMC Syst Biol*, **7**(Suppl 2):S4, 2013.