

# Supplementary Information

---

## Guided Visual Exploration of Genomic Stratifications in Cancer

**Marc Streit<sup>1,6</sup>, Alexander Lex<sup>2,6</sup>, Samuel Gratzl<sup>1</sup>, Christian Partl<sup>3</sup>, Dieter Schmalstieg<sup>3</sup>, Hanspeter Pfister<sup>2</sup>, Peter J Park<sup>4</sup> & Nils Gehlenborg<sup>4,5</sup>**

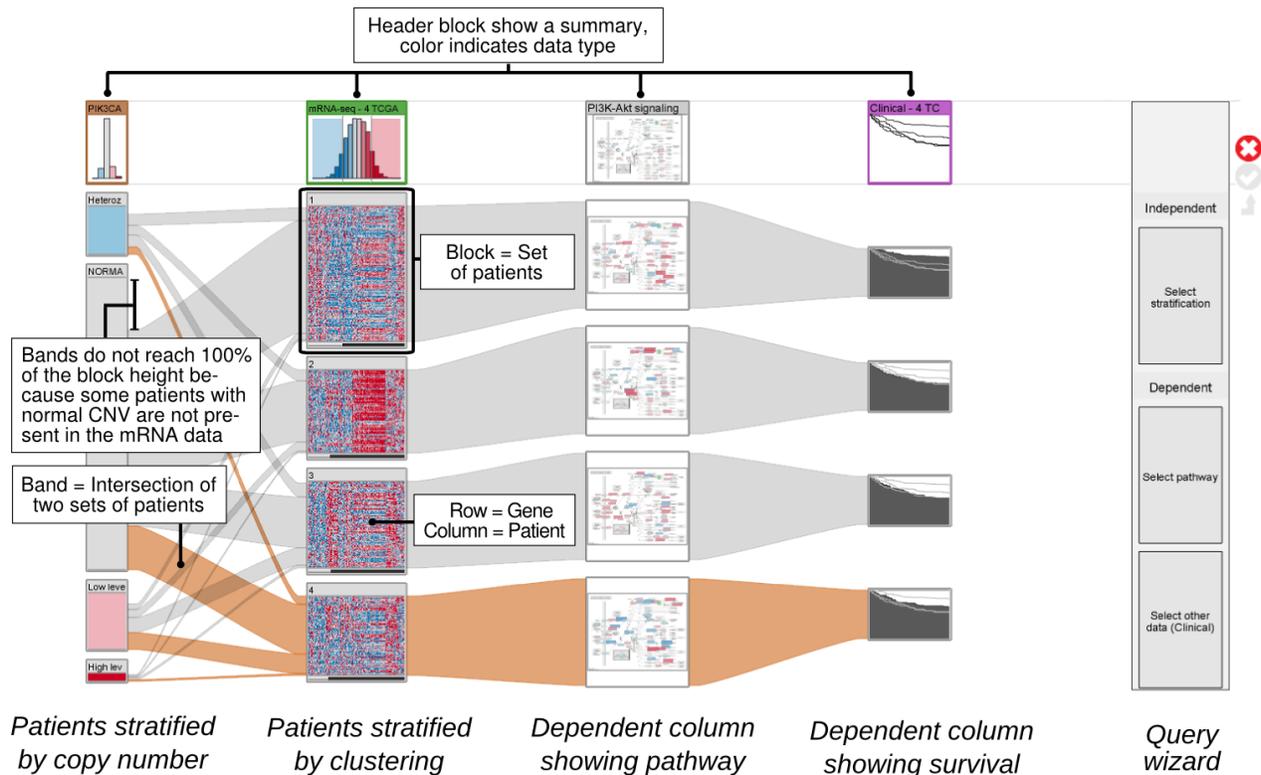
*<sup>1</sup>Institute of Computer Graphics, Johannes Kepler University Linz, Linz, Austria. <sup>2</sup>School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. <sup>3</sup>Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria. <sup>4</sup>Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. <sup>5</sup>Cancer Program, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA. <sup>6</sup>These authors contributed equally to this work*

Email: [nils@hms.harvard.edu](mailto:nils@hms.harvard.edu) or [peter\\_park@hms.harvard.edu](mailto:peter_park@hms.harvard.edu)

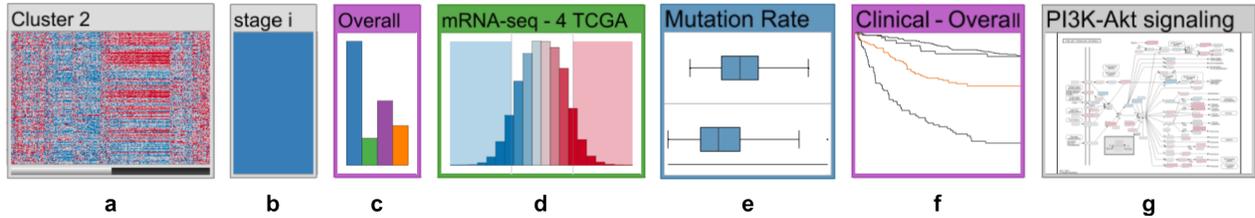
## Table of Contents

Supplementary Figures	2
Supplementary Tables	21
Supplementary Discussion: StratomeX and Related Approaches	24
Supplementary Note: Clear Cell Renal Carcinoma Case Study	27
Supplementary Methods	34

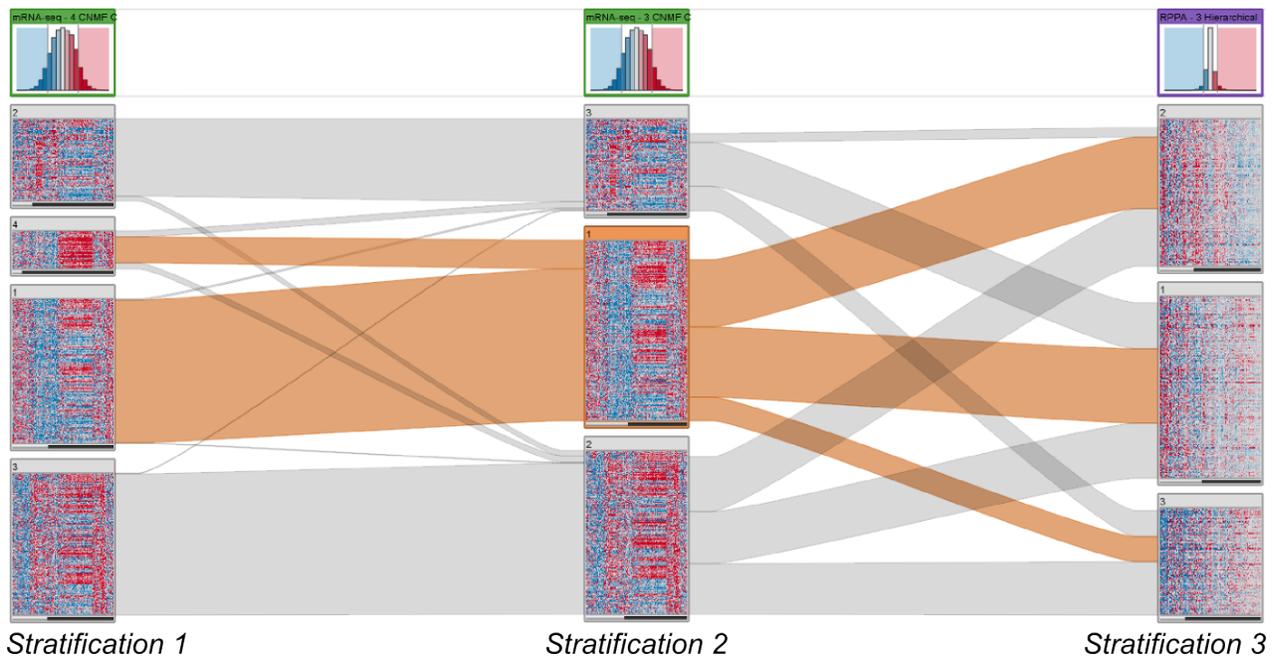
## Supplementary Figures



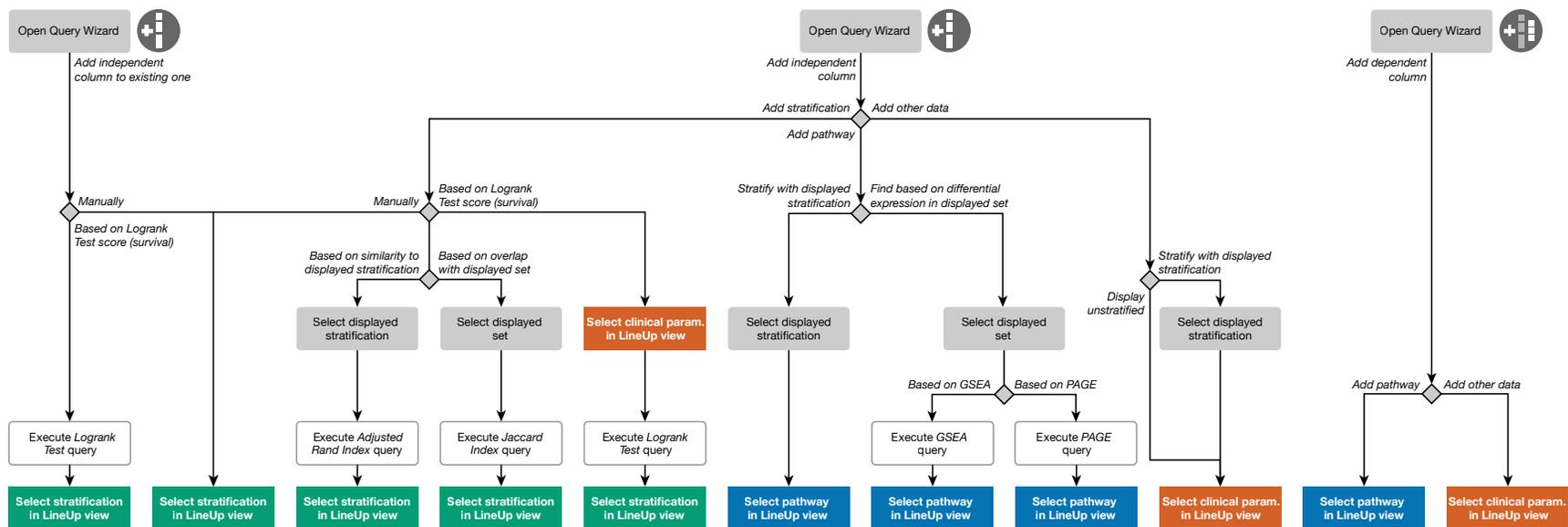
**Supplementary Figure 1. StratomeX user interface.** Stratifications are represented as columns of stacked blocks. Bands between columns visualize the overlap of patients between adjacent stratifications (see **Supplementary Fig. 3** for more details). Here, the first column from the left stratifies the patients by the copy number status of a gene (*PIK3CA*) while the second column groups the patients by the result of a clustering algorithm applied to mRNA expression profiles. Depending on the data type, different visualizations are used to present the data associated with a block (i.e., a group of patients) (**Supplementary Fig. 2**). While the header block at the top summarizes the data of all patients in a given column, the visualization in each block below only represents the data of the patients from the corresponding patient subset. The height of the blocks is scaled to be proportional to the number of patients they contain, if such scaling can be applied to the corresponding visualization technique. Column 3 and 4 are *dependent columns*, meaning that they use the same stratification as the column that they depend on. In this case, the third and fourth column use the stratification of the second column, but apply the stratification to a different dataset. In the third column, the average mRNA expression of all four groups from the second column is color-coded onto the KEGG PI3K-Akt signaling pathway (hsa04151). The fourth column shows survival data using the same stratification as the second column. The column on the right shows the 'query wizard', which is an assistive user interface that supports users in the process of adding new columns to StratomeX (see **Supplementary Fig. 5** for all options that the wizard supports). A detailed description of the user interface and features can be found at <http://help.caleydo.org>.



**Supplementary Figure 2. Block visualizations in StratomeX.** Various visualization techniques are provided for visualizing the data associated with a block: (a) heatmaps for tabular data, (b) uniformly colored blocks for categorical data, (c) histograms for tabular data, (d) histograms for categorical data, (e) box plots for numerical data, (f) Kaplan-Meier plots, and (g) pathways with average values mapped onto the nodes. The height of the visualizations in (a) and (b) is scaled proportionally to the number of patients they represent, while the height of the other visualizations is constant.



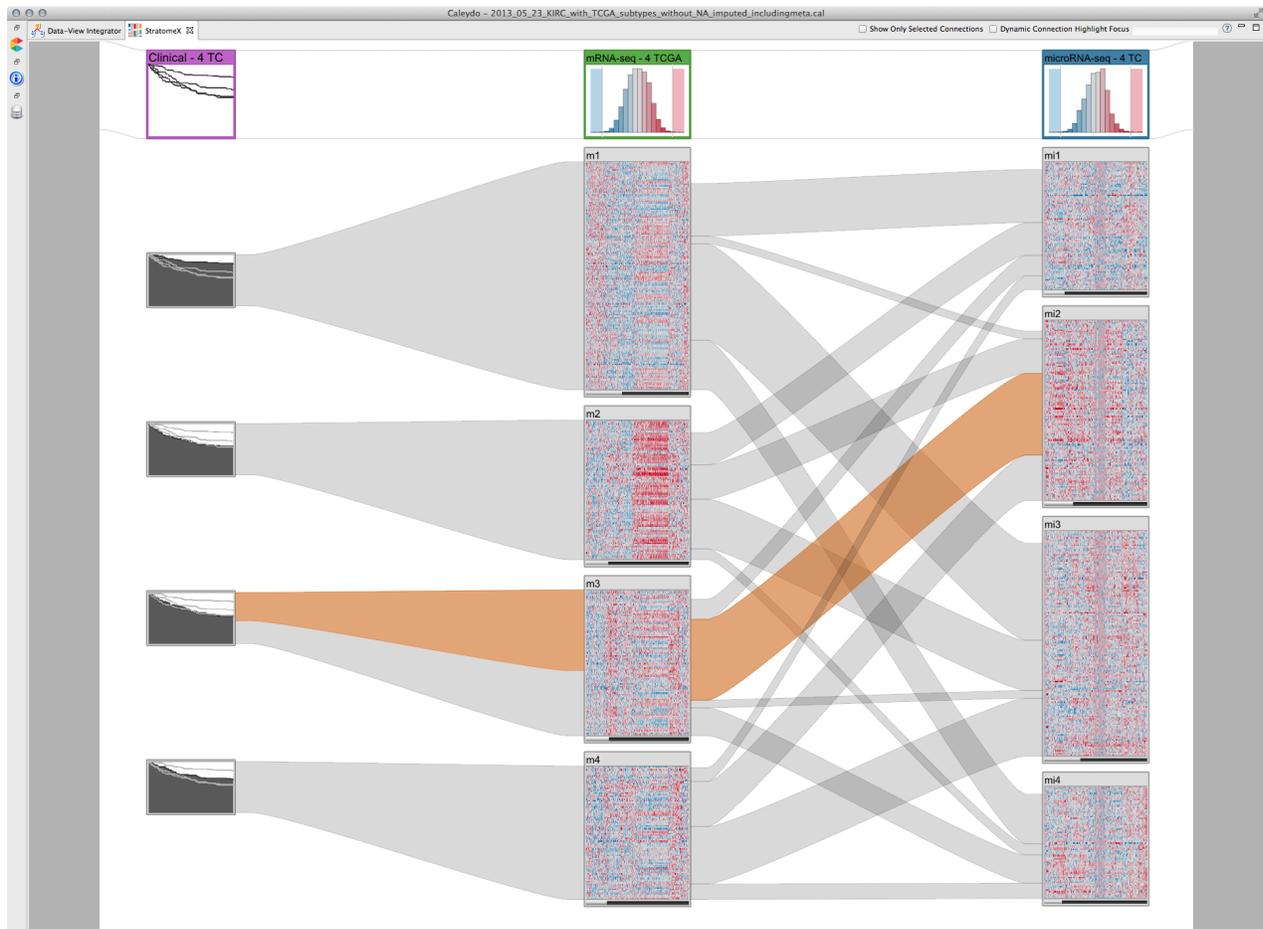
**Supplementary Figure 3. Correlation between stratifications.** The width of the bands denotes the overlap between subsets of adjacent stratifications. While there is a strong correlation between stratifications 1 and 2, stratifications 2 and 3 are much more dissimilar. The subsets in stratifications 1 and 2 are almost identical, except for two subsets (clusters 1 and 4) in stratification 1 that are merged into a single larger one (cluster 1) in stratification 2. In contrast, the bands between stratification 2 and 3 fan out almost equally, which is an indicator for weak correlation.



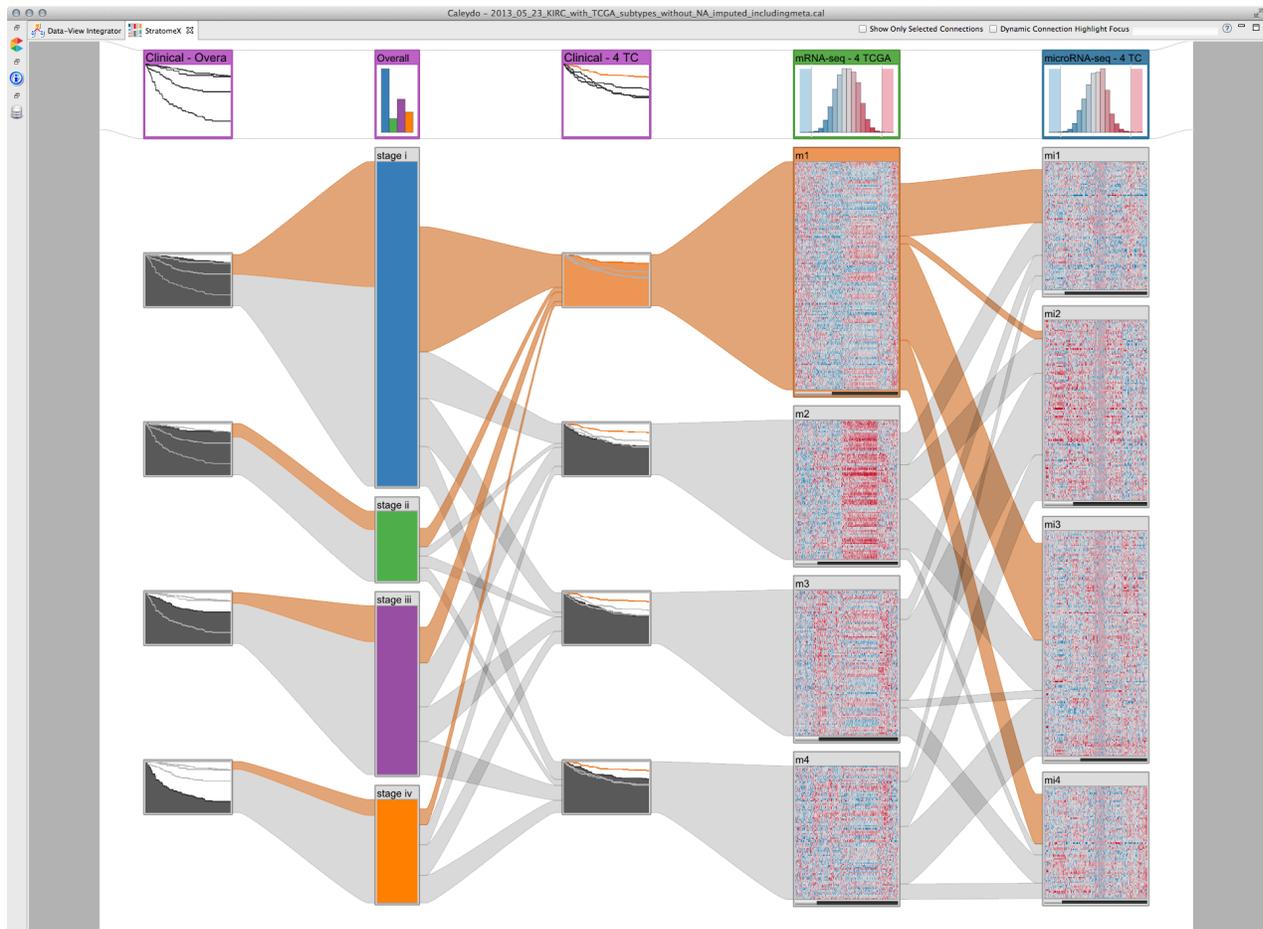
**Supplementary Figure 4. Flow chart of query options and the query wizard menu items used to trigger them.** Diamond nodes denote decision points within the wizard that require user input. Edges describe the options the user can choose from. Gray boxes represent actions taken by the user in the StratomeX view, white boxes with gray outlines mark actions executed by the system, and colored boxes indicate actions taken by the user in the LineUp view. Depending on the current step, the user either needs to select an option from a list, or is instructed to take actions in the user interface, for example, to select a stratification (column) or a set of patients (block) in StratomeX (**Supplementary Fig. 1**) or to perform actions in the LineUp view (**Supplementary Fig. 5**). Once this workflow is successfully completed, a new column will be added to StratomeX.



**Supplementary Figure 5. LineUp view with multi-attribute ranking and filtering.** The interactive ranking visualization allows users to filter and order stratifications according to a single attribute or a combination of multiple attributes, such as the sum or the maximum of attributes, associated with the stratifications. In this example, each row corresponds to a gene for which a series of attributes is available. Attributes can be (1) simple general metrics such as the number of groups or patients in a stratification, (2) data specific metrics, such as the mutation rate, (3) computed scores based on queries triggered by the user, such as the similarity between two patient subsets, or (4) imported scores or groupings that have been computed using an external tool, such as the Mutation Significance (MutSig) q-value of the genes. By selecting a row, the stratification will be added to StratomeX as a new column (see **Supplementary Fig. 1**). Additionally, analysts can search for stratifications of interest by typing their name. The list of available datasets on the left allows analysts to select the subset of stratifications that will be scored by the query and incorporated into the ranking. In this example, only gene mutation calls are included. If the analyst would also select ‘Methylation’, for example, all stratifications (clusterings) defined on DNA methylation data would be added to the ranked list as well. The ‘memo pad’ on the right allows analysts to store attributes that are not of immediate interest, but might become relevant later in the analysis and can then be re-added to the ranking visualization using drag-and-drop. A detailed description of the user interface and features can be found at <http://help.caleydo.org>.



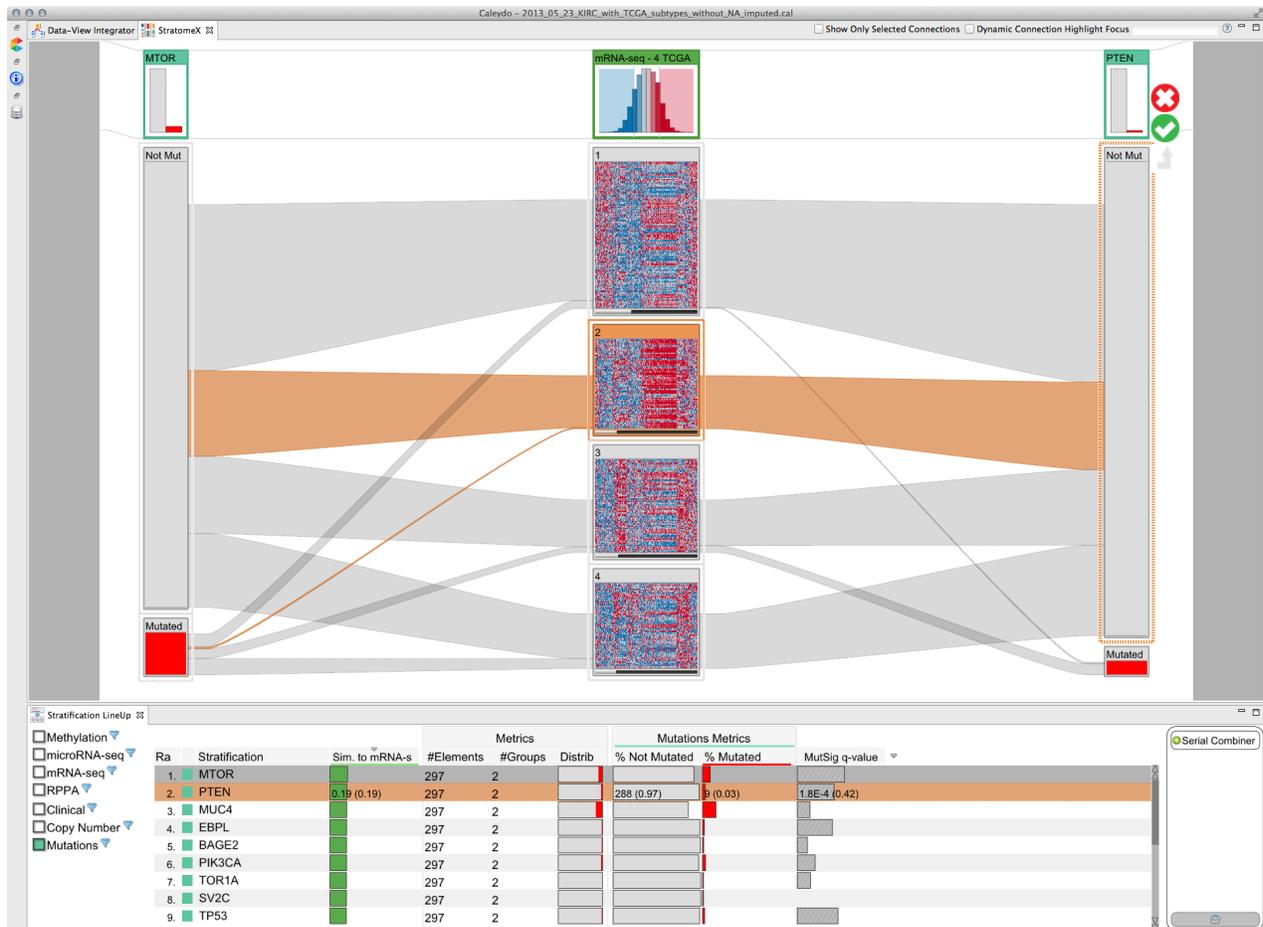
**Supplementary Figure 6a. Kaplan-Meier plots for patient survival ('Days to Death') stratified by mRNA clusters.** The stratification applied to the Kaplan-Meier plots (1st column from left) is reported in the TCGA ccRCC marker paper. The overlap between mRNA (2nd column) and microRNA clusters (3rd column) is shown in addition (see also **Supplementary Video 1** from 0:32 to 3:16). The band pattern between the mRNA and microRNA columns indicates low overall correlation between the clusterings. Two wide bands, however, stand out between mRNA cluster m3 and microRNA cluster mi2 and between mRNA cluster m1 and microRNA cluster mi3. The former band, representing the intersection of the two patient sets, is selected and therefore highlighted in orange. Part of the band between mRNA cluster m3 and the corresponding Kaplan-Meier plot is highlighted in orange as well. It represents the subset of patients in the selected band. The color of the header blocks indicates the data type of the column, e.g., green for mRNA data, blue for microRNA data, and purple for clinical data. The survival curves (left column) for the patients stratified by mRNA clusters (center column; m1, m2, m3, m4 from top to bottom) indicate that survival of patients in cluster m1 is better than survival of those in the other clusters.



**Supplementary Figure 6b. Survival curves for patients stratified by tumor stage and mRNA clusters.** StratomeX view illustrating the correlation between patient survival ('Days to Death') stratified by tumor stage ('Overall Stage') (1st and 2nd column from left) and mRNA clusters (3rd and 4th column), as well as correlation with microRNA clusters (5th column) (see also **Supplementary Video 1** from 3:34 to 4:27). Patient survival stratified by tumor stage indicates that advanced tumor stages are strongly correlated with worse outcomes. Furthermore, the wide band between tumor stage set 'stage I' and the Kaplan-Meier plot corresponding to mRNA cluster m1 shows that the majority of patients in mRNA cluster m1 (63%) also are in the 'stage I' set. Cluster m1 is selected and the patients in this cluster are represented by the orange highlight shown in the bands between all columns. This highlighting emphasizes that only very few patients in cluster m1 are also in microRNA cluster mi2.



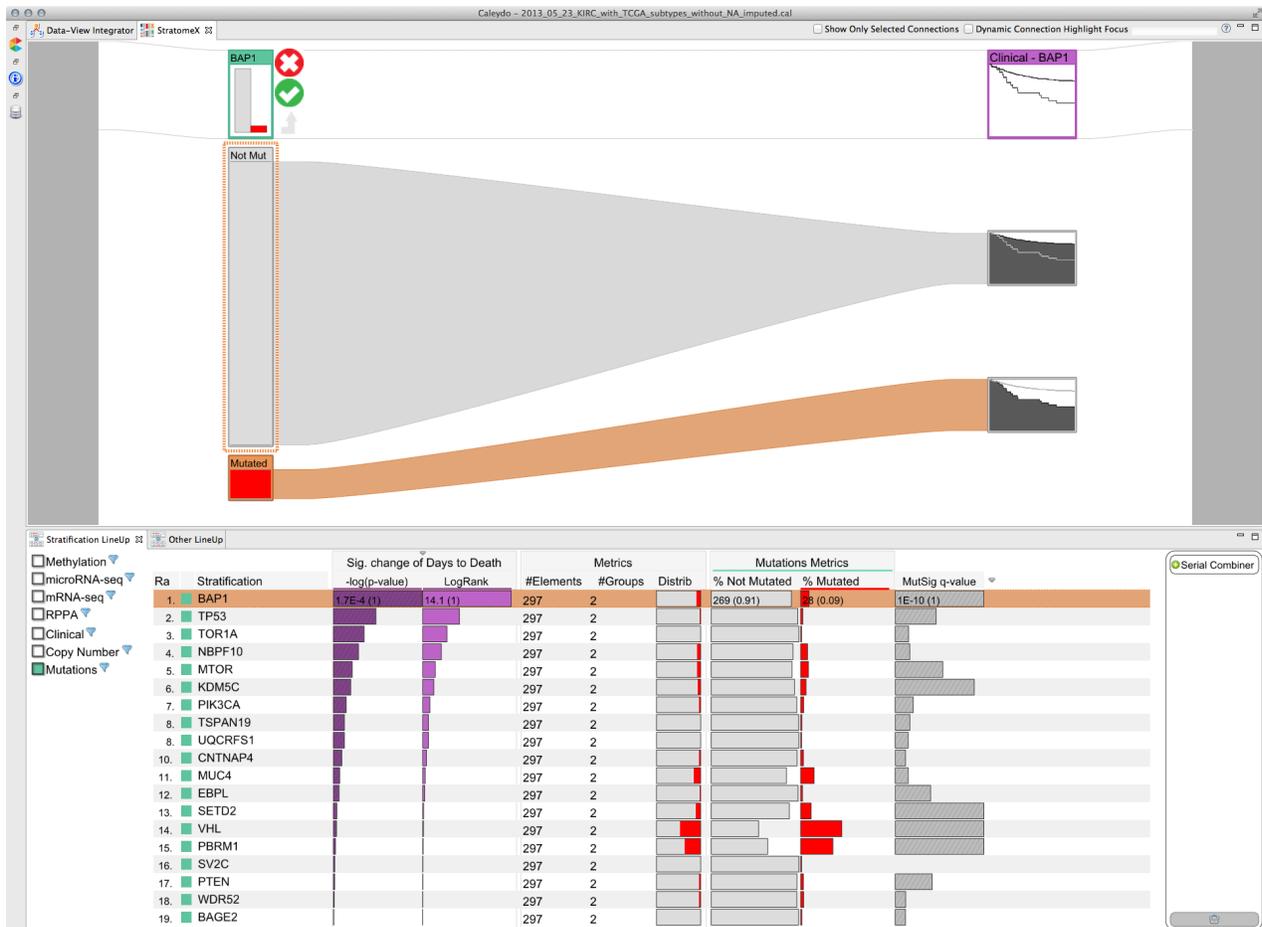
**Supplementary Figure 7. Differential expression of ribosomal genes between mRNA clusters reported by the TCGA ccRCC marker paper.** See also **Supplementary Video 1** from 4:29 to 5:28. The results of a gene set enrichment query against mRNA cluster m4 are shown in the LineUp view at the bottom. The table shows KEGG pathways ranked by the PAGE enrichment scores (light green bars). Additionally, PAGE p-values are shown next to the enrichment scores (the diagonal hatch pattern indicates an 'inverted' mapping, i.e., long bars correspond to low p-values), along with the absolute number of genes in the mRNA data set that could be mapped to the pathway and the percentage of mapped genes. The KEGG Ribosome pathway (hsa003010) is selected, as indicated by the orange highlighting in the table. The pathway is also shown in preview mode in the StratomeX view and the query cluster m4 is highlighted in orange. Almost all ribosomal genes have higher than average expression levels in cluster m4, as indicated by the large red boxes, while they have lower than average expression levels in cluster m1, as indicated by the blue boxes. Expression levels are more diverse in clusters m2 and m3, and in both clusters the ribosomal genes appear to exhibit similar expression patterns.



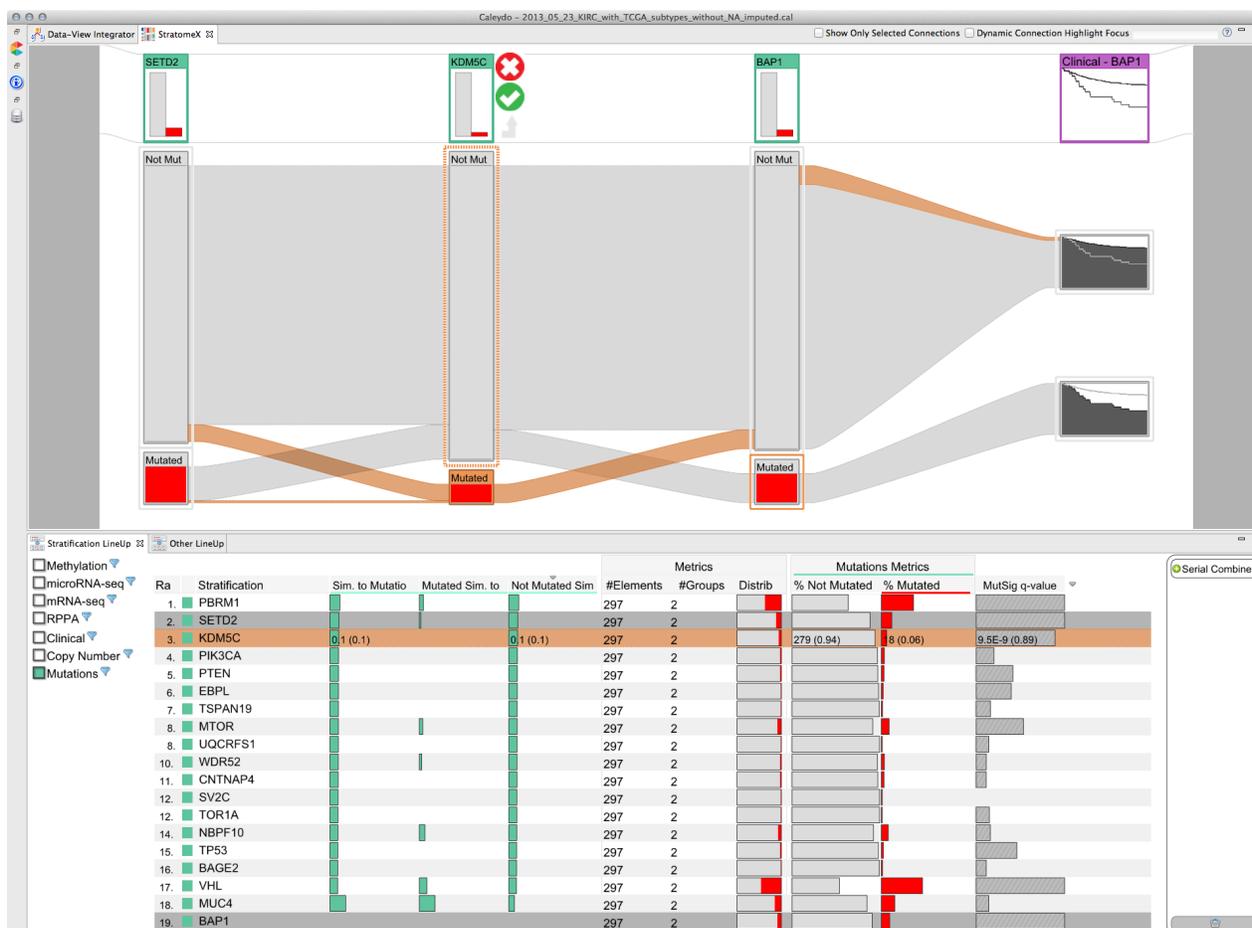
**Supplementary Figure 8. Characterization of an mRNA expression cluster with gene mutation status.** Absence of *MTOR* mutations (1st column from left) and almost exclusive presence of *PTEN* mutations (3rd column) characterize mRNA clusters m2 and m3 (2nd column), respectively. The LineUp view at the bottom shows significantly mutated genes ranked by the overlap of their mutated or not mutated patient set with mRNA cluster m2. The top hits for the query are *MTOR* and *PTEN*, which have also been added to the StratomeX view. The query cluster m2 is selected, as indicated by the orange highlighting. The green bars in the LineUp view encode the Jaccard Index representing the overlap of the mutated or not mutated patient set with mRNA cluster m2, which was computed for all patient sets with a minimum size of 10. The gray bars in the adjacent column represent the MutSig q-values ranging between 0 and 0.1. The diagonal hatch pattern indicates an 'inverted' mapping of q-values. Additional columns to the right show metrics such as the distribution of the mutated and not mutated categories across patients for which a mutation status of the corresponding gene is available.



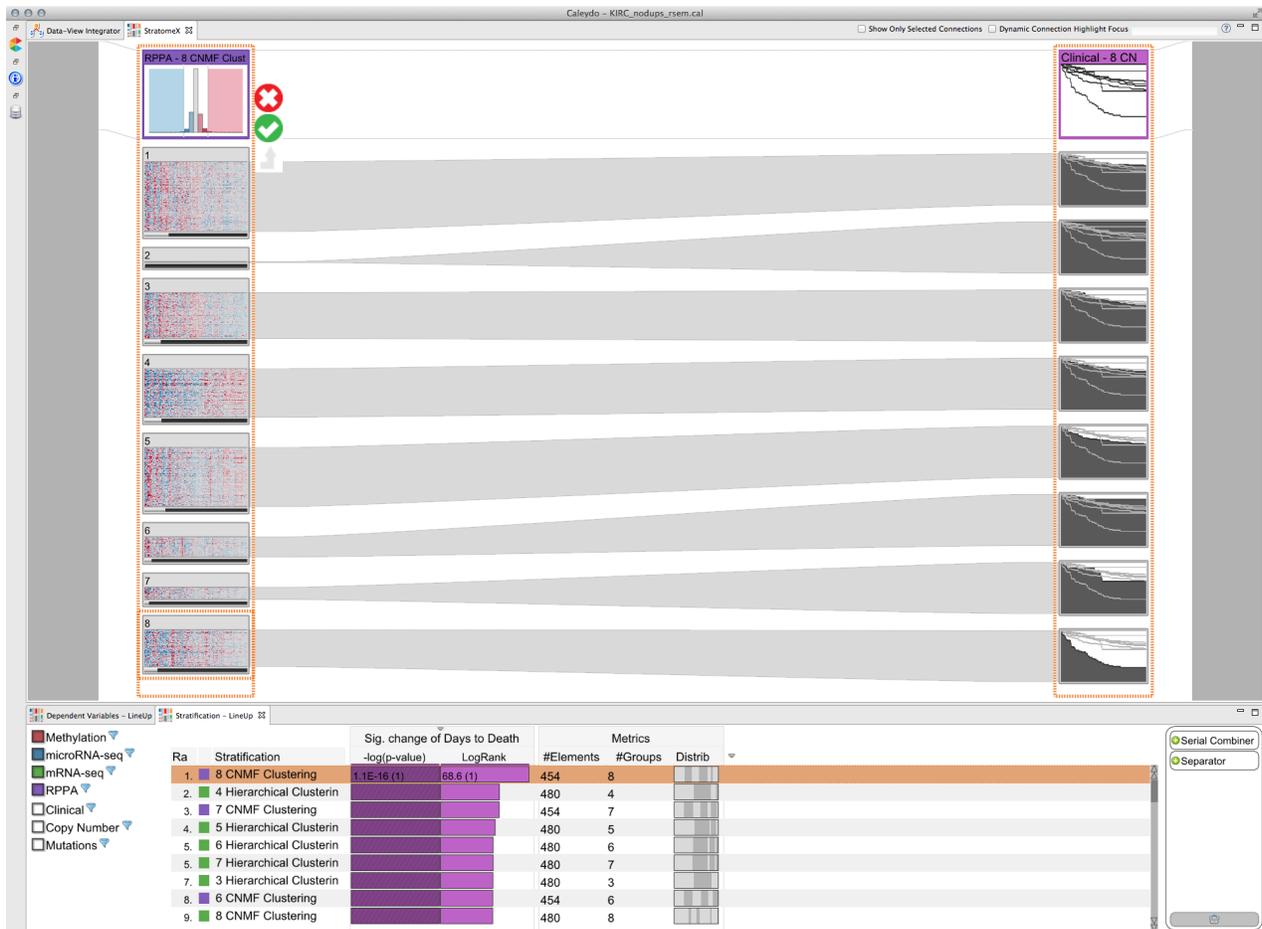
**Supplementary Figure 9. StratomeX view showing correlation between mRNA expression clusters and copy number status of a tumor suppressor gene.** TCGA mRNA cluster m3 (left column) is correlated with deletions of tumor suppressor gene *CDKN2A* (right column) (see also **Supplementary Video 1** from 5:44 to 7:40). The LineUp view at the bottom shows a ranking of genes classified as tumor suppressor genes (indicated by 'TSG' label in the last column of the LineUp view) based on the overlap of the mRNA m3 cluster with the set of patients with a deletion of the given gene. The score is the Jaccard Index, which is represented by the green bars. The minimum patient set size for which scores were computed was 10. The top 5 genes in the results list have very similar scores. Additional metrics show the distribution of copy number events for the given genes across the patient cohort. *CDKN2A* is selected in the results table and shown in the StratomeX view. Bands representing patients in cluster m3 with either a homozygous or a heterozygous deletion of *CDKN2A* are selected and highlighted in orange. The orange bands connecting to the block representing mRNA cluster m3 indicate that about half of all patients in that cluster have a *CDKN2A* deletion. In the overall cohort, only about a third of all patients have either a homozygous (dark blue block) or heterozygous (light blue block) deletion, very few have a low level amplification (light red block) and the remainder has two copies (white block), as indicated by the block heights in the *CDKN2A* copy number stratification.



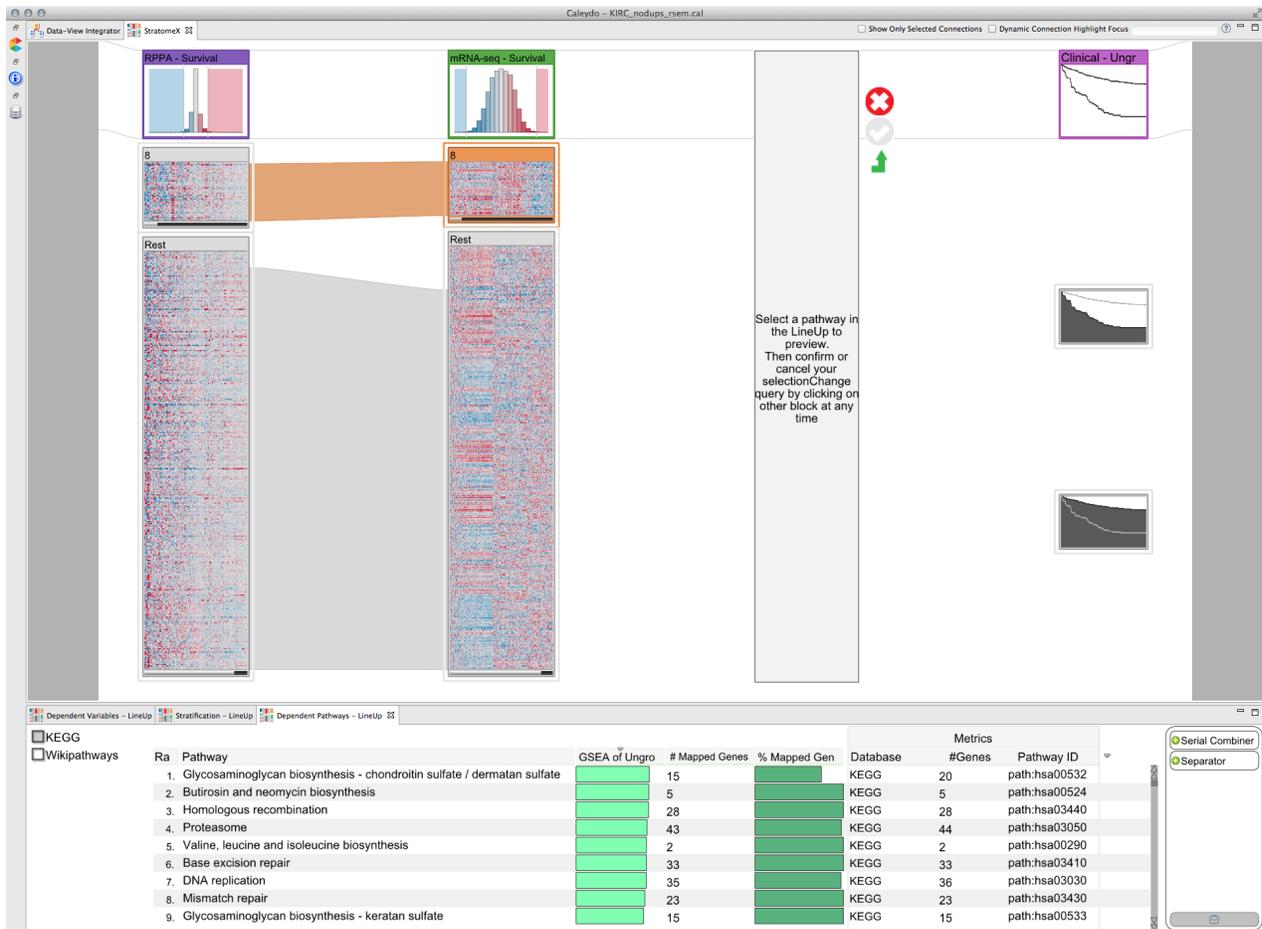
**Supplementary Figure 10. Querying for mutations correlated with survival.** A query was triggered using the StratomeX query wizard preview column showing a Kaplan-Meier plot of the 'Days to Death' variable for the overall cohort (see also **Supplementary Video 1** from 7:48 to 9:05). The result of the query is shown in the LineUp view at the bottom and represents a ranking of 19 significantly mutated genes based on the logrank score (purple bars) and corresponding p-values (dark purple bars, diagonal hatch pattern indicates 'inverted' mapping with  $-\log(p)$ ), obtained by applying the corresponding stratifications to the 'Days to Death' variable. The minimum patient set size for which scores were computed was 10. Additional columns show common metrics for mutated genes, as well as the MutSig q-values ranging between 0 and 0.1. The diagonal hatch pattern indicates an 'inverted' mapping. The top hit *BAP1* has been selected in the LineUp view and is shown in the StratomeX view (left column) next to the Kaplan-Meier plots for 'Days to Death' stratified by *BAP1* mutation status (right column) (see also **Supplementary Video 1** from 9:05 to 9:14). The plots indicate that patients with a mutation in *BAP1* have worse outcomes than those without a *BAP1* mutation. The logrank score is 14.095 - rounded to 14.1 - and the p-value is  $p = 0.00017$ .



**Supplementary Figure 11. StratomeX view showing mutually exclusive mutations.** Mutually exclusive mutation patterns of *SETD2* (1st column from left), *KDM5C* (2nd column), and *BAP1* (3rd column), as well as patient survival stratified by *BAP1* (4th column). Pairwise mutually exclusive mutations are easily identified by the distinctive 'X' band crossings visible at the bottom of the StratomeX view. The LineUp view shows the list of 19 significantly mutated genes as defined by a MutSig q-value of  $q \leq 0.1$  (last column of the LineUp view) ranked by the overlap of the patient set in which they are not mutated with the patient set that has mutations in *BAP1*, as indicated by the orange outline around the red *BAP1* mutated block (3rd column, bottom block). The top three hits are *PBRM1*, *SETD2*, and *KDM5C*. The rank is derived from the Jaccard index score shown in the third column of cyan colored bars, as indicated by the small arrow in the column header ("Not Mutated Sim."). The second column of cyan colored bars ("Mutated Sim. to") shows the scores for the overlap between the set of mutated patients in both the query gene *BAP1* and the corresponding gene in the set of selected stratifications, while the first column ("Sim. to Mutation") shows the score of the second or third column for the given gene, depending on which one is higher.

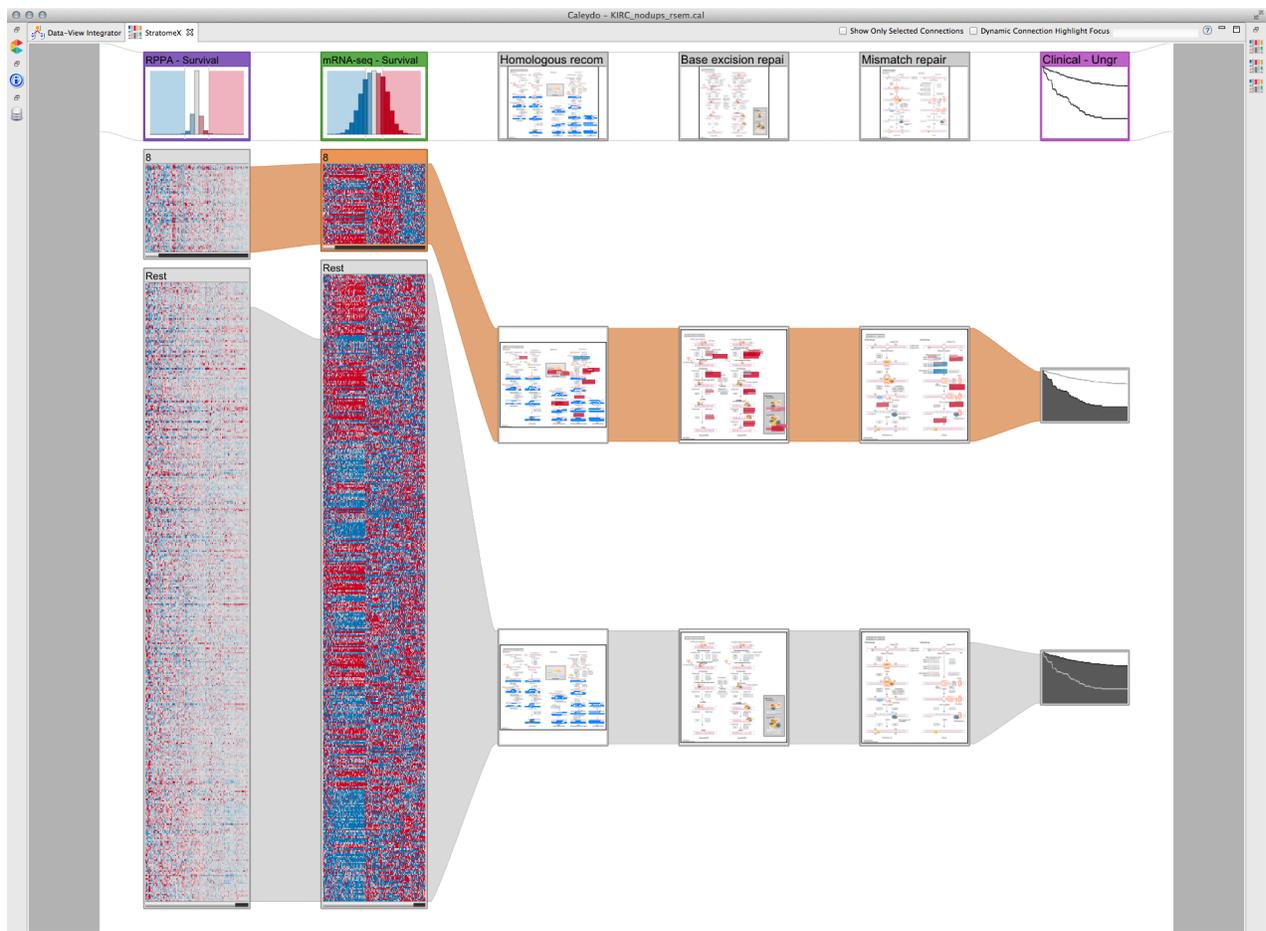


**Supplementary Figure 12. RPPA clusters and expression patterns and Kaplan-Meier plots showing survival data for the patients in the clusters defined by the RPPA expression levels. The LineUp view at the top bottom shows the results of the 'log rank' query of mRNA, microRNA, RPPA and methylation clustering results against the 'Days to Death' variable.**

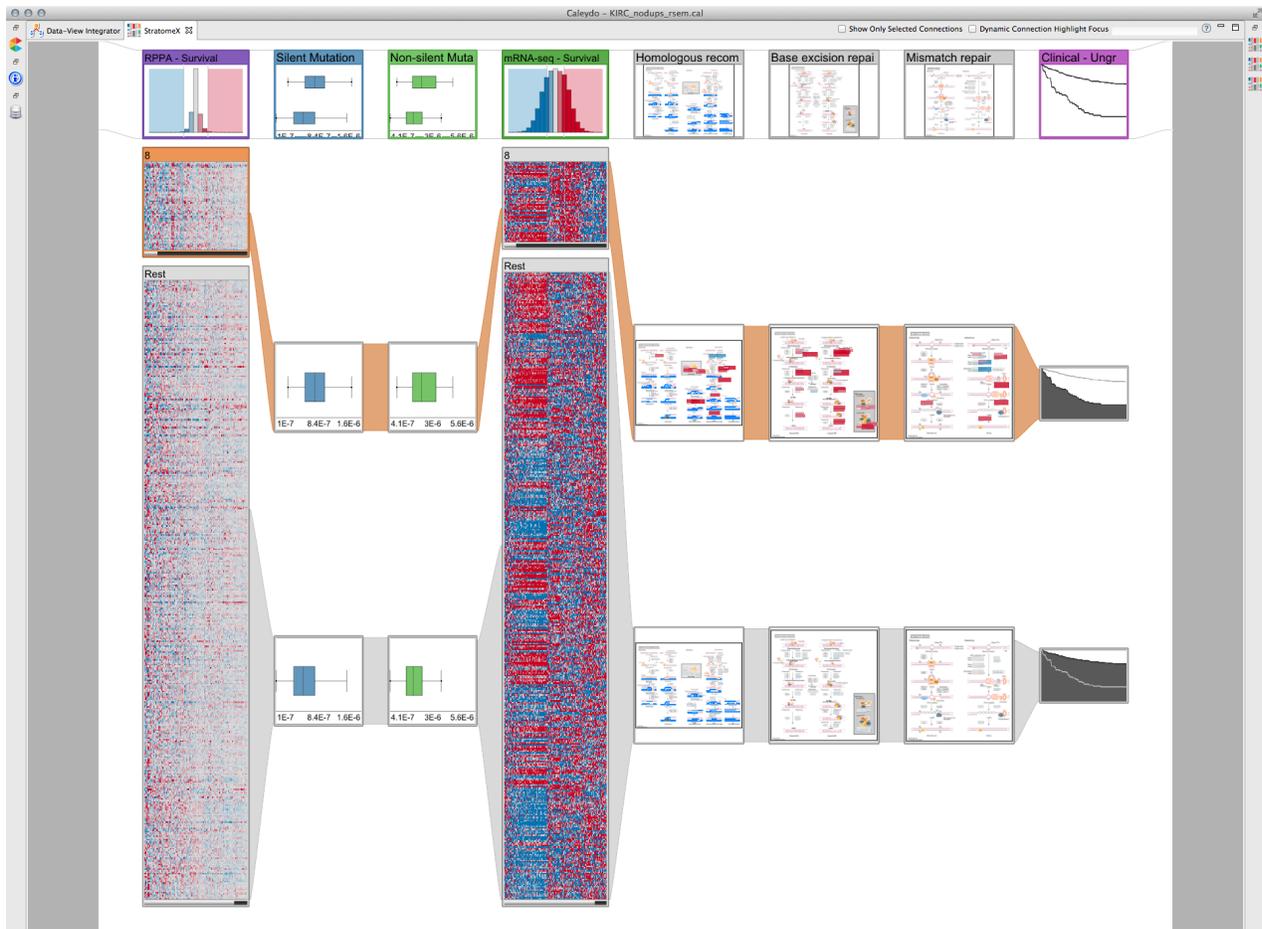


**Supplementary Figure 13. Newly created two-class stratification of patients with RPPA data separating patients in cluster 8 from the rest and associated survival curves.** RPPA data is shown in the 1st column from the left. The orange selection is highlighting the patients in cluster 8. The 2nd column is showing mRNA-seq data using the stratification derived from the RPPA data.

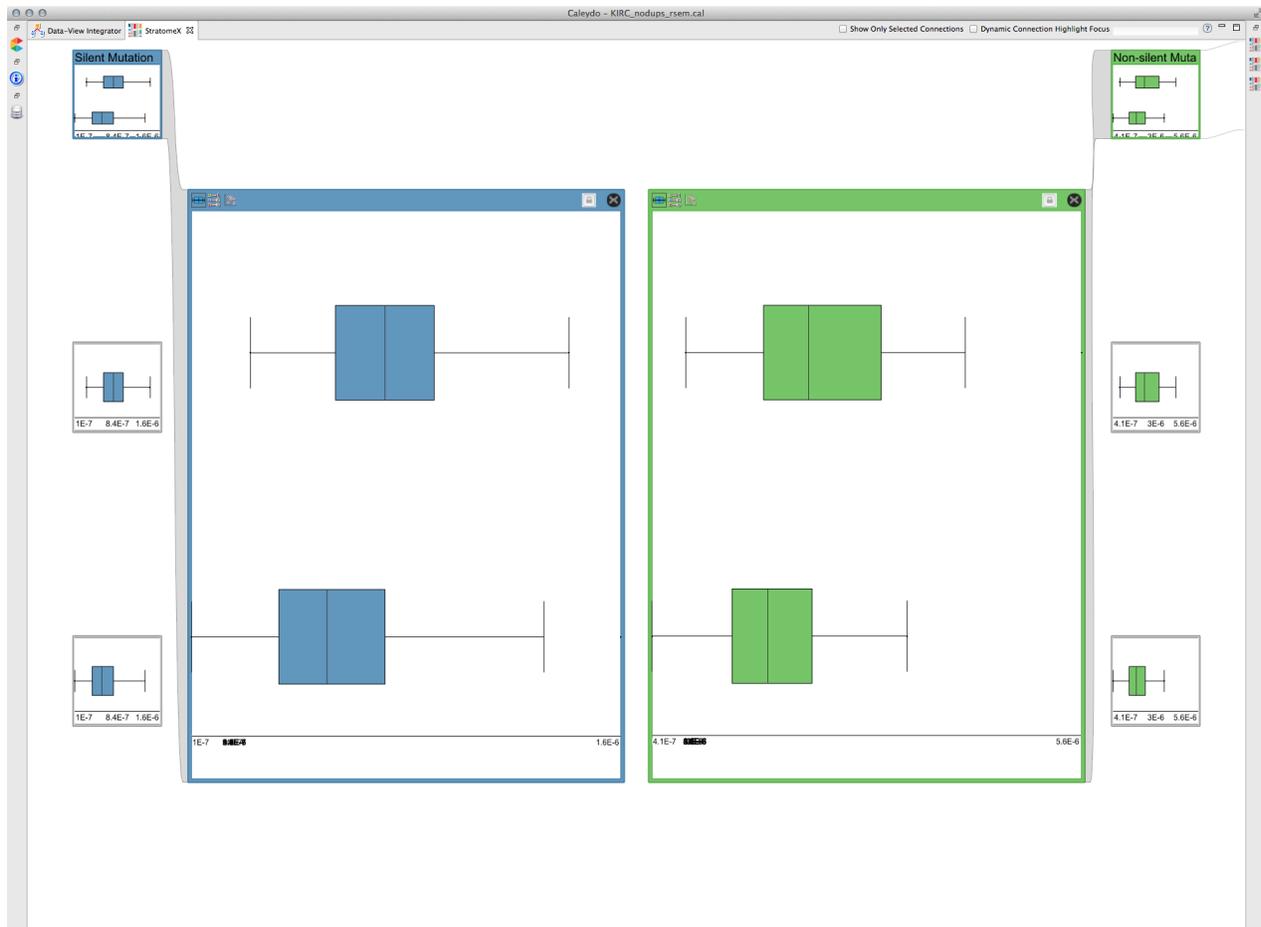
The results of a GSEA query to identify pathways with differential activation in cluster 8 based on mRNA-seq expression levels are shown in the LineUp view at the bottom. The top 8 hits include homologous recombination, base excision repair, and mismatch repair.



**Supplementary Figure 14. StratomeX view showing mRNA expression mapped onto the homologous recombination, base excision repair, and mismatch repair pathways.** The 2nd column from the left shows mRNA expression levels. Pathway maps are shown in columns 3 through 5. Red boxes indicate expression levels higher than the cohort average and blue boxes indicate expression levels lower than the cohort average.



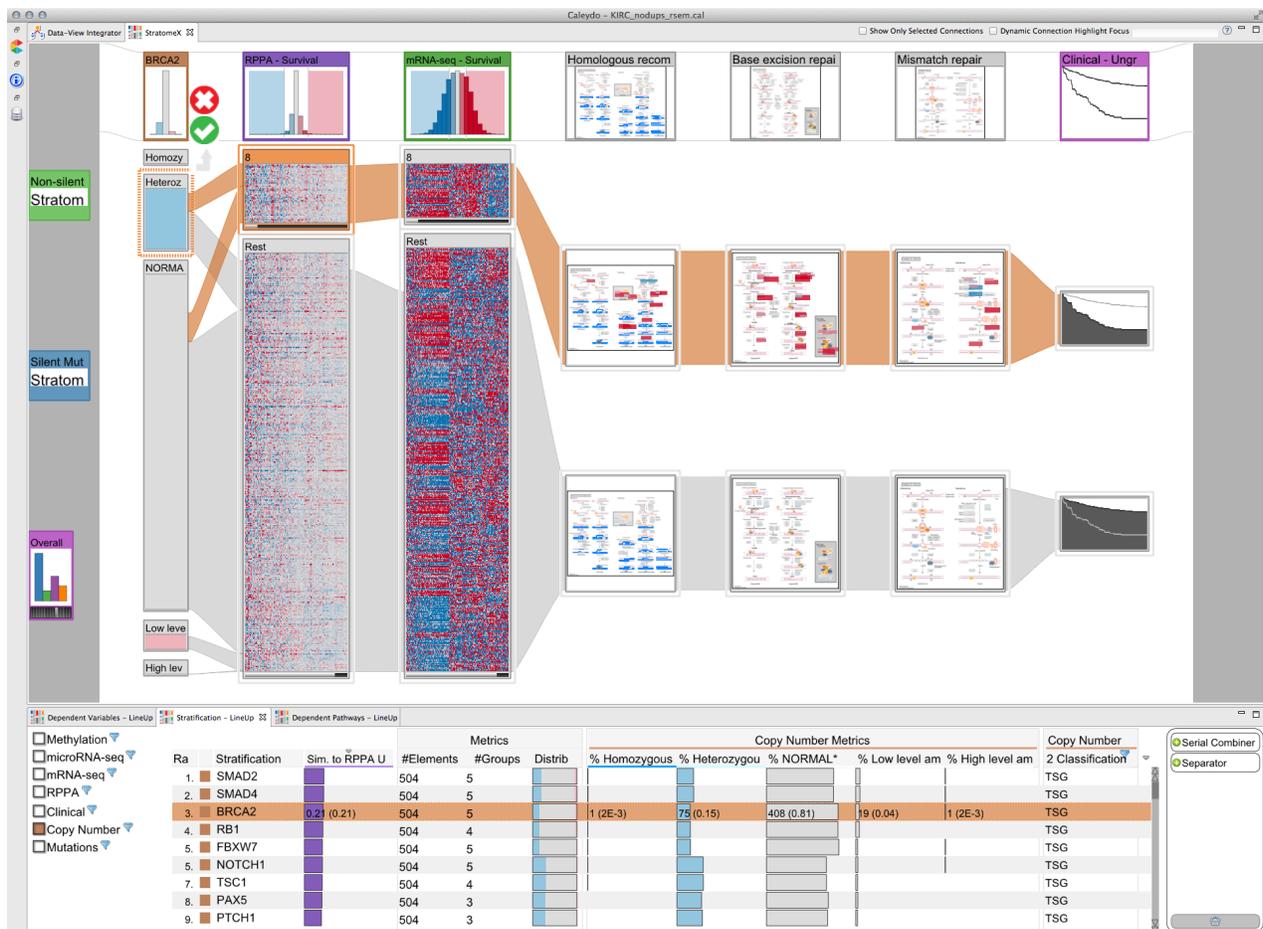
**Supplementary Figure 15a. StratomeX view showing box plots for silent and non-silent mutation rates for cluster 8 compared to the remaining patients. Silent and non-silent mutation rates are shown in columns 3 and 4 from the left, respectively. Whiskers in box plots end at +/- 1.5 IQR from the 3rd quartile and the 1st quartile, respectively.**



**Supplementary Figure 15b. Detail view of header blocks for silent and non-silent mutation rates for cluster 8 compared to the remaining patients.** Silent mutation rates are shown in blue on the left and non-silent mutation rates are shown in green on the right. Whiskers in box plots end at  $\pm 1.5$  IQR from the 3rd quartile and the 1st quartile, respectively. A tendency towards higher rates in cluster 8 (top) is visible for both silent and non-silent mutation rates.

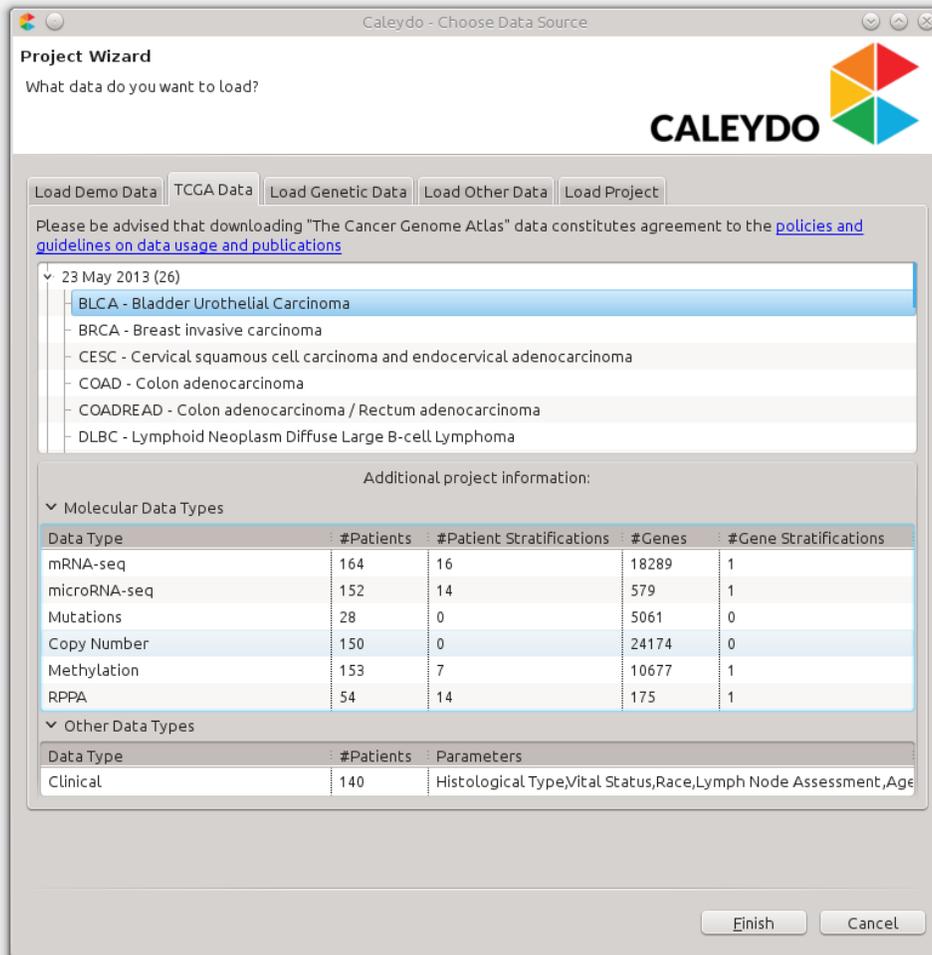


**Supplementary Figure 16. StratomeX view summarizing the findings of the characterization of RPPA consensus NMF cluster 8.** Patients in cluster 8 are highlighted by the orange band. The rightmost column shows stage information for the patients in the cohort and the notable overlap of patients with stage III and stage IV tumors with cluster 8. The LineUp view at the bottom shows the results of the Jaccard Index query of cluster 8 against categorical clinical variables.



**Supplementary Figure 17. StratomeX view illustrating the overlap between patients with *BRCA2* heterozygous deletions with patients in cluster 8.** *BRCA2* copy number status is shown in the 1st column from the left and cluster 8 is shown in the 2nd column. The LineUp view at the bottom shows the results of the Jaccard Index query of cluster 8 against tumor suppressor genes (TSG) with deletions where *BRCA2* is ranked third.

The two mutation rate columns and the overall survival columns have been (temporarily) dragged from the view onto the gray area on the left. They can be added back to the main view by dragging the thumbnail representations next to any of visible columns.



**Supplementary Figure 18. Selection of TCGA packages in the Caleydo Project Wizard.** The wizard is shown when Caleydo launches and the TCGA Data tab provides access to public TCGA datasets prepared for use with Caleydo as described in Supplementary Methods. The upper half of the TCGA data tab provides an overview of the available data packages grouped by analysis date and tumor type. Once the user has selected a data package, information about the package contents is shown in the lower half of the tab. When the user clicks the 'Finish' button, the corresponding data package will be downloaded from our server and opened in Caleydo or opened directly from the local file cache.

## Supplementary Tables

**Supplementary Table 1. Comparison of StratomeX and other cancer subtype analysis techniques.** We distinguish between knowledge-driven (K) and data-driven (D) approaches. The former represents verification of hypotheses that were generated based on the knowledge of the analyst and the latter describes the identification of correlations and patterns based on the data without prior knowledge of the analyst.

Technique	Description	Approach	How are cluster overlaps explored/confirmed?	What visualization support is available?	Strengths	Weaknesses
<b>manual (ad hoc with generic tools)</b>	correlation testing with simple statistics (Jaccard index, adjusted Rand index, etc.)	K	interpretation of numerical scores and/or static plots	<ul style="list-style-type: none"> <li>• static R/MatLab plots</li> <li>• Excel charts</li> <li>• etc.</li> </ul>	<ul style="list-style-type: none"> <li>• flexibility</li> </ul>	<ul style="list-style-type: none"> <li>• hypothesis required</li> <li>• scripting skills required</li> <li>• time consuming (due to generic nature of tools, that are not focused on the task)</li> </ul>
<b>algorithmic approaches</b>	(unsupervised) clustering, “clusters of clusters”, network-based stratification [1]	D	interpretation of numerical scores and/or static plots	<ul style="list-style-type: none"> <li>• static R /Matlab plots</li> <li>• Excel charts</li> <li>• etc.</li> </ul>	<ul style="list-style-type: none"> <li>• automation possible</li> <li>• comprehensive statistics to evaluate significance of findings possible</li> </ul>	<ul style="list-style-type: none"> <li>• difficult interpretation of results</li> <li>• no user input or interaction possible</li> <li>• scripting skills required</li> </ul>
<b>matrix-based (heatmap) visualizations</b>	matrix with mixed data types	K	interpretation of static plots	<ul style="list-style-type: none"> <li>• heatmaps (clustered; mixing multiple data types)</li> </ul>	<ul style="list-style-type: none"> <li>• good for presentation of confirmed hypotheses (the plots are widely used in papers)</li> </ul>	<ul style="list-style-type: none"> <li>• hypothesis required</li> <li>• sorting of rows and columns can only be determined by a single stratification, making it challenging to see correlation between multiple data types</li> </ul>
<b>original StratomeX visualization technique</b>	visualization technique for comparison of multiple clusterings [2]	K	interpretation of interactive visualizations	<ul style="list-style-type: none"> <li>• StratomeX only</li> </ul>	<ul style="list-style-type: none"> <li>• intuitive interpretation of correlations</li> </ul>	<ul style="list-style-type: none"> <li>• hypothesis required</li> </ul>
<b>StratomeX w/ guided visual exploration</b>	(described in this manuscript)	K + D	interpretation of interactive visualizations and numerical scores	<ul style="list-style-type: none"> <li>• StratomeX for comparative visualization of stratifications</li> <li>• LineUp for ranking of query results</li> </ul>	<ul style="list-style-type: none"> <li>• intuitive interpretation of correlations</li> <li>• visualization supported by analytical queries</li> <li>• user interface combines visualization with queries</li> <li>• efficient due to focus on subtype exploration</li> </ul>	<ul style="list-style-type: none"> <li>• no user-integration of novel statistical approaches</li> </ul>

## Supplementary Table 2. Molecular data characteristics.

<i>Assay</i>	<i>Patients</i>	<i>Measurements</i>	<i>Type</i>
mRNA-seq expression	480	18,327 genes	continuous matrix
microRNA-seq expression	481	455 microRNAs	continuous matrix
RPPA protein expression	454	123 proteins	continuous matrix
DNA methylation	294	2,093 genes	continuous matrix
Copy Number status	504	24,174 genes	categorical
Mutation	297	10,749 genes	categorical

## Supplementary Table 3. Clinical data characteristics. Summary statistics were rounded to the nearest integer.

<i>Parameter</i>	<i>Patients</i>	<i>Summary</i>	<i>Type</i>
Age	501	years; min=27, median=61, max=90	continuous
Age at Diagnosis	502	years; min=26, median=61, max=90	continuous
Days to Death	159	days; min=2, median=735, max=2830	continuous
Days to Last Follow Up	498	days; min=0, median=1043, max=3377	continuous
Ethnicity	352	hispanic/latino=24, not hispanic/latino=328	categorical
Gender	502	female=173, male=329	categorical
Histological Type	502	kidney clear cell renal carcinoma=502	categorical
Lymph Node Assessment	495	no=364, yes=131	categorical
Race	495	asian=8, black=21, white=466	categorical
Overall Stage	502	stage I=244, stage II=52, stage III=127, stage IV=78	categorical
M Stage	502	m0=425, m1=77	categorical
N Stage	502	n0=234, n1=18, nx=250	categorical
T Stage	502	t1=22, t1a=125, t1b=102, t2=55, t3a=120, t3b=52, other=26	categorical
Tumor Tissue Site	502	kidney=502	categorical
Vital Status	502	deceased=160, living=342	categorical

## References for Supplementary Tables

[1] Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T., Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108-1115 (2013)

[2] Lex, A. *et al.*, StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. *Comput Graph Forum* **31**, 1175-1184 (2012)

# Supplementary Discussion

## StratomeX and Related Approaches

Data analysis and visualization methods for cancer subtype analysis have three distinct application areas: (1) data exploration, i.e., discovery of novel insights, (2) hypothesis confirmation, i.e., finding supporting evidence for or against a working theory, and (3) presentation, i.e., communicating findings to others. Unlike other approaches, our guided visual exploration approach aims to address all three areas, with a focus on data exploration and hypothesis confirmation.

### **Data Exploration**

Within the data exploration area, we identify three primary tasks: (a) the creation of novel and improved stratifications, (b) judging the quality of stratifications, and (c) reasoning about stratifications. Stratifications are created using, e.g., clustering algorithms based on mRNA patterns [1] or network based stratification [2]. Our approach employs such methods, i.e., enables analysts to run various clustering algorithms, or to import the result of such algorithms. In addition, StratomeX enables analysts to manually refine stratifications, e.g., by splitting clusters based on a clinical variable.

The quality of stratifications can be judged based on algorithmically derived measures, such as Dunn's index [3], or silhouette values [4], or visually, either by visualizing the content of clusters in, e.g., cluster heatmaps [5], or by visualizing differences between alternative clustering results [6]. Our approach is the first to integrate all of these methods: scores can be loaded as supplemental data for stratifications, which can then be used to judge and rank stratifications. More importantly, StratomeX integrates both, the visualization of cluster content and the analysis of cluster differences in a single concise visualization. Finally, our approach enables analysts to reason about stratifications, e.g., to identify supporting evidence in clinical or other data, by dynamically exploring the whole space of the stratome using targeted queries. This makes it easy for analysts to quickly check large quantities of candidate stratifications for mutual support.

The deep integration of analytical methods and visual exploration distinguishes the method described here from our previously published visualization-only approach [7]. The original method enables only a *knowledge-driven approach*, i.e., the confirmation and communication of existing hypothesis based on the analyst's knowledge of the dataset. By integrating methods to identify and rank stratifications, clinical variables, and pathways, we enable a *data-driven approach* that does not rely on an analyst's prior knowledge of the dataset to cancer subtype analysis. Such a data driven approach is necessary for data exploration in large datasets to discover novel insights.

### **Confirmatory Analysis**

The data-driven approach, however, also plays a major role in confirmatory analysis. StratomeX makes it possible to efficiently put candidate stratifications in context of other data types, such as clinical outcomes, to judge effects of different stratifications, or pathways, to speculate about causes and effects of a particular cancer subtype. While we employ algorithms such as gene set enrichment analysis [8] to identify pathways, and logrank tests to identify interesting stratifications

based on clinical variables, it is the deep integration of these analytic processes with the interactive visualization that accelerates the analytical workflow. This enables analysts to explore a larger number of hypotheses in less time and allows them to perform a deeper analysis of the data than possible with other approaches in the same amount of time.

### ***Presentation***

Finally, StratomeX is also well suited for the presentation of results. While it is not the goal of StratomeX to produce publication-ready figures, our visual representation is suitable to efficiently convey important characteristics of candidate subtypes. The visual encoding used by StratomeX is easy to understand and visually appealing. Also, StratomeX can be used to communicate among distributed teams, either by exporting figures from StratomeX, or by passing along project files that contain all the data as well as the analysis setup.

### ***Comparison with other Approaches***

We have summarized the core features of common approaches for subtype identification and characterization in **Supplementary Table 1**. In addition to listing alternative approaches, we distinguish between the original, visualization-only StratomeX and the extended StratomeX described here. We emphasize that there is a spectrum of approaches that range from pure algorithmic to pure (static) visualization approaches and that the extended StratomeX technique combines the strengths of tools from both ends of this spectrum.

In particular, the key features that are distinguishing the extended StratomeX described here from the original publication is the deep integration of analytical and visual methods to enable data exploration. Specifically, we integrated the following: (1) integrated algorithms for querying a database of stratifications, pathways, and clinical variables (see also **Supplementary Methods**): Jaccard Index, Adjusted Rand Index, logrank Test, Gene Set Enrichment Analysis (GSEA), and Parametric Assignment of Gene Set Enrichment (PAGE); (2) a query interface directly built into the visualization that provides step-by-step instructions ('query wizard'); (3) integration of the LineUp visualization to show query results; (4) support for columns that show categorical (clinical) variables such as tumor staging; (5) new block visualizations like box plot and histogram to support numerical (clinical) variables such as mutation rates (see **Supplementary Fig. 2**).

### ***Usability***

Analysis of large, heterogeneous cancer genomics data sets for the identification and characterization of subtypes is without doubt a complex undertaking that requires sophisticated tools and expertise. Any tool or approach used for this purpose will require some training for new users. We argue that StratomeX is more accessible to users without advanced computational skills than the other approaches discussed below, since it (a) offers immediate visual feedback, (b) it does not require the scripting skills that are essential for most alternatives and (c) includes a visual 'query wizard' that provides step-by-step instructions to help users define complex queries.

## References for Supplementary Discussion

- [1] Verhaak, R., *et al.*, Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98-110 (2010)
- [2] Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T., Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108-1115 (2013)
- [3] Curtis, C. *et al.*, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352 (2012)
- [4] Tan, T. Z. *et al.* Functional genomics identifies five distinct molecular subtypes with clinical relevance and pathways for growth control in epithelial ovarian cancer. *EMBO Mol. Med.* **5**, 983-998 (2013)
- [5] Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863-14868 (1998)
- [6] Lex, A., Streit, M., Partl, C., Kashofer, K. & Schmalstieg, D. Comparative Analysis of Multidimensional, Quantitative Data. *IEEE Trans. Vis. Comput. Graph.* **16**, 1027-1035 (2010)
- [7] Lex, A. *et al.*, StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. *Comput Graph Forum* **31**, 1175-1184 (2012)
- [8] Subramanian, A., *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545-15550 (2005)

# Supplementary Note

## Clear Cell Renal Carcinoma Case Study

A comprehensive integrative study of molecular alterations in clear cell renal carcinoma (ccRCC) published by The Cancer Genome Atlas (TCGA) consortium [1], reported subsets of patients defined by unsupervised clustering of mRNA and microRNA profiles. Furthermore, these potential tumor subtypes were characterized in terms of somatic genomic alterations, differentially activated pathways, patient outcomes, and additional criteria.

By reproducing findings of the TCGA ccRCC paper using publicly available TCGA data, we demonstrate that the extended StratomeX is a powerful and efficient approach to discover biologically meaningful features that characterize tumor subtypes. We created a Caleydo data package containing molecular profiling data, clinical parameters, and automated analysis results for clear cell renal carcinoma (known as 'KIRC' within TCGA) using the output of the TCGA Firehose pipeline maintained by the Broad Institute as of 23 May 2013 as described in **Supplementary Methods**. Additionally, we extracted the microRNA- and mRNA-based patient subtype assignments reported in the TCGA consortium paper from the supplementary tables and included them in the case study dataset (**Supplementary Dataset 1**). Furthermore, we obtained a list of significantly mutated genes and their q-values [2] (**Supplementary Dataset 2**) generated by the Firehose MutSig v2.0 [3] module and a tumor suppressor gene and oncogene classification [4] (**Supplementary Dataset 3**). Characteristics of the molecular and clinical data used in this case study are summarized in **Supplementary Tables 2** and **3**. The case study was conducted with Caleydo 3.1.3, available for Windows, Linux, and Mac OS X computers at <http://www.caleydo.org>.

### ***Characterizing mRNA and microRNA Clusters***

We started our exploration with the mRNA and microRNA subtypes reported in the TCGA ccRCC paper (**Supplementary Dataset 1**). The paper describes four subtypes for each of the two data types, named m1 - m4 for mRNA subtypes and mi1 - mi4 for microRNA subtypes, respectively. We looked up the two corresponding stratifications and corresponding data matrices in the LineUp view and added them to the StratomeX view as heatmap columns (**Supplementary Fig. 6a**, see also **Supplementary Video 1** from 0:32 to 2:43). While there is little overall correlation between the two stratifications, two cluster pairs appear to overlap more than the others. These pairs are m1/mi3 and m3/mi2, which have also been reported in the TCGA paper as having a significantly higher overlap than the other clusters.

Next, we used the query wizard to add patient survival times ('Days to Death') stratified by the mRNA clusters to the StratomeX view. We observed notable differences in outcomes across the clusters (**Supplementary Fig. 6a**, see also **Supplementary Video 1** from 2:43 to 3:16). This is also in line with the survival analysis reported by the TCGA paper, which found that patients in cluster m1 have the best outcomes, while patients in m2 and m3 have the shortest survival times. Furthermore, we added a stratification based on tumor staging (clinical variable 'overall stage') as well as patient survival times for the four staging groups (**Supplementary Fig. 6b**, see also **Supplementary Video 1** from 3:34 to 4:27). Patient outcomes get worse in later tumor stages, which is expected. We also found that the m1 cluster, which has the best survival times, consists

of over 60% of patients with *Stage I* tumors, which might play a role in the good survival times observed for that cluster.

We were also interested in whether there are pathways that are enriched in the mRNA clusters. Using the query wizard, we queried the KEGG pathway collection [5] for pathways enriched in cluster m4 (**Supplementary Fig. 7**, see also **Supplementary Video 1** from 4:29 to 5:28), for which the TCGA marker paper reported overexpression of ribosomal gene sets. The PAGE gene set enrichment analysis algorithm [6] was applied to the mRNA expression levels of the patients in cluster m4 relative to the expression levels of the union of all patients in m1, m2, and m3 by selecting m4 as the query set in the query wizard. The resulting list of pathways includes both the proteasome (hsa03050) and the ribosome (hsa03010) pathways among the top 10 results. Even though the p-values are not significant, visual inspection using the LineUp 'preview mode' shows that almost all genes in both of these pathways are expressed at much higher levels in m4 than in the other three clusters. In the case of the ribosome pathway, this is particularly striking (**Supplementary Fig. 7**, see also **Supplementary Video 1** from 5:23 to 5:39) and in accordance with the findings of the TCGA marker paper, which used Gene Set Analysis [7] to identify differentially expressed gene sets. We confirmed our findings by performing the same query using the Gene Set Enrichment Analysis (GSEA) algorithm [8], which ranks the ribosome and the proteasome pathways 4th and 5th, respectively.

We then proceeded to further describe the mRNA subtypes by identifying characteristic presence or absence of gene mutations. First, we loaded MutSig q-values from Firehose [2] (**Supplementary Dataset 2**) as an additional attribute for the gene mutation stratifications and applied an 'inverted' mapping defined by  $-\log(q)$ , so that lower q-values are recognized as 'better' results by LineUp, i.e. represented by longer bars, and ranked higher. By applying a cutoff of q-value  $< 0.1$ , we obtained a filtered list of 19 significantly mutated genes. Using these stratifications as input, we queried for overlap between significantly mutated genes and the four mRNA clusters using the Jaccard Index. When querying against cluster m2, we found that the top results are *PTEN* and *MTOR*, which are mutated in zero and one patient in m2, respectively. Adding both genes to the StratomeX view revealed that *MTOR* is mutated in 5.4% to 11.6% of patients in m1, m3, and m4, but only in 1.1% of patients in m2, i.e. in one patient. Furthermore, we found that apart from a single case in m1, *PTEN* is mutated only in patients in cluster m3 (**Supplementary Fig. 8**). The *PTEN* mutation observed only in cluster m3 is also highlighted in the TCGA marker paper.

Following this characterization based on significantly mutated genes, we further investigated the patients in cluster m3 for distinctive copy number changes. For the purpose of this case study, we focused on deletions of known tumor suppressor genes that overlap with the mRNA clusters. Therefore, we loaded a classification of genes into tumor suppressor genes and oncogenes [4] (**Supplementary Dataset 3**) and associated this classification with the gene copy number stratifications. This allowed us to filter these stratifications based on the classification of the corresponding genes, and to remove all oncogenes and genes without classification, resulting in a set of 71 tumor suppressor genes. We then queried the copy number stratifications associated with these genes for overlap with the m3 cluster and additionally limited the query to those patients who have a homozygous or a heterozygous deletion of a tumor suppressor gene. To limit the query, we deselected the other options ('NORMAL', 'Low level amplification', 'High level amplification') in the dataset-level filter. The top 5 genes returned by the query are *PTCH1*,

*PAX5*, *NOTCH1*, *TSC1*, and *CDKN2A*. Among these, *CDKN2A* stood out, since in addition to 44% of patients in m3 having a heterozygous deletion in this gene, an additional 9% have a homozygous deletion of this tumor suppressor gene (53% with any deletion), while in clusters m1, m2, and m4 the percentage of patients with any deletion is 16% (m1), 30% (m2), and 36% (m4), respectively (**Supplementary Fig. 9**, see also **Supplementary Video 1** from 5:44 to 7:40). Like the *PTEN* mutation, the TCGA marker paper describes the *CDKN2A* deletion of patients in the m3 cluster as a notable feature. The fairly large number of *CDKN2A* deletions in m4 and m2, however, was not reported in the TCGA paper. This discrepancy could be caused by the increased number of samples in our case study compared to the TCGA paper (504 vs. 417 patients with copy number calls) (**Supplementary Table 2**).

### **Exploring Gene Mutations**

After characterizing the mRNA and microRNA expression clusters, our next aim was to identify significantly mutated genes that affect survival. The list of 19 significantly mutated genes was queried against the 'Days to Death' survival variable with the 'logrank query' of the query wizard. The top result (p-value of  $p = 0.00017$ , logrank = 14.095) is *BAP1*. The Kaplan-Meier curves indicate that patients with mutations in *BAP1* have notably poorer outcomes, which is in concordance with the findings of the TCGA marker paper. We added both the Kaplan-Meier plots and the mutation stratification for *BAP1* to the StratomeX view (**Supplementary Fig. 10**, see also **Supplementary Video 1** from 7:48 to 9:14). Next, we performed a mutually exclusive mutation query of *BAP1* against the 19 significantly mutated genes. This query, which ranks genes based on the overlap of the patients without a mutation in the corresponding gene against the patients with *BAP1* mutations, returned *PBRM1*, *SETD2*, and *KDM5C* as top hits. Like the *BAP1* protein, the products of all three genes are involved in chromatin remodeling. The mutually exclusive nature of these mutations could be an indicator for the role of epigenetic changes in ccRCC, which are also noted in the TCGA marker paper. *KDM5C* is perfectly mutually exclusive to *BAP1*, which is illustrated by the characteristic 'X' pattern of the bands connecting the two stratifications (**Supplementary Fig. 11**).

### **Identification of a Patient Set with Poor Survival Times**

To demonstrate that StratomeX supports the generation and refinement of novel hypotheses using a combination of data-driven and knowledge-driven queries, we further explored the clear cell renal carcinoma data set. Unlike the findings in previous parts of this case study, the results reported here have not been reported by the TCGA marker paper [1] and to our knowledge they also have not been reported elsewhere.

As discussed earlier in this case study, we found that patients with mutations in *BAP1* have much worse outcomes compared to patients without such mutations (see also **Supplementary Fig. 10**). While this observation was made in sequence-level data, we were also interested in whether we could find such patient sets based on patterns in functional data, such as the expression levels of microRNAs, mRNAs, or proteins.

We queried a total of 51 clustering results - 16 for mRNA-seq data [9,10], 14 for microRNA-seq data [11,12], 14 for protein expression data (reverse-phase protein array, RPPA) [13,14] and 7 for DNA methylation data [15] - obtained from the 23 May 2013 Firehose analysis run against the 'Days to Death' survival variable with the 'logrank query' of the query wizard.

The top result (logrank test score = 68.6, p-value = 1.1e-16,) is a clustering of RPPA data into 8 clusters found using a consensus non-negative matrix factorization clustering approach [13]. The Kaplan-Meier curves indicate that the 57 patients in cluster 8 have notably poorer outcomes (see **Supplementary Fig. 12**). Since the outcomes of the patients in the other groups are all fairly similar, we decided to study the patients in cluster 8 (n = 57) relative to the remaining patients (n = 397) and created a new stratification of the RPPA data with only two groups: “cluster 8” and “rest” (see **Supplementary Fig. 13**).

### **Characterization of the Patient Set**

Our goal was to characterize the cluster 8 patient set that we identified based on protein expression profiles and the poorer survival times of the patients in that set. Using the Jaccard Index query of the query wizard, we searched all gene mutations for overlap with cluster 8, but found that even frequently mutated genes such as *MTOR* and *BAP1* are mutated in only 5 and 6 of the 57 patients in cluster 8, respectively. Due to the low frequency of mutations in cluster 8, they are likely not responsible for the overall poor survival times.

Next, we used gene set enrichment analysis to identify KEGG pathways that exhibit differential activation between cluster 8 and the rest of the patients. Since there are only 123 unique proteins in the RPPA data set, we applied the cluster 8 vs rest cluster assignment to the mRNA-seq expression matrix (see **Supplementary Fig. 13**) and ran the GSEA query provided by the query wizard on the two newly created mRNA expression clusters representing cluster 8 and the rest of the patients. The top 8 results contained three DNA repair mechanisms (see **Supplementary Fig. 13**): homologous recombination (hsa03440; rank 3), base excision repair (hsa03410; rank 6), and mismatch repair (hsa03420; rank 8). We added these pathways to the StratomeX view (see **Supplementary Fig. 14**). Study of the pathway images revealed that generally genes in the homologous recombination, base excision repair, and mismatch repair pathway are expressed in cluster 8 at levels higher than the cohort average and are therefore shown in red. Notable exceptions are *BRCA2* and genes of the MRN complex in the homologous recombination pathway and members of the MutL-homolog (MLH) and MutS-homolog (MSH) families in the mismatch repair pathway, which in cluster 8 are expressed at levels lower than the cohort average and are therefore shown in blue. Down-regulation of mRNA expression levels of MSH and MLH family genes relative to normal samples has been associated with renal cell carcinoma in an RT-PCR-based study [16].

In addition to the heterozygous deletion of *BRCA2* in close to 40% of patients in cluster 8, the differential expression of several DNA repair pathways in the same cluster is a further indicator that the DNA of these patients might be harboring more mutations than those of other patients. To test this hypothesis, we downloaded the silent and non-silent mutation rates for all available patient genomes (24 patients in cluster 8 and 250 in the rest) from the MutSig 2.0 Firehose pipeline run of 23 May 2013 [2] (file “KIRC-TP.patients.counts\_and\_rates.txt”, **Supplementary Dataset 4**). We visualized these mutation rates as box plots based on our 2-class stratification of the patients representing cluster 8 and the rest (see **Supplementary Fig. 15a**). The header block detail views for the silent and the non-silent mutation rate columns shows that both silent and non-silent mutation rates tend to be higher in cluster 8 (see **Supplementary Fig. 15b**). Using the data export function of StratomeX, we exported the stratified mutation rates and computed Welch’s two-sample *t*-test for both the silent and non-silent mutation rates in R. In both cases the

difference in mutation rates is significant (p-value = 0.007967 for non-silent, p-value = 0.009021 for silent). This result corroborates our previous observation that DNA repair mechanisms might be interrupted in the tumors of patients contained in cluster 8.

Next, we queried cluster 8 against categorical clinical variables and found that it has notable overlap with patients whose tumors are classified as stage III (42.11% of patients in cluster 8, 12.85% of patients in the rest) and stage IV (42.11% of patients in cluster 8, 23.17% of patients in the rest) (see **Supplementary Fig. 16**). With the present data it is not possible to discern, however, whether the increased mutation rates and differential activation of DNA repair pathways are an effect of the advanced tumor stages of the patients in clusters 8 or if these patients present with advanced tumors due to more aggressive cancers caused by high mutation rates resulting from defects in DNA repair mechanisms.

Finally, we looked for overlap between cluster 8 and copy number changes in known tumor suppressor genes and oncogenes. Using the cancer gene classification by Vogelstein *et al.* [4] introduced above, and the Jaccard Index query, we searched for amplifications of oncogenes and deletions of tumor suppressor genes that overlap with cluster 8. The third best hit of the query for tumor suppressor gene deletions is *BRCA2*, which is heterozygously deleted in 38.6% of patients (n = 22) in cluster 8, which also correspond to 29.3% of all patients with a *BRCA2* deletion (see **Supplementary Fig. 17**). *BRCA2* is heterozygously deleted in only 11.84% (n = 47) of the remaining 397 patients, which corresponds to 67.62% of patients with a *BRCA2* deletion. *BRCA2* is well known for its involvement in breast and ovarian cancers and its role in DNA repair mechanisms such as homologous recombination [17]. This supports our earlier observation that DNA repair mechanisms is affected in the patients in cluster 8, which might play a role in their poor outcomes.

In summary, we identified a set of 57 patients in the clear cell renal carcinoma cohort with significantly poorer survival times than the remaining patients and more than 85% advanced stage tumors that have differentially activated DNA repair pathways and significantly increased mutation rates. These are new observations that were not reported in the TCGA marker paper publication on clear cell renal carcinoma. Given the evidence found by our exploration of the data with StratomeX, a more detailed analysis of this patient set is likely to reveal additional information about the molecular changes underlying the observations discussed in this case study.

## References for Supplementary Note

- [1] The Cancer Genome Atlas Research Network, Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-49 (2013)
- [2] Broad Institute TCGA Genome Data Analysis Center, Kidney Renal Clear Cell Carcinoma: Mutation Analysis (MutSig v2.0) May 2013, *Broad Institute of MIT and Harvard* doi:10.7908/C1RF5S2P (2013)
- [3] Lawrence, M., *et al.*, Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013)
- [4] Vogelstein, B., *et al.*, Cancer Genome Landscapes. *Science* **339**, 1546-1558 (2013)
- [5] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M., KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.* **40**, D109-D114 (2012)
- [6] Kim, S.-Y., and Volsky, D.J., PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics* **6**, 144 (2005)
- [7] Efron, B. and Tibshirani, R., On Testing the Significance of Sets of Genes. *Ann. Appl. Stat.* **1**, 107-129 (2007)
- [8] Subramanian, A., *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545-15550 (2005)
- [9] Broad Institute TCGA Genome Data Analysis Center, Kidney Renal Clear Cell Carcinoma: Clustering of mRNAseq gene expression: consensus NMF May 2013, *Broad Institute of MIT and Harvard*, doi:10.7908/C12R3PQ3 (2013)
- [10] Broad Institute TCGA Genome Data Analysis Center, Kidney Renal Clear Cell Carcinoma: Clustering of mRNAseq gene expression: consensus hierarchical May 2013, *Broad Institute of MIT and Harvard*, doi:10.7908/C1Z03662 (2013)
- [11] Broad Institute TCGA Genome Data Analysis Center, Kidney Renal Clear Cell Carcinoma: Clustering of miRseq mature expression: consensus NMF May 2013, *Broad Institute of MIT and Harvard*, doi:10.7908/C1JS9NGH (2013)
- [12] Broad Institute TCGA Genome Data Analysis Center, Kidney Renal Clear Cell Carcinoma: Clustering of miRseq mature expression: consensus hierarchical May 2013, *Broad Institute of MIT and Harvard*, doi:10.7908/C1F18WSJ (2013)
- [13] Broad Institute TCGA Genome Data Analysis Center, Kidney Renal Clear Cell Carcinoma:

Clustering of RPPA data: consensus NMF May 2013, *Broad Institute of MIT and Harvard*, doi:10.7908/C1KP806Z (2013)

[14] Broad Institute TCGA Genome Data Analysis Center, Kidney Renal Clear Cell Carcinoma: Clustering of RPPA data: consensus hierarchical May 2013, *Broad Institute of MIT and Harvard*, doi:10.7908/C1G15XWG (2013)

[15] Broad Institute TCGA Genome Data Analysis Center, Kidney Renal Clear Cell Carcinoma: Clustering of Methylation: consensus NMF May 2013, *Broad Institute of MIT and Harvard*, doi:10.7908/C10Z71B8 (2013)

[16] Deguchi, M., Shiina, H., Igawa, M., Kaneuchi, M., Nakajima, K., Dahiya, R., DNA mismatch repair genes in renal cell carcinoma, *J. Urol.* **169**, 2365 (2003)

[17] Tutt, A. N., van Oostrom, C. T., Ross, G. M., van Steeg, H., Ashworth, A., Disruption of Brca2 increases the spontaneous mutation rate in vivo: synergism with ionizing radiation, *EMBO Rep.* **3**, 255 (2002)

## Supplementary Methods

### Scoring Queries

In addition to basic browsing, filtering, and ranking of stratifications, pathways, and clinical variables, StratomeX supports a series of advanced query methods to find additional stratifications and pathways based on patterns identified in the StratomeX view (**Supplementary Fig. 4**).

Some of the queries implemented in StratomeX are based on hypothesis tests for which p-values are provided along with the test scores. The results of these queries, however, must not be interpreted as statistically reliable results, since correction for multiple hypothesis testing is not provided in the current implementation although some queries involve thousands of tests. Generally, the scores are provided to guide the user to stratifications or pathways that provide additional insight into patterns observed in the StratomeX view and to generate new hypotheses.

#### ***Scoring stratifications based on similarity to a selected stratification***

This query is useful for finding stratifications that are similar to a currently displayed stratification. The Adjusted Rand Index [1] is used to compare each stratification in the collection against the query stratification selected by the user.

#### ***Scoring stratifications based on overlap with a selected patient set***

In contrast to the Adjusted Rand Index, which quantifies similarities between stratifications, this type of query is designed to identify stratifications that contain sets similar to a query set in a displayed stratification. The score for a set is the Jaccard Index describing its similarity to the query set and computed for all sets in every stratification in the collection of stratifications, but only the best score for each stratification will be reported. In addition, if the query is triggered from a binary stratification, such as mutations, a mutual exclusivity score is computed per set, which can be used to identify genes that are mutated in non-overlapping sets of patients.

#### ***Scoring stratifications based on logrank test for patient survival***

This query identifies sets of patients that exhibit altered survival times compared to the rest of the patients in the same stratification. It uses the logrank test (Mantel-Haenszel test) to score the stratifications and assigns larger scores to more extreme differences in survival. Similar to the previous method, the score is computed for each considered set of stratifications and the best result per stratification is presented to the user. A p-value for the best result is provided as guidance.

#### ***Scoring pathways based on gene set enrichment for a selected patient set***

This type of query is designed to identify pathways that are over- or underexpressed in a patient set relative to the rest of the cohort. It takes a set of patients as input and computes differential gene expression levels for patients in the query set against the rest of the patients in the same stratification. The differential expression levels are used to score pathways using either Gene Set Enrichment Analysis (GSEA) [2] or Parametric Assignment of Gene Set Enrichment (PAGE) [3]. Additional meta information, such as the number and percentage of mapped genes, are shown to

allow filtering operations, such as exclusion of pathways with too many or few genes with expression levels.

### ***Importing externally computed scores***

Any external score associated with stratifications or pathways can be imported using the data import wizard, and used for exploration of the data, as demonstrated in the case study presented in the **Supplementary Note**.

## **TCGA Data Package Generation**

Caleydo StratomeX is designed to operate on large and heterogeneous data sets that integrate multiple molecular profiling techniques with clinical parameters and various analysis results, such as clustering results, copy number calls and mutation calls. Since there is no common file format to describe such integrated data sets, data matrices and analysis results are typically distributed as individual files, which have to be downloaded and imported into StratomeX one by one. This makes it tedious for users of StratomeX to create comprehensive datasets, in particular when the data is frequently updated. Caleydo addresses this issue by providing a binary format for project files to store and share such datasets. The creation of these data packages can be also be automated, for example, to support project file generation for large studies in batch mode.

The Cancer Genome Atlas (TCGA) project is the most comprehensive source for integrative cancer genomics data sets to date. Due to the incremental collection and processing of tumor samples, the datasets for the over twenty tumor types studied by the project are changing frequently. An automated analysis pipeline called Firehose (<http://gdac.broadinstitute.org>) has been developed at the Broad Institute of MIT and Harvard to preprocess and perform comprehensive automated analyses on each tumor cohort without human intervention. The outputs of this pipeline are made publicly available as a community resource and represent the basis for many integrative analyses performed by TCGA analysis teams.

We have developed a data packaging tool to assemble Caleydo project files based on multiple input sources. We use this tool to generate project files based on the output of the Firehose analysis pipeline for all TCGA tumor types processed by Firehose. Our tool takes advantage of the standardized output format and directory structure used by all Firehose workflows. In the current implementation up to 24 data files from 18 Firehose workflows are included in the data package for each tumor type. The data packages are generated for each public Firehose analysis run.

### ***Data Matrices and Analysis Results extracted from Firehose***

Since package generation is performed for multiple tumor types and Firehose pipelines runs, the following variables are used:

<i>Variable</i>	<i>Description</i>
<analysis-date:format>	Firehose analysis run date, e.g. 2013-05-23. In addition, a specific date format can be given, e.g. '<analysis-date:YYYYMMDD>' resolves to '20130523'.
<data-date:format>	Firehose data run date, e.g. 2013-05-23. See above for formatting.
<tumor-base>	Tumor type, e.g. KIRC, GBM
<tumor-subset>	Tumor type including the sample type, e.g. KIRC-TP; By default, we extend <tumor-base> with '-TP' (primary tumor) unless <tumor-base> is SKCM, which is mapped to SKCM-TM (metastatic tumor) or LAML, which is mapped to LAML-TB (blood).
<profile>	The molecular data type, e.g. mRNA, microRNA

In addition, due to the evolution of the Firehose pipeline itself and missing data, fallback files are used. By default, data files containing full matrices with all genes/microRNAs/proteins are used. If they are not available, data files containing only the 1500 most variable genes/microRNAs/proteins, or another Firehose-provided subset, are used. If a data package cannot be found at all, it will be ignored by the data packager.

All package locations given below are relative to this base URL: <http://gdac.broadinstitute.org/>

### **mRNA Data Matrices**

#### **Default**

<i>Package</i>	runs/analyses__<analysis-date:YYYY_MM_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org_<tumor-subset>.mRNA_Preprocess_Median.Level_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz
<i>File</i>	<tumor-subset>.medianexp.txt

#### **Fallback**

<i>Package</i>	runs/analyses__<analysis-date:YYYY_MM_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org_<tumor-subset>.mRNA_Clustering_CNMF.Level_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz
<i>File</i>	outputprefix.expclu.gct
<i>Notes</i>	Contains only the most variable genes.

### **mRNA-seq Data Matrices**

#### **Default**

<i>Package</i>	runs/data__<data-date:YYYY_MM_DD>/data/<tumor-base>/<data-date:YYYYMMDD>/gdac.broadinstitute.org_<tumor-subset>.mRNAseq_Preprocess.Level_4.<data-date:YYYYMMDD>00.0.0.tar.gz
<i>File</i>	<tumor-base>.uncv2.mRNAseq_RSEM_normalized_log2.txt
<i>Altern. File 1</i>	<tumor-base>.uncv1.mRNAseq_RPKM_log2.txt
<i>Altern. File 2</i>	<tumor-base>.mRNAseq_RPKM_log2.txt

## Fallback

*Package* runs/analyses\_\_<analysis-date:YYYY\_MM\_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.mRNAseq\_Clustering\_CNMF.Level\_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz

*File* outputprefix.expclu.gct

*Notes* Contains only the most variable genes.

## microRNA Data Matrices

### Default

*Package* runs/data\_\_<data-date:YYYY\_MM\_DD>/data/<tumor-base>/<data-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.miR\_Preprocess.Level\_4.<data-date:YYYYMMDD>00.0.0.tar.gz

*File* <tumor-subset>.miR\_expression.txt

## Fallback

*Package* runs/analyses\_\_<analysis-date:YYYY\_MM\_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.miR\_Clustering\_CNMF.Level\_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz

*File* outputprefix.expclu.gct

*Notes* Contains only the most variable microRNAs.

## microRNA-seq Data Matrices

### Default

*Package* runs/analysis\_\_<analysis-date:YYYY\_MM\_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.miRseq\_Preprocess.Level\_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz

*File* <tumor-base>.uncv2.miRseq\_RSEM\_normalized\_log2.txt

*Altern. File 1* <tumor-base>.mRNAseq\_RPKM\_log2.txt

## Fallback

*Package* runs/analyses\_\_<analysis-date:YYYY\_MM\_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.miRseq\_Clustering\_CNMF.Level\_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz

*File* outputprefix.expclu.gct

*Notes* Contains only the most variable microRNAs.

## DNA Methylation Data Matrices

### Default

*Package* runs/analyses\_\_<analysis-date:YYYY\_MM\_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.Methylation\_Clustering\_CNMF.Level\_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz

*File* outputprefix.expclu.gct

*Notes* Contains only the most variable genes.

## Reverse Phase Protein Array (RPPA) Data Matrices

### Default

*Package* runs/analyses\_\_<analysis-date:YYYY\_MM\_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.RPPA\_Clustering\_CNMF.Level\_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz

*File* outputprefix.exclu.gct

*Notes* Contains only the most variable proteins.

## Patient Clustering by Consensus Non-Negative Matrix Factorization

For mRNA(-seq), microRNA(-seq), DNA Methylation, and RPPA matrices we obtain the following clustering results:

*Package* runs/analyses\_\_<analysis-date:YYYY\_MM\_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.<profile>\_Clustering\_CNMF.Level\_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz

*File* cnmf.membership.txt

## Patient Clustering by Consensus Non-Negative Matrix Factorization

For mRNA(-seq) and microRNA(-seq) matrices we obtain the following clustering results:

*Package* runs/analyses\_\_<analysis-date:YYYY\_MM\_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.<profile>\_Clustering\_Consensus.Level\_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz

*File* <tumor-subset>.allclusters.txt

## Copy Number Calls

*Package* runs/analyses\_\_<analysis-date:YYYY\_MM\_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.CopyNumber\_Gistic2.Level\_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz

*File* all\_thresholded.by\_genes.txt

The copy number calls are represented as ordinal categorical data and we apply the following mapping:

<i>Value</i>	<i>Label</i>
-2	Homozygous deletion
-1	Heterozygous deletion
0	NORMAL
1	Low level amplification
2	High level amplification

## Mutation Calls

*Package* runs/analyses\_\_<analysis-date:YYYY\_MM\_DD>/data/<tumor-base>/<analysis-date:YYYYMMDD>/gdac.broadinstitute.org\_<tumor-subset>.MutSigNozzleReport2.0.Level\_4.<analysis-date:YYYYMMDD>00.0.0.tar.gz

*File* <tumor-subset>.final\_analysis\_set.maf

Only binary status for mutations ('not mutated' or 'mutated') is currently supported. We parse the MAF file row by row and for each row <x>, we assign status 'mutated' to gene <x.HUGO\_Symbol> in patient <x.Tumor\_Sample\_Barcode> and 'not mutated' if there is no such row in the MAF file.

## Clinical Parameters

*Package* runs/data\_\_<data-date:YYYY\_MM\_DD>/data/<tumor-base>/<data-date>/gdac.broadinstitute.org\_<tumor-subset>.Merge\_Clinical.Level\_1.<data-date>00.0.0.tar.gz

*File* <tumor-base>.clin.merged.txt

We use the 'patient.bcrpatientbarcode' field in the extracted clinical data file to identify patients and then map the following clinical parameters (depending on availability):

<i>Label</i>	<i>Field (prefix 'patient.')</i>	<i>Type</i>
Gender	gender	categorical
Ethnicity	ethnicity	nominal categorical
Race	race	nominal categorical
Age (days)	daystobirth	natural number
Days to Death	daystodeath	natural number
Vital Status	vitalstatus	nominal categorical
Age At Initial Pathologic Diagnosis	ageatinitialpathologicdiagnosis	natural number
Days To Last Follow Up	daystolastfollowup	natural number
Histological Type	histologicaltype	nominal categorical
Tumor Tissue Site	tumortissuesite	nominal categorical
Radiation Risk Exposure	personlifetimeriskradiationexposureindicator	categorical
Lymph Node Assessment	primarylymphnodepresentationassessment	ordinal categorical
Focus Type	primaryneoplasmfocustype	nominal categorical

<i>Label</i>	<i>Field (prefix 'patient.stageevent.')</i>	<i>Type</i>
Overall Stage	pathologicstage	ordinal categorical
T Stage	tnmcategories.pathologiccategories.pathologicct	ordinal categorical
N Stage	tnmcategories.pathologiccategories.pathologiccn	ordinal categorical
M Stage	tnmcategories.pathologiccategories.pathologiccm	ordinal categorical

### **Preprocessing Steps performed by the Package Builder**

The following preprocessing is performed by the package builder on the extracted data matrices containing mRNA(-seq), microRNA(-seq), DNA Methylation, and RPPA measurements.

1. Missing values in gene/microRNA/protein expression matrices and DNA methylation matrices are imputed using a k-nearest neighbors (kNN) imputation algorithm [4]. We chose  $k = 10$  and determine the distance between a gene/microRNA/protein expression profile or DNA methylation profile  $X$  with missing values and all other profiles  $Y_i$  by computing  $d(X, Y) = \text{mean}((X_i - Y_i) * (X_i - Y_i))$  (squared and normalized Euclidean distance) for all  $i$ , based on all non-missing values. The missing value for a patient  $p$  in  $X$  is replaced by the average value for the given patient  $p$  across the  $k$  nearest  $Y_i$ . If a gene  $X$  contains more than 50% missing values, the missing values are imputed using the global patient mean, instead of the mean of the  $k$  nearest neighbors mean, because it is unlikely to find appropriate neighbors in such cases. The patient mean is also used in situations where all neighbors have missing values for the patient for which a missing value is to be imputed.
2. The matrices are z-score normalized unless the fallback options are used (which are already z-score normalized), where each entry  $x$  is replaced with  $(x - \text{mean})/\text{sd}$ , where  $\text{mean}$  and  $\text{sd}$  correspond to the gene/microRNA/protein expression profile mean and standard deviation, respectively.
3. If a full gene matrix is available for a given data type, a sampled version is created for visualization purposes. The sampled matrix contains the 1500 most variable genes, according to their median absolute deviation (MAD). Genes with more than 80% missing values are discarded.
4. For sampled matrices, we apply hierarchical clustering with average linkage using Euclidean distance to the mRNA/microRNA/protein dimension of the matrix. This step is performed for improved visualization of the data in heatmaps, which display only the sampled data.

## References for Supplementary Methods

- [1] Hubert, L. and Arabie, P., Comparing partitions. *Journal of Classification* **2**, 193-218 (1985)
- [2] Subramanian, A., *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545-15550 (2005)
- [3] Kim, S.-Y., and Volsky, D.J., PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics* **6**, 144 (2005)
- [4] Troyanskaya, O., *et al.*, Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525 (2001)