**How does it work?**

GeneIntelligence (GI) provides a fully automatic tool for exploratory data analysis purposes. The main concept behind GI software is to figure out, and rank associations between measurements (genes expression or methylation levels) and the phenotype of interest (cancer grade, treatment response). Unlike other methods (such as linear regression) our algorithms have no direct assumptions about input data. Thus, they are not biased when, for example, the assumption of normality is not met.

**How to use GI?**

To work with our software You need to sign up, and then prepare data (for more details about supported data types see **data sources** section) in a specific but simple manner. GI software input must contain two files in **csv** (comma-separated values) format. First one is a **dataset** containing samples and measurements (Fig. 1), please note that orientation of the dataset matrix is crucial for correct data interpretation.

| | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 | ... | Sample 96 | Sample 97 | Sample 98 | Sample 99 | Sample 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Marker 1** | 0.399331 | 0.641422 | 0.451378 | 0.586302 | 0.761094 | ... | 0.530739 | 0.193134 | 0.550380 | 0.747962 | 0.829768 |
| **Marker 2** | 0.284235 | 0.776740 | 0.754289 | 0.025583 | 0.457787 | ... | 0.007735 | 0.093815 | 0.468623 | 0.684118 | 0.671044 |
| **Marker 3** | 0.160823 | 0.146710 | 0.249721 | 0.339304 | 0.283308 | ... | 0.436064 | 0.386162 | 0.548084 | 0.177627 | 0.778889 |
| **Marker 4** | 0.453052 | 0.647405 | 0.150433 | 0.172332 | 0.952193 | ... | 0.184721 | 0.023693 | 0.069511 | 0.020453 | 0.469305 |
| **Marker 5** | 0.586594 | 0.615811 | 0.268058 | 0.551970 | 0.093838 | ... | 0.657936 | 0.738773 | 0.189183 | 0.749221 | 0.213258 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **Marker 996** | 0.583810 | 0.072024 | 0.138708 | 0.272905 | 0.320847 | ... | 0.568011 | 0.674850 | 0.569462 | 0.317151 | 0.257352 |
| **Marker 997** | 0.330525 | 0.066787 | 0.549399 | 0.848706 | 0.781252 | ... | 0.259799 | 0.063520 | 0.800575 | 0.114661 | 0.778248 |
| **Marker 998** | 0.526365 | 0.770025 | 0.972812 | 0.008097 | 0.469279 | ... | 0.653814 | 0.473823 | 0.220452 | 0.213908 | 0.386036 |
| **Marker 999** | 0.983433 | 0.661838 | 0.136511 | 0.856487 | 0.857294 | ... | 0.475064 | 0.417092 | 0.965862 | 0.024642 | 0.196885 |
| **Marker 1000** | 0.775953 | 0.265396 | 0.374706 | 0.789065 | 0.616747 | ... | 0.954421 | 0.083763 | 0.896978 | 0.620914 | 0.944289 |

1000 rows × 100 columns

Fig 1: Thumbnail of example dataset, rows contain samples, columns contain measurements e.g. expression or methylation levels per sample.

Second necessary file is **POI** phenotype-of-interest (Fig. 2) and should include sample ID (the same as in the dataset file) and information about the phenotype. Additionally POI may contain covariates (i.e. age, sex or BMI) to adjust analysis. And in case of high heterogeneity tissue (such as blood) information about cell proportions (details in section cell fraction correction).

| | POI | Sex |
|---|---|---|
| **Sample 1** | Target | 1 |
| **Sample 2** | Target | 1 |
| **Sample 3** | Control | 0 |
| **Sample 4** | Control | 0 |
| **Sample 5** | Target | 0 |
| **...** | ... | ... |
| **Sample 96** | Target | 1 |
| **Sample 97** | Target | 1 |
| **Sample 98** | Control | 0 |
| **Sample 99** | Control | 1 |
| **Sample 100** | Target | 1 |

100 rows × 2 columns

Fig 2: Thumbnail of example POI, rows contains samples, columns contains information about sample phenotype, covariate contains additional information about sample (1 - male, 0 - female)

When files are prepared to use, sign in into account and navigate to new analysis section, enter basic information and then press **next** button (Fig. 3)

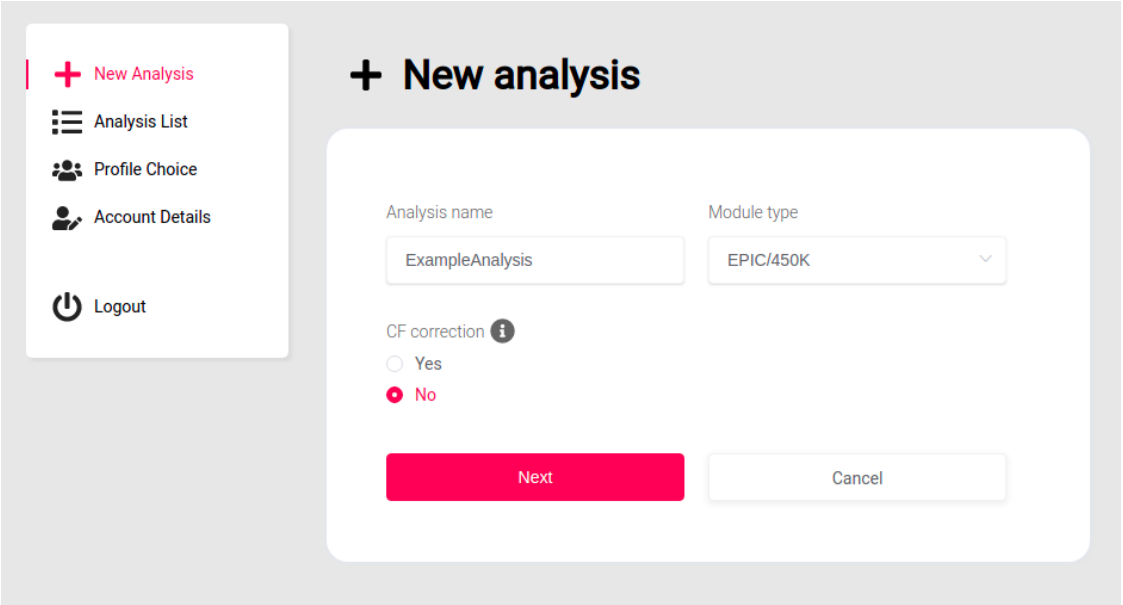


Fig 3: In **new analysis** section, set analysis name (all analysis are stored in **analysis list** section), module type (**EPIC/450K** for methylation and **RNA-seq** for expression data [currently not available]), turn on/off correction for cell fraction proportion (this option requires additional data in POI file)

Now You should see a new window (Fig. 4) here You can upload **dataset** and **poi** files, when files transfer finish press **run** button**.**

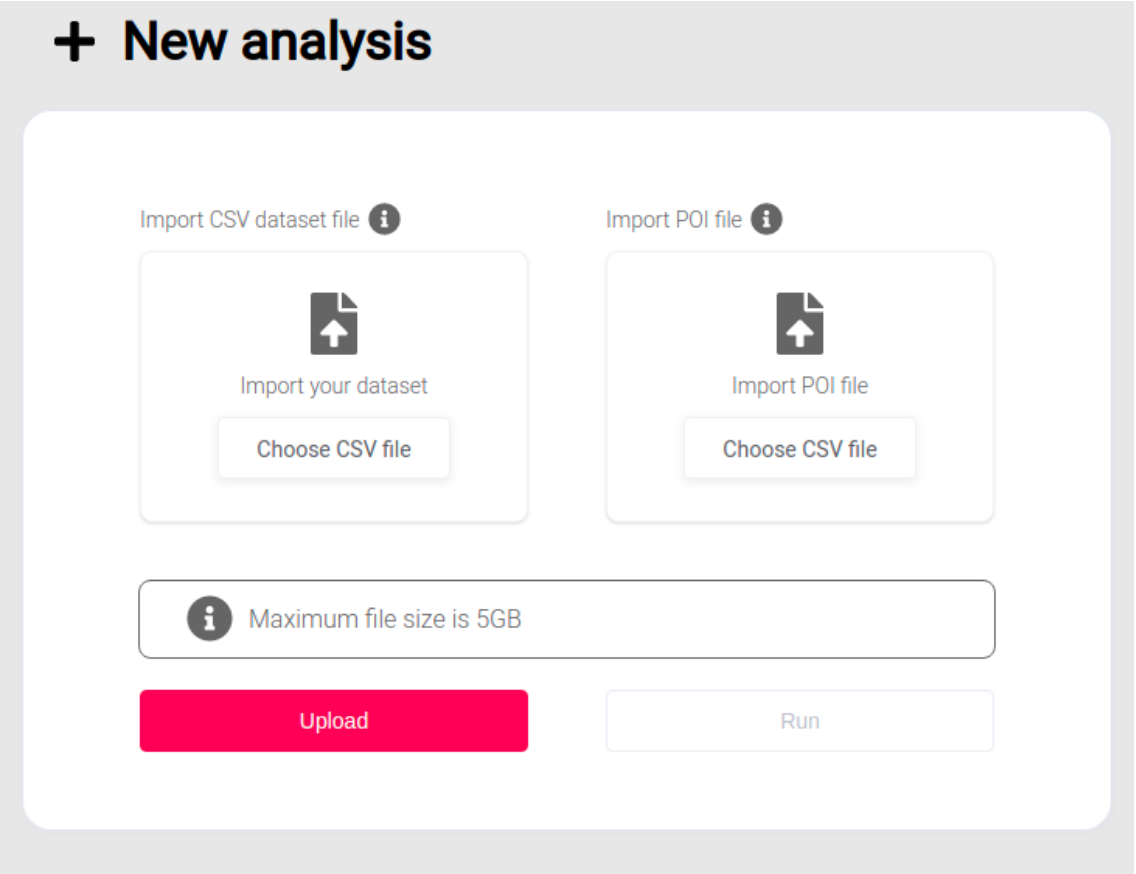


Fig 4: In this section You can set and upload files to our infrastructure. When files transferring process finish, the **run** button will activate**.**

Analysis process may take up to several hours depending on the dataset size. When the process ends we will send You notification via email and analytical report will be available in **analysis section** (Fig. 5).

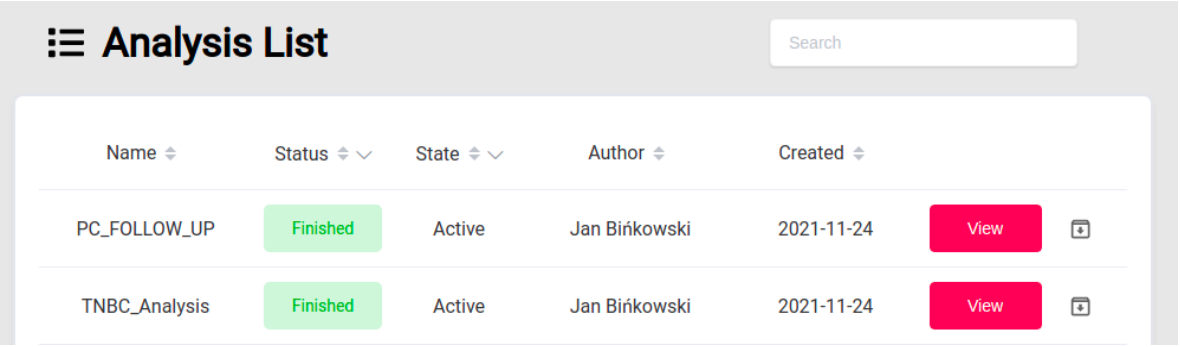


Fig 5: Here You can search, sort, archive and view prepared analytical reports.

**Important notes and tips about files preparation:**

- Covariates in POI file must be numerical (at least on ordinal scale)
- In case of multi-class study (e.g. when poi is tumor stage) phenotypes must be encoded to numerical, ordered form 1 - stage 1, 2 - stage 2, 3 - stage 3 etc.
- Samples ID must perfect match between dataset and POI files
- If cell-fraction correction is turned on, POI file must contain data about cell proportion in separate columns (columns used for correction should starts with CF_ prefix e.g. CF_CD4T, CF_Monocytes etc.)
- Currently we can process volumes up to 5gb, if your datasets exceed this limit please contact us

If provided files do not fulfill those conditions, our validations systems will stopp analysis and inform You about the reason via email.

**Data sources**

For now GI software can handle data from EPIC/450K Illumina BeadChips containing methylation levels expressed as beta-values. Future versions of GI software will provide modules to work with gene expression or mixed-type / custom data. If you have a specific dataset currently not supported by the GI online version, please contact us.

**Covariates**

To adjust analyse, users may provide covariates in the POI file, they will be considered in markers importance evaluation.

**Cell fraction correction**

In case of complex tissue such as blood, there is a need to adjust measurements for cell fraction proportions. This is due to the fact that different cell types vary in specific gene methylation (or expression) levels. It causes that variation in cell proportion may result in differences in observed measurements thus false positive/negative results.
This method requires data about the cell fractions in the POI file in separate columns with **CF_** prefix. Please note that this is not a necessary step of analysis. If You do not know how to predict cell fraction proportions from expression or methylation data, please contact us.

**Analytical report**

1) Tabular report provides numerical parameters useful to estimate marker importance. **Marker** is a variable name from a dataset file, if **biased** is equal to False **Adj. p-value** display p-value adjusted for provided covariates, **biased** gives an information if this specific variable can not be adjusted (if POI file does not contain any additional variables or biased == True Adj. p-value is calculated using parametric or nonparametric test for differences among means). **Separation** gives information how well marker can distinguish studied groups (separation = 1 means perfect, linear separation). Factor is a metric that allows to order markers (the strongest have the highest factor value). Please note that this value is study-specific and can not be directly compared between different analyses. In multiclass cases additional post hoc-tests are performed for each marker.
2) Clustermap is a heatmap ordered by an unsupervised clustering algorithm . This type of analysis allows us to figure out patterns specific for each poi.
3) Principal component analysis (PCA) and TSNE allows projecting high dimensional data into lower dimensional space.
4) Trees are graphical way to represent as simple as possible how to disninghuis poi using selected markers.