SNPsetter

User guide

Table of Contents

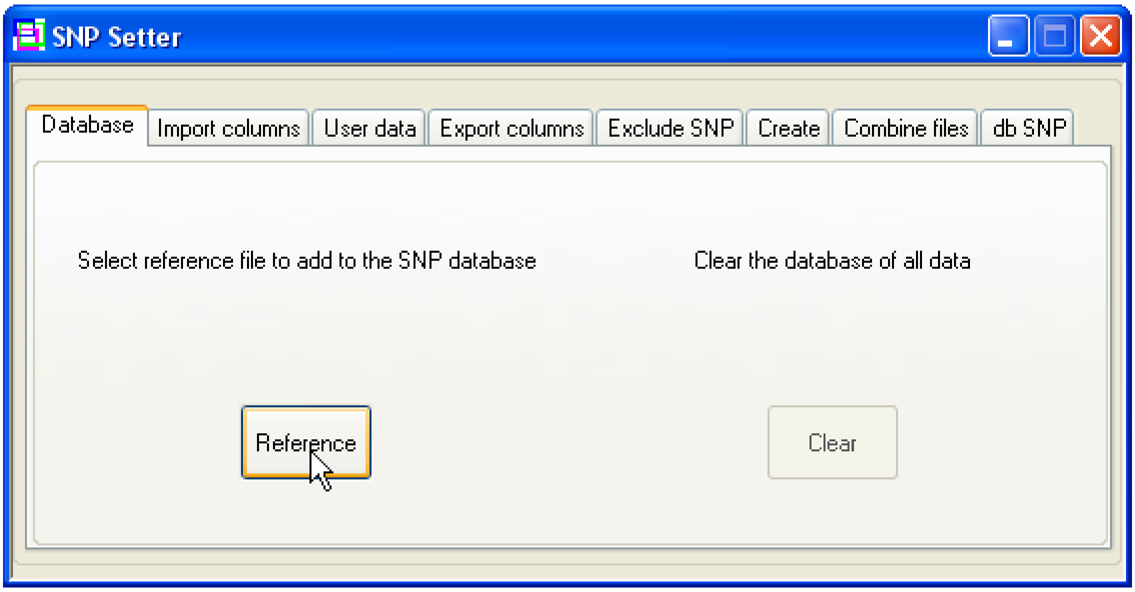
1. Description

SNPsetter is a utility that will process a set of Affymetrix genotype files in such a way as to ensure that all files have a common SNP dataset, linked to standardized annotation data (thus ensuring, for example, that map positions are uniform). The standardized files can be generated from input files in either the usual Affymetrix ***.xls*** format or the newer BRLMM text format.

2. Creating the reference database

The SNPsetter interface (Figure 1) consists of a set of eight tabs, each of which corresponds to a particular set of processes in the standardization pipeline. The Database tab initially provides access to a single function, since the first requirement will be to select a reference SNP data file.

Pressing the Reference button allows the user to select a file containing SNP data which will be subsequently be used as the reference file. This can be either an Affymetrix data file (***.xls***), an Affymetrix annotation file (***.csv***) or a SNP position file from the NCBI dbSNP website (see Creating SNP position files from NCBI dbSNP below). [The time taken by this importing process varies considerably according to the nature of the SNP data set and computer hardware speed; *e.g.* 10k SNP data files take <2 s on a 3 GHz computer with 1GB of RAM, compared to 1 min for a 250k annotation file.] Finally, the Clear button removes all data from the database.



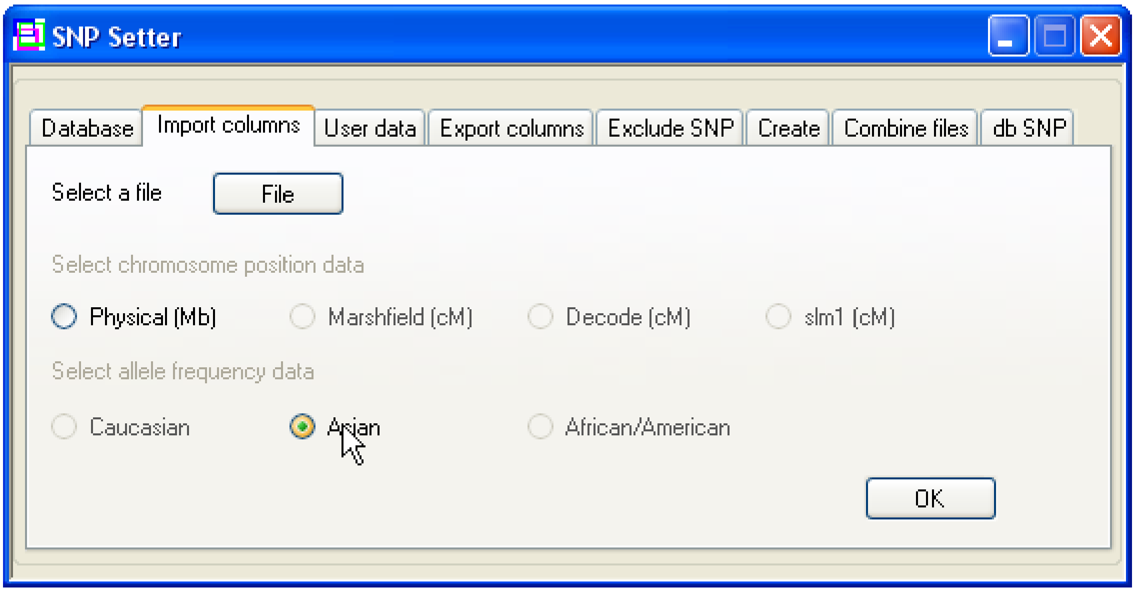
Select reference file to add to the SNP database

Clear the database of all data

Figure 1: Initial SNPsetter screen

3. Importing data from other Affymetrix files

If the selected reference file does not contain all the data columns requiredin the final set of files, it is possible to add or update them using the Import columns tab (Figure 2). Its File button allows a single Affymetrix *.xls* text data file to beselected; this file is then read, and the radio buttons corresponding to the available data columns areenabled. (In the example shown in Figure 2, the file only contained data columns for ***Asian*** allele frequencies and ***Physical*** position.)



Select a file

Select chromosome position data

○ Physical (Mb) ○ Marshfield (cM) ○ Decode (cM) ○ slm1 (cM)

Select allele frequency data

○ Caucasian ● Asian ○ African/American

Figure 2: Importing data columns

A single required data column can be selected, and entered by clicking OK. [***N.B.*** If this selected data column was not present at all in the original reference file, the new data will simply be added to the database; otherwise, that column of the original reference file will be completely discarded and the new data entered instead. For any SNP that is represented in the database but not in the imported dataset, the value of the imported column field is set to zero. This procedure can be repeated to import or update additional columns as desired.]

4. Importing user-defined data

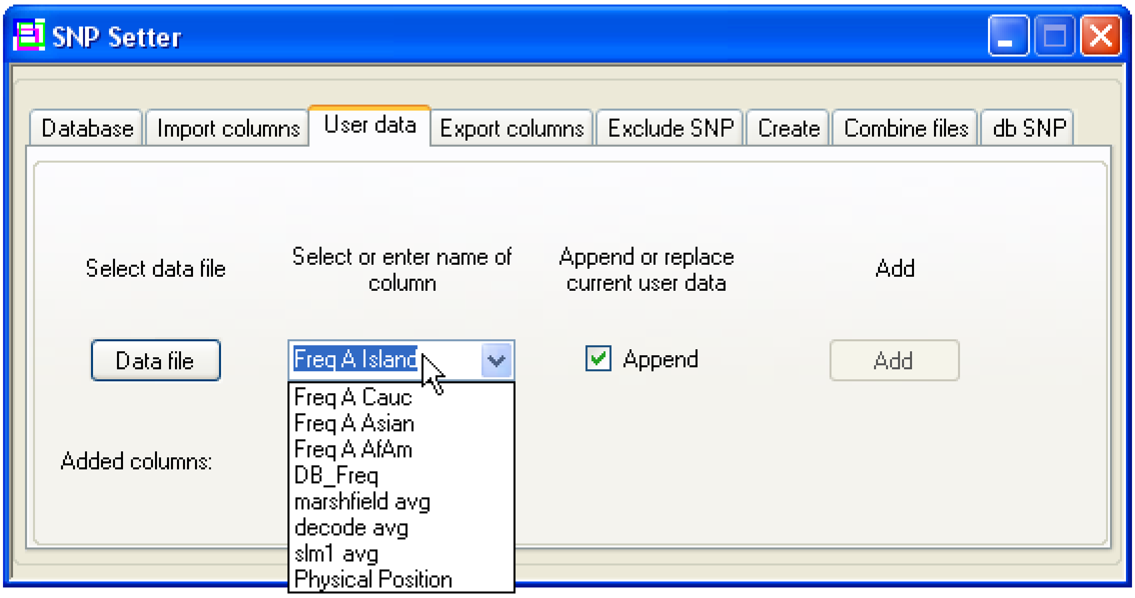


Figure 3: Adding user-defined data

The User data tab (Figure 3) allows user-defined data columns to be added to the files. The additional data must be in a plain text file (**.txt*), formatted as shown in Table 1.

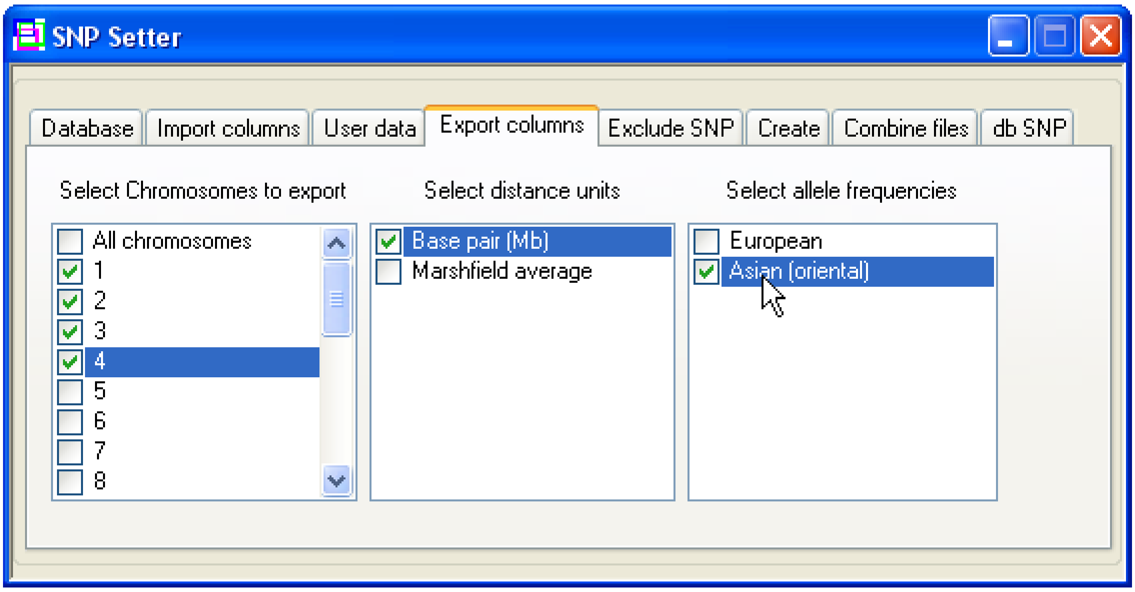
SNP name	Data
rs713055	0.44
rs949459	0.64
rs713298	0.53
rs713419	0.57
rs1071932	0.6
rs997173	0.61
rs997238	0.67
rs997897	0.8
rs713968	0.72
rs1072378	0.84

Table1: Format for user defined data files; each line consists of a SNP name and the data value, separated by a tab character.

No title row is needed, and the SNP name (*e.g.* rs713055) and its data value should be separated by a tab character. The new data can either be added as an extra column or used to replace the data in an Affymetrix-defined column. The title of the column can be selected from the drop down list (Affymetrix predefined columns) or entered by typing in the list box (new column). In Figure 3, the data would thus be added as a new column named "Freq A Island" and appended to the new files. (If the Append checkbox is not ticked, any user-defined data column that was previously added during the current session would be discarded.) The new data are loaded when the Add button is clicked.

5. Selecting data columns to export

Once all the desired data columns have been loaded into the programme, the user must select the types of information that it is desired to export; this is done via the Export columns tab.



Select Chromosomes to export
- □ All chromosomes
- ☑ 1
- ☑ 2
- ☑ 3
- ☑ 4
- □ 5
- □ 6
- □ 7
- □ 8

Select distance units
- ☑ Base pair (Mb)
- □ Marshfield average

Select allele frequencies
- □ European
- ☑ Asian (oriental)

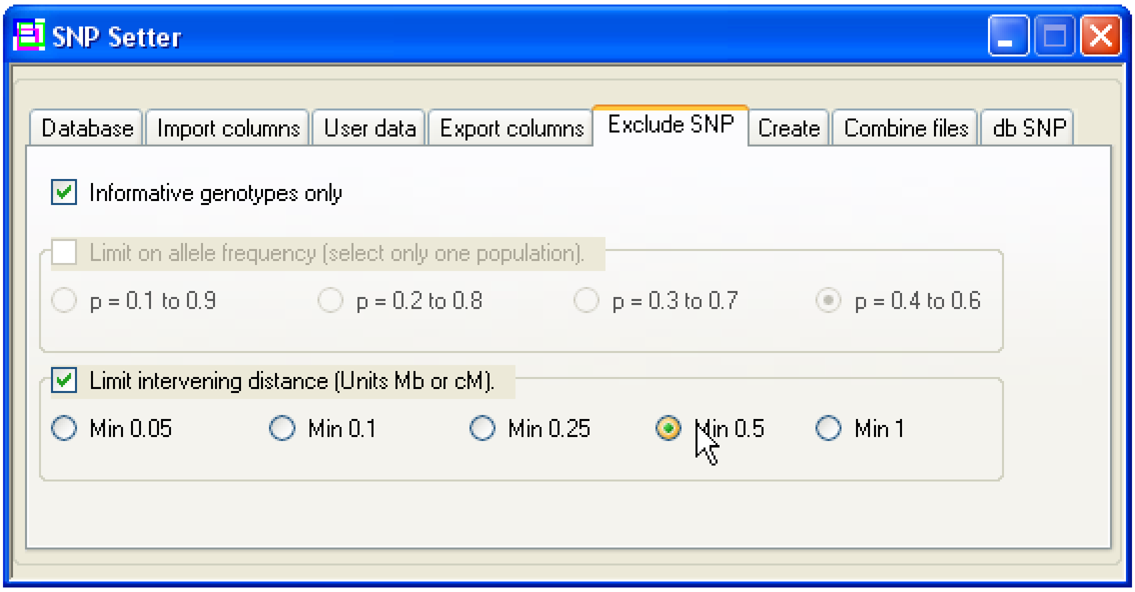
Figure 4: Selecting data for export

This page allows selection of subsets of chromosomes for export, and a choice from the data columns that are represented in the database. (In Figure 4, only SNPs on chromosomes 1 to 4 will be exported, and only the physical ***Mb*** position, and ***Asian*** frequencies data columns will be exported. The absence of the ***slm1*** and ***Decode*** checkboxes from the distance units list indicates that they were not in the original reference file and have also not been added using the Import columns or User data tabs.)

6. Rule-based SNP exclusion

Using the Exclude SNP tab (Figure 5), subsets of SNPs can be selected for exclusion, based on their genotype, allele frequency or distance from the last SNP included in the file.

Also, if it is intended that SNP data from multiple individuals will be loaded in parallel, the Informative genotypes only checkbox can be used. When this is ticked, SNPs that are homozygous for the same genotype in all the entered data files will be excluded from the final standardized files. Clearly, since this requires SNPsetter to know all the SNP genotypes, this function cannot be used if it is intended to read and process the input data files sequentially (see Creating standardized files).



☑ Informative genotypes only

□ Limit on allele frequency (select only one population).

○ p = 0.1 to 0.9 ○ p = 0.2 to 0.8 ○ p = 0.3 to 0.7 ● p = 0.4 to 0.6

☑ Limit intervening distance (Units Mb or cM).

○ Min 0.05 ○ Min 0.1 ○ Min 0.25 ● Min 0.5 ○ Min 1

Figure 5: Filtering exported data

The inclusion of SNPs can also be filtered on the basis of position (or allele frequency), but only if a single type of positional data (or allele frequency) is first selected on the Export columns tab. For example, in the Selecting data columns to export section above, since two allele frequency sets were selected, it will not now be possible to exclude SNPs by allele frequency; contradictory sets of markers would be excluded by each of the two allele frequency sets. Contrariwise, since only "Base pair (Mb)" was chosen from the map position data options, excluding SNPs by position will be allowed. The Limit intervening distance option allows sparser, more evenly distributed subsets of SNPs to be exported, which may be desirable for some downstream linkage analysis purposes.

As the output files are created, SNPsetter checks for the "informative status" of each SNP. If the SNP is informative (or Informative genotypes only has not been selected) the programme then checks that SNP's allele frequency. If the frequency is within the selected limits, its position from the last SNP added to the file is calculated; if this exceeds the threshhold selected under Limit intervening distance, the SNP will be included. (The units used for the last comparison will be those previously chosen for export, and so could be either Mb or cM.)

7. Creating standardized files

The final step uses the Create tab to load the test data files and generate the standardized output files.

Selecting either the xls files (Parallel) or BRLMM files (Parallel) button prompts the user to choose a folder containing multiple test data files. (Files may also be selected from a number of different folders by repeated use of these buttons.) These files are then all read and their SNP genotype data are stored in the internal database. [***N.B.*** with a large number of files, or when using the 250k SNP datasets, this option may require a large amount of computer RAM. If the memory is insufficient, this batch-loading of data will take an extended period of time (>5 minutes for a set of thirteen SNP files). In such situations, it may be preferable to analyse the data one file at a time, using the other buttons on the Create tab (see below). However, if it is desired to use the option of excluding uninformative SNPs (see Rule-based SNP exclusion), batch-loading of the input files using the (Parallel) buttons is required.]

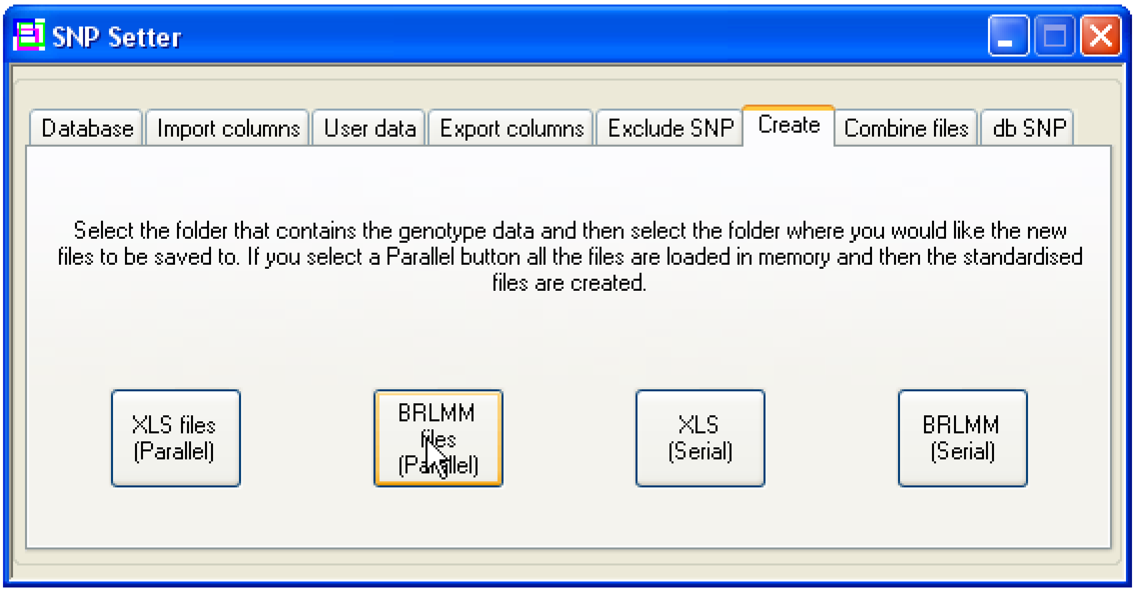

Select the folder that contains the genotype data and then select the folder where you would like the new files to be saved to. If you select a Parallel button all the files are loaded in memory and then the standardised files are created.

Figure 6: Creating final data files

If, on the other hand, the user wishes to create the standardized files without loading all the data from multiple input files, a folder can be selected for sequential reading, by clicking the xls (Serial) button (for Affymetrix ***.xls** tab-delimited text files) or BRLMM (Serial) (for BRLMM files).

In all cases, the output files will all be formatted as Affymetrix files, and since this step requires intensive reading and writing to the computer hard disk, it can be slow.

8. Supplementary functions

8.1. Combining files

Two different Affymetrix files (*e.g.* a 50k *Hin*dIII and a 50k *Xba*I file) can be combined using the Combine files tab (Figure 7).

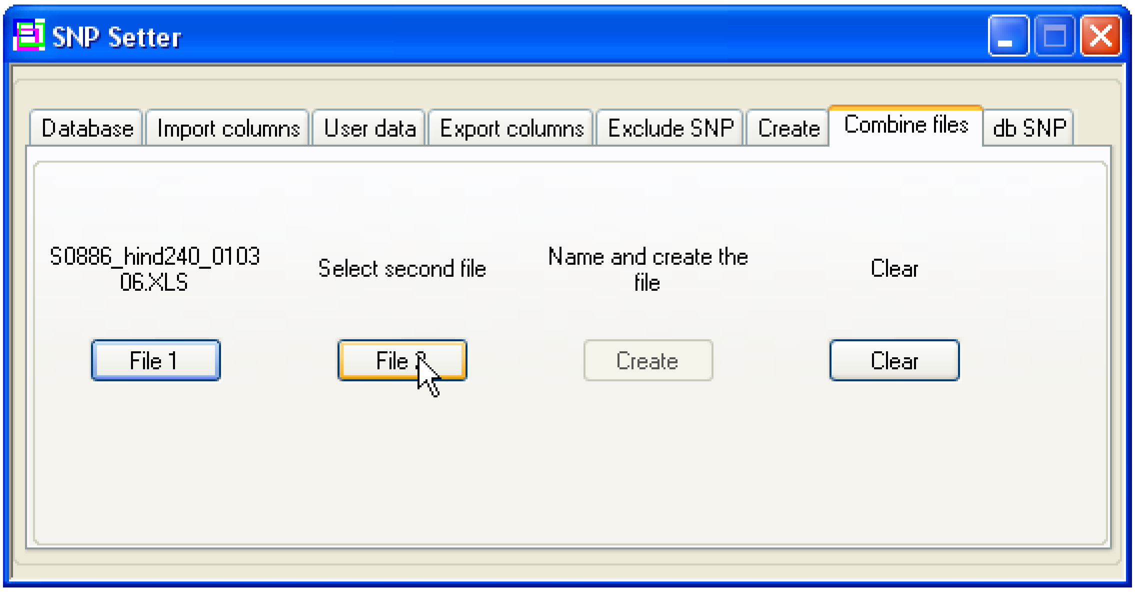


Figure 7: Combining two SNP files

This allows the resultant file to be analysed as a single set of data by other software applications. One application of this would be to combine two different types of file into a single reference file to be used by SNPsetter. This would be useful, for example, if the patient data consisted of two mutually exclusive sets of different file types. By combining one file of each type into a reference file, the internal database will include all the SNPs from both sets, and would be able to accept patient files of either type. When each output file is then generated, SNPs that were not in the original input file are included, with a genotype of “NoCall”. These output files can then be used as input for other programmes (*e.g.* IBDfinder or AutoSNPa), allowing all the patients to be processed as one group.

However, whenever such file-combining operations are carried out, it must be verified that both types of input file have SNP annotation data derived from the same genome freeze. If this is not the case, the resultant combined set of SNPs could be incorrectly ordered.Also, for this appproach to be useful, the downstream analysis programme must be able to accept datasets that have a large number of “NoCall” SNPs.

When combining two files of SNP data from the same patient, some SNPs may be in common between both data sets. In this case, the following rules are applied in order to resolve any genotyping or annotation conflicts between the two data sets:

- If the genotyping data conflict, the genotype for that SNP is set to “NoCall”.
- If the SNP annotation differs between files, any annotation present in the first file is used. (If the first file does not include an annotation for that SNP, the field is set to “0”, irrespective of any annotation in the second file.

The two files to be combined are selected using the File 1 and File 2 buttons. Once this is done, the Create button will prompt for the name of the output file. To clear the input files and create a new combined file, the Clear button is used.

8.2. Creating SNP position files from the NCBI dbSNP web page

An up-to-date set of SNP position data can be retrieved from the NCBI web site. To do this, you will need a web browser open at http://www.ncbi.nlm.nih.gov/projects/SNP/dbSNP.cgi?list=rslist. Start SNPsetter and go to the dbSNP tab (Figure 8). Select the Create button and choose an Affymetrix data file that contains a column of SNP ***rs*** names. A text file will be generated in the same folder, containing a list of all the ***rs*** names in the data file.

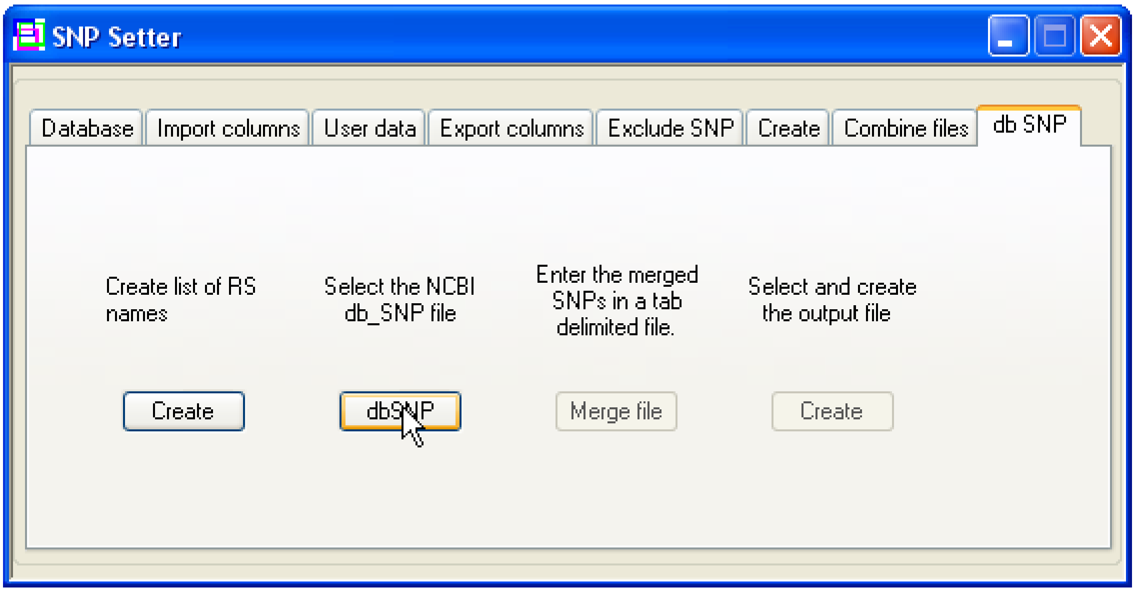


Figure 8: Retrieving position data from dbSNP

Open this file, select and copy all the text, and paste it into the text box on the NCBI web page. Enter an email address, select the output format as “CHROMOSOME RPT” and press “Submit” to send the web page. The NCBI server will then send an email containing a link to the results file, which has been compressed with ***gzip***. (The email also contains a link to http://www.gzip.org/ which contains information on a number of utilities which can decompress the ***.gz*** file.) Once the NCBI results file has been downloaded and decompressed,return to SNPsetter and select it using the dbSNP button. The new reference file can then be generated by by clicking Create. This new output file contains the up-to-date positions of the SNPs and can be used as a reference file or for importing physical position data (see Importing user-defined data).