Published online 6 December 2019

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 1 **1**
doi: 10.1093/nargab/lqz019

Omics Playground: a comprehensive self-service platform for visualization, analytics and exploration of Big Omics Data

Murodzhon Akhmedov[1,2,3], Axel Martinelli[1,3], Roger Geiger[1] and Ivo Kwee[1,3,*]

[1]Institute for Research in Biomedicine, Faculty of Biomedical Sciences, Università della Svizzera Italiana, 6500 Bellinzona, Switzerland, [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland and [3]BigOmics Analytics, 6500 Bellinzona, Switzerland

Received November 08, 2019; Editorial Decision November 12, 2019; Accepted November 19, 2019

ABSTRACT

As the cost of sequencing drops rapidly, the amount of 'omics data increases exponentially, making data visualization and interpretation—'tertiary' analysis a bottleneck. Specialized analytical tools requiring technical expertise are available. However, consolidated and multi-faceted tools that are easy to use for life scientists is highly needed and currently lacking. Here we present Omics Playground, a user-friendly and interactive self-service bioinformatics platform for the in-depth analysis, visualization and interpretation of transcriptomics and proteomics data. It provides a large number of different tools in which special attention has been paid to single cell data. With Omics Playground, life scientists can easily perform complex data analysis and visualization without coding, and significantly reduce the time to discovery.

INTRODUCTION

The current progress in sequencing technologies is leading to an exponential increase in the amount of high throughput data generated. In particular gene expression data, in the shape of microarrays and RNA-seq, are now abundant. As technologies have become affordable, the bottleneck is not anymore the availability of the data but the analysis of it.

The current landscape of bioinformatics tools consist of a plethora of free software packages and stand-alone web services that provide a specific bioinformatic analysis. Bioinformaticians would juggle their data between websites and create custom scripts to glue the packages together, but because this required some programming skills, it was not the realm of the biologists. However, in recent years, the development of easy-to-use self-service bioinformatics (SSB) platforms, targeted at biologists with no previous bioinformatics experience, have made such processes increasingly more accessible.

As big omics data continues to grow and more bioinformatic analysis is needed, the demand for easy-to-use SSB platforms will expand. To this end, we have developed the Omics Playground, a self-service bioinformatics platform for the visualization, analysis and exploration of big omics data.

Currently available platforms. Self-service bioinformatics (SSB) platforms are typically those that (i) target biologists with no or little bioinformatics skills as users, (ii) provide an integrated solution for end-to-end analysis and (iii) have a high degree of interactivity and visualization. A number of SSB platforms have been developed over the years to address the analysis of not only RNA-seq data, but also other type of omics data, such as DNA-seq, CHIP-seq and proteomics data (Table 1). Among the free SSB platforms, DEBrowser (1) and Biojupies (2) are good examples of integrated platforms for RNA-seq data analysis, and, in the case of BioJupies, provides the possibility to access and analyze previously published data sets through the GEO (3). The WIlsON platform (4) is an example of a versatile platform, which is agnostic in the type of data supporting multiple types of omics data including genomic, transcriptomic, metabolomic and proteomic data (Table 1). Among the commercial platforms, Genialis (genialis.com) and Rosalind (onramp.bio/rosalind) support various types of omics data, while AIR (transcriptomics.sequentiabiotech.com) is solely focused on RNA-seq data. A particular case is represented by Paintomics3 (5), which not only supports various types of omics data but also, uniquely among the platforms in the list, integrates multi-omics data from the same experiment in a graphical display of KEGG pathways.

Self-service platforms can also be characterized based on three levels of analysis: (i) mapping and quantification (primary analysis), (ii) statistical testing (secondary analysis) and (iii) data visualization and interpretation (tertiary analysis). In contrast to the typically linear workflow of first and

*To whom correspondence should be addressed. Tel: +41 91 820 0368; Email: kwee@bigomics.ch

Table 1. The feature comparison of Omics Playground with available platforms in the literature. The '✓' symbol represents the availability of the feature

		Playground	DEBrowser	DEapp	BioJupies	WIlsoN	Network An.	ASAP	Paintomics3	Shiny NGS	Rosalind	Genialis	AIR
General features	Development	R	R	R	R, Python	R	R, Javascript, HTML5	R, Java, Python	R, Python	R	NA	NA	NA
	Supported DE algorithms	7	3	3	1	0	3	4	0	0	1	3	4
	Supported species	2	8	0	2	0	17	69	58	NA	16	3	45000
	Gene set databases supported	68	3	0	143***	0	5	3	1	9	4	1	1
	Single Cell RNA-seq support	✓						✓					
	Proteomics	✓				✓			✓				
Primary an.	Fastq Input	✓*			✓		✓**				✓	✓	✓
	Low Counts filtering	✓*	✓		✓		✓	✓					
	Batch effect correction	✓*	✓				✓	✓					
	Density plots	✓	✓				✓						
Secondary analysis	IQR Plots	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	
	PCA/MDS	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓
	tSNE	✓					✓	✓					
	Combine DE algorithm results	✓*											✓
	Multiple comparisons	✓*					✓	✓	✓				
	Meta-analysis (multp. expr. tables)	✓*	✓				✓				✓		
	Scatter Plots	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓
	Heatmaps	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
	Gene information hyperlinks	✓			✓		✓	✓	✓	✓		✓	
Thirtiary analysis	KEGG graphical display	✓							✓				
	Gene Set Barcode Plot	✓								✓			
	CMap access	✓			✓								
	Public datasets links	✓*			✓								
	Interaction Networks visualisation	✓					✓		✓				
	Time-Series						✓		✓				
	Intersection between contrasts	✓								✓	✓	✓	
	Regulatory Omics			✓	✓	✓			✓	✓	✓		
	Region-based Omics		✓	✓		✓			✓	✓	✓	✓	
	Biomarker selection	✓											
	Survival analysis	✓											
	Combine multi-omics analyses								✓				

* via separate R script ** via Galaxy (galaxyproject.org) *** via Enrichr (amp.pharm.mssm.edu/Enrichr)

second level analysis, tertiary analysis needs a high level of interactivity. Commercial platforms, such as Genialis, Rosalind and AIR, tend to offer pre-defined pipelines for primary and secondary analysis, while open source platforms, such as DEBrowser, DEapp (6) and ASAP (7) focus more on customizable secondary and tertiary analysis. Consequently, most of the commercial platforms in Table 1 readily accept raw FASTQ files as input and automate downstream count matrix preparation. Open source platforms, on the other hand, often require a flat text file containing a count matrix of the features being analysed, which needs to be produced separately. The WIlsON platform requires a specific input file format called CLARION. An exception is represented here by Biojupies and Network Analyst (8), which can accept raw sequence files (in the FASTQ format) as input and produce count matrices based on them (although via a Galaxy platform in the case of Network Analyst). Finally, platforms devoted exclusively to visualization, such as Paintomics3, focus on tertiary analysis only.

Along with data sets getting larger (i.e. more samples), the number of possible contrasts grows quadratically with respect to the number of conditions. Support for multiple contrasts is not a common feature, with most platforms either focusing on single pairwise comparisons. Network Analyst stands out in the list by offering a more thorough comparison analysis, extending to nested comparisons, as well as supporting time series (which can also be visualized in Paintomics3).

Other distinguishing features of SSB platforms are the number of supported species, the number of available expression analysis algorithms, and the number of gene data sets for enrichment score analysis supported by each platform. Platforms that focus on fewer, popular species (usually mouse and human) and for which more gene data sets are available, tend to also offer more gene enrichment analysis options. Biojupies and DEBrowser are good examples of this strategy and include multiple pathway databases, the GTEx database (9), as well as comparing expression profiles against those of perturbagens through the Connectivity Map (CMap) database (10). Platforms that support more species, such as AIR, Rosalind, Paintomics3 or ASAP, tend to focus on fewer pathways (usually the KEGG pathways and GO terms).

With the advent of single cell sequencing, support for single cell RNA-seq data sets is also becoming an increasingly desirable feature for SSB platforms. These are explicitly supported only by ASAP at the moment among the platforms considered (Table 1).

The Omics Playground. The Omics Playground platform offers a unique combination of features that distinguishes it from the other SSB platforms currently available (Supplementary Figure S1). We believe that data preprocessing (primary analysis) and statistical testing (secondary analysis) are now well established, and the most challenging task is currently data interpretation (tertiary analysis) that often takes the longest time but where actual insights can be gained. Therefore, the Omics Playground focuses strongly on tertiary analysis while providing good support for secondary analysis.

The Omics Playground currently handles gene expression microarray, RNA-seq and LC-MS/MS proteomics data, and supports two species, human and mouse. The Omics Playground has been in particular devised to also support single cell RNA-seq data (like the ASAP platform), as well as traditional gene expression experiments.

The platform combines the differential expression analysis with up to seven different algorithms, including the popular limma (11), edgeR (12) and DEseq2 (13) packages. The enrichment of more than 50 000 gene sets from vari-

ous databases is computed using multiple methods including Fisher's exact test, ssGSEA, GSVA, Spearman correlation, camera and fry (14–18). Omics Playground offers the second largest number of gene set databases within the platforms in Table 1, with only BioJupies offering more. Multiple statistical methods are combined using meta-analysis, providing a list of highly reliable hits identified across algorithms. A similar feature is provided only by the commercial AIR platform among the ones presented in the table.

Furthermore, Omics Playground offers a graphical display of individual gene expression profiles on KEGG pathways images, a feature that is only present in Paintomics3. Like ShinyNGS (github.com/pinin4fjords/shinyngs), the platform also displays gene sets barcode plots (Table 1). Similarly to BioJupies, a drug connectivity map (CMap) provides a visual tool for comparing expression profiles to find potentially relevant similar and opposing signature across more than 5000 perturbagens from the L1000 database. The platform also has a special module for immune cell profiling and, uniquely among all the other platforms, modules performing both biomarker selection and survival analysis.

MATERIALS AND METHODS

Implementation

The current version of our platform (1.0) is implemented in R (19) using the Shiny web application framework (20). The overview of the platform is shown in Figure 1 and consists of two main components. The first component addresses the data importing and precomputation tasks offline, while the second component hosts an interface framework that supports real-time visualization and interaction with users. The following sections describe the features of each component.

Data import and precomputation

The data import and precomputation involves preparing the input data through filtering, normalizing and precomputing statistics for some analyses and importing it into the platform. The data cleaning and precomputation is performed offline to support real-time interaction by minimizing user interface latency (Supp. Doc. Chapters 3 and 4).

Data import. Users can import their transcriptomics or proteomics data to the platform by either uploading the data through the interface or preparing an input object using scripts. For uploading, the platform requires the counts, samples information, genes information and contrasts tables in CSV format. Users can provide their own counts or download the relevant data from repositories such as GEO (3), and arrange other files accordingly. On the other hand, an input object can be prepared using scripts from different types and formats of data, including FASTQ, where users can implement their preferred alignment or quantification methods (21–23). With scripts it is also possible to do more detailed data cleaning, filtering, normalization and preprocessing. The platform contains necessary example scripts for an input object preparation.

Filtering. The data preprocessing includes some filtering criteria, such as filtering of genes based on variance, the expression across the samples, and the number of missing values. Similarly, samples can also be filtered based on the read quality, total abundance, unrelated phenotype, or an outlier criterion.

Normalization. The raw counts are converted into counts per million (CPM) and log2. Depending on the data set, a quantile normalization can be applied. Known batches in the data can be corrected with limma (11) or ComBat (24). Other unknown batch effects and unwanted variation can be further removed using surrogate variable analysis in the sva package (25).

Offline computation. Statistics for the differentially expressed genes (DEG) and gene set enrichment (GSE) analyses are precomputed to accelerate the visualization on the interface.

Omics Playground interface

The interface of the platform is subdivided into basic and expert modes. Basic mode includes fundamental analysis modules such as data view, clustering, differential expression, gene set enrichment, intersection and functional analyses, while expert mode includes additional modules such as signature, biomarker and single-cell profiling. Users can choose the interface mode according to their level of expertise. The main purpose of having two different modes is to provide a customizable experience suited to each users background.

The platform contains nine main functional modules. After selecting and loading the data from the home page, users can proceed with any analysis on desire. There is no specific order between the analysis modules that users should follow, as most of the statistics are precomputed offline in the previous steps. A brief description and functionality of each module is provided below.

Home. The platform starts running from the `home` panel. Basically, this module contains general information about all available data sets. For each data set, this tab reports a brief description as well as the total number of samples, genes, gene sets (or pathways), the corresponding phenotypes and the collection date. Users can choose the interface mode, select and load the public data of their interest, or upload their own data and start the analysis from here (Supp. Doc. Chapter 6).

Data view. For the selected data set, the `data view` module provides a descriptive statistical analysis at a gene level with visualizations (Supp. Doc. Chapter 7). For a gene specified by the user, the `plot` section displays figures related to the expression level of the gene, correlation with other genes, and average expression ranking within the data set. It also correlates the gene with other gene expressions in data sets such as ImmProt (26) and HPA (27), and plots the cumulative correlation. Furthermore, tissue expression for a selected gene is displayed using the GTEx database. For further information from the literature, hyperlinks are provided to link the selected gene to databases like OMIM (28),

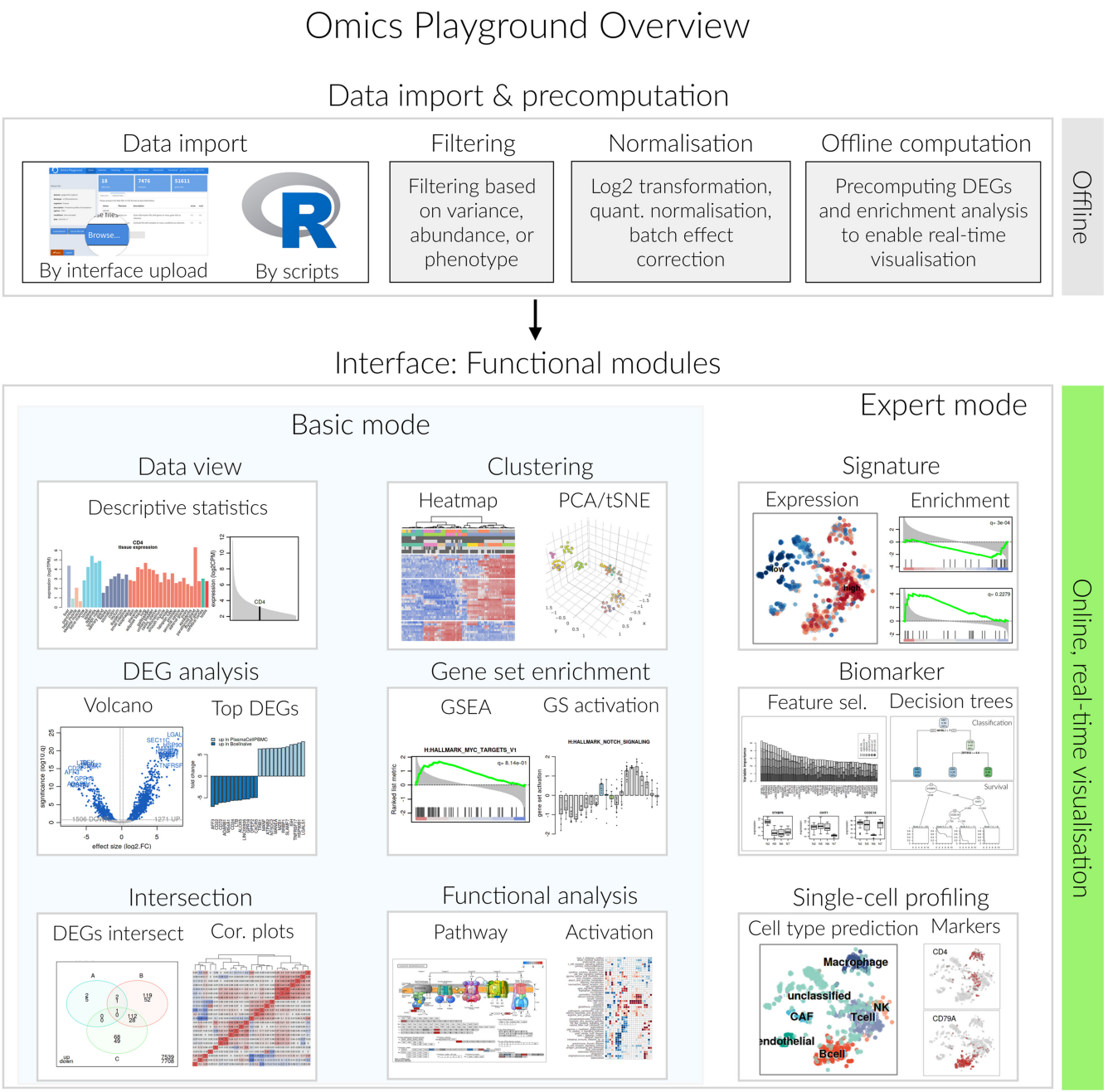

Figure 1. An overview of the Omics Playground. The platform consists of data cleaning and preprocessing and a user interface. Data preprocessing is handled offline to enable real-time visualization and interaction on the interface.

KEGG (29), and GO (30). In the visual analysis, users can filter out some samples or collapse the samples by phenotype class. It is also possible to visualize the information on a raw count level (CPM) instead of a log2 level (logCPM).

The total number of counts (abundance) per sample and their distribution among the samples are displayed in the `counts` section. For each sample, the user can also see the percentage of counts for major gene types, such as CD molecules, kinases or RNA-binding motifs.

Further correlation analysis across, the samples can be performed under the `gene table` section, where genes are ordered in the table according to the correlations with the selected gene. The gene-wise average expression of samples per phenotype classes is also presented in the table. More detailed information about the samples is reported in the `sample table`.

Clustering. The clustering module performs a holistic clustering analysis of the samples (Supp. Doc. Chapter 8). The main output of this feature is 2-fold: (i) It generates a `heatmap` of samples and (ii) It also provides a `PCA/tSNE`

plot of samples obtained by principal components analysis or t-distributed stochastic embedding algorithms (31,32).

The heatmap analysis can be performed on a gene level expression or gene set level expression in which, for each gene set (or pathway), an average expression is computed from the gene expression data using summary methods such as GSVA and ssGSEA (15). During the heatmap generation, users have various option that they can select, such as splitting the samples by a phenotype class provided in the data (e.g. tissue, cell type or gender). In addition, users have to specify the top $N = 50$, 500 features to be used in the heatmap for hierarchical clustering. The criteria to select the top features are: (i) sd - features with the highest standard deviation across all the samples, (ii) specific features that are overexpressed in each phenotype class compared to the rest, (iii) pca—principal components computed by the irlba package (31). The top features in the heatmap are then divided into five clusters based on their expression profiles. For each cluster, the platform provides a functional annotation under the annotate cluster section using more than 42 published reference databases, including but not limited to well-known databases such as MSigDB, KEGG and GO (29,30,33).

PCA and t-SNE plots can be found in the PCA/tSNE tab, which shows the relationship between samples in 2D as well as in 3D space for visual analytics. Users can customize the PCA/tSNE plot using a phenotype class provided in the data.

Differential expression analysis. The expression module contains a differentially expressed genes (DEG) analysis between contrasts (e.g. tumor versus control) (Supp. Doc. Chapter 9). The analysis begins with the selection of a contrast. There are further options to filter out some genes by functional families, logarithmic fold change (logFC) and false discovery rate (FDR).

DEG analysis is performed using four commonly accepted methods, namely: *t*-test (standard, Welch), limma (no trend, trend, voom), edgeR (QLF, LRT), and DESeq2 (Wald, LRT) (11–13). For each selected contrast, the results of these methods are combined and reported under the table section, where meta.q for a gene represents the highest q value among the methods and the number of starts indicates how many methods have significant q values ($q < 0.05$). Users can sort genes by logFC, meta.q, or average expression in an interactive table. By clicking on a gene, it is possible to see which gene sets include that gene, and check the status of the differential expression in other comparisons from the plots section. The section can also display volcano and MA plots. Furthermore, for the top 10 DEGs within the selected comparison, average expression plots across the samples are displayed in the top genes section.

Another important feature of this module is the simultaneous visualization of volcano plots for all comparisons under the volcano (all) section.

Gene set enrichment analysis. This module visualizes a differential expression analysis at a gene set level (Supp. Doc. Chapter 10). Expression analysis for each gene set (or pathway) is computed from gene expression data using summary methods such as GSVA and ssGSEA (15). The platform has more than 50 000 gene sets and pathways in total, which are divided into 30 gene set collections such as Hallmark, MSigDB, KEGG and GO (29,30,34).

Users specify which contrast they want to visually analyze using a particular gene set collection. To ensure statistical reliability, the platform performs GSE analyses using seven different methods, including Spearman rank correlation, GSVA, ssGSEA, Fisher's exact test, GSEA, camera and fry (14–18). The results are combined and gene sets can be optionally filtered by logFC and FDR thresholds before being visualized in an interactive and sortable table under the enrichment table menu. For each gene set, a meta.q value and stars qualifier are calculated as described previously and volcano plots of its genes and barplots of expressions per phenotype class are displayed (under plots). Additionally, the list of genes in that gene set are visualized in a separate table and for every gene it is possible to see the barplot of expressions per phenotype class and a scatter plot of gene to gene set expressions. Individual gene sets expression profiles can be visualized against all available contrasts (compare tab). Under the volcano (all) tab, volcano plots for all contrasts are displayed.

Functional analysis. This module provides higher level functional and visual analysis of the contrast space using the KEGG and GO graph structures (Supp. Doc. Chapter 11). Given the profile of a particular contrast, it also searches for the closest drug profiles from the L1000 drug expression database (10).

Within the KEGG graph section, each pathway is scored for the selected contrast profile and reported in an interactive table. The scoring is performed by considering the total number of genes in the pathway (n), the number of genes in the pathway supported by the contrast profile (k), the ratio of k/n, and the ratio of lupregulated or downregulated genesl/k. Additionally, the table contains the list of the upregulated and downregulated genes for each pathway and a q value from the Fisher's test for the overlap. Pathway maps can be summoned from the interactive table, with individual genes colored according to their differential expression (upregulation: red; downregulation: blue). Another important feature is an activation-heatmap including the comparison of activation levels of pathways (or pathway keywords) across multiple contrast profiles.

All the features described under the KEGG graph tab, such as scoring the gene sets and drawing an activation-heatmap, can be performed for the GO database under the GO graph tab. Instead of pathway maps, an annotated graph structure provided by the GO database is plotted for every selected gene set.

The drug connectivity map (Drug C-Map) section correlates the selected contrast profile with more than 5000 known drug profiles from the L1000 database (10), and shows the top 10 similar and opposite profiles by running the GSEA algorithm (17) on the contrast-drug profile correlation space. It also provides an activation-heatmap for drugs across multiple contrast profiles. Users can perform the contrast-drug profile correlation analysis in mono (single drug) or combo (combination of two drugs) mode.

Intersection analysis. The intersection analysis module enables users to compare multiple contrasts by intersecting the genes of profiles. Its main goal is to identify contrasts showing similar profiles (Supp. Doc. Chapter 12).

For the selected contrasts, the platform provides volcano plots and pairwise correlation plots between the profiles under the `pairs` tab. Simultaneously, it plots a Venn diagram with the number of intersecting genes between the profiles in the `venn diagram` section. The list of intersecting genes with further details is also reported in an interactive table, where users can select and remove a particular contrast from the intersection analysis. In addition, it is possible to check a scatter plot of two profiles as well as the correlation-heatmap of multiple profiles under the `two-pairs` and `correlation` tabs, respectively. The `connectivity` graph tab constructs a network, in which nodes represent contrasts and edges are obtained from the pairwise-correlation of corresponding profiles.

Signature analysis. In this module, users can test gene signatures by calculating an enrichment score. They can use a sample list provided on the platform or upload their own gene list. Instead of a short list, a profile can also be selected, which is a complete gene list derived from one of the contrasts in the analysis (Supp. Doc. Chapter 13).

After uploading a gene list, the `markers` section produces a t-SNE plot of samples for each gene, colored by expression levels (upregulation: red; downregulation: blue). The `enrichment` tab performs the enrichment analysis of the gene list against all contrasts by running the GSEA algorithm (17) and plots enrichment outputs. The `signature c-map` section associates the provided signature list or contrast profile with similar profiles of other experiments, obtained from ten published data sets. Finally, under the `overlap/similarity` tab, users can compare their gene list with all the gene sets and pathways in the platform through statistics such as the total number of genes in the gene set (K), the number of intersecting genes between the list and the gene set (k), the overlapping ratio of k/K, as well as the p and q values by the Fisher's test for the overlap test.

Biomarker analysis. This module performs the biomarker selection that can be used for classification or prediction purposes (Supp. Doc. Chapter 14). To better understand which genes, mutations, or gene sets influence the final phenotype the most, Playground calculates a variable importance score for each feature using state-of-the-art machine learning algorithms, including LASSO (35), elastic nets (36), random forests (37), and extreme gradient boosting (38), and provides the top 50 features according to cumulative ranking by the algorithms. By combining several methods, the platform aims to select the best possible biomarkers. The phenotype of interest can be multiple categories (classes) or patient survival data. Instead of choosing a phenotype, users can also specify a particular contrast from the analysis and perform biomarker selection.

The platform also provides a heatmap of samples based on identified top features. In addition, it generates a classification tree or survival tree using top features and provides expression boxplots by phenotype classes for features present in the tree.

Cell profiling. The cell profiling module is used to infer cell types in a sample using prediction methods and reference data sets from the literature. Currently, we have implemented a total of eight methods and nine reference data sets to predict immune cell types (four data sets), tissue types (two data sets), cell lines (two data sets) and cancer cell types (one data set). Although this feature is very suitable for a single-cell sequencing data, it provides useful information about the proportion of different cell types in samples obtained by the bulk sequencing method (Supp. Doc. Chapter 15).

For each gene pair combination, the platform can generate a cytometry-like plot of samples under the `cytoplot` tab. The aim of this feature is to observe the distribution of samples in relation to the selected gene pairs. For instance, when applied to single-cell sequencing data from immunological cells, it can mimic flow cytometry analysis and distinguish T helper cells from other T cells by selecting the CD4 and CD8 gene combination.

The `markers` section provides potential marker genes, that is the 36 genes with the highest standard deviation within the expression data across the samples. For every gene, it produces a t-SNE plot of samples, with samples colored in red when the gene is overexpressed in corresponding samples. Users can also restrict the marker analysis by selecting a particular functional group. There are in total 89 such functional groups, including chemokines, transcription factors, genes involved in immune checkpoint inhibition, and so on.

It is also possible to perform a gene copy number variation (CNV) analysis under the `CNV` tab. The copy number is estimated from gene expression data by computing a moving average of the relative gene expression along the chromosomes. A heatmap of samples versus chromosomes is generated, where samples can be annotated further with a phenotype class provided in the data.

RESULTS

To illustrate the use cases of the Omics Playground, we reanalyzed some publics data sets. For single-cell RNA-seq data, we downloaded the melanoma data set GSE72056 of (39). Our platform recapitulates well the original findings of the paper. The t-SNE clustering (Figure 2A) separates the different cell types. Figure 2B and C show the volcano plot, MA plot and most differentially expressed genes between malignant and non-malignant cells. The CNV map (Figure 2D) confirms the major chromosomal copy number variations found in the malignant cells. Figure 2E shows high enrichment of a immune checkpoint signature, particularly concentrated in the T cells. The biomarker heatmap (Figure 2F) highlights the marker genes for each cell type. Each gene cluster is furthermore automatically annotated with the most correlated gene sets (Figure 2G).

To elucidate the mechanism of action of a new drug, or for the intention of drug repurposing, it is often useful to find other drugs that have similar or opposing signatures compared to some given fold change profile. As an example, using data from GSE114716 (40), Figure 2F shows the top ranked drugs with most similar or most opposing signatures to Ipilimumab, a novel monoclonal an-

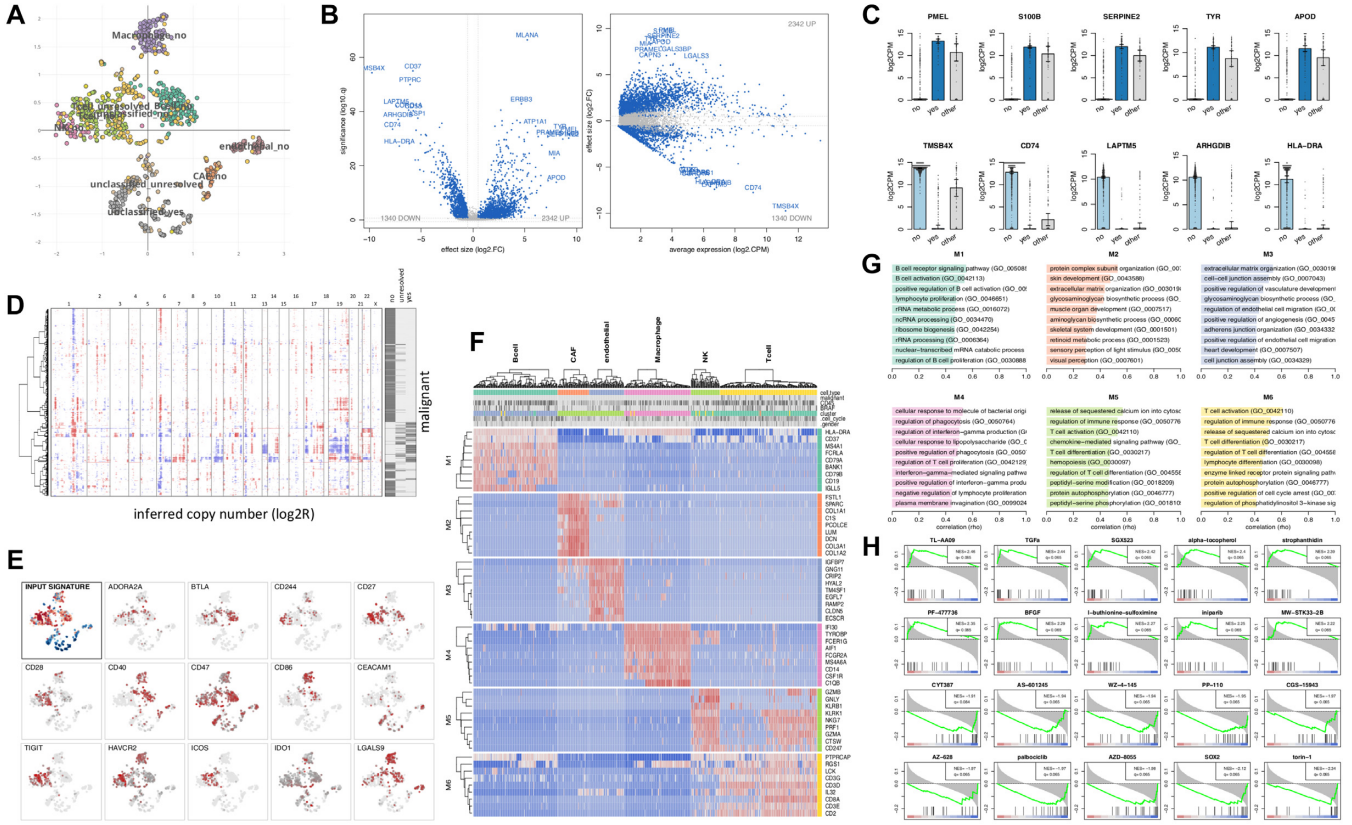

Figure 2. Analysis and visualization using the Omics Playground of public data sets: melanoma scRNA-seq GSE72056 (**A**–**H**) and Ipilimumab GSE114716 (H). (A) t-SNE clustering with cell type annotation. (B) Volcano and MA plot for the malignant versus non-malignant contrast. (C) Corresponding differentially expressed genes. (D) Inferred copy number for sample Cy80. (E) Enrichment distribution for an immune checkpoint signature showing high enrichment in T and B cells. (F–G) Biomarker heatmap and corresponding enrichment for non-malignant cells. (H) Drug enrichment profiles for most similar and opposing drugs compared to Ipilimumab treatment.

tibody targeting CTLA-4 used in tumor therapy to stimulate the immune system. The complete list contains several compounds that stimulate the immune system, such as alpha-tocopherol (41), but also highlights compounds that are not commonly associated with the modulation of immune responses, such as strophanthidin, an intropic drug that has recently been shown to display pro-inflammatory activities (43).

Figure 3A shows the volcano plots corresponding to eight different statistical tests comparing time-dependent activation of T cells at 48 h versus 12 h (42). We see that both standard *t*-test and the Welch *t*-test show much less power to detect significant genes compared to the other methods. The result from edgeR-QLF is close to those of the two limma based methods, while edgeR-LRT is very similar to the results of DESeq2-Wald. Figure 3B shows a microarray gene expression data set, GSE10846 (44), of diffuse large B-cell lymphoma (DLBCL). Figure 3C and D show the variable importance plot and a survival tree on the overall survival of the DLBCL patients, respectively.

With larger data sets, often the number of contrasts increases and complicates the overall analysis. For example, the proteomics data set of (26) comprises 26 populations of seven major immune cell types, measured during resting and activated states. There are more than 300 possible comparisons to make. To gain a better overview, the gene set activation matrix (Figure 3E) helps to visualize the similarities between multiple contrasts on a functional level. Alternatively, similarities can be visualized as a contrast correlation heatmap (Figure 3F). For the same data set, Figure 3G shows a computed partition tree that classifies the major cell types.

DISCUSSION

The Omics Playground is a platform that can currently visualise results from transcriptomic and proteomic data sets, with support for single-cell RNA-seq data. It is based on the R-shiny architecture, providing a modular structure that can be easily updated and expanded. The platform can visualize both standard and more sophisticated omics analysis results, thus providing a highly versatile platform for both novice and more advanced users. We demonstrated its range of uses by re-analyzing previously published data sets of various origins in order to reproduce the main findings and show additional information that can be evinced through the platform.

Future plans for the platform include adding support for other omics data types (such as miRNA and Chip-seq), as well as integration of various omics data sets akin to what

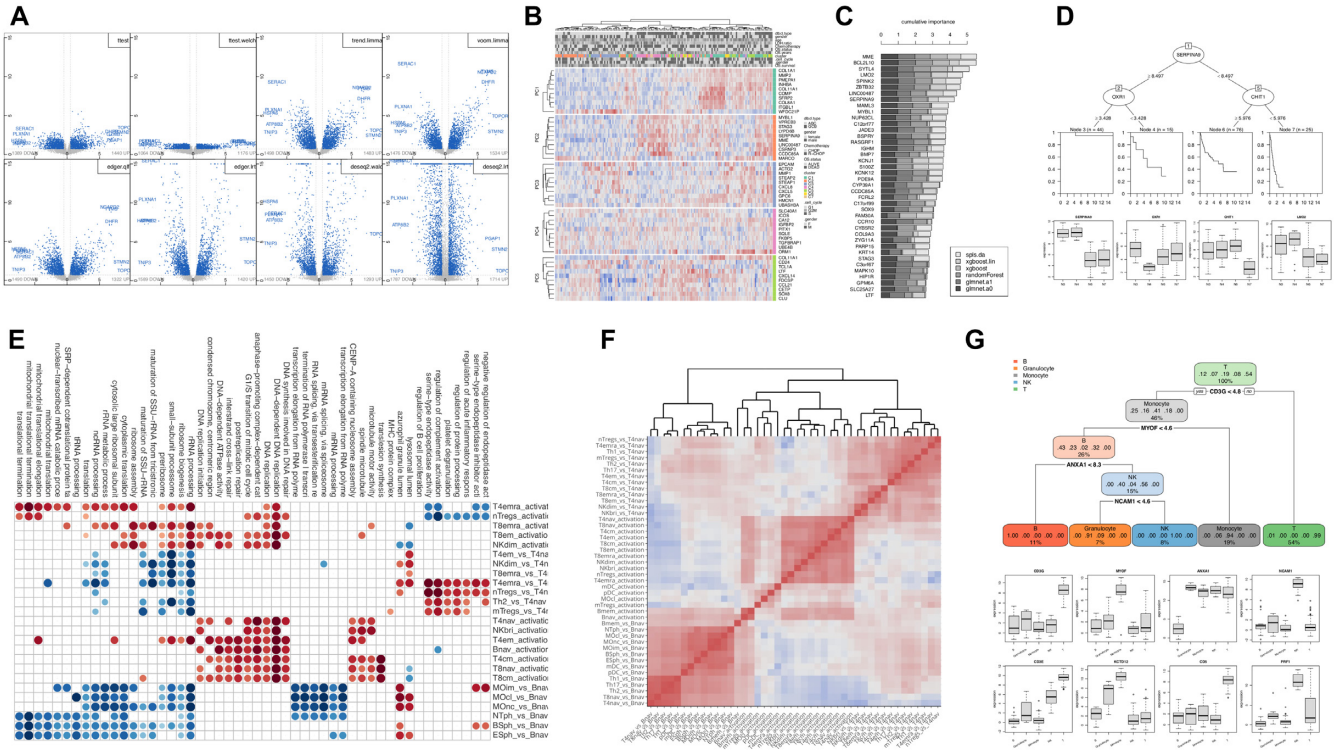

Figure 3. Analysis and visualization of public data sets using the Omics Playground. (**A**) Volcano plots corresponding to eight different statistical methods comparing time-dependent expression of T cell activation at 48h vs. 12h (42). (**B–D**) Hierarchical cluster heatmap, variable importance plot and survival tree for the diffuse large B-cell lymphoma data set GSE10846. (**E–G**) Gene Ontology activation matrix, contrast heatmap and classification tree for the immune cell data set of (26).

currently offered by Paintomics3. Support for time series data set analysis is also planned. Finally, we would like to offer support for isoform specific analysis, which is currently still a rather uncommon feature among omics platforms. While the biological relevance of protein isoforms is still controversial (45–47), there is increasing interest in measuring expression differences at the isoform level, as well as following isoform switching in time-series studies, and various tools are available for that purpose (48–51) that can be integrated within the Omics Playground platform.

The Omics Playground empowers the average life science user with an easy-to-use integrated software environment for self-service analytics of big omics data. The platform also provides a unique combination of tools for more sophisticated analysis normally only available to experienced bioinformaticians. To cope with the ever-growing amount of omics data, it is important to make self-service omics analytics available to non-specialists.

DATA AVAILABILITY

The source code of the Omics Playground is available on GitHub at https://github.com/bigomics/omicsplayground, free for academic purposes. Users also can download the docker image of the platform on Docker hub at https://hub.docker.com/r/bigomics/omicsplayground or find the online documentation on Read-the-docs at https://omicsplayground.readthedocs.io.

GLOSSARY

Signature: a list of selected genes (e.g. by significance or fold change); Condition: a specific phenotype group (e.g. tumor or control); Contrast: a comparison between two conditions (e.g. tumor versus control); Profile: a vector of fold changes corresponding to a certain comparison; Hierarchical clustering: a method that groups similar samples into groups; *q* value: an FDR-adjusted *P* value; Biomarker: a biological feature (gene, mutation or gene set) that characterizes a specific physiological or pathological process.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank the research groups of the Institute for Research in Biomedicine (IRB) for their constructive feedback, which has contributed to the improvement of our platform.

FUNDING

This work was supported by BigOmics Analytics.
Conflict of interest statement. M.A. and I.K are founders of BigOmics Analytics.

REFERENCES

1. Kucukural,A., Yukselen,O., Ozata,D.M., Moore,M.J. and Garber,M. (2019) DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics*, **20**, 6.
2. Torre,D., Lachmann,A. and Maayan,A. (2018) BioJupies: automated generation of interactive notebooks for RNA-Seq data analysis in the cloud. *Cell Syst.*, **7**, 556–561.
3. Clough,E. and Barrett,T. (2016) The gene expression omnibus database. *Stat. Genom.*, **1418**, 93–110.
4. Schultheis,H., Kuenne,C., Preussner,J., Wiegandt,R., Fust,A., Bentsen,M. and Looso,M. (2018) WIlsON: Web-based Interactive Omics VisualizatioN. *Bioinformatics*, **35**, 1055–1057.
5. Hernndez-de-Diego,R., Tarazona,S., Martinez-Mira,C., Balzano-Nogueira,L., Furio-Tari,P., Pappas,G.J. Jr and Conesa,A. (2018) PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.*, **46**, W503–W509.
6. Li,Y. and Andrade,J. (2017) DEApp: an interactive web interface for differential expression analysis of next generation sequence data. *Source Code Biol. Med.*, **12**, 2.
7. Gardeux,V., David,F.P., Shajkofci,A., Schwalie,P.C. and Deplancke,B. (2017) ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics*, **33**, 3123–3125.
8. Zhou,G., Soufan,O., Ewald,J., Hancock,R.E., Basu,N. and Xia,J. (2019) NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.*, **47**, doi:10.1093/nar/gkz240.
9. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
10. Subramanian,A., Narayan,R., Corsello,S.M., Peck,D.D., Natoli,T.E., Lu,X., Gould,J., Davis,J.F., Tubelli,A.A., Asiedu,J.K. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.
11. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
12. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
13. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
14. Kofler,R. and Schltterer,C. (2012) Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics*, **28**, 2084–2085.
15. Hnzelmann,S., Castelo,R. and Guinney,J. (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.
16. Fisher,R.A. (1922) On the interpretation of χ 2 from contingency tables, and the calculation of P. *J. Royal Stat. Soc.*, **85**, 87–94.
17. Sergushichev,A. (2016) An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv doi: https://doi.org/10.1101/060012, 20 June 2016, preprint: not peer reviewed.
18. Wu,D. and Smyth,G.K. (2012) Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.*, **40**, e133.
19. Team R.C. (2013) R: A language and environment for statistical computing. http://www.R-project.org.
20. Chang,W., Cheng,J., Allaire,J., Xie,Y. and McPherson,J. (2015) Shiny: web application framework for R. R package version 1.2.0. https://cran.r-project.org/package=shiny.
21. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417.
22. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525.
23. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
24. Johnson,W.E., Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.
25. Leek,J.T., Johnson,W.E., Parker,H.S., Jaffe,A.E. and Storey,J.D. (2014) sva: Surrogate Variable Analysis R package version 3.10. 0. https://bioconductor.org/packages/release/bioc/html/sva.html.
26. Rieckmann,J.C., Geiger,R., Hornburg,D., Wolf,T., Kveler,K., Jarrossay,D., Sallusto,F., Shen-Orr,S.S., Lanzavecchia,A., Mann,M. *et al.* (2017) Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat. Immunol.*, **18**, 583.
27. Uhlen,M., Oksvold,P., Fagerberg,L., Lundberg,E., Jonasson,K., Forsberg,M., Zwahlen,M., Kampf,C., Wester,K., Hober,S. *et al.* (2010) Towards a knowledge-based human protein atlas. *Nat. Biotechnol.*, **28**, 1248.
28. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
29. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
30. Gene Ontology Consortium. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
31. Witten,D.M., Tibshirani,R. and Hastie,T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.
32. Van Der Maaten,L. (2014) Accelerating t-SNE using tree-based algorithms. *J. Machine Learn. Res.*, **15**, 3221–3245.
33. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
34. Liberzon,A., Birger,C., Thorvaldsdttir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The molecular signatures database hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
35. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Software*, **33**, 1.
36. Candes,E. and Tao,T. (2007) The Dantzig selector: Statistical estimation when p is much larger than n. *Annals Stat.*, **35**, 2313–2351.
37. Breiman,L. (2001) Random Forests, *Machine Learn.*, **45**, 5–32.
38. Chen,T. and Guestrin,C. (2016) Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM International Conference on Multimedia*. pp. 785–794.
39. Tirosh,I., Izar,B., Prakadan,S.M., Wadsworth,M.H., Treacy,D., Trombetta,J.J., Rotem,A., Rodman,C., Lian,C., Murphy,G. *et al.* (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, **352**, 189–196.
40. Goswami,S., Apostolou,I., Zhang,J., Skepner,J., Anandhan,S., Zhang,X., Xiong,L., Trojer,P., Aparicio,A., Subudhi,S.K. *et al.* (2018) Modulation of EZH2 expression in T cells improves efficacy of antiCTLA-4 therapy. *J. Clin. Invest.*, **128**, 3813–3818.
41. Morel,S., Didierlaurent,A., Bourguignon,P., Delhaye,S., Baras,B., Jacob,V. and Van Mechelen,M. (2011) Adjuvant System AS03 containing -tocopherol modulates innate immune response and leads to improved adaptive immunity. *Vaccine*, **29**, 2461–2473.
42. Geiger,R., Rieckmann,J.C., Wolf,T., Basso,C., Feng,Y., Fuhrer,T., Kogadeeva,M., Picotti,P., Meissner,F., Mann,M. *et al.* (2016) L-arginine modulates T cell metabolism and enhances survival and anti-tumor activity. *Cell*, **167**, 829–842.
43. Kara,K., Sakowska,A., Walczak-Drzewiecka,A., Ryba,K., Dastych,J., Bachorz,R.A. and Ratajewski,M. (2018) The cardenolides strophanthidin, digoxigenin and dihydroouabain act as activators of the human ROR/RORT receptors. *Toxicol. Letters*, **295**, 314–324.
44. Lenz,G., Wright,G., Dave,S.S., Xiao,W., Powell,J., Zhao,H., Xu,W., Tan,B., Goldschmidt,N., Iqbal,J. *et al.* (2008) Stromal gene signatures in large-B-cell lymphomas. *New Engl. J. Med.*, **359**, 2313–2323.
45. Nilsen,T.W. and Graveley,B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457–463.
46. Kalsotra,A. and Cooper,T.A. (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genetics*, **12**, 715–729.
47. Tress,M.L., Abascal,F. and Valencia,A. (2017) Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci.*, **42**, 98–110.

48. Leng,N., Dawson,J.A., Thomson,J.A., Ruotti,V., Rissman,A.I., Smits,B.M., Haag,J.D., Gould,M.N., Stewart,R.M. and Kendziorski,C. (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**, 1035–1043.
49. Zhang,C., Zhang,B., Lin,L.L. and Zhao,S. (2017) Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, **18**, 583.
50. Guo,W., Calixto,C.P., Brown,J.W. and Zhang,R. (2017) TSIS: an R package to infer alternative splicing isoform switches for time-series data. *Bioinformatics*, **33**, 3308–3310.
51. Vitting-Seerup,K. and Sandelin,A. (2017) The landscape of isoform switches in human cancers. *Mol. Cancer Res.*, **15**, 1206–1220.