GenomeGems

User Manual

Ver. 3.0

March 19, 2012

Table of Contents

1. Introduction

1.1 Rationale

Organizing the great amount of sequences generated by Deep Sequencing so that mutations, which might possibly be biologically relevant, are easily identified is a difficult task. Yet, for this assignment only limited automatic tools exist. *GenomeGems* comes to gap this need by evaluating variability in Deep Sequencing generated genetic data in a simple tabular depiction, graphical representation and visualization for comparing multiple sequencing datasets. As such, via automatic, clear and accessible presentation of processed Deep Sequencing data, this tool aims to facilitate ranking of genomic variance calling.

GenomeGems runs on a local PC and is freely available at:

http://www.tau.ac.il/~nshomron/GenomeGems

1.2 Software Work-Flow

A basic implementation of *GenomeGems* enables the user, to analyze and visualize the input data according to the following flow chart:

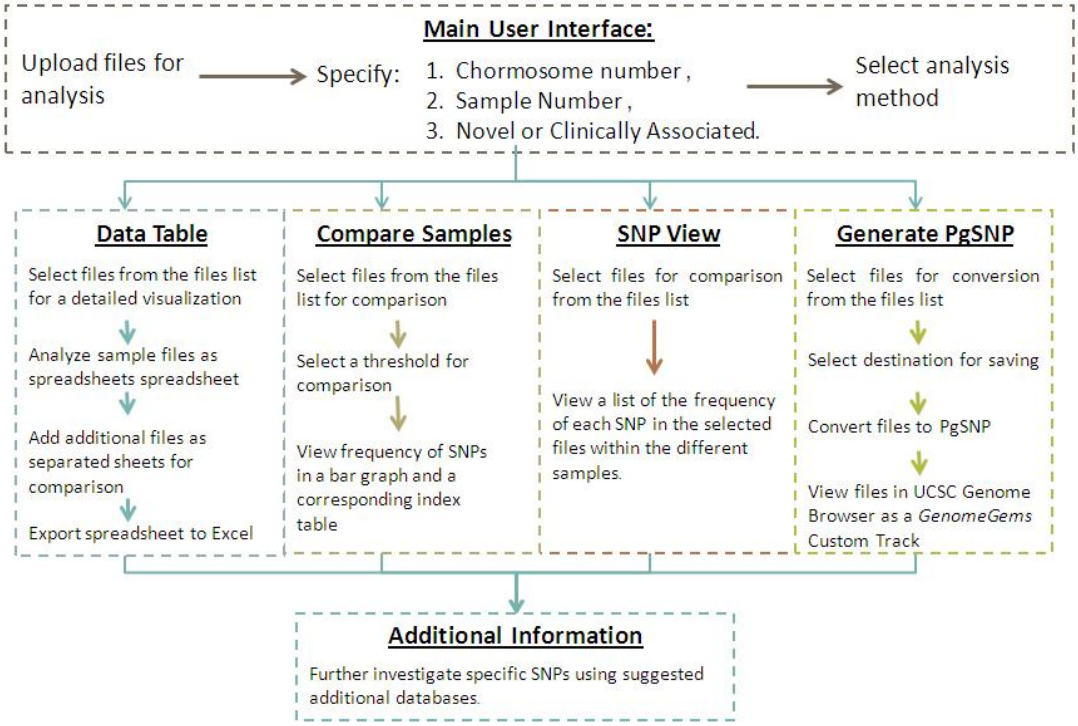

Figure 1- An illustration of the organization of *GenomeGems* in a schematic workflow.

GenomeGems provides different analysis functions which are described in the above schematic workflow. The user uploads the SNP files in the pre-determined format and chooses the form of analysis required: translation to a PG-SNP file format for UCSC visualization, visualization via data table, sample comparison via bar graph or table. In addition, more information about investigated SNPs can be obtained by using the suggested links to external databases.

1.3 New in *GenomeGems* 3.0

Version 3.0 of GenomeGems, released in March 2012, includes validation of data input. This involves validation in the main screen of GenomeGems, as well as in the different analysis screens. An error will appear in case any information required for analysis is missing.

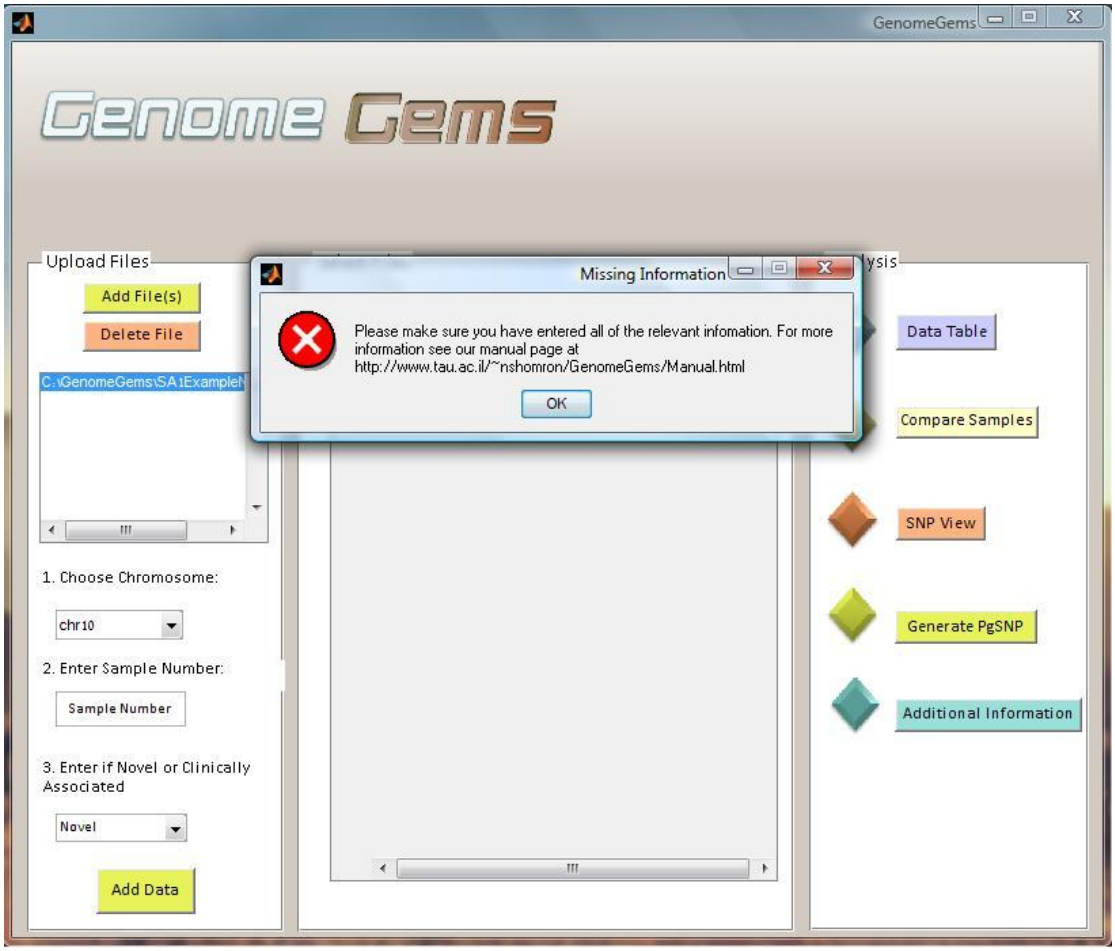

Figure 2 – Example of an error message as a result of missing information in the Main Screen (user did not enter Sample Number)

2. Software Description

2.1 Key Features

The key design feature underlying *GenomeGems* application is facilitating the final steps of Deep Sequencing data analysis via organizing and allowing accessible presentation of the data, thus leading to a rapid shift to the next step of experimental mutation detection. A sample processing pipeline is presented in Figure 3.

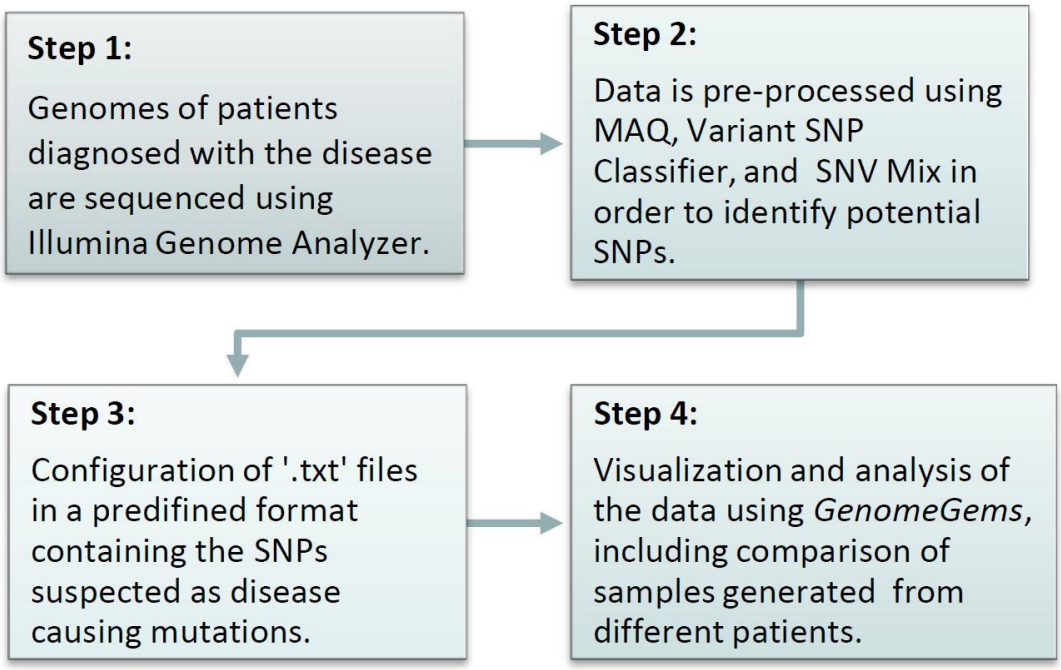

Figure 3- A flow chart describing the process performed on a sample data.

2.2 Development Environment

GenomeGems was developed using MATLAB 2009 functions and MATLAB's GUI tools. In addition, *GenomeGems* integrates well with the University of California Santa Cruz (UCSC) Genome Browser for the purpose of SNP visualization within investigated chromosomes.

3. Installation

3.1 Software and System Requirements

1. Windows operating system XP or Vista (*GenomeGems* currently is not supported by Windows 7). To identify your operating system: Click **Start**, right-click **My Computer**, and then click **Properties**. The edition of Windows you're running is displayed.

2. MCR 7.10 (URL LINK: **h**ttp://www.math**works.com/h**elp/toolbox/compi**ler/f12-99**9353.html)
3. ActiveX—MS Spreadsheet 10 (URL LINK: http://www.microsoft.com/download/en/details.aspx?displaylang=en&id=9468)

Alternatively, MCR 7.10 and ActiveX components can be downloaded directly from the following link:

http://www.tau.ac.il/~nshomron/GenomeGems/SystemReqs.html

3.2 Installing the Software

GenomeGems is freely available at http://www.tau.ac.il/~nshomron/GenomeGems.

Installation is currently available for Windows XP and Vista:

1. Press the 'System Requirements' link on the left panel of the website home page.
2. Install MCR 7.10
3. Install ActiveX—MS Spreadsheet 10
4. Download and run the *GenomeGems* application[1]

4. Getting Started with *GenomeGems*

Upload files containing a list of SNPs after analysis by MAQ (or other software) in a pre-determined format. The files must be in ".txt" format with columns separated by tabs. The files must contain the following data (in this specific order):

1. Chromosome number
2. SNP Position
3. Consensus Nucleotide
4. SNP nucleotide
5. Score of the SNP
6. Number of reads of each nucleotide

If any information is missing use "0". Other optional information that can be submitted: Gene Name, SNP Novel or Known, CDS (Coding Sequence) or Non-Coding, Synonymous or

[1] **NOTE**: In case the main user interface of *GenomeGems* does not open after running the application, there may be a problem with the MCR installation. Please confirm whether you have installed MCR 7.10 version correctly. If you still have a problem please contact us at: GenomeGem@gmail.com

Non-Synonymous, Amino Acid Replacement, SNP ID for known SNPs and so on. See Figure 4 for an example of a sample file. Additionally, you can download a few sample datasets from the *GenomeGems* website, at http://www.tau.ac.il/~nshomron/GenomeGems.

Data input is supplied by uploading the files that are to be analyzed, and choosing the chromosomes relevant for each file. This list of files and chromosomes is saved, and is later accessed throughout the employment of the tool.

```
chr10 18429624  C  A 184 A:35  C:16  C->A  CACNB2 novel Non Coding
chr10 73573082  T  C  27 C:3 T:4 T->C  CDH23 novel CDS [9729-9730 3243-3244] subst_NONSYNONYMOUS[TCC:S cCC:P U]
chr10 78709061  C  T 255 T:163  C:164  C->T  KCNMA1 novel CDS [2469-2470 823-824]  subst_NONSYNONYMOUS[GTC:V aTC:I c]
chr10 79397459  A  C  14 0 0 A->C  KCNMA1 novel Non Coding
chr10 79397521  A  C  4 0 0 A->C  KCNMA1 novel Non Coding
chr10 79397523  C  G  4 0 0 C->G  KCNMA1 novel Non Coding
chr10 79397525  C  G  4 0 0 C->G  KCNMA1 novel Non Coding
chr10 79397536  C  T  0 0 0 C->T  KCNMA1 novel Non Coding
chr10 87359674  T  A 110 A:33  T:57  T->A  GRID1 novel Non Coding
chr10 87359697  A  C  67 C:27  A:85  A->C  GRID1 novel Non Coding
```

Figure 4– Example of the input file format required for *GenomeGems*. The file must contain data from one single sample, and must not contain a heading line. The file may contain one single chromosome or all chromosomes, but in both cases the user must specify the chromosome for analysis. The data in the file must be separated into columns using tabs, and must contain the first 7 columns: chromosome number, SNP position, consensus nucleotide, SNP nucleotide, score of the SNP, number of reads for each nucleotide, as shown in the figure. The file may include any additional data in the following columns, also separated by tabs.

5. *GenomeGems* Step by Step

5.1 Uploading Data Files

The main user interface contains three distinct panels as seen in Figure 5: (A) Upload Files, (B) Select Files, and (C) Analysis.

The Upload Files panel contains a list into which input files are uploaded, chromosome on which the analysis will be performed is selected, and the sample number and whether the data is of novel or clinically associated SNPs are specified. After this information is specified for each sample uploaded, press the 'Add Data' button and the data file that has been just added will appear in the 'Select Files' table.

Multiple files containing multiple samples may be uploaded, but each file must be of one single sample. In addition, multiple chromosomes on which the later analysis will be performed can be chosen.

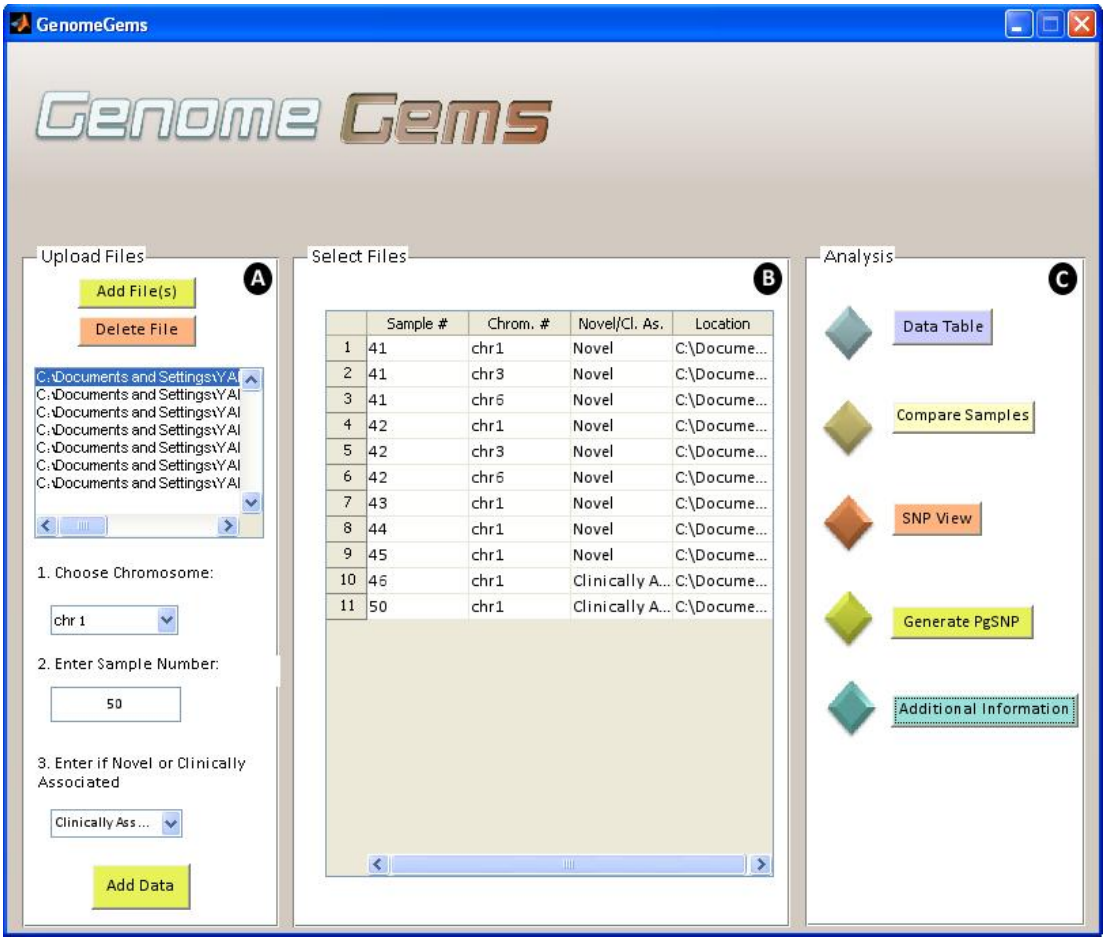

Figure 5 - The *GenomeGems* main user interface contains three distinct panels, (A) Upload Files, (B) Select Files and (C) Analysis. The user may upload an unlimited number of samples and chromosomes that will later be available from each of the analysis tools.

The selected files, with a specified sample number, chromosome number, novel or clinically associated and location appear in the 'Select Files' panel (marked as B) as a list. This list of files must include all of the files that are required for the later analysis. At any stage the user may return to the main user interface in order to add more files for analysis.

The 'Analysis' panel (marked as C) contains the different functions available for analysis. At the moment, the tool contains five options for analysis: Data Table, Compare Samples, SNP View, Generate PgSNP, and Additional Information.

5.2 Analysis via Data Table

Data Table user interface enables analysis of the data uploaded by the user inside the actual tool in addition to fast export to Excel using Microsoft Office Spreadsheet ActiveX Control component. To visualize the data associated with the chromosome that was selected initially, press the number of the sample you would like to view as an Excel Worksheet and the 'Show File' button[2]. Table contents can be cleared by pressing on the 'Clear Sheet' button. The Data Table user interface is able to show a number of samples and chromosomes simultaneously as different sheets.

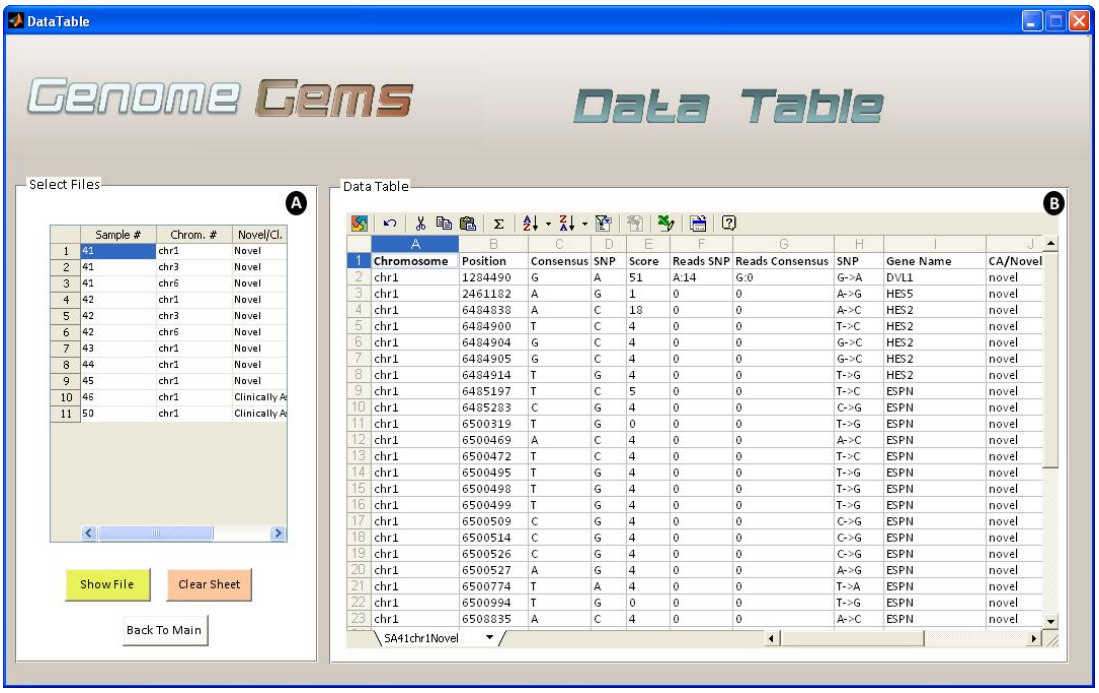

Figure 6 - The Data Table analysis interface enables the user to (A) select the files for viewing, one at a time and (B) view the data in a clear and familiar MS Spreadsheet environment, allowing easy export to Excel. Multiple files may be shown as separated sheets.

[2] **NOTE**: In case the table does not properly appear in the 'Data Table' window there may be a problem with the ActiveX Control installation. Please check whether you have installed the ActiveX MS Spreadsheet 10 correctly. If the problem persists please contact us at: GenomeGem@gmail.com

5.3 Comparing SNPs between samples

When searching for a disease causing mutation, multiple samples are sequenced from a population which is either related or is diagnosed with the specific disease. In case several samples are uploaded into *GenomeGems*, to the user may compare and calculate the frequency of appearance of each SNP in the different samples.

In order to compare between samples follow these steps:

1. Select the samples you would like to compare (appearing in the 'Select Files' panel at the 'SampComp' window) using the CTRL button for choosing multiple samples. **Make sure that all the samples selected contain the same chromosome number.**
2. Choose the threshold for comparison such that SNPs that appear in a number of samples that is lower than or equal to the threshold will be filtered.
3. Press the 'Show Bar Graph' button to view the frequencies of each SNP which surpass the threshold value selected, along with a corresponding table which serves as an index.

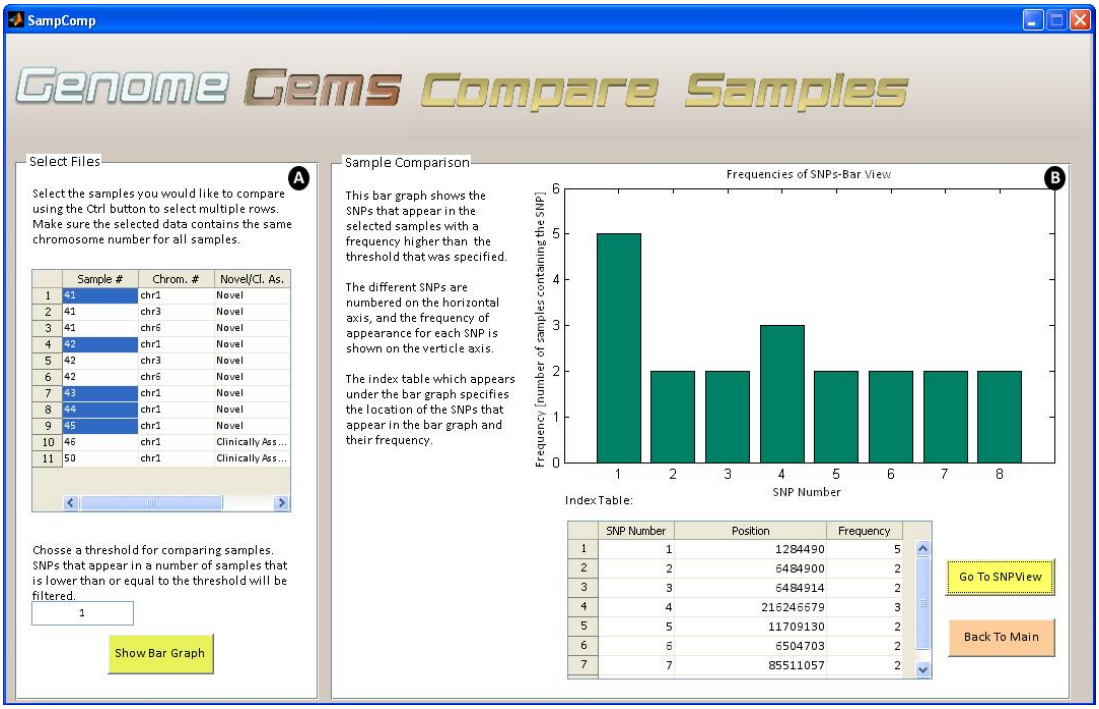

Figure 7 - The Compare Samples interface allows the user to (A) select files for comparison and choose a threshold for minimal SNP frequency and (B) view the results in a bar graph and a corresponding index table.

You can also visualize these results directly via SNPView analysis tool by pressing the 'Go To SNPView' button.

Upon selection of desired files for analysis, the 'SNP View' interface displays a table containing the sample numbers that include each SNP in the specific chromosome defined formerly. This data may be useful for further analysis, and can be easily exported to Microsoft Excel.

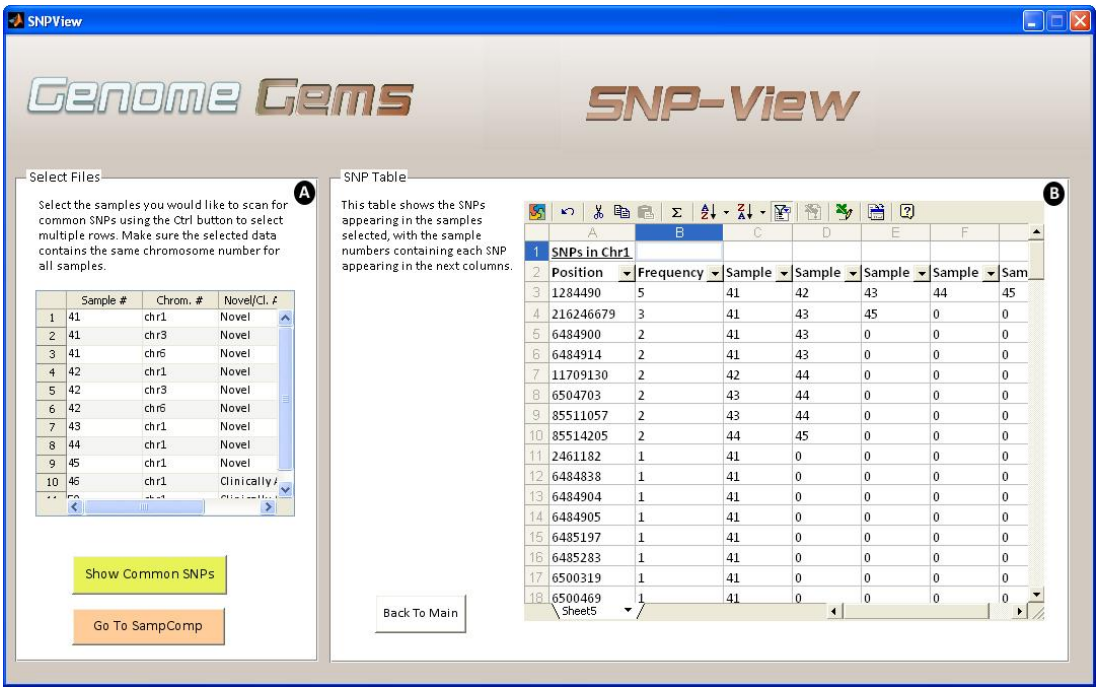

Figure 8 - The SNP-View interface allows the user to (A) select sample files for comparison containing the same chromosome number and (B) view a list of SNPs appearing in the selected samples, in the specified chromosome, with a list of the samples in which each SNP appears. The list may be easily exported to Excel for further analysis.

To get a comparative visualization via SNPView follow these steps:

1. Select the samples you would like to compare (appearing in the 'Select Files' panel at the 'SNPView' window) using the CTRL button to select multiple samples. **Make sure that all the samples selected contain the same chromosome number.**
2. Press the 'Show Common SNPs' button to get the 'SNP table' which shows all the SNPs appearing in the samples selected with the sample numbers containing each SNP[3].

[3] **NOTE**: In case the table does not properly appear in the SNPView' window there may be a problem with the ActiveX Control installation. Please check whether you have installed the ActiveX MS Spreadsheet 10 correctly. If the problem persists please contact us at: GenomeGem@gmail.com

5.4 Visualization of SNPs using UCSC Genome Browser

GenomeGems uses an algorithm for generating PgSNP format files from the original data files, which can then be uploaded as a custom track in the UCSC Genome Browser. To learn more about the PgSNP file format please see:
http://genome.ucsc.edu/FAQ/FAQformat.html#format10

A display of the SNPs uploaded by the user is consequently created and supplementary information supplied by UCSC can be viewed. The additional information supplied by UCSC is the context of the SNP – CDS or Intron, and the properties of the changed amino acid – polarity, acidity and hydropathy.

To get a sample data visualization via the UCSC using *GenomeGems* follow these steps:

1. In the main user interface choose the 'Generate PgSNP' analysis tool (a PgSNP new window will be opened)
2. From the 'Select File' panel at the 'PgSNP' window, choose the sample file you would like to convert to a PgSNP file format.
3. Press the' Browse' button and specify the location where you want to save the PgSNP file.
4. Press the 'Create PgSNP File' button to finally create the new file.
5. Follow the 5 simple steps appearing in the 'Upload Custom Track To UCSC' panel which will instruct you in uploading and visualizing (via UCSC Genome Browser) the PgSNP file you have just created using *GenomeGems*.

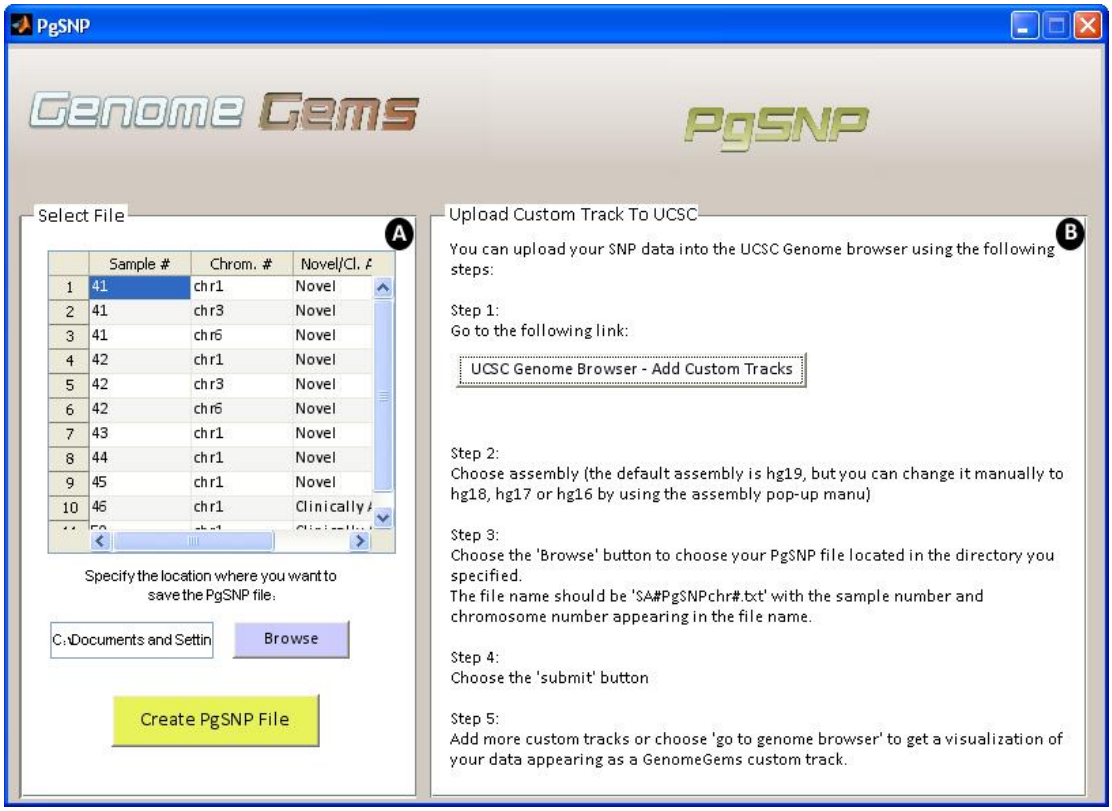

Figure 9 - The PgSNP interface allows the user to (A) choose a file for conversion to PgSNP format and specify the location where the file will be saved, and (B) instructs the user how to upload the file to UCSC as a Custom Track in five simple steps.

5.5 Using external links for additional useful information

For further investigation and annotation of specific SNPs and of the impacts of amino acid changes encoded by the mutant gene on a human protein, *GenomeGems* suggests additional external useful links:

- Polymorphism Phenotyping v2 (PolyPhen-2, URL: http://genetics.bwh.harvard.edu/pph2)
- Server of the Identification of Functional Regions in Proteins (ConSurf Server, URL: http://consurf.tau.ac.il/index_nucleotides.php)
- Prediction of Transmembrane Regions and Orientation (TMpred, URL: http://www.ch.embnet.org/software/TMPRED_form.html)
- Online Mendalian Inheritance in Man (OMIM, URL: http://www.ncbi.nlm.nih.gov/omim)
- University of California Santa Cruz (UCSC, URL: http://genome.cse.ucsc.edu)

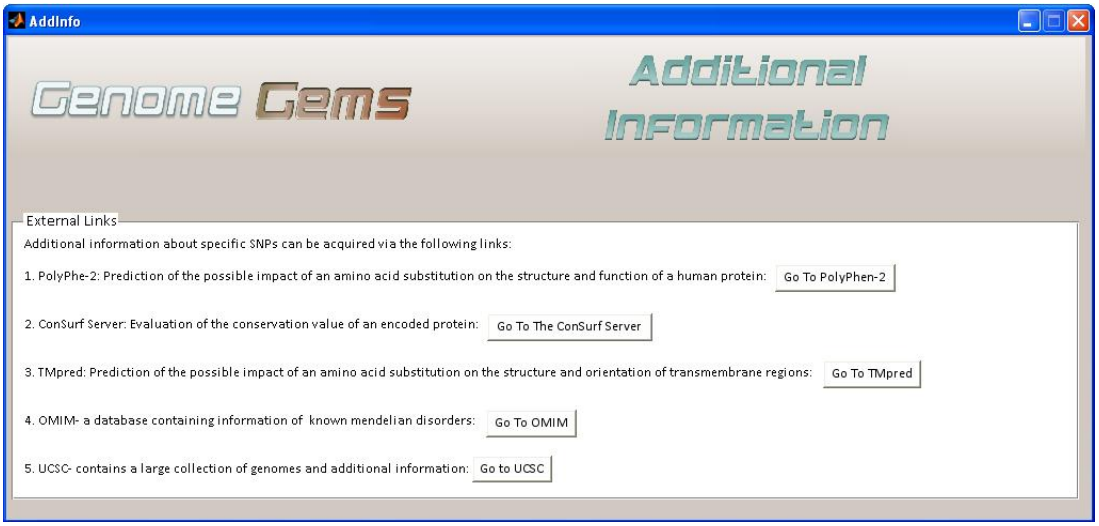

Figure 10 – The Additional Information interface enables quick transfer to suggested additional databases for further analysis of SNPs.

6. Summary

GenomeGems enables researchers to identify variances in genetic sequences which are potentially disease-causing in an efficient manner.

GenomeGems' main advantages are its:

1. Applicability for any Deep Sequencing data (given the correct input file generated)
2. Ability to run on a standard Personal Computer with a Windows Operating System
3. Integration with the UCSC Genome Browser and Microsoft Excel
4. Possible comparison and analysis of a large number of samples

GenomeGems' main virtues allow:

1. Reducing variability in selecting which mutations should be further investigated
2. Facilitating genomic research via clear and accessible presentation of processed Deep Sequencing data
3. Assisting rapid turnover of information and a quick lead to further experimental mutation detection

In addition to the currently implemented features of *GenomeGems*, development of additional tools for further analysis is underway.

7. A Quick Guide to GenomeGems:

Installing

Go to the *GenomeGems* website (http://www.tau.ac.il/~nshomron/GenomeGems) and follow these steps:

- Press the 'System Requirements' link on the left panel of the website home page.
- Install MCR 7.10
- Install ActiveX—MS Spreadsheet 10
- Download and run the *GenomeGems* application

Uploading Data:

In the main user interface upload the input files via the 'Upload Files' left panel, select a chromosome on which the analysis will be performed, specify the sample number and specify whether the data is of novel or clinically associated SNPs. Then, press the 'Add Data' button, and you will see the data file you have just added in the 'Select Files' table.

Analyzing Data:

In the 'Analysis' right panel at the main user interface, choose the form of analysis required:

- **Data table**- Choose the sample you would like to view as an Excel Worksheet and the 'Show File' button. You can erase the table contents easily by pressing on the 'Clear Sheet' button.

- **Compare Samples**- Select the samples you would like to compare (appearing in the 'Select Files' panel at the 'SampComp' window) using the CTRL button for choosing multiple samples. **Make sure that all the samples selected contain the same chromosome number.** Choose the threshold for comparison and press the 'Show Bar Graph' button to view the frequencies of each SNP which surpass the threshold value selected, along with a corresponding table which serves as an index.

- **SNPView**- Select the samples you would like to compare (appearing in the 'Select Files' panel at the 'SNPView' window) using the CTRL button to select multiple samples . **Make sure that all the samples selected contain the same chromosome number.** Press the 'Show Common SNPs' button to get the 'SNP table' which shows all the SNPs appearing in the selected samples with the sample numbers containing each SNP.

- **Generate PgSNP**- From the 'Select File' panel at the 'PgSNP' window, choose the sample file you would like to convert to a PgSNP file format. Press the' Browse' button and specify the location where you want to save the PgSNP file. Press the 'Create PgSNP File' button to finally create the new file. Follow the 5 simple steps appearing in the 'Upload Custom Track To UCSC' panel which will instruct you in uploading and visualizing (via UCSC Genome Browser) the PgSNP file you have just created using *GenomeGems*.

  Additional Information- *GenomeGems* suggests additional external useful links: PolyPhen-2, ConSurf Server, TMpred, OMIM and UCSC.

Schematic Summary: