City Research Online

City St George's, University of London

Citation: Turkay, C., Lex, A., Streit, M., Pfister, H. & Hauser, H. (2014). Characterizing Cancer Subtypes Using Dual Analysis in Caleydo StratomeX. IEEE Computer Graphics and Applications, 34(2), pp. 38-47. doi: 10.1109/mcg.2014.1

This is the unspecified version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/3642/

Link to published version: https://doi.org/10.1109/mcg.2014.1

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Characterizing Cancer Subtypes using the Dual Analysis Approach in Caleydo

Cagatay Turkay, Alexander Lex, Marc Streit, Hanspeter Pfister, and Helwig Hauser

Abstract—The comprehensive analysis and characterization of cancer subtypes is an important problem to which significant resources have been devoted in recent years. In this paper we integrate the dual analysis method, which uses statistics to describe both the dimensions and the rows of a high dimensional dataset, into *StratomeX*, a *Caleydo* view tailored to cancer subtype analysis. We introduce *signi cant difference plots* for showing the elements of a candidate cancer subtype that differ significantly from other subtypes, thus enabling analysts to characterize cancer subtypes. We also enable analysts to investigate how samples relate to the subtype they are assigned and to the other groups. Our approach gives analysts the ability to create well-defined candidate subtypes based on statistical properties. We demonstrate the utility of our approach in three case studies, where we show that we are able to reproduce findings from a published cancer subtype characterization.

Index Terms—Cancer Subtypes, Biological Data Visualization.

1 Introduction

Cancer is one of the most-common causes of death and virtually everyone is or will be either directly or indirectly affected by it. While there has been significant progress in the diagnosis, prevention, and treatment of cancer, there are still many open questions to be answered, methods to be improved, and drugs to be developed. While cancer is a multifactorial disease, involving environmental factors and lifestyle choices, it has a strong genetic component. In the post-genomic age research on cancer is largely conducted using methods of molecular biology to record and analyze the genetic alterations responsible for cancer. One important field in cancer research is the analysis and characterization of cancer subtypes. While cancers are colloquially referred to by the tissue they originate from (e.g., lung cancer because it occurs in the lung), there are in fact significant differences between cancers from the same tissue, which are characterized by various biomolecular properties. These different forms of cancer are called subtypes. Large scale research projects such as *The Cancer Genome Atlas (TCGA)*[1] elicit comprehensive genomic and clinical datasets with the goal of characterizing the molecular alterations responsible for cancer; and of identifying and characterizing cancer subtypes.

Due to next-generation sequencing and micro-array technology, these projects can utilize large and heterogeneous datasets capturing more aspects of the complex process from the genomic information to the functional consequences than ever before. However, deriving insight from these complex datasets remains a challenging task. Current analysis largely relies on custom scripts to find interesting genes or clusters of patients in these datasets. To remedy this, we have developed *Caleydo StratomeX* [1], an interactive visualization method to analyze and discover relationships within large and heterogeneous biomolecular datasets. StratomeX can be used to evaluate overlaps and relationships of patient *stratifications*, i.e., groupings or clusterings of patients.

However, StratomeX does not enable analysts to identify the characteristic genes of candidate subtypes, nor does it communicate how patients relate to a given subtype. The former is important since the characteristic genes are also potentially causally involved in a subtype and thus may be a target for a therapeutic or diagnostic approach. The latter, investigating how samples relate to a subtype, can be used to estimate the quality of candidate subtypes and to build a deeper characterization of a subtype.

In this paper, we address these limitations by integrating the *dual analysis approach* [2], a general high-dimensional data analysis methodology, into StratomeX. Our primary contribution is the embedded use of dual analysis views and *significant difference plots*, a novel visual representation of the differences between data subsets, within StratomeX. This approach enables domain scientists to (1) discover genes that are distinctive for specific subtypes, and (2) observe the properties of the member samples of a cluster and compare how they behave in different

- *Cagatay Turkay is both with the Department of Informatics, University of Bergen, Norway, and the giCentre at City University, London, UK. E-mail: Turkay.Cagatay.1@city.ac.uk*
- *Alexander Lex and Hanspeter Pfister are with School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA E-mail: {alex, pfister}@seas.harvard.edu*
- *Marc Streit is with Institute of Computer Graphics, Johannes Kepler University Linz, Linz, Austria E-mail: marc.streit@jku.at*
- *Helwig Hauser is with the Department of Informatics, University of Bergen, Norway E-mail: Helwig.Hauser@uib.no*

1. http://cancergenome.nih.gov

datasets and clusters. With these, we provide a deeper understanding of the stratifications of heterogeneous genomics datasets. As a secondary contribution, we investigate the potential of the dual analysis approach to interactively generate patient stratifications in StratomeX.

We demonstrate our application in three case studies with data from TCGA and validate our findings against those published by the TCGA consortium.

2 Biological Background and Analysis Tasks

Modern cancer subtype analysis is based on a variety of biomolecular datasets that capture different aspects of the process of life, starting with the information stored in the genome to the functional products that trigger biochemical reactions in the cells. Projects such as TCGA capture information on gene activity, on factors influencing the process of expression, and on the actual structure and sequence of the genome. An example for gene activity data is mRNA data ("gene expression"), which measures the abundance of mRNA in the cell. mRNA is translated into proteins, which are the functional products. Methylation and miRNAs influence the process of gene expression in various ways and thus are an important factor in many processes and diseases.

All these processes play a role in the development of certain cancers, and consequently, a comprehensive analysis solution needs to take all these datasets, in addition to meta-data, such as clinical data about patients, into account. In this paper, we demonstrate our method by investigating mRNA, miRNA, and methylation data. However, in a comprehensive analysis one would also incorporate other datasets, for instance, related to structural variations occurring on various scales in the genome.

In previous work, we have elicited analysis tasks for cancer subtype analysis [1]. These tasks are concerned with finding and evaluating stratifications of patients based on multiple datasets. We recently revisited these requirements in collaboration with domain scientists and found the need to supplement them with the following tasks to characterize the stratifications further:

T1 **Find Distinctive Elements**
Identifying distinctive elements of clusters in a stratification provides a deeper understanding of why a particular cluster exists and how it relates to other clusters within the analysis. Distinctive elements are also good candidates to investigate as diagnostic markers or may even be causally involved in the disease.

T2 **Compare Samples**
Investigating the characteristics of the samples over several datasets and in comparison to other stratifications is important in building a more complete picture of the properties of a group of samples. One can observe how strongly the members of a cluster are related and explore whether they show similar properties in a dataset that is different than the one used for clustering.

T3 **Create Clusters**
Analysts should be able to create clusters in an exploratory manner and interactively compare the intermediate results to meta-data such as clinical data. Moreover, this manual clustering process should enable analysts to merge observations made in different datasets. The thus created clusters are well defined in terms of statistical properties and richer in terms of the sources of information included in the construction phase.

Combined with the previously elicited ones, these tasks make it possible to analyze, create, and characterize cancer subtypes based on multiple datasets.

3 Methodological Building Blocks

Our solution that enables the aforementioned tasks is based on an integration of two visual analysis methodologies, *Caleydo StratomeX* and the *Dual Analysis Approach*. Before introducing the details of how we improve these methodologies by joining their strengths, we provide brief descriptions of them.

3.1 Caleydo and StratomeX

Caleydo[2] is an open-source visualization framework focused on biomolecular data analysis. Caleydo provides rich functionality for loading and handling multiple heterogeneous datasets as well as stratifications defined on the data. A core strength of Caleydo is the ability to slice datasets into meaningful subsets and to flexibly combine multiple small visualizations of these subsets, using views such as histograms or heat maps, to a fully integrated composite visualization [3]. Caleydo is one of examples of a visual method that improve the analysis of genomics data, other well-known tools are the Hierarchical Cluster Explorer [4] and Mayday [5].

StratomeX is a comparative visualization technique that makes use of the slicing concept. It enables analysts to investigate the relationships between multiple stratifications (patient groupings) which are represented as columns. Each column consists of multiple stacked "blocks", where each brick corresponds to a group of patients in the column's stratification. Ribbons with varying width visualize the overlap between groups of neighboring stratifications, resulting in an overall appearance similar to *Parallel Sets* [6] or *Sankey Diagrams* [7]. Wide ribbons indicate a strong overlap between two groups and thin or absent ribbons correspond to only a few or no shared patients. Each brick contains a visualization showing the data of the patients in that group. Analysts can switch

2. http://www.caleydo.org

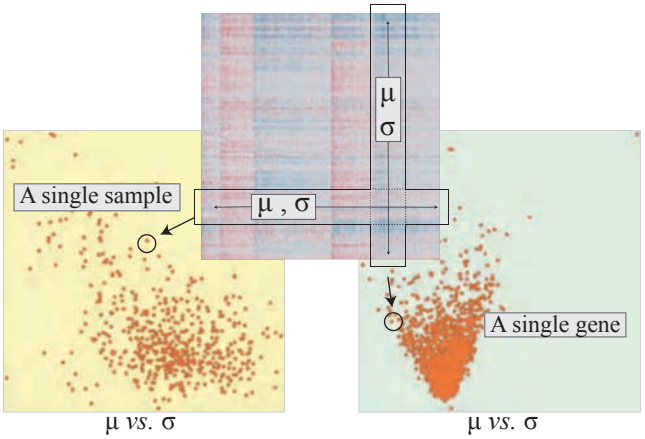


Fig. 1. Setting up dual analysis views where the data is depicted as a 2D heatmap for illustration. Samples and genes are visualized in separate views over statistical measures. In order to construct a view that depicts samples (yellow background), statistics for each sample (μ and σ in this case) are computed using a *row* of the data. A visualization for the genes (light-green background), on the other hand, is constructed with statistics computed over a *column* of the data.

between different types of visualizations on-demand. For numerical data we use clustered heat maps as default views within the blocks in StratomeX, since they are very effective for communicating global trends and patterns in the data.

3.2 The Dual Analysis Approach

The dual analysis approach [2] was shown to be effective in the analysis of high dimensional data. In this method, the visual analysis is carried out in parallel on both the data items and the dimensions. This duality is achieved by using statistics computed both over the rows and the columns of a dataset.

As an example, consider an mRNA gene expression dataset given as a 2D data table with n rows and p columns, where each row corresponds to a single sample (patient) and each column to a single gene. The expression values are contained in the cells of the matrix.

After appropriate normalization is applied to the data, we calculate the central tendency (μ or $median$) and the spread (standard deviation σ or inter-quartile range IQR) using each one of the n samples and p genes separately. Notice that we calculate the robust counterparts of statistical moments to increase the resistance of the statistics to outlier values. Since experts are often accustomed to using non-robust versions of the statistics (e.g., μ or σ), we incorporate such measures in our system. This helps users to quickly familiarize themselves with the information communicated in the views and at any point during an analysis, the experts have the flexibility to modify the set of statistics used. Figure 1 illustrates how the dual analysis views are constructed. Notice that visualizations of samples have a yellow background with each point representing a sample, and visualizations of genes have a light-green background with each point depicting a gene. The location of a single point in a scatterplot is determined by the computed statistics. The analysis process can be elaborated through the use of statistics other than the first two statistical moments. For the analyses carried out in this paper, we also compute the *skewness* (*skew*) that indicates how asymmetric a distribution of values is (and also in which direction) and the *kurtosis* (*kurt*) that characterize the "peakedness". How these measures are utilized is demonstrated in the case studies.

4 Characterizing Cancer Subtypes through Visual Analysis

To facilitate the characterization of cancer subtypes in heterogeneous genomic and clinical datasets, we introduce a visual analysis methodology that makes use of the dual analysis approach to construct specialized views that represent clusters in Caleydo. We achieve this by incorporating two different visualizations as *blocks* in StratomeX: (1) dual analysis based scatterplots depicting either the genes or the samples, and (2) *significant difference plots*. In addition, we also use these visualizations as separate linked views to enhance the interactive visual exploration process and achieve tasks such as manual creation of clusters (Task **T3** in Section 2).

4.1 Embedded dual analysis views

In this work, we extend the visualization options for *blocks* in StratomeX with scatterplots of either the genes or the samples constructed using the dual analysis approach. The embedded dual analysis views in StratomeX can be seen in Figure 2. If the embedded scatterplot is a visualization of the samples (having a yellow background), it only displays those samples that are members of the represented cluster (see Columns 1 and 2 in Figure 2). On the other hand, if a scatterplot of genes is preferred, the brick displays the statistics for all the genes computed using only the members of the cluster being represented.

We enhance the interactive exploration functionalities by enabling a selection mechanism that is linked with all the views in StratomeX. It is possible to select both samples (selection in the second cluster in the second column of Figure 2) and genes (selection in the second cluster in the third column in Figure 2) at the same time. Also note that the ribbons in StratomeX highlight the selection of the samples in Figure 2.

4.2 Significant difference plots

Since the comparison of subsets is one of the fundamental tasks in tumor subtype analysis, we facilitate

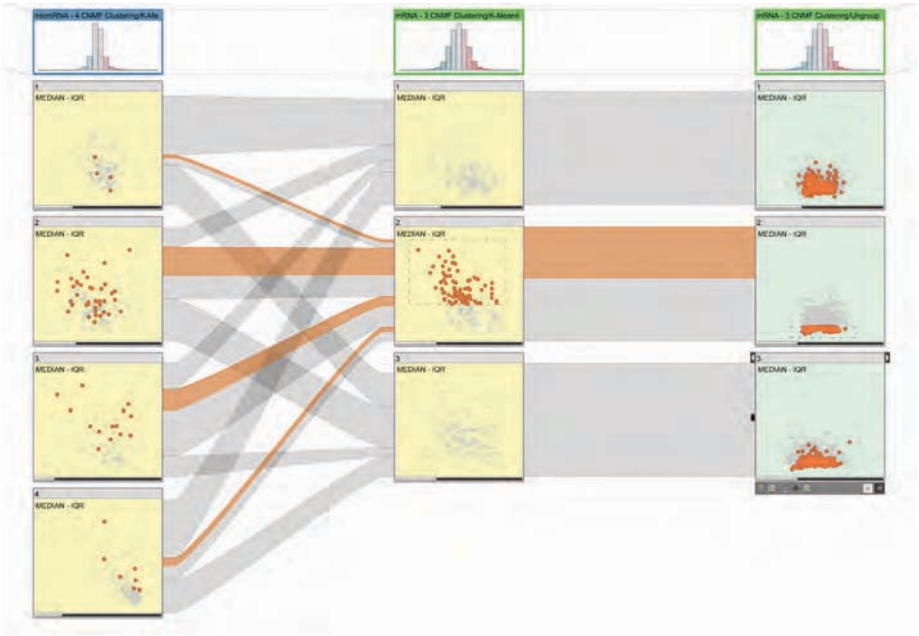


Fig. 2. Embedded dual analysis views in StratomeX. The first column shows a 4-cluster stratification for a microRNA dataset. The scatterplots show median versus inter-quartile-range for the samples in the cluster. The second column shows a 3-cluster stratification for a mRNA dataset, again showing samples. The third column uses the same 3-cluster stratification for the same dataset, but shows genes instead of samples. The scatterplots of samples (yellow background) depict the statistical characteristics of the members of each cluster and the scatterplots of genes (light-green background) depict statistics computed for the genes using only the samples from the cluster represented by the brick. The selection of samples is highlighted in the first two columns and also in the ribbons. The selection of the genes makes it possible to investigate the distribution of expression values for the genes for different clusters in a stratification.

the visual comparison of subsets with a novel visualization called *significant difference plot*. In previous work, we used similar plots to effectively display the changes in statistical computations in response to a selection made by the user [8]. In this paper, we extend this approach with the determination and the communication of the significance of the differences being visualized.

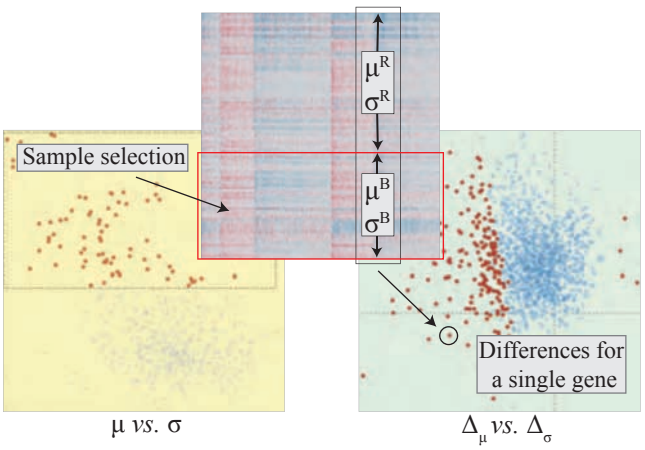


Fig. 3. Significant difference plot. A set of samples is selected. The differences of the selected samples (B) compared to the not-selected samples (R) is plotted for the genes. Genes that show significant differences are depicted in red and all others in blue.

Figure 3 illustrates how significant difference plots (or, shortly difference plots) are constructed. The user first selects (brushes) a subset of samples (we denote the set of selected samples as B and the rest as R). In response, the system automatically calculates the μ and σ values for each gene using only the set of selected samples B (μ^B and σ^B) and the rest of samples R (μ^R and σ^R) separately. We then compute the differences between the values with:

$$\Delta_\mu = \mu^R - \mu^B \quad , \quad \Delta_\sigma = \sigma^R - \sigma^B \tag{1}$$

Note that Δ_μ and Δ_σ are both data vectors of size p, the number of genes. The difference plot then visualizes these values for all the p genes. When there is no difference for the expression values of a gene for subsets S and R, it is placed at the origin $(0, 0)$.

The difference plot in Figure 3 (right) displays the distribution of the differences in the statistic computations in response to the (sample) selection in the scatterplot (Figure 3 left). Notice that in this example most genes have lower μ values for the selected items, i.e., are placed to the left of the y-axis.

Communicating signi cance – One very important consideration when differences between two subsets are analyzed is the notion of *statistical significance*, i.e., whether the difference is likely to occur by chance or not. As in many other domains, *statistical hypothesis test*s are employed to test for significance in the analysis of genomic data [9]. In this work we enhance difference plots with the integrated use of statistical hypothesis testing.

In order to compute the significance, we utilize the *two-sample Welch's t-test* as the integrated hypothesis testing procedure [10]. We choose this test since it does not assume that the two subsets have equal variance, which makes it more suitable for our application. We perform the statistical test on the two subsets B and R (as introduced above), and test against the (*null*) hypothesis that these two subsets have equal central tendencies. We compute the t statistic and the degrees of freedom $d.f.$ with:

$$t = \frac{\overline{\mu}_B - \overline{\mu}_R}{\sqrt{\frac{s_B^2}{N_B} + \frac{s_R^2}{N_R}}} \quad (2)$$

$$d.f. = \frac{(s_B^2/N_B + s_R^2/N_R)^2}{(s_B^2/N_B)^2/(N_B - 1) + (s_R^2/N_R)^2/(N_R - 1)} \quad (3)$$

where $\overline{\mu}_i$ is the sample mean, s_i^2 is the sample variance and N_i is the sample size of subsets B and R.

We then use these values together with the t-distribution and test the null hypothesis with a significance level of 0.05 and using a two-tail strategy. This test is performed for all the p genes in the data. For each gene, we store whether it shows a significant difference between the two subsets B and R. We communicate this significant difference information by modifying the color of each gene in the difference plot. Genes that have significant differences are colored red, while the others are shown in blue, as can be seen in Figure 3 (right). This enhancement to the difference plot enables analysts to get immediate feedback on the significance of differences. Based on this initial assessment, analysts can employ more advanced routines to confirm the significance of the changes between the two subsets.

Difference plots as blocks – Similar to scatterplots, we also embed difference plots as blocks in StratomeX. While constructing the difference plots as blocks, we again compute the Δ_μ and Δ_σ values for each of the genes using Equation 1. Here, however, B corresponds to the samples that are members of the cluster being represented while R corresponds to the rest of the samples in the dataset. In addition, we also compute the significance of the differences and color the visualization accordingly. The resulting difference plot blocks communicate which genes are more distinctive for each cluster. Moreover, the selection mechanism enables the analyst to compare these distinctive genes between different clusters. For an example of this feature, refer to the first part of Section 5.

5 Case Studies

We demonstrate the effectiveness of our approach through an analysis of a comprehensive breast invasive carcinoma (BRCA) dataset collected by the TCGA consortium. We use the mRNA expression data, miRNA sequencing data, and DNA methylation data from over 800 breast cancer patients. The goal of the case studies is to demonstrate how the proposed visual analysis approach enables analysts to execute the three tasks described in Section 2. To begin with, we load the BRCA data which is available prepackaged for Caleydo. In addition to the raw data, we load a recently published stratification of samples [11] that will serve as a basis for comparisons.

5.1 T1 Case Study: Find Distinctive Elements

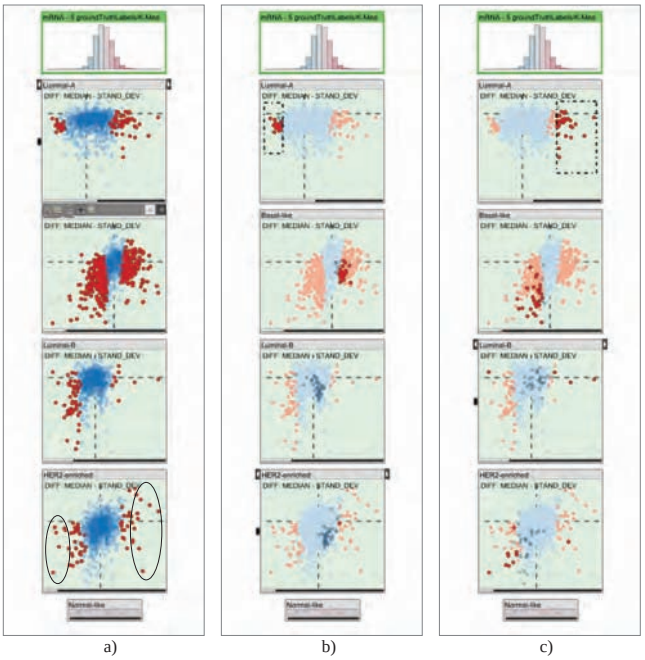
We start our analysis by comparing the significantly distinctive genes that are suggested by our computations and those that have been identified in the aforementioned article. The 4 subtypes that are reported in the reference study are: *Luminal-A*, *Basal-like*, *Luminal-B*, and *HER2-enriched*, as shown in Figure 4-a). The reference study identified a list of genes that are differentially expressed for the *HER2-enriched* subtype by using unsupervised clustering (refer to supplementary Table 7 of the BRCA study [11]). We select the 7 most significantly under-expressed genes[3] and 10 most significantly over-expressed genes[4] as marked in Figure 4-a. 7 out of the 7 under-expressed and 6 out of 10 over-expressed genes are identical to the ones found in the reference study. This match demonstrates that our interactive visual analysis approach quickly yields relevant results in determining descriptive genes.

We continue our analysis with the investigation of distinctive genes between particular subtypes (see task **T1**). We focus our attention on the *Luminal-A* subtype and explore the expression characteristics of distinctive genes for *Luminal-A* in comparison to the other subtypes. We first select the significantly under-expressed genes[5] for the *Luminal-A* subtype in Figure 4-b. We observe that the *significantly under-expressed genes for Luminal-A are often over-expressed for the Basal-like subtype*. This leads to the conclusion that *these genes are good markers to distinguish the Luminal-A from the Basal-like subtype*. Similarly, when the over-expressed genes are selected for the *Luminal-A* subtype (Figure 4-c), we observe that these genes

3. *AGR3, ESR1, GFRA1, NPY1R, PGR, SERPINA3, SUSD3*

4. *ABCA12, CALML5, CLCA2, CRYM, DCD, GLYATL2, MUCL1, NXPH1, PNMT, SOX11*

5. *AQP9, FAM83D, GGH, MCM10*, and *MMP1* being some of the lowest



a) b) c)

Fig. 4. Using embedded difference plots to find descriptive genes. (a) Descriptive genes are marked for the *HER2-enriched* subtype. A comparison to the reference study shows the relevance of the marked genes. (b) Under-expressed genes for the *Luminal-A* subtype are selected and we observe that they show over-expression for the *Basal-like* subtype, i.e., constitute good features to discriminate these two subtypes. (c) The over-expressed genes for Luminal-A could also be considered good discriminators for this subtype but show similar expression profiles for *Basal-like* and *HER2-enriched* subtypes.

are under-expressed for *Basal-like* subtype. However unlike the previous set, these genes also show similar expression profiles for the *HER2-enriched* subtype. Consequently, these genes carry less distinctive characteristics compared to the previous set.

5.2 T2 Case Study: Compare Samples

In the second case study we investigate how certain properties of samples from a particular subtype, for instance outliers or trends, are shared among different datasets (**T2**).

We start with an investigation of the characteristics of samples from the *Basal-like* subtype by considering the mRNA, microRNA, and DNA methylation datasets. We bring up a StratomeX view with the subtypes from the reference study as the first column and unstratified versions of the datasets mRNA, microRNA, and methylation from left to right, as shown in Figure 5-a. When all samples from the *Basal-like* subtype are selected, we observe the following that further characterizes this subtype: samples from the *Basal-like* subtype have *lower expression values with a high variance in mRNA* and have *higher expression values in the microRNA* dataset. When looking at their *DNA methylation values, however, we do not observe any dominant characteristics*.

We use the same approach to determine the characteristics of a cluster that is computed as a result of an unsupervised clustering of the mRNA dataset (first column in Figure 5-b). We select the "core members" of the second cluster, i.e., those that have similar expression values and variance. We observe that these samples do not show any dominant characteristics in an unsupervised clustering of microRNA data (second column in Figure 5-b). However, when considering the reference subtypes from the BRCA study (third column in Figure 5-b), we observe that the selected samples constitute a subgroup of the *Luminal-A* subtype. We can also see that these samples are the over-expressed *Luminal-A* members with a lower variance. Based on this observation, we can claim that *cluster-2* from the mRNA stratification can be utilized to determine a subgroup of *Luminal-A*.

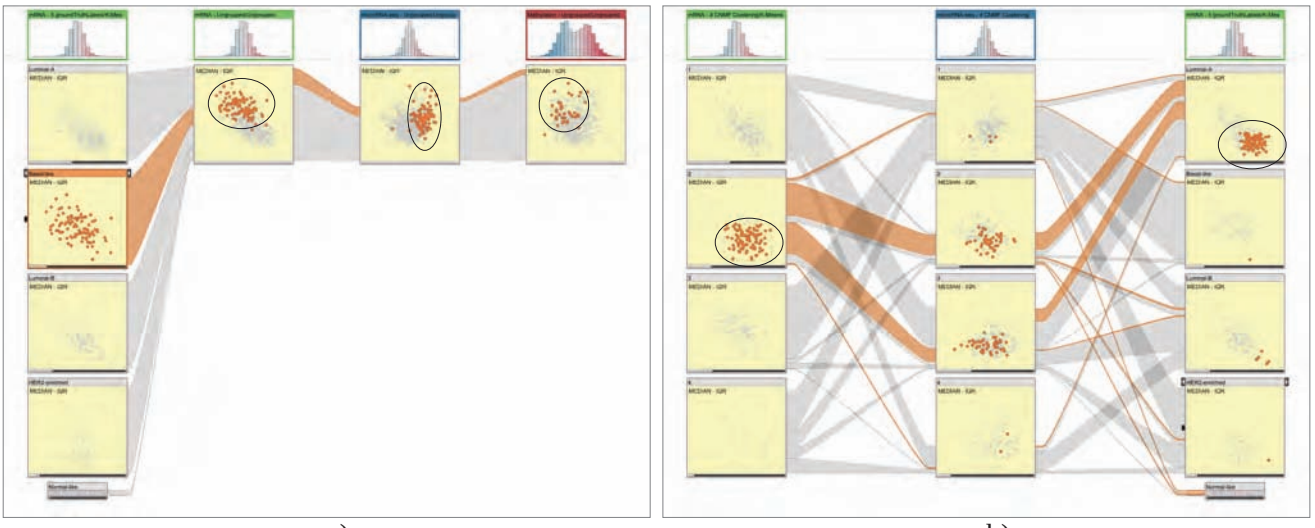
5.3 T3 Case Study: Create Clusters

In certain cases in tumor subtype analysis, the stratification information is not readily available. In these cases, we make use of the dual analysis methodology to manually create stratifications as an alternative to automated methods (**T3**). This mechanism enables the analyst to discover structures through different views of multiple datasets and represent these structures as a stratification.

For demonstrate such a manual clustering process on the BRCA data. In this process, we use dual analysis views as separate linked views rather than embedded in StratomeX, i.e., the selections in any of the views are highlighted in the others. We bring up two linked views of the mRNA dataset: *skew* vs. *kurt* visualization of the genes (Figure 6-a) and a difference plot for the samples for Δ_μ vs. Δ_σ (Figure 6-b). Also, we add two other views of the mRNA-seq dataset: *median* vs. *IQR* visualization of the genes (Figure 6-c,e) and a difference plot for the samples for Δ_μ vs. Δ_σ (Figure 6-d,f).

We start by marking an unstratified mRNA dataset as the target for the manual clustering (through a user interface not shown in the images) and the clustering process is then as follows:

Step-1: Here, we make use of the skewness of the distribution of the values for the genes. High skewness indicates that a gene has non-uniform expression levels over the samples and thus is a good candidate to be a discriminator between subtypes. Therefore in this example, we select the genes that are left-skewed (negative skew values) (Figure 6-a) and select a group of samples that are visually separated from the rest (left of the difference plot Figure 6-b). At this point, we mark this subset of samples as a stratification of the



a) b)

Fig. 5. (a) Investigating the sample profiles for *Basal-like* subtype (column 1) over three different datasets (left-to-right: mRNA, microRNA, and methylation). The subtype contains samples with lower values and high variance for mRNA data and usually higher values in the microRNA data. In the methylation data, however, no dominant characteristic is observed. (b) "Core" members of a cluster from an unsupervised stratification of mRNA data are selected (marked, left) and visualized with a microRNA stratification (column 2) and the subtypes. We observe that the selected members correspond to a subgroup in the *Luminal-A* subtype (marked, right).

mRNA dataset (the first cluster in the first column of StratomeX in Figure 6-g). This operation is performed through the UI which is not shown in the image.

Step-2: We now switch to the mRNA-seq dataset and select those genes that have higher expression values and higher variety (Figure 6-c). The difference plot is updated automatically and we select those samples that have higher expression values and lower variance (Figure 6-d). Notice that we make use of the difference plot here and select those in the lower-right quadrant of the view, i.e., high values and variety. Also note that the selection here is guided by the axes of the visualization rather than the observed visual structures as in the first step – this amounts to another strategy to make interesting selections. We make this selection due to the fact that one would expect to see higher variance and higher values for the samples in response to the selection of genes in Figure 6-c. We finish this step by marking the selection of samples as a second cluster.

Step-3: Without updating the selection of genes, we move on by selecting the samples that have higher variety but smaller mRNA-seq values for the selected genes (Figure 6-f). Notice that the selection here corresponds to the upper-left quadrant of the difference plot, i.e., lower values, higher variety. This last selection of samples is marked as the third cluster in the data. The rest of the samples are left as an unclustered set.

In order to evaluate our custom stratification, we compare it against the classification from the reference study (Figure 6-g). We observe that the cluster made in Step-1, characterized with genes that have negative skewness, has almost a complete overlap with the *Basal-like* subtype. The second cluster from Step-2 largely corresponds to a subgroup of *Luminal-A* subtype. Finally, more than half of the samples from the third cluster belong to the *HER2-enriched*. This overlap between the manually created clusters and the reference subtypes show that the manual clustering leads to relevant results.

Our interactive approach enables analysts to consider different data sources in the manual clustering steps (mRNA and mRNA-seq in this case). This makes it possible to merge interesting structures observed in several datasets using different perspectives on the data, i.e., using a *skew* vs. *kurt* view for the mRNA and a *median* vs. *IQR* for the mRNA-seq dataset. This flexibility leads to outcomes that are not so straightforward to generate through automated methods. Moreover, the manual clustering process provides a mechanism to externalize the findings of the analysis. Manually generated clusters become parts of the analysis that can be compared with the automatically computed results, e.g., manually built clusters vs. hierarchical clustering results.

6 Conclusion

In this paper, we integrate dual analysis views and significant difference plots within Caleydo StratomeX, a state-of the art cancer subtype visualization tool. Our approach facilitates the characterization of cancer subtypes by enabling an investigation of them over both the samples and the genes. Such a duality in

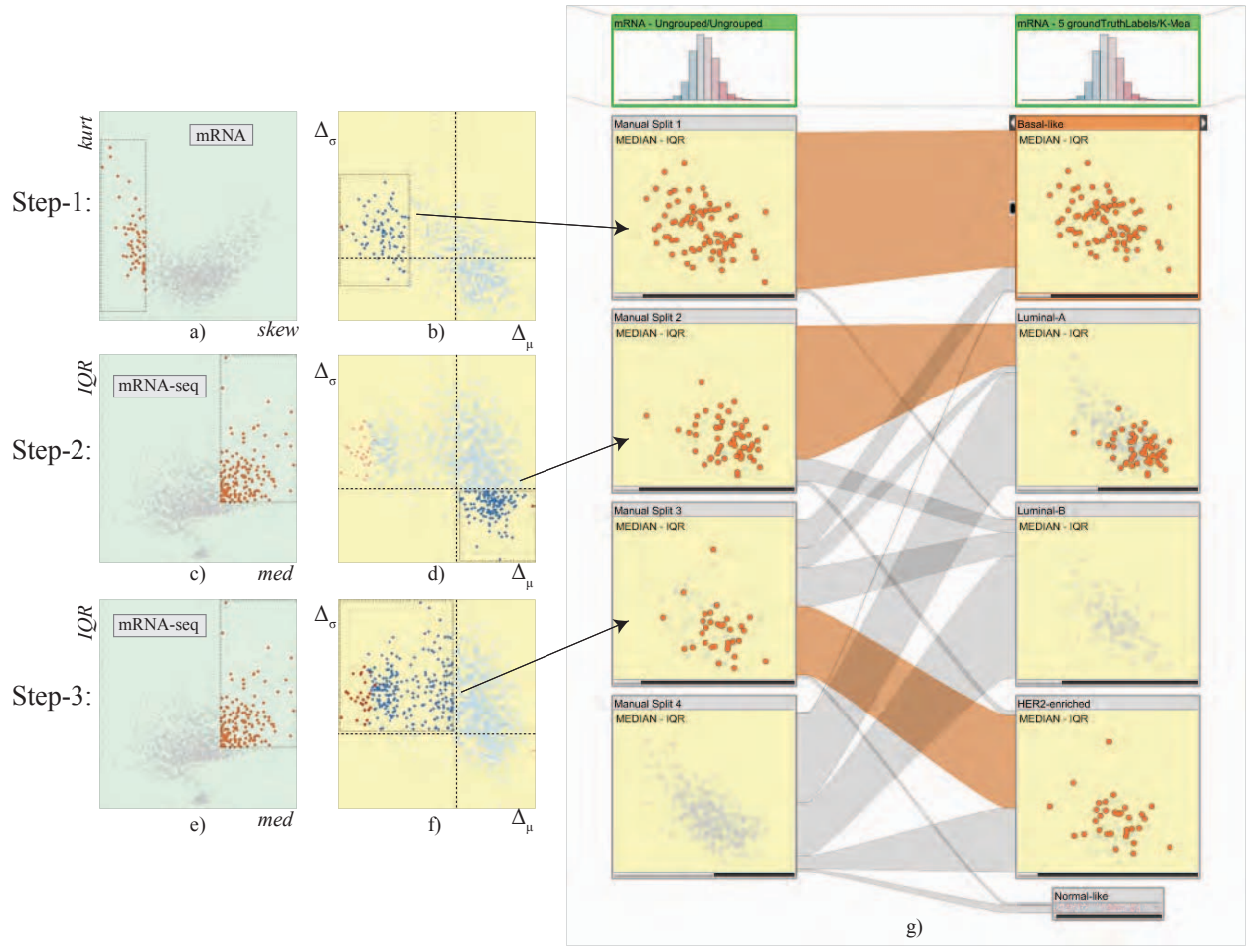


Fig. 6. Manual clustering of unstratified mRNA dataset using dual analysis views. Negatively skewed genes are selected through $skew$ vs. $kurt$ visualization (a) and the difference plot for the samples is updated automatically (b) where we observe a group of samples with lower values and mark them as our first cluster (b). We then switch to the mRNA-seq dataset and select genes that are higher-expressed with a large variety within the values (c,e). We identify two groups and mark them as clusters 2 (d) and 3 (f). For validation, we compare our stratification with the subtypes from the reference study and observe a significant overlap with the subtypes.

representing stratifications provides deeper insight on the characteristics of subtypes. Using Caleydo's multi-dataset capabilities we are able to generate such insights based on different datasets, as demonstrated in **T2** in Section 5.

We also demonstrate how the dual analysis approach can be used to create clusters based on statistical properties and merge structures from different datasets, a challenging task to achieve through automated methods. We show the utility of our approach in three case studies. In concert with the existing StratomeX functionality, we have observed that we have created a powerful tool for experts to analyze and characterize cancer subtypes.

In the future, we aim to integrate advanced statistical tests and procedures, such as the analysis of variance (ANOVA), Bonferroni correction [9] and dimension reduction methods. We plan to include these methods through the integration of the statistical computing environment R [12]. We also consider to extend the capability of difference plot to depict the comparison of more than two groups. Furthermore, instead of comparing one cluster to all the other elements, we plan to implement mechanisms to compare clusters with each other.

Acknowledgments

We thank Nils Gehlenborg, Samuel Gratzl, and Christian Partl for their input. This work is supported by the grant J 3437-N15 by the Austrian Science Fund (FWF) and the Air Force Research Laboratory and DARPA grant FA8750-12-C-0300.

References

[1] A. Lex, M. Streit, H. Schulz, C. Partl, D. Schmalstieg, P. J. Park, and N. Gehlenborg, "StratomeX: visual analysis of Large-Scale heterogeneous genomics data for cancer subtype characterization," *Computer Graphics Forum (EuroVis '12)*, vol. 31, no. 3, pp. 1175–1184, 2012.

[2] C. Turkay, P. Filzmoser, and H. Hauser, "Brushing dimensions – a dual visual analysis model for high-dimensional data," *IEEE Transactions on Visualization and Computer Graphics (InfoVis '11)*, vol. 17, no. 12, pp. 2591 –2599, 2011.
[3] A. Lex, H.-J. Schulz, M. Streit, C. Partl, and D. Schmalstieg, "VisBricks: multiform visualization of large, inhomogeneous data," *IEEE Transactions on Visualization and Computer Graphics (InfoVis '11)*, vol. 17, no. 12, pp. 2291–2300, 2011.
[4] J. Seo and B. Shneiderman, "Interactively exploring hierarchical clustering results," *Computer*, vol. 35, no. 7, pp. 80–86, 2002.
[5] J. Dietzsch, N. Gehlenborg, and K. Nieselt, "Mayday – a microarray data analysis workbench," *Bioinformatics*, vol. 22, no. 8, pp. 1010–1012, 2006.
[6] R. Kosara, F. Bendix, and H. Hauser, "Parallel sets: Interactive exploration and visual analysis of categorical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 4, pp. 558–568, 2006.
[7] P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive sankey diagrams," in *Proceedings of the IEEE Symposium on Information Visualization (InfoVis '05)*. IEEE, 2005, pp. 233–240.
[8] C. Turkay, J. Parulek, and H. Hauser, "Dual analysis of dna microarrays," in *Proceedings of the Conference on Knowledge Management and Knowledge Technologies (i-KNOW '12)*, 2012.
[9] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, vol. 7, no. 1, pp. 55–65, 2006.
[10] G. D. Ruxton, "The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test," *Behavioral Ecology*, vol. 17, no. 4, pp. 688–690, 2006.
[11] D. Koboldt, R. Fulton, M. McLellan, H. Schmidt, J. Kalicki-Veizer, J. McMichael, L. Fulton, D. Dooling, L. Ding, E. Mardis *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.
[12] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org

Cagatay Turkay is a faculty member and lecturer in Applied Data Science at the giCentre at Department of Computer Science at City University London, United Kingdom. He has a Ph.D. degree from University of Bergen, Norway and MSc.from Sabanci University, Istanbul, Turkey. His research mainly focuses on the tight integration of interactive visualizations, data analysis techniques and supporting exploratory knowledge and capabilities of experts. He has a special interest in high-dimensional, temporal data from bioinformatics and the biomolecular modelling domain.

Alexander Lex is a post-doctoral visualization researcher at the Visual Computing Group at the Harvard School of Engineering and Applied Sciences. He did his PhD, Master's and Undergraduate studies at at the Institute for Computer Graphics and Vision at the Graz University of Technology, Graz, Austria. His primary research interests are data visualization especially applied to molecular biology, and human computer interaction. His focus is his work on Caleydo, which is both, software that can be used by life science experts to visualize biomolecular data and pathways, but also a platform for implementing prototypes of radical visualization ideas.

Marc Streit is assistant professor at the Institute of Computer Graphics at the Johannes Kepler University Linz in Austria, where he is leading the visualization group. He finished his PhD at Graz University of Technology in early 2011 and moved to Linz later that year. His scientific areas of interest include Information Visualization, Visual Analytics, and Biological Data Visualization, where he is particularly interested in the integrated analysis of large heterogeneous data. Dr. Streit has won multiple best paper awards at major visualization conferences. He is recognized for his work on cancer subtype analysis and the visualization of pathways. His research is embedded in the Caleydo project (www.caleydo.org), where he is one of the project leaders and founding-members.

Hanspeter Pfister is Gordon McKay Professor of Computer Science in the School of Engineering and Applied Sciences at Harvard University. His research in visual computing lies at the intersection of visualization, computer graphics, and computer vision. It spans a wide range of topics, including bio-medical visualization, 3D reconstruction, GPU computing, and data-driven methods in computer graphics. Before joining Harvard he worked for over a decade at Mitsubishi Electric Research Laboratories where he was most recently Associate Director and Senior Research Scientist. Dr. Pfister has a Ph.D. in Computer Science from the State University of New York at Stony Brook and an M.S. in Electrical Engineering from ETH Zurich, Switzerland. He is the recipient of the 2010 IEEE Visualization Technical Achievement award. He has authored over 40 US patents and over 70 peer-reviewed publications and book chapters, including 18 ACM SIGGRAPH papers, the premier forum in Computer Graphics. He is co-editor of the first textbook on Point-Based Computer Graphics, published by Elsevier in 2007.

Helwig Hauser is professor at the Informatics Department of the University of Bergen, Norway, where he is leading a research group on visualization since 2007. His research interests are diverse in visualization, including interactive visual analysis, flow visualization, illustrative visualization, and the application of visualization to the fields of medicine, geosciences, biology, fluid dynamics, etc. Before moving to Norway and since 2003, Helwig Hauser was the scientific director of the VRVis Research Center in Vienna, Austia. Earlier, he was assistant professor at the Vienna University of Technology, Austria, from which he also received his graduate and doctoral degrees (in 1994 and 1998) as well as his habilitation (2003), before he then, in 2000, became key researcher in visualization at the then newly founded VRVis. Helwig Hauser received several awards, including the biannual Heinz-Zemanek Award in computer science from OCG in 2006 (for his habilitation work on generalizing focus+context visualization) and the Dirk Bartz Prize for visual computing in medicine in 2013 (for pioneering high-quality visualization of sonographic medical data). Helwig Hauser is an active member of the international visualization community, for example, providing his editorial services as paper chair of central conferences (recently IEEE InfoVis 2014 and 2013, PacificVis 2012, EuroVis 2011, etc.) as well as associate editor of major journals (recently IEEE TVCG, Computer Graphics Forum, etc.). He is also member of several committees, including the EuroVis Steering committee.