SOFTWARE TOOL ARTICLE

shinySISPA: A web tool for defining sample groups using gene sets from multiple-omics data [version 1; referees: 1 approved with reservations]

Bhakti Dwivedi [1], Jeanne Kowalski[1,2]

[1]Winship Cancer Institute , Emory University, Atlanta, GA, 30322, USA
[2]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, 30322, USA

v1 **First published:** 22 Feb 2018, **7**:213 (doi: 10.12688/f1000research.13934.1)
Latest published: 15 Jun 2018, **7**:213 (doi: 10.12688/f1000research.13934.2)

Abstract

As opposed to genome-wide testing of several hundreds of thousands of genes on very few samples, gene panels target as few as tens of genes and enable the simultaneous testing of many samples. For example, some cancer gene panels test for 50 genes that can affect tumor growth and potentially identify treatment options directed against the genetic mutation. The increasing popularity of gene panel testing has spurred the technological development of panels that test for diverse data types such as expression and mutation. Once samples are tested, there is the desire to examine clinical associations based on the panel and for this purpose, one would like to identify, among the samples tested, which show support for a molecular profile (e.g., presence of mutation with increased expression) versus those samples that do not among the genes tested. With user-specified molecular profile of interest, and gene panel data matrices (e.g., gene expression, variants, etc.) that define the profile, shinySISPA (Sample Integrated Set Profile Analysis) is a web-based shiny tool to define two sample groups with and without profile support based on our previously published method from which clinical associations may be readily examined. The shinySISPA can be accessed from http://shinygispa.winship.emory.edu/shinySISPA/.

Keywords

Sample profile, Gene set profile, Genomics, Integrated analysis

Open Peer Review

Referee Status: ✔

	Invited Referees 1
REVISED **version 2** published 15 Jun 2018	✔ report
version 1 published 22 Feb 2018	? report

1 **Yun Zhang**, University of Rochester, USA

Discuss this article

Comments (0)

Corresponding author: Jeanne Kowalski (jeanne.kowalski@emory.edu)

Author roles: Dwivedi B: Data Curation, Formal Analysis, Investigation, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kowalski J**: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work is funded by the Georgia Research Alliance Scientist Award (Jeanne Kowalski); Biostatistics and Bioinformatics Shared Resource of Winship Cancer Institute of Emory University and NIH/NCI [Award number P30CA138292, in part]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Dwivedi B and Kowalski J. **shinySISPA: A web tool for defining sample groups using gene sets from multiple-omics data [version 1; referees: 1 approved with reservations]** *F1000Research* 2018, **7**:213 (doi: 10.12688/f1000research.13934.1)

First published: 22 Feb 2018, **7**:213 (doi: 10.12688/f1000research.13934.1)

Introduction

Unlike gene set profiling, sample profiling is a challenge due to the heterogeneity between, and within the tumor patient samples. Identification of homogenous groups of samples or molecular subtypes is commonly approached using clustering methods (e.g., Handl *et al.*, 2005; Kowalski *et al.*, 2016; Monti *et al.*, 2003; Şenbabaoğlu *et al.*, 2014; Verhaak *et al.*, 2010). Whether or not the sample groups are meaningful and the clustering stable, requires additional testing, is highly subjective, limited to examining changes in a single data type, and often require removal of genes or samples to obtain the desired results. While clustering tools such as TNBC subtyping (Chen *et al.*, 2012; Lehmann *et al.*, 2011) are convenient for subtype discovery and sample classification, they are restricted to studying a specific cancer tissue and data type, within the context of established expression signature profiles. Although these methods may prove useful in certain case, there is a need for a basic tool that can identify sample groups using any combination of genomic data types based on a gene or gene set and molecular profile of interest. Some examples of gene sets may be derived from a specific biological process, network, gene enrichment analysis, a gene panel, etc. A molecular profile is a series of increasing or decreasing changes among diverse data types operating on a given gene set. For example, a gene mutation with expression is a molecular profile of increased variant support with increased levels of expression. The shinySISPA is a web tool developed to implement the novel method, SISPA (Kowalski *et al.*, 2016), for defining samples with similar molecular profiles based on a user input gene set and data types. SISPA does not impose analytical distribution assumptions on the data, and is scalable to define samples that support a general profile defined by any combination of genomic data types applied to any number of genes.

Methods

Implementation

SISPA is written in the R programming language (R project) and the shiny web application framework is implemented using the Shiny R package (Chang *et al.*, 2016).

The tool is hosted on a 64bit CentOS 6 server (http://shinygispa.winship.emory.edu/shinySISPA/) running the Shiny Server program designed to host R Shiny applications. This tool has been extensively tested on Windows 7 and Mac Pro 10 operating system with firefox and google chrome browser. Given a dataset of 377 samples and 16 genes under the two-feature analysis, it took three seconds to obtain shinySISPA defined sample groups and less than a second to generate the waterfall plot. The time it takes to generate sample profile diagnostic plots depends on the number of genes in a set; it took less than 10 seconds for 16 genes in both sets of a two-feature analysis. As a note, speed at which results are generated is also dependent on the internet connectivity.

Operation

The tool workflow consists of four basic inputs as shown in Figure 1:

(1) *Selecting the analysis type*. User selects a single-, two-, or three-feature analysis, where a feature corresponds to a specific data type (e.g., expression, methylation, mutation, copy number variation) and thus, a single-feature analysis refers to use of a single data type, while a two-feature uses a combination of two data types and so forth.

(2) *Uploading the data.* User inputs the data for each feature containing the genes and samples of interest. The same samples are required for each feature, though the gene sets may differ between features.

(3) *Specifying a molecular profile.* A molecular profile is a series of increasing ("up") or decreasing ("down") genomic changes within each feature. In Figure 1, a profile of decreased expression with decreased copy number is input.

(4) *Selecting the number of breaks to define sample groups*. User can specify the change point detection method (Killick *et al.*, 2016) for finding optimal break points in the distribution of computed composite (among features) z-scores within samples (Kowalski *et al.*, 2016; see Supplementary File 1).

The results are output in four separate tabs:

(1) *Input Data.* Summarizes the user input data in terms of the input number of genes, number of samples, and box plot distribution by data type.

(2) *SISPA Results.* Outputs the table of defined sample groups with their gene set enrichment score for the selected analysis type and molecular profile of interest. The scatter plot on the right displays all the change points detected within the data-set, samples falling in the topmost change point are the samples with the profile activity. The frequency plot at the rightmost bottom represents the distribution of the number of samples with and without the profile activity.

(3) *Waterfall Plot.* To visualize the sample groups that correlate with the profile of interest. Samples with the profile activity have the highest score and are shown in orange filled bars, while samples without the profile activity are shown with grey-filled bars.

(4) *Sample Profile.* Represents the diagnostic plots to visualize the distribution of the user-input data overall by the identified sample groups. It also allows the users to view data distribution for a selected gene in the set within each data type to assess what genes in particular satisfy the profile versus samples without profile.

All results generated during the process are directly downloadable on the user's local computer. A detailed manual with tool settings are provided in the Supplementary File 1. Upon forming such sample groups, one may readily examine the effect of a profile on various clinical and biological clinical outcomes.

Use case

We applied shinySISPA to profile newly diagnosed multiple myeloma (MM) patients for decreased gene expression and copy number based on a GISPA (Gene Integrated Set Profile Analysis)-derived gene set characterizing the IgH translocation in

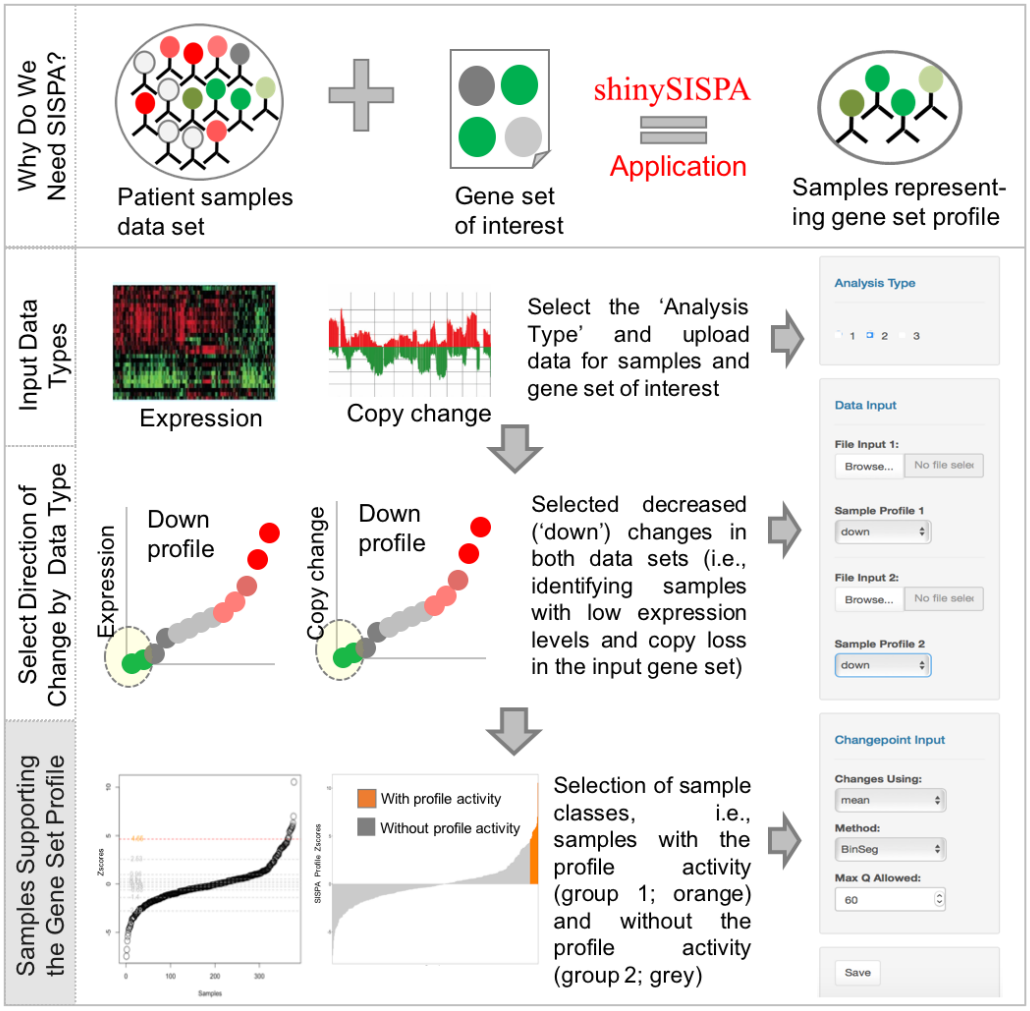

Figure 1. A schematic representation of shinySISPA workflow for a two-feature analysis. Here, we define samples supporting the molecular profile of decreased gene expression and copy loss. The tool requires user selection of analysis type, user upload of data types on samples and gene sets, and specification of a profile to output the samples supporting that profile (shown in grey). The samples are selected based on a change point model applied to composite (among features and genes), within-sample z-scores. A waterfall plot of profile activity is output with samples selected in orange as showing the most support for the profile.

the MM cell lines (Kowalski *et al.*, 2016). The t(14;16) translocation is known to be associated with poor prognosis in MM. By applying the shinySISPA tool to the t(14;16) characterized gene set profile, we were able to translate cell line profiles to patient profiles. Using the IA6 release of Multiple Myeloma Research Foundation (MMRF) CoMMpass study, we downloaded data from 377 newly diagnosed patients at pre-treatment with available clinical outcomes, RNA-Seq expression, and DNA-copy number variations from the MMRF Research Gateway portal. Based on our two-feature analysis, 7 of the 300 MM patients were defined with profile activity (Figure 1) by identifying changes in variance using change point v2.2.2. Furthermore, we used CASAS (Rupji *et al.*, 2017) to compare survival curves of the identified two sample groups for downstream clinical interpretation. We found seven samples with profile activity to be significantly (P<0.0001) associated with poor survival as compared to the 300 samples without the profile activity (HR = 9.81; 95% CI = (3.39, 28.37)).

Conclusion

We have demonstrated the utility of our shinySISPA tool in translating cell line characterized gene sets molecular profile to patient profiling (Kowalski *et al.*, 2016); however, one can use any *a priori*-defined gene sets with any combination of molecular data for identifying samples with a similar gene set profile. The introduction of a change point model to select samples with

profile support offers a more objective approach than with clustering methods. With only a gene set and a combination of data types from the same samples, the tool is widely applicable to many settings. For example, shinySISPA may be used to define patients based on known drug targets and pathways, or to identify patients that may be at risk for poor prognosis based on known prognostic markers.

Data and software availability

The example sample data used to demonstrate shinySISPA workflow is available on the web-version and with the package source code at: https://github.com/BhaktiDwivedi/shinySISPA.

The shinySISPA software is available at http://shinygispa.winship.emory.edu/shinySISPA/

The stand-alone version of SISPA is available at https://www.bioconductor.org/packages/SISPA/.

Archived source code as at the time of publication: https://doi.org/10.5281/zenodo.1164284 (Dwivedi & Kowalski, 2018)

License: shinySISPA is available under the GNU public license (GPL-3)

Competing interests

No competing interests were disclosed.

Grant information

This work is funded by the Georgia Research Alliance Scientist Award (Jeanne Kowalski); Biostatistics and Bioinformatics Shared Resource of Winship Cancer Institute of Emory University and NIH/NCI [Award number P30CA138292, in part]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We would like to thank the Cancer Informatics Core of the Winship Cancer Institute of Emory University for supporting the CentOS server. Special thanks to Kenneth Buck for his assistance with building and configuring the server for hosting shinySISPA. Special thanks to Manali Rupji for her assistance on clinical survival analysis.

Supplementary material

Supplementary File 1: The shinySISPA manual

Click here to access the data.

References

Chang W, *et al.*: **shiny: Web Application Framework for R.** *R package version 0.13.2.* 2016.

Chen X, Li J, Gray WH, *et al.*: **TNBCtype: A Subtyping Tool for Triple-Negative Breast Cancer.** *Cancer Inform.* 2012; **11**: 147–156.
PubMed Abstract | Publisher Full Text | Free Full Text

Dwivedi B, Kowalski J: **shinySISPA: A web tool for defining sample groups using gene sets from multiple omics data (Version 1.0).** *Zenodo.* 2018.
Data Source

Handl J, Knowles J, Kell DB: **Computational cluster validation in post-genomic data analysis.** *Bioinformatics.* 2005; **21**(15): 3201–3212.
PubMed Abstract | Publisher Full Text

Killick R, Haynes K, IA E: **changepoint: An R package for changepoint analysis.** *R package version 2.2.1.* 2016.

Kowalski J, Dwivedi B, Newman S, *et al.*: **Gene integrated set profile analysis: a context-based approach for inferring biological endpoints.** *Nucleic Acids Res.* 2016; **44**(7): e69.
PubMed Abstract | Publisher Full Text | Free Full Text

Lehmann BD, Bauer JA, Chen X, *et al.*: **Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies.** *J Clin Invest.* 2011; **121**(7): 2750–2767.
PubMed Abstract | Publisher Full Text | Free Full Text

Monti S, Tamayo P, Mesirov JP, *et al.*: **Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data.** *Mach Learn.* 2003; **52**(1–2): 91–118.
Publisher Full Text

Rupji M, Zhang X, Kowalski J: **CASAS: Cancer Survival Analysis Suite, a web based application [version 2; referees: 2 approved].** *F1000Res.* 2017; **6**: 919.
PubMed Abstract | Publisher Full Text | Free Full Text

Şenbabaoğlu Y, Michailidis G, Li JZ: **Critical limitations of consensus clustering in class discovery.** *Sci Rep.* 2014; **4**: 6207.
PubMed Abstract | Publisher Full Text | Free Full Text

Verhaak RG, Hoadley KA, Purdom E, *et al.*: **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*.** *Cancer Cell.* 2010; **17**(1): 98–110.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ?

Version 1

Referee Report 16 April 2018

doi:10.5256/f1000research.15147.r32660

? **Yun Zhang**
Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA

The authors introduce a Shiny application that are developed as a user-friendly tool for conducting omics-data analysis with their published methodology work. The article provides detailed description of the inputs and outputs for the shiny application, and shows working example for the use of this application. This article is index-able with some necessary modifications.
The following are my major comments.

1. Before introducing the shiny application, the authors should provide a brief summary of their methodology work, including clear definitions of important terminologies and a description of their model. It would greatly help the readers to understand the contents followed subsequently.
    1. With a glimpse on the methodology paper, I found the terms such as “molecular profile”, “feature” are defined misleadingly in this article.
    2. What is a change point model?
2. For the input element (4), where to select the number of breaks? Is it the max Q allowed? What is the max Q? What are the options for “Changes Using” in the bottom right of Figure 1?
3. For the output element (2), it is better to include a figure with the description.
4. In the “Use case” section,
    1. Why GISPA is used? What’s the relation of GISPA and SISPA?
    2. How many genes are used? Is it 16?
    3. Where are the 7 patients in Figure 1? Are they the orange bars?
    4. What is “using change point v2.2.2”?
    5. “We found seven samples with profile activity to be significantly (P<0.0001) associated with poor survival as compared to the 300 samples without the profile activity (HR = 9.81; 95% CI = (3.39, 28.37)).” Previously mentioned, there are 377 patients. Why are 70 samples missing?
5. For the reproducibility of the work, please upload a default dataset in the shiny application, so that the readers/users can replicated the analysis for the first time.

Also, I provide my minor comments below.

1. Please consider a thorough revision of the English writing for this scientific article. There are a number of grammatical errors, and sentences do not flow smoothly. Please avoid using colloquial words in scientific writing.
2. Please modify the legend of Figure 1. In the brackets “(shown in grey)”, it is hard to locate where it is referring to since there are many grey colors in the figure. Also, please add labels to the subfigures that are referred in the text.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 01 Jun 2018

Bhakti Dwivedi, Emory University, USA

Referee Report 16 Apr 2018
Yun Zhang, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA
Approved with Reservations
The authors introduce a Shiny application that are developed as a user-friendly tool for conducting omics-data analysis with their published methodology work. **The article provides detailed description of the inputs and outputs for the shiny application**, and shows working example for the use of this application. This article is index-able with some necessary modifications.
The following are my major comments.

1) Before introducing the shiny application, the authors should provide a brief summary of their methodology work, including clear definitions of important terminologies and a description of their model. It would greatly help the readers to understand the contents followed subsequently.

Response: We have provided a reference to our methods paper that includes detailed information on the SISPA approach. In this paper, our focus is upon introducing the application of the method in terms of tool development and implementation by providing detailed examples and information on data input and output/results. Considering this focus, along with space constraints, we have opted to not repeat the already published method description and instead, reference it.

1. With a glimpse on the methodology paper, I found the terms such as “molecular profile”, “feature” are defined misleadingly in this article.

Response: We have intentionally used the terms “molecular profile” and “feature” in the same context as in the published methods papers, whereby “molecular profile” refers to change of either increase (“up”) or decrease (“down”) and “feature” refers to a specific data type (e.g., expression, methylation, copy number change). We have also provided examples of the term “molecular profile” to further clarify the context.

In this paper:

“A “feature” corresponds to a specific data type (e.g., expression, methylation, mutation, copy number variation) and thus, a single-feature analysis refers to use of a single data type, while a two-feature uses a combination of two data types and so forth.” (pg# 3 last paragraph, under *Selecting the analysis type*)

“A “molecular profile” is a series of increasing (“up”) or decreasing (“down”) genomic changes within each feature…” (pg# 4 second paragraph, under *Specifying a molecular profile)*

In the published methodology paper:

“The Cancer Genome Atlas (TCGA) nomenclature (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) that references a specific data type (RNA-Seq expression, DNA CpG methylation, etc.) as a feature.”

“A profile is defined by specifying *a priori*, a change of either increase or decrease within each of the features..”

2. What is a change point model?

Response: A changepoint model is a method for identifying changepoints within data. It is a published method. Please see below references to the changepoint method and the changepoint R package for details:

- Killick R and Eckley IA (2014). “changepoint: An R Package for Changepoint Analysis.” *Journal of Statistical Software*, **58**(3), pp. 1–19. http://www.jstatsoft.org/v58/i03/.
- Killick R, Haynes K and Eckley IA (2016). *changepoint: An R package for changepoint analysis*. R package version 2.2.2, https://CRAN.R-project.org/package=changepoint.

2) For the input element (4), where to select the number of breaks? Is it the max Q allowed? What is the max Q? What are the options for “Changes Using” in the bottom right of Figure 1?

Response: Yes, the allotted maximum number of breaks is specified using the “Max Q Allowed”. “In input element (4), users can modify the Changepoint Input (Killick R, *et al*., 2016) to find the optimal break points within the estimated profile sample score (Kowalski, *et al*., 2016). The changes can be found using mean(“mean”), variance (“var”) or both (“meanvar”) with the user-specified changepoint method (“AMOC”, “BinSeg”, “PELT”, or “SeqNeigh”) given the allotted maximum number of change points (“Max Q allowed”). Note that the number of change points identified may differ for the same dataset depending on the change point R package version installed on the system. Currently we are running changepoint version 2.2.2 on our hosting server...”

Please see pg# 6 of the supplementary file 1 for the details.

3) For the output element (2), it is better to include a figure with the description.

Response: The output elements (“Input Data” “SISPA results”, “Waterfall Plot” and “Sample Profile”) screenshot including the figures are explained in detail in the supplementary file 1. Please see pg# 7-11 under “Result” Section.

4) In the “Use case” section,

1. Why GISPA is used? What’s the relation of GISPA and SISPA?

Response:

GISPA (Gene Integrated Set Profile Analysis) is a method designed to define gene sets with similar, a priori specified molecular profile. While, SISPA (Sample Integrated Set Profile Analysis) is a method designed to define sample groups with similar gene set a priori specified molecular profile. Both GISPA and SISPA method are published in Nucleic Acid Research (Kowalski et al., 2016; PMID: 26826710).

GISPA was used to identify genes with decreased expression and decreased copy change molecular profile in a multiple myeloma cell line with IgH translocation. This gene set is published in the methodology paper Nucleic Acid Research (Kowalski et al., 2016; PMID: 26826710).

Here, we extracted RNA-seq expression, and copy number change data for GISPA derived gene set characterizing the IgH translocation on 377 newly diagnosed patients enrolled in the coMMpass clinical trial to define samples with a similar gene set profile, i.e., decreased expression with copy loss. This example data is provided with this paper. Pg# 4 last paragraph under “Use case” describes the use of application of SISPA using GISPA derived gene sets.

2. How many genes are used? Is it 16?

Response: Yes. The number of genes in the expression and copy number variation data is 16.

3. Where are the 7 patients in Figure 1? Are they the orange bars?

Response: The patients with and without profile activity are highlighted in “Samples Supporting the Gene Set Profile” labeled section of the Figure 1. Yes, the 7 patients are highlighted in orange-filled bars.

4. What is “using change point v2.2.2”?

Response: “using change point v2.2.2”? means that we have used changepoint R package version 2.2.2.

5. “We found seven samples with profile activity to be significantly (P<0.0001) associated with poor survival as compared to the 300 samples without the profile activity (HR = 9.81; 95% CI = (3.39, 28.37)).” Previously mentioned, there are 377 patients. Why are 70 samples missing?

Response: We have corrected the typo, please see page# 4 and 5, last paragraph. It is 7 of 370 patients.

“Based on our two-feature analysis, 7 of the 370 MM patients were defined with profile activity (Figure 1) by identifying changes in variance using change point v2.2.2. Furthermore, we used CASAS (Rupji *et al.*, 2017) to compare survival curves of the identified two sample groups for downstream clinical interpretation. We found seven samples with profile activity to be significantly (P<0.0001) associated with poor survival as compared to the 370 samples without the profile activity (HR = 9.81; 95% CI = (3.39, 28.37)).”

5) For the reproducibility of the work, please upload a default dataset in the shiny application, so that the readers/users can replicated the analysis for the first time.

Response: All users can access and analyze the example dataset (i.e., default dataset) used in the paper by choosing the “Example data” from the Upload Input option on the web-interface. The data is also available to download from GitHub (https://github.com/BhaktiDwivedi/shinySISPA).

Users are able to obtain the same exact results, i.e., samples with and without profile activity using the current default settings implemented in shinySISPA web tool.

Also, I provide my minor comments below.

1. Please consider a thorough revision of the English writing for this scientific article. There are a number of grammatical errors, and sentences do not flow smoothly. Please avoid using colloquial words in scientific writing.
Response: We have reviewed the manuscript and do not identify any such problem. If the reviewer still feels strongly about it, we kindly request specific examples to be cited from the text.

2. Please modify the legend of Figure 1. In the brackets "(shown in grey)", it is hard to locate where it is referring to since there are many grey colors in the figure. Also, please add labels to the subfigures that are referred in the text.
Response: We have addressed and incorporated these changes. Please see updated Figure 1 and Figure legend.

- Is the rationale for developing the new software tool clearly explained?

Yes

- Is the description of the software tool technically sound?

Partly

- Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

No

- Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

- Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.
I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.