Check for updates

SOFTWARE TOOL ARTICLE

REVISED GNOSIS: an R Shiny app supporting cancer genomics survival analysis with cBioPortal [version 2; peer review: 2 approved]

Lydia King[1,2], Andrew Flaus[3], Simone Coughlan[1,2], Emma Holian[2], Aaron Golden[2]

[1]SFI Centre for Genomics Data Science, National University of Ireland, Galway, H91 TK33, Ireland
[2]School of Mathematical & Statistical Sciences, National University of Ireland, Galway, H91 TK33, Ireland
[3]Centre for Chromosome Biology, School of Natural Sciences, National University of Ireland, Galway, H91 TK33, Ireland

v2 **First published:** 18 Jan 2022, **5**:8
https://doi.org/10.12688/hrbopenres.13476.1
Latest published: 12 Sep 2022, **5**:8
https://doi.org/10.12688/hrbopenres.13476.2

Abstract

Exploratory analysis of cancer consortia data curated by the cBioPortal repository typically requires advanced programming skills and expertise to identify novel genomic prognostic markers that have the potential for both diagnostic and therapeutic exploitation. We developed GNOSIS (GeNomics explOrer using StatistIcal and Survival analysis in R), an R Shiny App incorporating a range of R packages enabling users to efficiently explore and visualise such clinical and genomic data. GNOSIS provides an intuitive graphical user interface and multiple tab panels supporting a range of functionalities, including data upload and initial exploration, data recoding and subsetting, data visualisations, statistical analysis, mutation analysis and, in particular, survival analysis to identify prognostic markers. GNOSIS also facilitates reproducible research by providing downloadable input logs and R scripts from each session, and so offers an excellent means of supporting clinician-researchers in developing their statistical computing skills.

Keywords

Cancer Genomics, cBioPortal, Precision Oncology, Statistical Analysis, Survival Analysis, Data Exploration, R, RShiny app

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 12 Sep 2022		
version 1 18 Jan 2022	✓ view	✓ view

1. **Jessica C. Mar**, The University of Queensland, Brisbane, Australia
2. **Tiago C. Silva**, University of Miami, Miami, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Lydia King (lydia.king@nuigalway.ie), Aaron Golden (aaron.golden@nuigalway.ie)

Author roles: King L: Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Flaus A**: Supervision, Writing – Review & Editing; **Coughlan S**: Supervision; **Holian E**: Methodology, Supervision, Writing – Review & Editing; **Golden A**: Conceptualization, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Science Foundation Ireland and National Breast Cancer Research Institute under Grant number [18/CRT/6214].
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: King L, Flaus A, Coughlan S *et al.* **GNOSIS: an R Shiny app supporting cancer genomics survival analysis with cBioPortal [version 2; peer review: 2 approved]** HRB Open Research 2022, **5**:8 https://doi.org/10.12688/hrbopenres.13476.2

First published: 18 Jan 2022, **5**:8 https://doi.org/10.12688/hrbopenres.13476.1

REVISED **Amendments from Version 1**

Introduction: In response to reviewer 1, the following sentence has been added to paragraph 2 of the Introduction: "While there are a number of tools readily available to carry out exploratory analysis, survival analysis, statistical analysis, copy number alteration (CNA) calling, annotation and visualisation and exploration of the CNA landscape with respect to survival, these tools are limited, self-contained and for CNA analysis, often require users to have access to the raw or segmented data.". This sentence points users to alternative tools to GNOSIS that may be useful, but also highlights the utility of GNOSIS.

Statistical and Survival Analysis: In response to reviewer 2, the following sentence has been added to paragraph 3 of the Statistical and Survival Analysis section: "Users also have the option to carry out a basic descriptive analysis by groups using the compareGroups package". Following reviewer comments, the option to use the compareGroups function in GNOSIS was added and this sentence alerts readers/users to this.

Statistical and Survival Analysis: In response to reviewer 1, the following sentence has been added to paragraph 3 of the Statistical and Survival Analysis section: "To aid users in this, information buttons containing links to useful resources are available throughout the app". This sentence highlights that links to useful resources are available throughout GNOSIS to help readers/users run statistically sound analyses.

Operation: In response to reviewer 2, the compareGroups function was added to GNOSIS. To document this change "compareGroups" was added to the R package list in paragraph 2 of the Operation section.

Throughout Manuscript: In response to reviewer 2, we have changed sentences containing ".png" and "PNG" to ".png and .svg" and "PNG and SVG" throughout the manuscript. These changes were necessary as following reviewer comments GNOSIS was updated to allow users to download plots in both formats.

Any further responses from the reviewers can be found at the end of the article

Introduction

Cancer diagnosis, classification and treatment generally follows an integrative approach combining clinical features and tissue-based biomarkers[1,2]. In recent years, there has been an increased interest in using genetic testing to guide treatment decisions, predict patient response and determine likely prognoses for cancers associated with specific pathogenic variants[3]. Such a precision oncology paradigm has been fostered by the extensive efforts of many cancer genomics consortia, yielding extraordinarily rich repositories of genomic and associated clinical data of hundreds to, in some cases, thousands of cancer patients[4,5].

Summary clinical and cancer genomic data are available from a number of consortia websites, with cBio Cancer Genomics Portal (cBioPortal)[6,7] offering one of the best known and regularly accessed consolidated curations for multiple consortia; cBioPortal provides both graphical user interface (GUI)-based and representational state transfer (RESTful) mediated means for researchers to explore clinical and genomics data. However, cBioPortal's exploratory capabilities have their limitations, requiring the implementation of a more sophisticated 'off site' analysis that typically requires significant prior programming experience. This remains arguably the greatest barrier for many clinician-researchers wishing to explore hypotheses in precision oncology. While there are a number of tools readily available to carry out exploratory analysis[8,9], survival analysis[10–12], statistical analysis[12], copy number alteration (CNA) calling, annotation and visualisation[13–15] and exploration of the CNA landscape with respect to survival[15], these tools are limited, self-contained and for CNA analysis, often require users to have access to the raw or segmented data.

An ideal solution to these limitations would be the availability of a software environment supporting the integration of cBioPortal-hosted data products, their visualisation and tractable manipulation using standard biostatistical methodologies. Such an environment would provide a convenient means of testing exploratory hypotheses, particularly those assessed in the context of survival analysis, in a way that would be both reproducible and interpretable. Based on our experience as part of a recent study[16] to investigate whether survival outcomes are associated with genomic instability in luminal breast cancers, we developed a software infrastructure using R to facilitate such exploratory work. Working with the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)[17] summary clinical and CNA data obtained from cBioPortal, we tailored this codebase to support GUI interactivity, and deployed it as an R Shiny app called GNOSIS (GeNomics explOrer using StatistIcal and Survival analysis in R).

GNOSIS leverages a number of R packages. The GUI front end employs 'tabs' for data upload, initial exploration, data subsetting and recoding, a range of visualisations, comprehensive survival analysis, association testing and mutation analysis. Furthermore, GNOSIS has a user-driven point-and-click interface that logs all user activity to facilitate reproducibility, and ultimately enables the statistical analysis and incorporation of multiple and diverse genomic features with patient data in a research or clinical setting.

GNOSIS provides a tractable means for clinician-researchers with a background in biostatistics to effectively engage with complex cancer genomics data, to experiment with exploratory hypotheses in a more intuitive way that would require greater expertise working at the command line, and to have a record of all activities from which subsequent, more focused and nuanced analyses can be based. GNOSIS also offers great potential in supporting the teaching of biostatistical methodologies relevant to clinical genomics applications. Given its open source basis and foundation in the R statistical programming environment, GNOSIS also offers a means for third parties to enhance and develop its functionality for broader clinical genomics.

Methods

Implementation

Overview. GNOSIS was initially developed to enable the exploration, visualisation and analysis of the METABRIC clinical and CNA summary data obtained from cBioPortal, as detailed in King *et al.* (2021)[16], and the following description of its operational capabilities have their basis in that study. Although GNOSIS accepts multiple file types, including comma-, semicolon- or tab-delimited, the default settings are suited to files downloaded from cBioPortal. If users wish to upload clinical or summary genomics data files from other sources, care should be taken to set appropriate default values. GNOSIS leverages a number of R packages, primarily shiny, tidyverse, ggplot2, survival, survminer, rpart, partykit and maftools[18–26]. A full list is provided in the Operation subsection. It allows users to carry out a comprehensive visual exploration and statistically robust survival analysis in a fast, simple and reproducible way, and we illustrate GNOSIS functionalities by referring to example screenshots of its operation at various relevant steps of the analysis documented in King *et al.* (2021)[16].

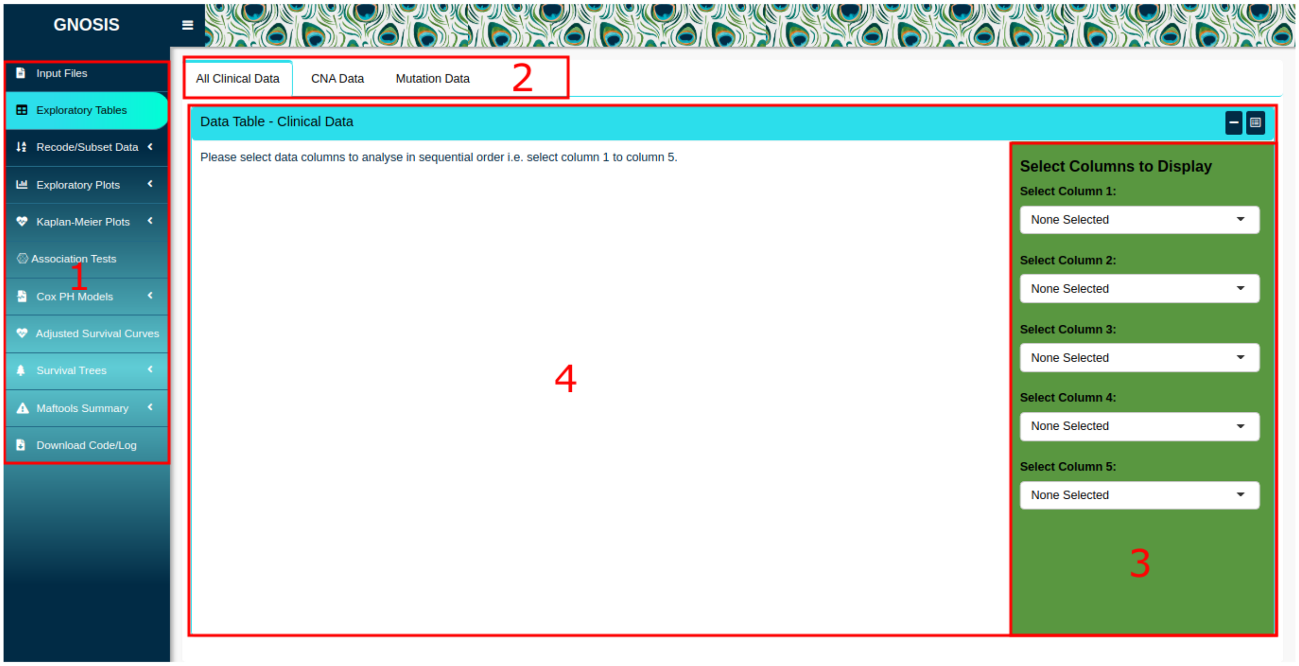
Data upload and formatting. In Figure 1 we show the GNOSIS front-end with the specific entry points and 'tabs' highlighted. GNOSIS accesses files locally on the user's file system, and in its default configuration, is optimised to use data files downloaded from cBioPortal. In the Input Files tab, users are provided with a space to upload the clinical patient and sample data, summary CNA data and mutation data. Whilst we have configured GNOSIS to work exclusively with both CNA and mutational data files, modification of the codebase allows users to reconfigure the GNOSIS GUI to import other genomic tracks from cBioPortal, as required. A preview of the data is provided in the GNOSIS viewing panel to ensure that the data has been read in correctly. It should be noted that the clinical patient and sample data should contain a column named "PATIENT_ID" and the CNA data should contain a column called "Hugo_Symbol". As these are core named data types for all subsequent analytics, warnings will be produced and downstream analysis will not be possible if they are missing.

Once the data is uploaded, further exploration of specific columns can be done using the Exploratory Tables tab, where up to five columns can be selected in the box sidebar and viewed. The columns should be selected in sequential order; if this is not adhered to an error will be displayed.

Before more extensive data exploration and analysis, users are encouraged to carry out pre-processing to ensure data is in the desired format using the Recode/Subset Data tab. Users are provided with a workspace to view the variables present in the data, their type and factor levels. Users can change variables to numeric or factors using the box sidebar, which contains a space to select relevant variables.

Subsequently, users can subset the data based on up to three categorical variables and carry out survival variable recoding. In cases where CNA data is uploaded, users may produce and segment CNA metrics for each patient, as well as select and extract specific genes for further analysis. After each operation, the space to explore variable information is updated. This allows users to confirm their alterations have been implemented correctly. These operations ensure that the data is in the correct format for downstream analysis.

To allow users to save their formatted file, a space within the tab is provided. If this exported data is uploaded to GNOSIS,



Please select data columns to analyse in sequential order i.e. select column 1 to column 5.

Select Column 1: None Selected

Select Column 2: None Selected

Select Column 3: None Selected

Select Column 4: None Selected

Select Column 5: None Selected

Figure 1. GNOSIS GUI with highlighted interface elements. (1) The Exploratory Tables tab is selected in the tab sidebar. (2) Within tab panels allowing multiple operations to be carried out and viewed in the one tab. (3) Box sidebar allowing users to select inputs, alter arguments and customise and export visualisations. (4) Viewing panel displaying output.

formatting of categorical data may have to be carried out again due to the default `stringsAsFactors` argument implemented when uploading data in R. In Figure 2 (A), we show how a given data file can be examined and filters applied to extract a subset, and in Figure 2 (B), the resulting subset following calculation of CNA metrics, including absolute CNA score, amplification score and deletion score for each patient, and subsequent quartile segmentation of the CNA scores is shown.

Data visualisation. Within the Exploratory Plots tab, GNOSIS provides users with a range of visualisations including boxplots, scatterplots, barplots, histograms and density plots. For example, in Figure 2 (C), a set of patient genome absolute CNA scores can be segmented - and labelled as such - into multiple equally-sized groups. These visualisations are implemented using the R package ggplot2[20]. This way, clinical and genomic data can be interrogated and visualised separately or in combination. For each visualisation, users can use the box sidebar to select which columns to interrogate, choose whether to include NA values in the plots, choose whether to display a legend and change the plot title, x- and y-axis titles and legend titles, among others. Further options available to users include the ability to produce boxplots where the sample size is reflected in the width of the boxplot, produce scatterplots coloured by an additional variable, and to produce plain, segmented and faceted histograms and density plots. Users can also download all the resulting plots as .pngs or .svgs in specified dimensions.

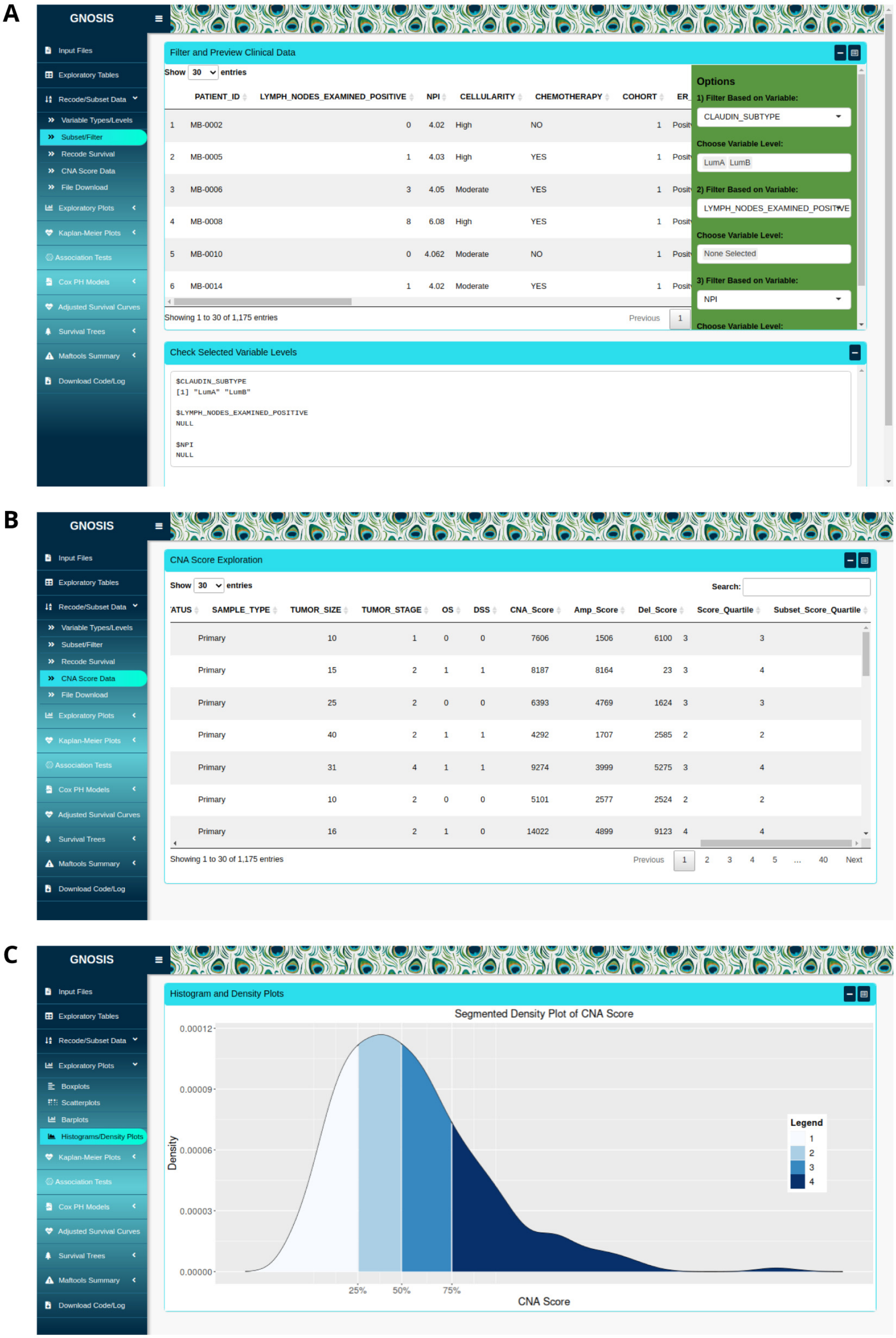
Statistical and survival analysis. The primary function offered by GNOSIS is statistically robust survival analysis. GNOSIS contains several step-wise tabs to provide a complete survival analysis of the data under investigation.

Initially the Kaplan-Meier (KM) Plots tab provides survival plots and the corresponding logrank tests to identify survival-associated categorical variables, both visually and statistically (Figure 3 (A)). The Kaplan-Meier Plots tab contains three sub-tabs which provide users with spaces to produce KM plots and logrank tests for selected clinical variables, for segmented CNA variables and for variables of interest split on treatment assignment (i.e. where patients received different treatments, e.g. split into patients who received radiotherapy and patients who did not). Within each sub-tab, the interface allows users to indicate which columns contain the survival time, event status (Overall Survival (OS) or Disease Specific Survival (DSS)) and the variable of interest. It should be noted that when producing KM curves for variables split by treatment assignment the selected treatment variable must be coded as a binary YES/NO. These KM curves can be customised and exported as .pngs or .svgs using the sidebar options.

The Association Tests tab uses association tests to identify variables that are linked to each other and enables users to identify potential confounding variables in the analysis. Variable selection is done within the box sidebar and statistical association tests available include the χ^2 test, Fisher's exact test, simulated Fisher's exact test, ANOVA, Kruskal-Wallis test, pairwise t-test and Dunn's test. The χ^2 test is used to assess the association between two categorical variables with sufficient cell sizes in the two-way table of categorical variables (Figure 3 (B)). Fisher's exact test can be used in the case where any cell size is sufficiently small. ANOVA can be used to test whether there is a difference in means between groups and the Kruskal-Wallis test may be used in the situation where the assumptions of the ANOVA test are violated. Pairwise comparisons using t-tests and Dunn's test are also available. In all cases, results of each individual association test are displayed alongside the adjusted p-values calculated using the Benjamini-Hochberg p-value adjustment. Users also have the option to carry out a basic descriptive analysis by groups using the compareGroups package[27]. It is important that users make sure they run the appropriate statistical tests for the question of interest, that all relevant assumptions are met and that the output is interpreted correctly. To aid users in this, information buttons containing links to useful resources are available throughout the app.

In the Cox Proportional Hazards (PH) models tab, users are provided with a workspace to produce both univariate and multivariable Cox models to identify survival-associated variables, and test the assumptions of these models using graphical diagnostics based on the scaled Schoenfeld residuals (Figure 3 (C)). The Cox PH model is a regression model commonly used to investigate the association between the survival time of patients and predictor variables. The Cox PH model works for both continuous and categorical variables and extends survival analysis methods to simultaneously assess the effect of several risk factors on survival time. To produce the univariate and multivariable Cox models, the box sidebar enables the selection of the columns that contain the survival time, event status and the variables to be included in the models. The output of each univariate Cox model is displayed along with a summary table containing the adjusted p-values calculated using the Benjamini-Hochberg p-value adjustment. The validity of the PH assumption of each multivariable model fitted can be assessed by producing visualisations based on the scaled Schoenfeld residuals. The Schoenfeld residuals are independent of time, and therefore a plot displaying a non-random pattern against time indicates that the PH assumption may be violated. Where a non-significant relationship between residuals and time is observed, the PH assumption is met. Again, these plots can be customised and exported in portable network graphics (PNG) format or scalable vector graphics (SVG) format.

Following multivariable Cox model selection, users are given the option to produce corresponding adjusted survival curves, which are survival curves adjusted for the covariates in the multivariable Cox model. Within the Adjusted Survival Curves tab, users are provided with a workspace to view the multivariable Cox model that was fitted in the previous tab. This will aid users when creating the new data frame needed to produce the adjusted survival curves. Users are provided with a space to set up the new data frame including the grouping variable, variable of interest and the variables to be kept constant (Figure 4 (A)). It should be noted that all variables included in the multivariable Cox model should be included in the new data frame. Plots displaying all the adjusted survival



Filter and Preview Clinical Data

Show 30 entries

	PATIENT_ID	LYMPH_NODES_EXAMINED_POSITIVE	NPI	CELLULARITY	CHEMOTHERAPY	COHORT	ER_
1	MB-0002	0	4.02	High	NO	1	Posit
2	MB-0005	1	4.03	High	YES	1	Posit
3	MB-0006	3	4.05	Moderate	YES	1	Posit
4	MB-0008	8	6.08	High	YES	1	Posit
5	MB-0010	0	4.062	Moderate	NO	1	Posit
6	MB-0014	1	4.02	Moderate	YES	1	Posit

Showing 1 to 30 of 1,175 entries

Options

1) Filter Based on Variable: CLAUDIN_SUBTYPE

Choose Variable Level: LumA LumB

2) Filter Based on Variable: LYMPH_NODES_EXAMINED_POSITIVE

Choose Variable Level: None Selected

3) Filter Based on Variable: NPI

Choose Variable Level:

Check Selected Variable Levels

```
$CLAUDIN_SUBTYPE
[1] "LumA" "LumB"

$LYMPH_NODES_EXAMINED_POSITIVE
NULL

$NPI
NULL
```

CNA Score Exploration

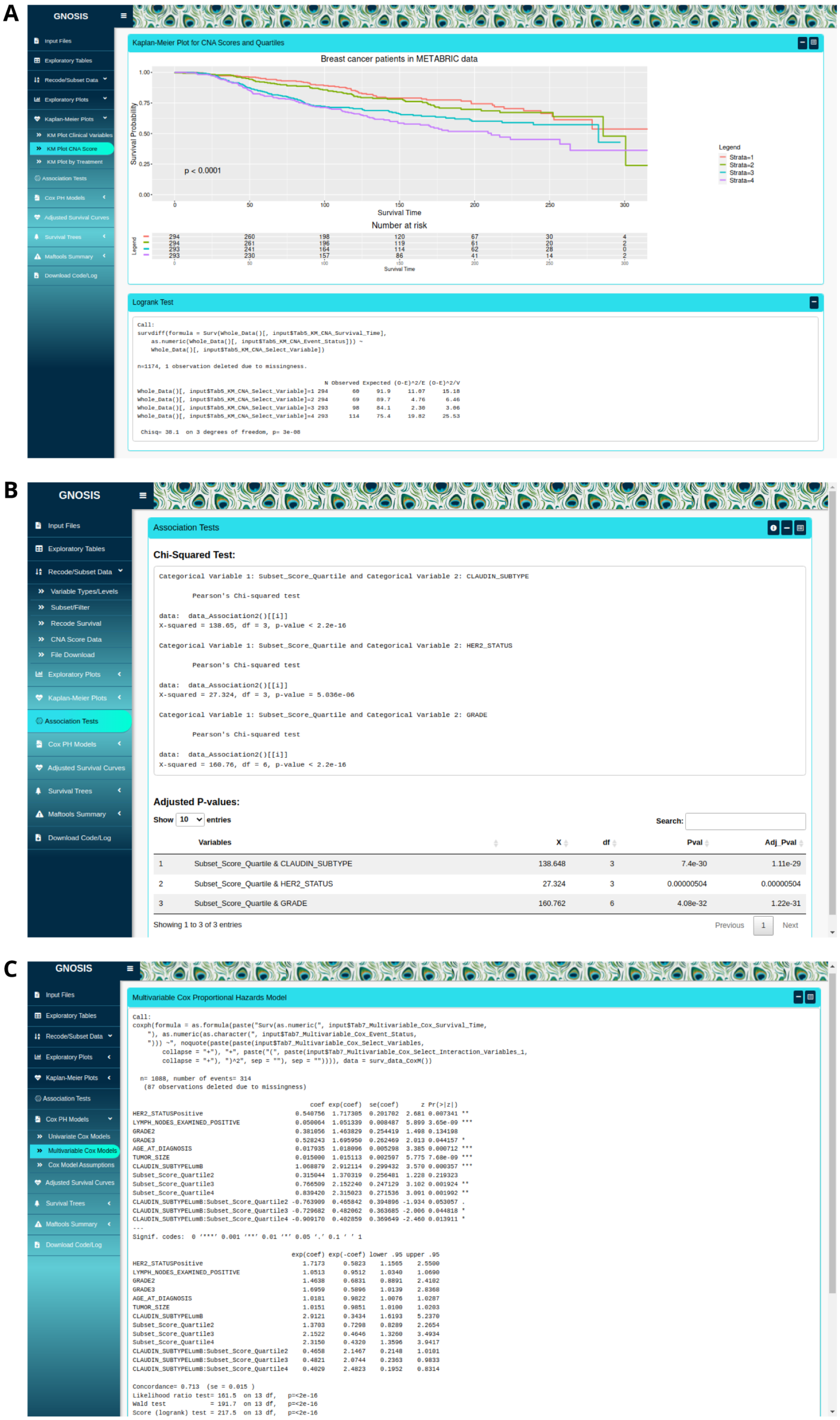
Show 30 entries

Search:

'ATUS	SAMPLE_TYPE	TUMOR_SIZE	TUMOR_STAGE	OS	DSS	CNA_Score	Amp_Score	Del_Score	Score_Quartile	Subset_Score_Quartile
	Primary	10	1	0	0	7606	1506	6100	3	3
	Primary	15	2	1	1	8187	8164	23	3	4
	Primary	25	2	0	0	6393	4769	1624	3	3
	Primary	40	2	1	1	4292	1707	2585	2	2
	Primary	31	4	1	1	9274	3999	5275	3	4
	Primary	10	2	0	0	5101	2577	2524	2	2
	Primary	16	2	1	0	14022	4899	9123	4	4

Showing 1 to 30 of 1,175 entries

Figure 2. (**A**) The Recode/Subset tab, where data is being subsetted based on subtype, luminal A and luminal B subtypes are selected. (**B**) The dataset after CNA metrics have been calculated and quartile segmented. (**C**) A density plot of the resulting quartile segmentation.



Legend	0	50	100	150	200	250	300
Strata=1	294	260	198	120	67	30	4
Strata=2	294	261	196	119	61	20	2
Strata=3	293	241	164	114	62	28	0
Strata=4	293	230	157	86	41	14	2

Logrank Test

```
Call:
survdiff(formula = Surv(Whole_Data()[, input$Tab5_KM_CNA_Survival_Time],
    as.numeric(Whole_Data()[, input$Tab5_KM_CNA_Event_Status])) ~
    Whole_Data()[, input$Tab5_KM_CNA_Select_Variable])

n=1174, 1 observation deleted due to missingness.

                                                   N Observed Expected (O-E)^2/E (O-E)^2/V
Whole_Data()[, input$Tab5_KM_CNA_Select_Variable]=1 294       60     91.9     11.07     15.18
Whole_Data()[, input$Tab5_KM_CNA_Select_Variable]=2 294       69     89.7      4.76      6.46
Whole_Data()[, input$Tab5_KM_CNA_Select_Variable]=3 293       98     84.1      2.30      3.06
Whole_Data()[, input$Tab5_KM_CNA_Select_Variable]=4 293      114     75.4     19.82     25.53

 Chisq= 38.1  on 3 degrees of freedom, p= 3e-08
```

Chi-Squared Test:

```
Categorical Variable 1: Subset_Score_Quartile and Categorical Variable 2: CLAUDIN_SUBTYPE

        Pearson's Chi-squared test

data:  data_Association2()[[i]]
X-squared = 138.65, df = 3, p-value < 2.2e-16

Categorical Variable 1: Subset_Score_Quartile and Categorical Variable 2: HER2_STATUS

        Pearson's Chi-squared test

data:  data_Association2()[[i]]
X-squared = 27.324, df = 3, p-value = 5.036e-06

Categorical Variable 1: Subset_Score_Quartile and Categorical Variable 2: GRADE

        Pearson's Chi-squared test

data:  data_Association2()[[i]]
X-squared = 160.76, df = 6, p-value < 2.2e-16
```

Adjusted P-values:

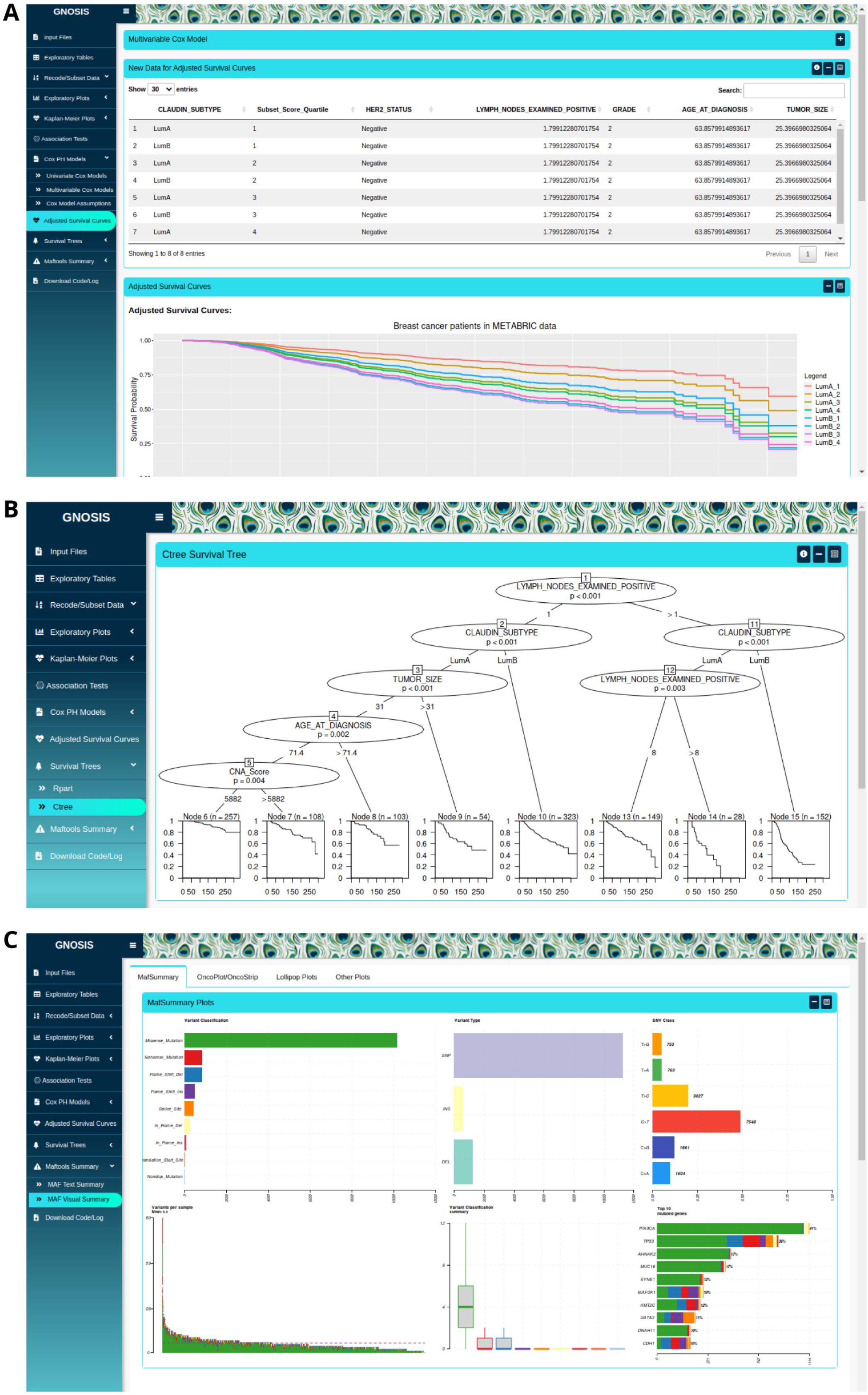
Show 10 entries Search:

	Variables	X	df	Pval	Adj_Pval
1	Subset_Score_Quartile & CLAUDIN_SUBTYPE	138.648	3	7.4e-30	1.11e-29
2	Subset_Score_Quartile & HER2_STATUS	27.324	3	0.00000504	0.00000504
3	Subset_Score_Quartile & GRADE	160.762	6	4.08e-32	1.22e-31

Showing 1 to 3 of 3 entries

```
Call:
coxph(formula = as.formula(paste("Surv(as.numeric(", input$Tab7_Multivariable_Cox_Survival_Time,
    "), as.numeric(as.character(", input$Tab7_Multivariable_Cox_Event_Status,
    "))) ~", noquote(paste(paste(input$Tab7_Multivariable_Cox_Select_Variables,
        collapse = "+"), "+", paste("(", paste(input$Tab7_Multivariable_Cox_Select_Interaction_Variables_1,
        collapse = "+"), ")^2", sep = ""), sep = "")))), data = surv_data_CoxM())

  n= 1088, number of events= 314
   (87 observations deleted due to missingness)

                                                coef exp(coef)  se(coef)      z Pr(>|z|)
HER2_STATUSPositive                         0.540756  1.717305  0.201702  2.681 0.007341 **
LYMPH_NODES_EXAMINED_POSITIVE               0.050064  1.051339  0.008487  5.899 3.65e-09 ***
GRADE2                                      0.381056  1.463829  0.254419  1.498 0.134198
GRADE3                                      0.528243  1.695950  0.262469  2.013 0.044157 *
AGE_AT_DIAGNOSIS                            0.017935  1.018096  0.005298  3.385 0.000712 ***
TUMOR_SIZE                                  0.015000  1.015113  0.002597  5.775 7.68e-09 ***
CLAUDIN_SUBTYPELumB                         1.068879  2.912114  0.299432  3.570 0.000357 ***
Subset_Score_Quartile2                      0.315044  1.370319  0.256481  1.228 0.219323
Subset_Score_Quartile3                      0.766509  2.152240  0.247129  3.102 0.001924 **
Subset_Score_Quartile4                      0.839420  2.315023  0.271536  3.091 0.001992 **
CLAUDIN_SUBTYPELumB:Subset_Score_Quartile2 -0.763909  0.465842  0.394896 -1.934 0.053057 .
CLAUDIN_SUBTYPELumB:Subset_Score_Quartile3 -0.729682  0.482062  0.363685 -2.006 0.044818 *
CLAUDIN_SUBTYPELumB:Subset_Score_Quartile4 -0.909170  0.402859  0.369649 -2.460 0.013911 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                           exp(coef) exp(-coef) lower .95 upper .95
HER2_STATUSPositive                           1.7173     0.5823    1.1565    2.5500
LYMPH_NODES_EXAMINED_POSITIVE                 1.0513     0.9512    1.0340    1.0690
GRADE2                                        1.4638     0.6831    0.8891    2.4102
GRADE3                                        1.6959     0.5896    1.0139    2.8368
AGE_AT_DIAGNOSIS                              1.0181     0.9822    1.0076    1.0287
TUMOR_SIZE                                    1.0151     0.9851    1.0100    1.0203
CLAUDIN_SUBTYPELumB                           2.9121     0.3434    1.6193    5.2370
Subset_Score_Quartile2                        1.3703     0.7298    0.8289    2.2654
Subset_Score_Quartile3                        2.1522     0.4646    1.3260    3.4934
Subset_Score_Quartile4                        2.3150     0.4320    1.3596    3.9417
CLAUDIN_SUBTYPELumB:Subset_Score_Quartile2    0.4658     2.1467    0.2148    1.0101
CLAUDIN_SUBTYPELumB:Subset_Score_Quartile3    0.4821     2.0744    0.2363    0.9833
CLAUDIN_SUBTYPELumB:Subset_Score_Quartile4    0.4029     2.4823    0.1952    0.8314

Concordance= 0.713  (se = 0.015 )
Likelihood ratio test= 161.5  on 13 df,   p=<2e-16
Wald test            = 191.7  on 13 df,   p=<2e-16
Score (logrank) test = 217.5  on 13 df,   p=<2e-16
```

Figure 3. (**A**) Kaplan-Meier (KM) plot for luminal breast cancer disease specific survival (DSS) for each CNA quartile group. The p-value associated with the logrank test and a risk table displaying the number of patients at risk at each time interval is displayed. (**B**) Example of a χ^2 analysis of the data, individual χ^2 tests displayed in top box and table with adjusted p-values displayed in bottom box. (**C**) Example of an implementation of a multivariable Cox model.



Show 30 entries Search:

	CLAUDIN_SUBTYPE	Subset_Score_Quartile	HER2_STATUS	LYMPH_NODES_EXAMINED_POSITIVE	GRADE	AGE_AT_DIAGNOSIS	TUMOR_SIZE
1	LumA	1	Negative	1.79912280701754	2	63.8579914893617	25.3966980325064
2	LumB	1	Negative	1.79912280701754	2	63.8579914893617	25.3966980325064
3	LumA	2	Negative	1.79912280701754	2	63.8579914893617	25.3966980325064
4	LumB	2	Negative	1.79912280701754	2	63.8579914893617	25.3966980325064
5	LumA	3	Negative	1.79912280701754	2	63.8579914893617	25.3966980325064
6	LumB	3	Negative	1.79912280701754	2	63.8579914893617	25.3966980325064
7	LumA	4	Negative	1.79912280701754	2	63.8579914893617	25.3966980325064

Showing 1 to 8 of 8 entries Previous 1 Next

Figure 4. (**A**) Output showing format of new data and adjusted survival curves corresponding to the previous multivariable Cox model. (**B**) Example output of a ctree survival tree analysis. (**C**) Sample output from use of the `maftools` package, a MafSummary plot is displayed.

curves and adjusted survival curves split based on grouping variable level are displayed for users to view, customise and download in PNG or SVG format.

In the case where the PH assumption of the multivariable Cox model is violated, users can apply recursive partitioning survival trees available in the Survival Trees tab. Users can use the rpart or ctree[23–25] algorithms with customised argument parameters to produce survival trees containing one or more variables along with the corresponding KM curves (Figure 4 (B)). Users are provided with a workspace to select the survival time, event status and variables of interest, and information on each of the customisable arguments is given by pressing the information button located at the top of the box. Similar to previous tabs, the survival trees and accompanying KM curves can be exported with specified plot width and height. It should be noted that the ctree algorithm will only work where the selected categorical variables are in factor form.

Mutation analysis. An additional function of GNOSIS is the ability to summarise, analyse and visualise mutation annotation format (MAF) files using maftools[26]. MAF files are used to store detected somatic variants and are usually provided as part of the cBioPortal downloads. The Maftools Summary tab in GNOSIS allows users to view the MAF summary, sample summary, gene summary and summary of the associated clinical data, if available. If clinical data are provided users need to make sure the column named "Tumor_Sample_Barcode" is present. These summaries provide a basic view of the uploaded MAF file and contain information on the number of mutations, type of mutations and genes affected by these mutations. The Maftools Summary tab also enables users to examine the mutational landscape of the tumours in a graphical way. The plots available include MAF summary plots which display the number of variants in each sample as a stacked barplot and variant types as a boxplot summarized by variant classification (Figure 4 (C)). GNOSIS also contains panels for oncoplots, oncostrips, graphs displaying transition and transversion rates, lollipop plots for up to three genes simultaneously, mutation load plots and somatic interaction plots, all derived from the original *maftools* package. Other functions of this tab include allowing users to customise and export these plots in PNG or SVG format with specific dimensions.

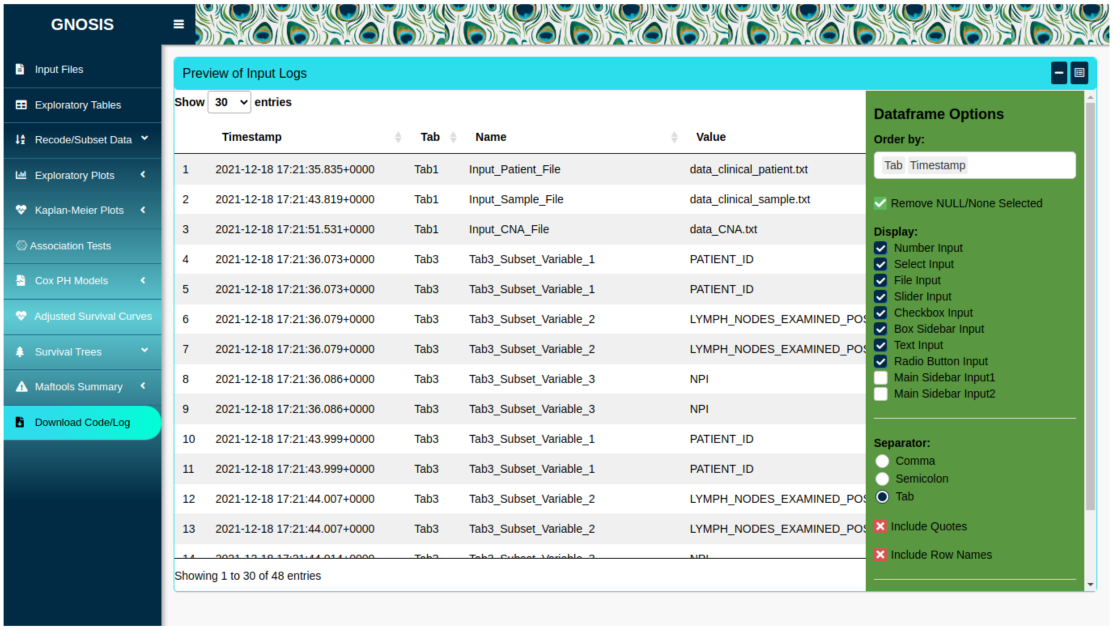
Reproducible research

GNOSIS facilitates reproducible research by providing a Download Code/Log tab where users can view and download a log containing information on all the inputs selected throughout the session, as well as downloading an R script containing code to reproduce the outputs displayed in the app (Figure 5).

Operation

GNOSIS is available on shinyapps.io and GitHub. This enables users to access GNOSIS via a web browser or run GNOSIS locally by downloading, extracting and launching the app manually in RStudio, or running the app in RStudio using: `shiny::runGitHub(repo='GNOSIS', username = 'Lydia-King',ref="main")`. The latter is recommended due to resource limitations imposed by the operators of the shinyapp.io website.

GNOSIS was originally developed using R version 3.5.1 but due to subsequent package updates, it now works on R versions



Preview of Input Logs

Show 30 entries

	Timestamp	Tab	Name	Value
1	2021-12-18 17:21:35.835+0000	Tab1	Input_Patient_File	data_clinical_patient.txt
2	2021-12-18 17:21:43.819+0000	Tab1	Input_Sample_File	data_clinical_sample.txt
3	2021-12-18 17:21:51.531+0000	Tab1	Input_CNA_File	data_CNA.txt
4	2021-12-18 17:21:36.073+0000	Tab3	Tab3_Subset_Variable_1	PATIENT_ID
5	2021-12-18 17:21:36.073+0000	Tab3	Tab3_Subset_Variable_1	PATIENT_ID
6	2021-12-18 17:21:36.079+0000	Tab3	Tab3_Subset_Variable_2	LYMPH_NODES_EXAMINED_POS
7	2021-12-18 17:21:36.079+0000	Tab3	Tab3_Subset_Variable_2	LYMPH_NODES_EXAMINED_POS
8	2021-12-18 17:21:36.086+0000	Tab3	Tab3_Subset_Variable_3	NPI
9	2021-12-18 17:21:36.086+0000	Tab3	Tab3_Subset_Variable_3	NPI
10	2021-12-18 17:21:43.999+0000	Tab3	Tab3_Subset_Variable_1	PATIENT_ID
11	2021-12-18 17:21:43.999+0000	Tab3	Tab3_Subset_Variable_1	PATIENT_ID
12	2021-12-18 17:21:44.007+0000	Tab3	Tab3_Subset_Variable_2	LYMPH_NODES_EXAMINED_POS
13	2021-12-18 17:21:44.007+0000	Tab3	Tab3_Subset_Variable_2	LYMPH_NODES_EXAMINED_POS
14	2021-12-18 17:21:44.014+0000	Tab3	Tab3_Subset_Variable_3	NPI

Showing 1 to 30 of 48 entries

Dataframe Options

Order by: Tab Timestamp

☑ Remove NULL/None Selected

Display:
☑ Number Input
☑ Select Input
☑ File Input
☑ Slider Input
☑ Checkbox Input
☑ Box Sidebar Input
☑ Text Input
☑ Radio Button Input
□ Main Sidebar Input1
□ Main Sidebar Input2

Separator:
○ Comma
○ Semicolon
● Tab

☒ Include Quotes

☒ Include Row Names

Figure 5. Dataframe containing log of inputs selected, which can be downloaded as a .txt file. Option to download R script containing code run in app also available.

≥ 4.0.0. GNOSIS depends on the packages, BiocManager, shiny, shinymeta, shinydashboard, dashboardthemes, shinydashboardPlus, shinyWidgets, shinycssloaders, shinylogs, fontawesome, DT, tidyverse, ggplot2, fabricatr, reshape2, operator.tools, rpart, rpart.plot, partykit, coin, survminer, survival, stats, rstatix, DescTools, car, compareGroups, R.utils, RColorBrewer and maftools[18–48], which are automatically installed and loaded when running GNOSIS manually from RStudio. Should GNOSIS be run using the `runGitHub()` function, shiny must be installed beforehand.

Use cases

GNOSIS was originally developed as part of a study to implement an exploratory and statistically robust survival analysis on the METABRIC luminal breast cancer cohort[16], and was pivotal in our ability to efficiently determine that CNAs reflecting genomic instability in luminal breast cancers are associated with survival. This work demonstrated both the utility and capability of this analytic ecosystem to facilitate oncogenomic analysis, and motivated us to make it available to the research community, to both use and further enhance, as appropriate.

The data utilised in the study[16] is available for download on cBioPortal as well as Zenodo (*Underlying data*[49]). Demonstration videos providing a walkthrough of GNOSIS are also provided on Zenodo (*Extended data*[50]). An Rmarkdown file and example R script containing the code to run the analysis presented are available on the project's GitHub.

We have also provided a subset of the METABRIC data used as part of King *et al.* (2021)[16] in the project's Zenodo repository to facilitate those users interested in exploring the capabilities of GNOSIS using the shinyapps.io app.

Conclusions

We have developed GNOSIS, an R Shiny app that supports the tractable and efficient exploratory analysis of cBioPortal clinical and genomic data products in a reproducible manner. Our experience with GNOSIS demonstrates its potential in helping researchers and clinicians in the analysis of archived consortia studies curated and accessible from cBioPortal, optimising the identification of variables and scores that have prognostic value and can aid in the identification of patients with a greater risk of lethal disease. Furthermore, GNOSIS' design and open-source basis makes it amenable to further development and enhancement by interested members of the community.

Data availability

Underlying data

Zenodo: Data associated with "Survival outcomes are associated with genomic instability in luminal breast cancers", https://doi.org/10.5281/zenodo.5791191[49] This project contains the following underlying data:

- data_clinical_patient.txt
- data_clinical_sample.txt
- data_CNA.txt
- data_CNA_subset_4000_genes.txt
- data_mutations_extended.txt

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Extended data

Zenodo: GNOSIS: an R Shiny app supporting cancer genomics survival analysis with cBioPortal, https://doi.org/10.5281/zenodo.5788544[50]

This project contains the following extended data:

- GNOSIS_Tab_1_Input_Files.mp4
- GNOSIS_Tab_1_Input_Files_with_Subtitles.mp4
- GNOSIS_Tab_2_Exploratory_Tables.mp4
- GNOSIS_Tab_2_Exploratory_Tables_with_Subtitles.mp4
- GNOSIS_Tab_3_Recode_Subset.mp4
- GNOSIS_Tab_3_Recode_Subset_with_Subtitles.mp4
- GNOSIS_Tab_4_Exploratory_Plots.mp4
- GNOSIS_Tab_4_Exploratory_Plots_with_Subtitles.mp4
- GNOSIS_Tab_5_KM_Plots.mp4
- GNOSIS_Tab_5_KM_Plots_with_Subtitles.mp4
- GNOSIS_Tab_6_Association_Tests.mp4
- GNOSIS_Tab_6_Association_Tests_with_Subtitles.mp4
- GNOSIS_Tab_7_Cox_Models.mp4
- GNOSIS_Tab_7_Cox_Models_with_Subtitles.mp4
- GNOSIS_Tab_8_Adjusted_Survival_Curves.mp4
- GNOSIS_Tab_8_Adjusted_Survival_Curves_with_Subtitles.mp4
- GNOSIS_Tab_9_Survival_Trees.mp4
- GNOSIS_Tab_9_Survival_Trees_with_Subtitles.mp4
- GNOSIS_Tab_10_Maftools.mp4
- GNOSIS_Tab_10_Maftools_with_Subtitles.mp4
- GNOSIS_Tab_11_Download_Code_Log.mp4
- GNOSIS_Tab_11_Download_Code_Log_with_Subtitles.mp4

Videos are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability

Source code available from: https://github.com/Lydia-King/GNOSIS

Archived source code as at time of publication: https://doi.org/10.5281/zenodo.6543416[51]

License: MIT

Acknowledgements

This manuscript makes use of data generated by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC). The authors wish to express their thanks to the patients and their families for having participated in the METABRIC study.

References

1. Russnes HG, Lingjærde OC, Børresen-Dale AL, *et al.*: **Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters.** *Am J Pathol.* 2017; **187**(10): 2152–2162.
PubMed Abstract | Publisher Full Text
2. Carbone A: **Cancer Classification at the Crossroads.** *Cancers (Basel).* 2020; **12**(4): 980.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Christofyllakis K, Bittenbring JT, Thurner L, *et al.*: **Cost-effectiveness of precision cancer medicine-current challenges in the use of next generation sequencing for comprehensive tumour genomic profiling and the role of clinical utility frameworks (Review).** *Mol Clin Oncol.* 2022; **16**(1): 21.
PubMed Abstract | Publisher Full Text | Free Full Text
4. International Cancer Genome Consortium, Hudson TJ, Anderson W, *et al.*: **International network of cancer genome projects.** *Nature.* 2010; **464**(7291): 993–998.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Tomczak K, Czerwińska P, Wiznerowicz M: **The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.** *Contemp Oncol (Pozn).* 2015; **19**(1A): 68–77.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Cerami E, Gao J, Dogrusoz U, *et al.*: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.** *Cancer Discov.* 2012; **2**(5): 401–404.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Gao J, Aksoy BA, Dogrusoz U, *et al.*: **Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.** *Sci Signal.* 2013; **6**(269): pl1.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Gassen J: **ExPanDaR: Explore Your Data Interactively.** R package version 0.5.3. 2020.
Reference Source
9. Meyer F, Perrier V: **esquisse: Explore and Visualize Your Data Interactively.** R package version 1.1.1. 2022.
Reference Source
10. Lánczky A, Győrffy B: **Web-Based Survival Analysis Tool Tailored for Medical Research (KMplot): Development and Implementation.** *J Med Internet Res.* 2021; **23**(7): e27633.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Korkmaz S, Goksuluk D, Zararsiz G, *et al.*: **geneSurv: An interactive web-based tool for survival analysis in genomics research.** *Comput Biol Med.* 2017; **89**: 487–496.
PubMed Abstract | Publisher Full Text
12. Zhou Y, Leung SW, Mizutani S, *et al.*: **MEPHAS: an interactive graphical user interface for medical and pharmaceutical statistical analysis with R and *Shiny*.** *BMC Bioinformatics.* 2020; **21**(1): 183.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Gu Z, Mullighan CG: **ShinyCNV: a Shiny/R application to view and annotate DNA copy number variations.** *Bioinformatics.* 2019; **35**(1): 126–129.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Ramesh RG, Bigdeli A, Rushton C, *et al.*: **CNViz: An R/Shiny Application for Interactive Copy Number Variant Visualization in Cancer.** *J Pathol Inform.* 2022; **13**: 100089.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Franch-Expósito S, Bassaganyas L, Vila-Casadesús M, *et al.*: **CNApp, a tool for the quantification of copy number alterations and integrative analysis revealing clinical implications.** *eLife.* 2020; **9**: e50267.
PubMed Abstract | Publisher Full Text | Free Full Text
16. King L, Flaus A, Holian E, *et al.*: **Survival outcomes are associated with genomic instability in luminal breast cancers.** *PLoS One.* 2021; **16**(2): e0245042.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Curtis C, Shah SP, Chin SF, *et al.*: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature.* 2012; **486**(7403): 346–352.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Chang W, Cheng J, Allaire JJ, *et al.*: **shiny: Web Application Framework for R.** 2020.
Reference Source
19. Wickham H, Averick M, Bryan J, *et al.*: **Welcome to the tidyverse.** *J Open Source Softw.* 2019; **4**(43): 1686.
Publisher Full Text
20. Wickham H: **ggplot2: Elegant Graphics for Data Analysis.** Springer-Verlag New York, 2016.
Reference Source
21. Therneau TM, Lumley T, Elizabeth A, *et al.*: **A Package for Survival Analysis in S.** version 2.38. 2015.
Reference Source
22. Kassambara A, Kosinski M: **survminer: Drawing Survival Curves using 'ggplot2'.** R package version 0.4.4. 2019.
Reference Source
23. Therneau T, Atkinson B: **rpart: Recursive Partitioning and Regression Trees.** R package version 4.1-15. 2019.
Reference Source
24. Hothorn T, Zeileis A: **partykit: A modular toolkit for recursive partytioning in R.** *J Mach Learn Res.* 2015; **16**: 3905–3909.
Reference Source
25. Hothorn T, Hornik K, Zeileis A: **Unbiased recursive partitioning: A conditional inference framework.** *J Comput Graph Stat.* 2006; **15**(3): 651–674.
Publisher Full Text
26. Mayakonda A, Lin D, Assenov Y, *et al.*: **Maftools: efficient and comprehensive analysis of somatic variants in cancer.** *Genome Res.* 2018; **28**(11): 1747–1756.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Subirana I, Sanz H, Vila J: **Building bivariate tables: The compareGroups package for R.** *J Stat Softw.* 2014; **57**(12): 1–16.
Publisher Full Text
28. Morgan M: **BiocManager: Access the Bioconductor Project Package Repository.** 2021.
Reference Source
29. Cheng J, Sievert C: **shinymeta: Export Domain Logic from Shiny using Meta-Programming.** 2021.
Reference Source
30. Chang W, Ribeiro BB: **shinydashboard: Create Dashboards with 'Shiny'.** 2021.
Reference Source
31. Lilovski N: **dashboardthemes: Customise the Appearance of 'shinydashboard' Applications using Themes.** 2021.
Reference Source
32. Granjon D: **shinydashboardPlus: Add More 'AdminLTE2' Components to 'shinydashboard'.** 2021.
Reference Source
33. Perrier V, Meyer F, Granjon D: **shinyWidgets: Custom Inputs Widgets for Shiny.** 2021.
Reference Source
34. Sali A, Attali D: **shinycssloaders: Add Loading Animations to a 'shiny' Output While It's Recalculating.** 2020.
Reference Source
35. Meyer F, Perrier V: **shinylogs: Record Everything that Happens in a 'Shiny' Application**. R package version 0.1.7. 2019.
Reference Source
36. Iannone R: **fontawesome: Easily Work with 'Font Awesome' Icons**. R package version 0.2.2. 2021.
Reference Source
37. Xie Y, Cheng J, Tan X: **DT: A Wrapper of the JavaScript Library 'DataTables'**. R package version 0.19. 2021.
Reference Source
38. Blair G, Cooper J, Coppock A, *et al.*: **fabricatr: ImagineYour Data BeforeYou CollectIt**. R package version 0.14.0. 2021.
Reference Source
39. Wickham H: **Reshaping data with the reshape package.** *J Stat Softw.* 2007; **21**(12): 1–20.
Publisher Full Text
40. Brown C: **operator.tools: Utilities for Working with R's Operators**. R package

version 1.6.3. 2017.
Reference Source

41. Milborrow S: **rpart.plot: Plot 'rpart' Models: An EnhancedVersionof 'plot. rpart'**. R package version 3.1.0. 2021.
Reference Source

42. Hothorn T, Hornik K, van de Wiel MA, *et al.*: **Implementing a class of permutation tests: The coin package.** *J Stat Softw.* 2008; **28**(8): 1–23.
Reference Source

43. R Core Team: **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2021.
Reference Source

44. Kassambara A: **rstatix: Pipe-Friendly Framework for Basic Statistical Tests**. R package version 0.7.0. 2021.
Reference Source

45. Andri S, Aho K, Alfons A, *et al.*: **DescTools: Tools for Descriptive Statistics**. R package version 0.99.43. 2021.
Reference Source

46. Fox J, Weisberg S: **An R Companion to Applied Regression**. Sage, Thousand Oaks CA, third edition, 2019.
Reference Source

47. Bengtsson H: **R.utils: Various Programming Utilities**. R package version 2.11.0. 2021.
Reference Source

48. Neuwirth E: **RColorBrewer: ColorBrewer Palettes**. R package version 1.1. 2014.
Reference Source

49. King L, Flaus A, Holian E, *et al.*: **Data associated with "Survival outcomes are associated with genomic instability in luminal breast cancers"**. 2021.
http://www.doi.org/10.5281/zenodo.5791191

50. King L, Flaus A, Coughlan S, *et al.*: **GNOSIS: an R Shiny app supporting cancer genomics survival analysis with cBioPortal**. 2021.
http://www.doi.org/10.5281/zenodo.5788544

51. King L: **Lydia-King/GNOSIS: GNOSIS (v1.0.3).** *Zenodo.* 2022.
http://www.doi.org/10.5281/zenodo.6543416

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 28 April 2022

https://doi.org/10.21956/hrbopenres.14690.r31939

✓ **Tiago C. Silva**
Division of Biostatistics, Department of Public Health Sciences, Miller School of Medicine, University of Miami, Miami, FL, USA

King *et al*. developed GNOSIS (GeNomics explOrer using StatistIcal and Survival analysis in R), an open-source R Shiny app that provides an easy non-coding way to analyze clinical and cancer genomic data. With data in the same format as the ones provided by cBioPortal, the users can perform EDA (Exploratory Data Analysis), survival analysis, and mutation analysis.

Overall, the graphical user interface is well designed and straightforward to use, and tutorial videos are available as documentation. Using the provided example dataset, I was able to reproduce most of its functionalities. Below are some comments and suggestions.

Major comments:

- Using the data_mutations_extended.txt file the Maftools summary tab is giving an error in the shinyapps.io version.
- If I upload the data and jump directly to exploratory plots, it asks to calculate CNA Scores. You should still be able to use this tab without calculating CNA Scores. Also, if you calculate as “single gene”, it seems this tab does not work.

Minor comments:

- In the Survival tree tabs and the corresponding survival curves panel, the plot could be improved by removing the gray background and the addition of the p-value.
- I was not able to understand the purpose of the “The exploratory tables” tab. Is it only selecting columns? If so, I believe this functionality is not useful and distracting, users could simply open it in spreadsheet software (MS. Excel, google sheets, open office) and have more freedom to filter.
- Providing a vectorized version of plots would be very useful since PNGs are not modifiable.
- Filtering the options in the selection of variables would make the GNOSIS easier to use.

- ○ The Association test tab could be improved with the addition of the compareGroups package.

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

Is the description of the software tool technically sound?
Yes

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

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?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Software engineering, bioinformatics, data analysis, cancer

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 10 Jun 2022

Aaron Golden, National University of Ireland, Galway, Ireland

Reviewer Point P 2.1 – *Using the data mutations extended.txt file the Maftools summary tab is giving an error in the shinyapps.io version.*

Reply: We thank the reviewer for bringing this to our attention, the issue is now fixed

Reviewer Point P 2.2 – If I upload the data and jump directly to exploratory plots, it asks to calculate CNA Scores. You should still be able to use this tab without calculating CNA Scores. Also, if you calculate as “single gene”, it seems this tab does not work.

Reply: We thank the reviewer for pointing this out. Users do have the ability to create exploratory plots for the clinical variables and CNA calls for each patient separately, once the respective files are uploaded on their own. However, when users upload the clinical and CNA data together, the two files are integrated and merged on the column PATIENT ID. In this case, rather than being able to select specific patients, the user is asked to select genes to interrogate. As a result, the number of options available for users to select increases significantly (for example from 2000 to ¿ 20000) and so to avoid overwhelming both the user and the App, users are asked to calculate CNA Scores for each patient or to choose a number of individual genes to interrogate. This means that instead of having upwards of 22,000 genes to choose from, users can choose to create a whole genome CNA Score or pick specific genes of interest to analyse. We hope the above explanation sufficiently explains why the user is asked to calculate CNA Scores/Select single genes before downstream analysis when both the clinical and CNA data are uploaded and addresses the reviewer's concerns.

Reviewer Point P 2.3 – *In the Survival tree tabs and the corresponding survival curves panel, the plot could be improved by removing the gray background and the addition of the p-value.*

Reply: We agree that the survival curve plots within the Survival Trees Tab, and indeed all plots throughout the App, could be made clearer by removing the gray background. As a result, the back ground colour for all plots has been changed to white. In terms of adding the p-value to the survival curves corresponding to survival tree node, this function is already available and can be implemented using the "Display P-value" button within the right box sidebar.

Reviewer Point P 2.4 – *I was not able to understand the purpose of the "The exploratory tables" tab. Is it only selecting columns? If so, I believe this functionality is not useful and distracting, users could simply open it in spreadsheet software (MS. Excel, google sheets, open office) and have more freedom to filter.*

Reply: We thank the reviewer for this feedback. The purpose of the "The exploratory tables" tab is to enable further exploration of user selected columns. While it technically is only selecting columns to view in a clearer, less cluttered way, and the users could indeed use spreadsheet software for this purpose, we think this tab is useful. It allows users to run all steps in their analysis in one place and gives them the opportunity to access the R code used to select columns via a downloadable R script. As a result, we have decided to keep this tab in the App.

Reviewer Point P 2.5 – *Providing a vectorized version of plots would be very useful since PNGs are not modifiable.*

Reply: We thank the reviewer for this extremely useful suggestion and in response changes have been made to the App. Users are now given the option to download the plots in both .png and .svg format.

Reviewer Point P 2.6 – *Filtering the options in the selection of variables would make the GNOSIS easier to use.*

Reply: Filtering of variables in the select boxes is already implemented. Users can remove the "None Selected" input in the select box and type in a letter or word to filter variable names. This makes it easier to use and means users can search for specific variables or reduce the number of variable names shown in select input boxes.

Reviewer Point P 2.7 – *The Association test tab could be improved with the addition of the compareGroups package*.

Reply: We thank the reviewer for this suggestion and have added a section where users can utilise the basic functions of the compareGroups package.

Competing Interests: No competing interests were disclosed.

Reviewer Report 15 March 2022

https://doi.org/10.21956/hrbopenres.14690.r31288

✓ **Jessica C. Mar**
Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, Qld, Australia

King *et al*. have developed an open-source tool using R Shiny that facilitates analyses on largescale cancer genomic data and related clinical variables. While the analyses that this tool implements are straightforward and routine, the utility of this tool is that it is easy to use and removes the need for advanced programming skills. Overall, this is a thoughtfully designed tool that succeeds in its aim to make this kind of data analysis accessible to everyone.

With regards to GNOSIS' role in generating robust and reliable statistical analyses, are there clear definitions or recommendations that users are directed to, to ensure that the statistics are being used appropriately with the input data? For example, on p7, the Chi-squared test is being used "with sufficient cell size" - is this defined somewhere in the tool? Similarly when working with survival analysis, it is necessary to ensure that the data meets regulatory assumptions, is there a note somewhere so that users can assess this?

Another aspect which the authors could comment on is what are some of the other alternatives or similar tools to GNOSIS that are currently out there? Either to appreciate the utility of GNOSIS or to give users options if they are looking for something that may sit outside of GNOSIS' scope?

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

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow

replication of the software development and its use by others?
Yes

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Computational Biology, Biostatistics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 10 Jun 2022

Aaron Golden, National University of Ireland, Galway, Ireland

Reviewer Point P 1.1 – *With regards to GNOSIS' role in generating robust and reliable statistical analyses, are there clear definitions or recommendations that users are directed to, to ensure that the statistics are being used appropriately with the input data? For example, on p7, the Chi-squared test is being used "with sufficient cell size" - is this defined somewhere in the tool? Similarly when working with survival analysis, it is necessary to ensure that the data meets regulatory assumptions, is there a note somewhere so that users can assess this?*

Reply: We thank the reviewer for bringing this up and agree that it is extremely important that the statistical tests are being used and interpreted appropriately. To aid users in deciding what test is most appropriate, if the assumptions of the test are met, how to interpret the results and also how to run a basic analysis in R we have included an information button in certain tabs which include links to helpful websites. These changes can be seen throughout the App, mainly in the Kaplan-Meier Plots tab, the Association Tests tab, the Cox PH Models tab and the Survival Trees tab, and we hope this will make it easier for users to perform a robust and statistically sound analysis.

Reviewer Point P 1.2 – *Another aspect which the authors could comment on is what are some of the other alternatives or similar tools to GNOSIS that are currently out there? Either to appreciate the utility of GNOSIS or to give users options if they are looking for something that may sit outside of GNOSIS' scope?*

Reply: We thank the reviewer for suggesting this and in response have amended the manuscript to include other alternatives or similar tools to GNOSIS that users may find helpful.

Competing Interests: No competing interests were disclosed.