Extracting Mouse Genetic Project raw phenotyping data

Introduction:

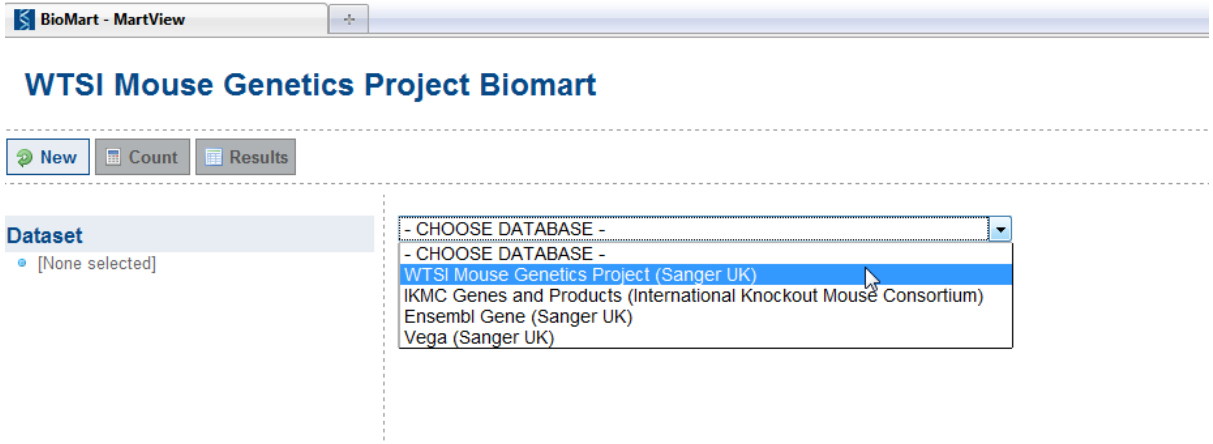
Raw phenotyping data is available through a publicly available BioMart[1] interface. This can be accessed at the following web address http://www.sanger.ac.uk/htgt/biomart. Mouse Genetics Project data which has passed all necessary internal quality control checks is automatically published to this BioMart within 24 hours. As of December 2011 this data set comprises approximately 9 x 10^6 data points and hence it is not recommended that users attempt to extract all data in one bolus. Rather, the query made to the BioMart should be carefully phrased to extract only the data pertinent to answering your specific scientific question. Below is a step-wise example of how to extract the data using the standard BioMart user interface. Further information on using a BioMart can be found at: http://bowtie-bio.sourceforge.net/recount/biomaRt.pdf.

Demonstration Example:

The following example, demonstrates the extraction of data associated with the Mysm1 knockout line.

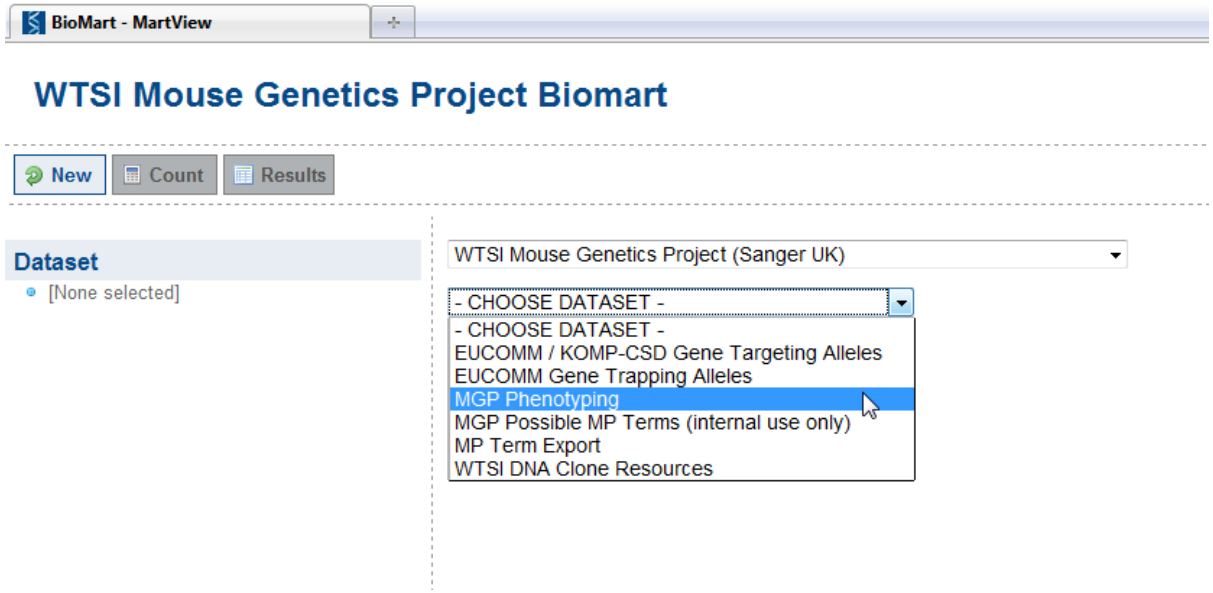
1. Open the BioMart by following the link: http://www.sanger.ac.uk/htgt/biomart.

2. Select "WTSI Mouse Genetics Project (Sanger UK) as the database (as shown below).

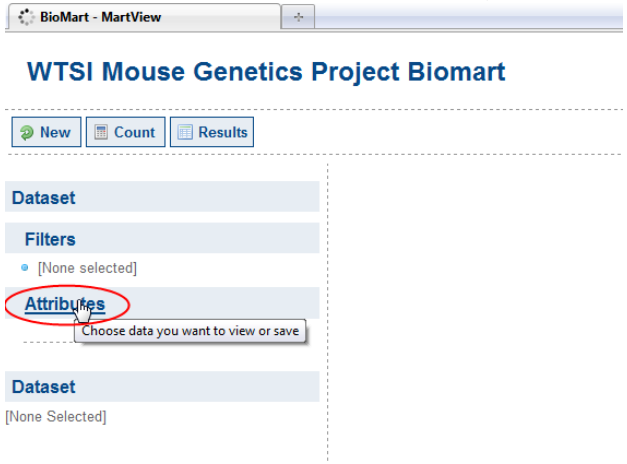


[1] BioMart is a generic data management system [Smedley *et al* "BioMart – biological queries made easy" (2009) BMC Genomics 10 (22)].

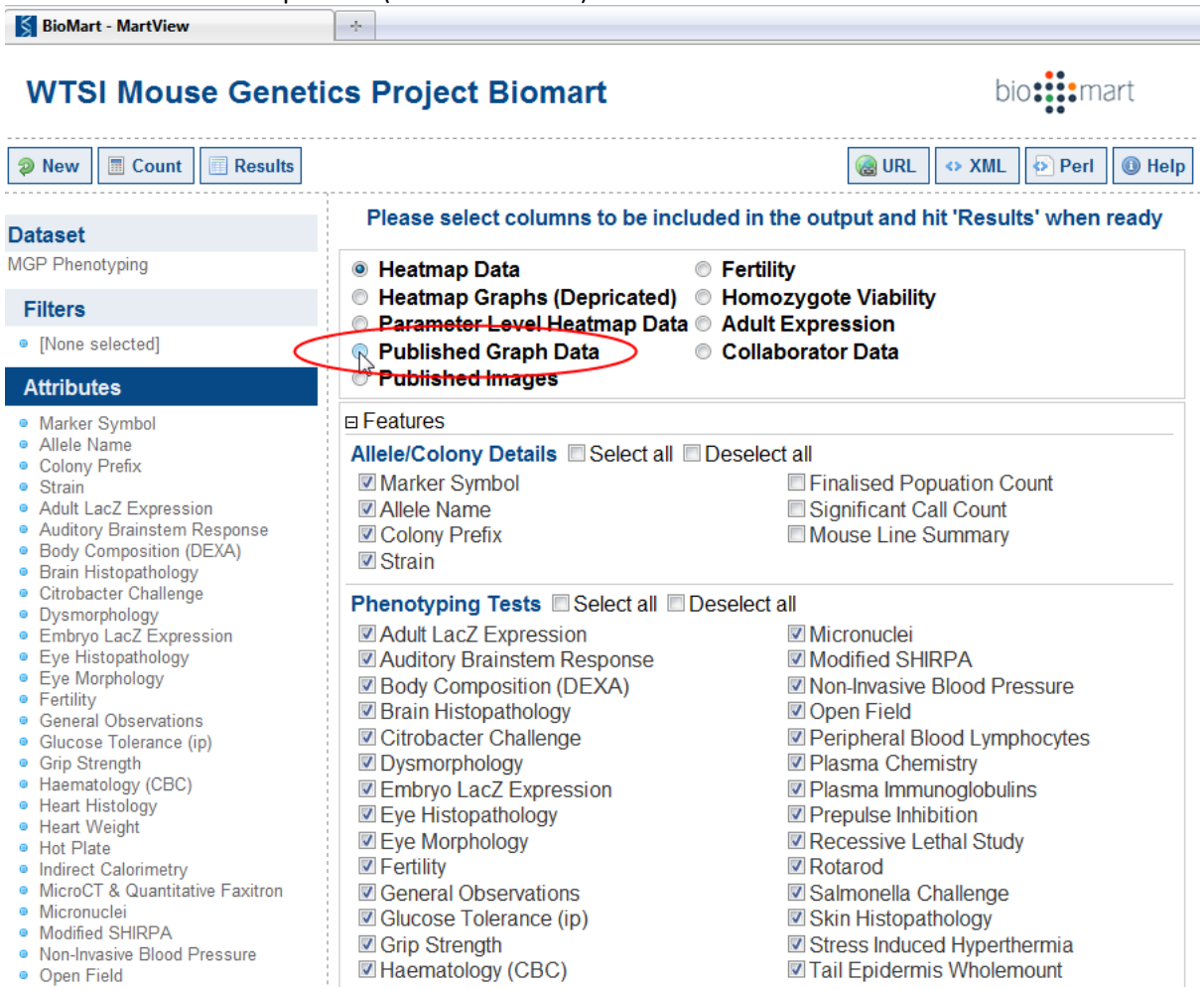
3. Select "MGP phenotyping" as the dataset (as shown below).



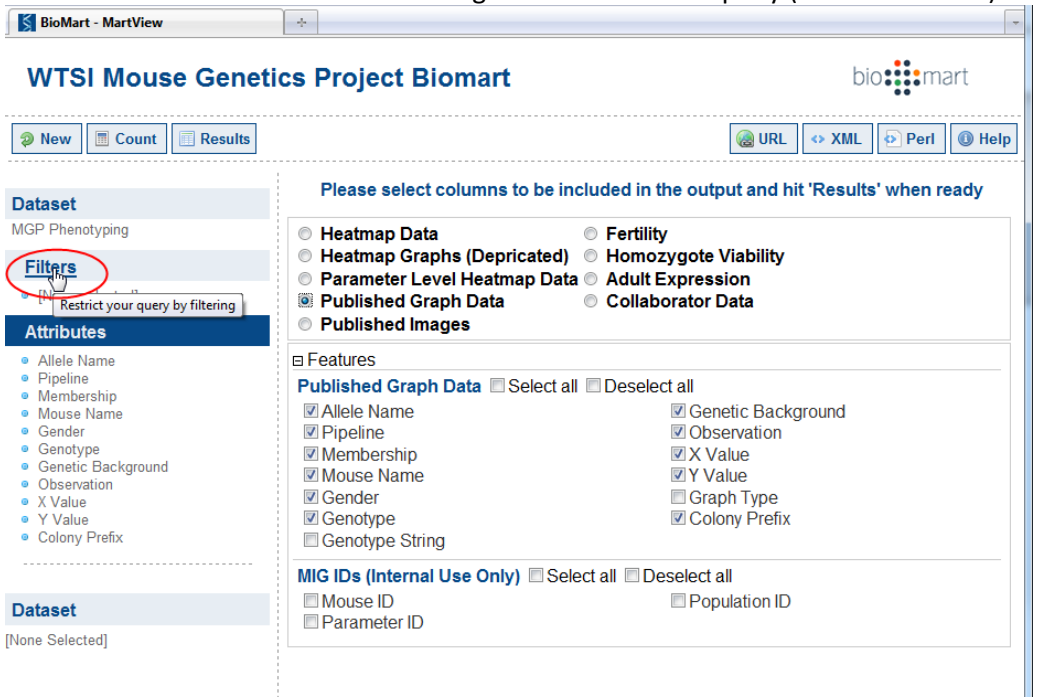
4. Click on the Attribute link on the left (as shown below).



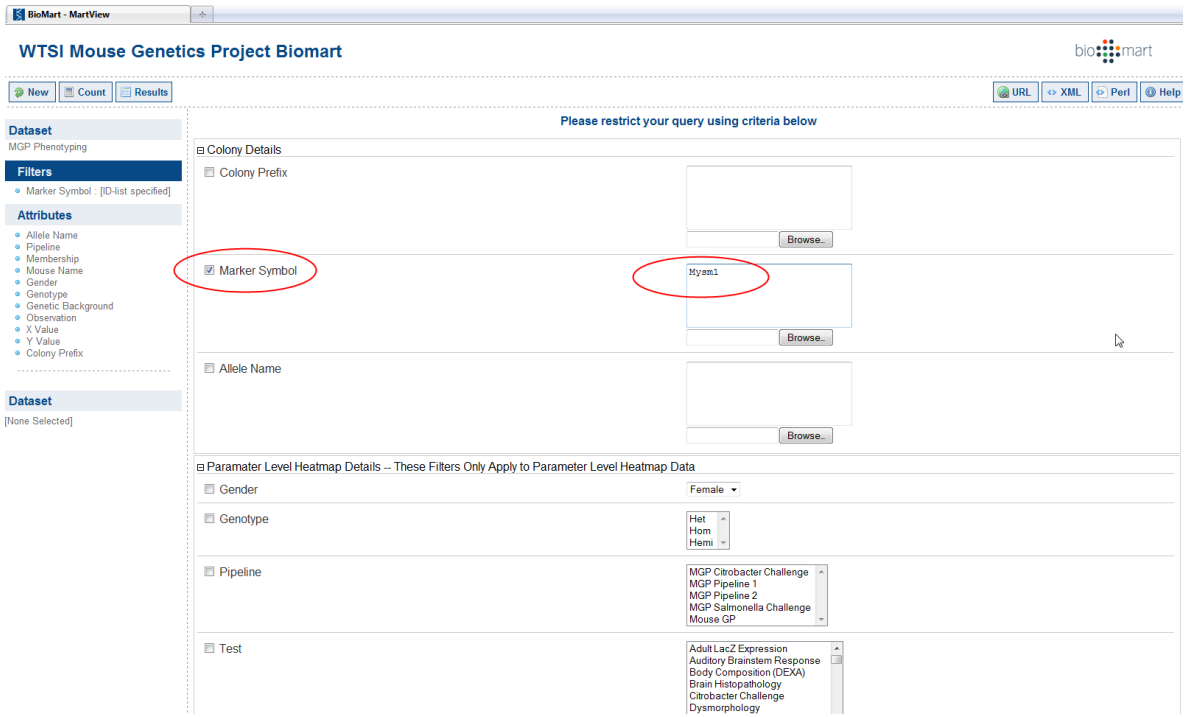
5. Select Published Graph Data (as shown below).



6. Now click on Filters to access a dialog box to restrict the query (as shown below).

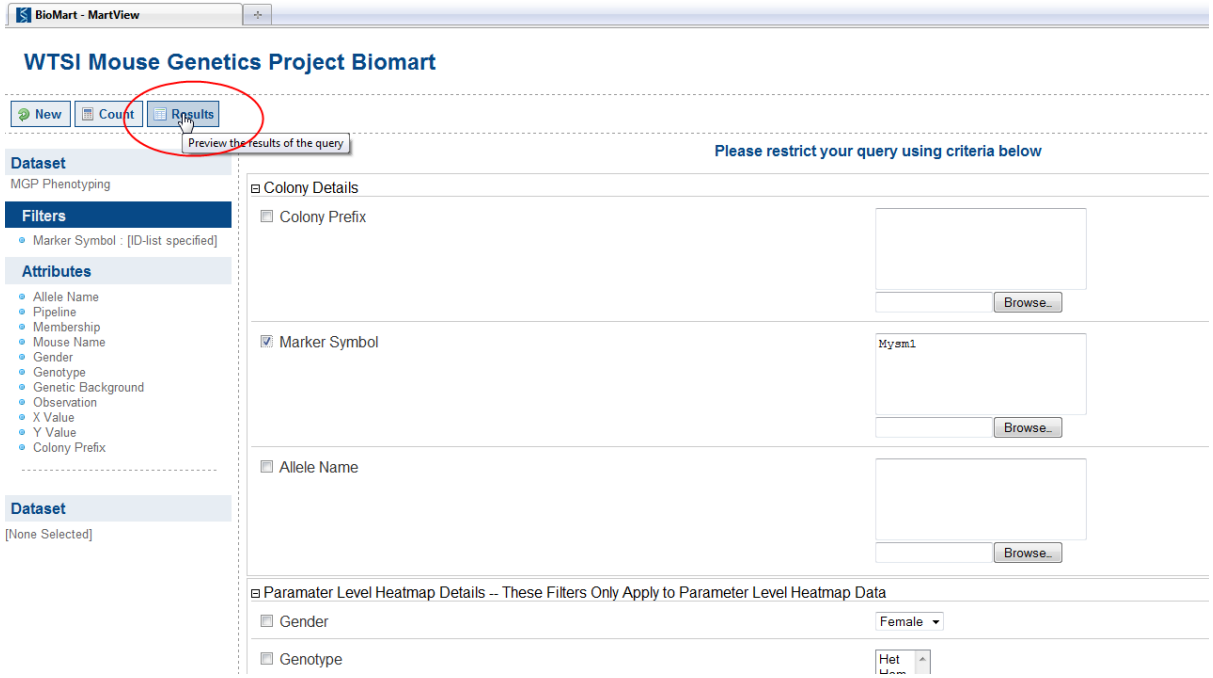


7. We can filter the data by a gene name e.g. Mysm1 (a.k.a. Marker symbol) (as shown below).

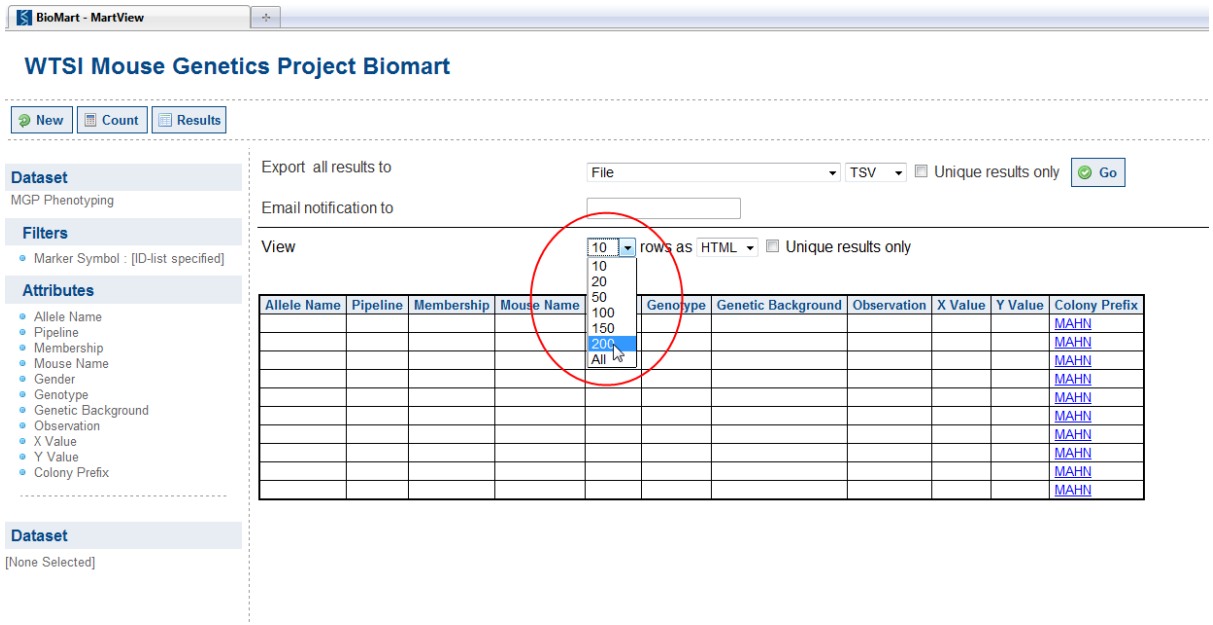


8. Click on the Results button (as shown below) to filter as specified.

Note if the program is running for a long time and no results appear, it is likely you have asked for too big a slice of the data in one query. Please close and reconsidered your filters. A poorly formulated query can cause so much data to be sent to your computer that the web browser (and potentially your computer) will become overwhelmed and stop responding. It should also be noted that the volume of data available through this interface is roughly two orders of magnitude greater than supported by Excel. As such careful pre-planning of queries is advised.

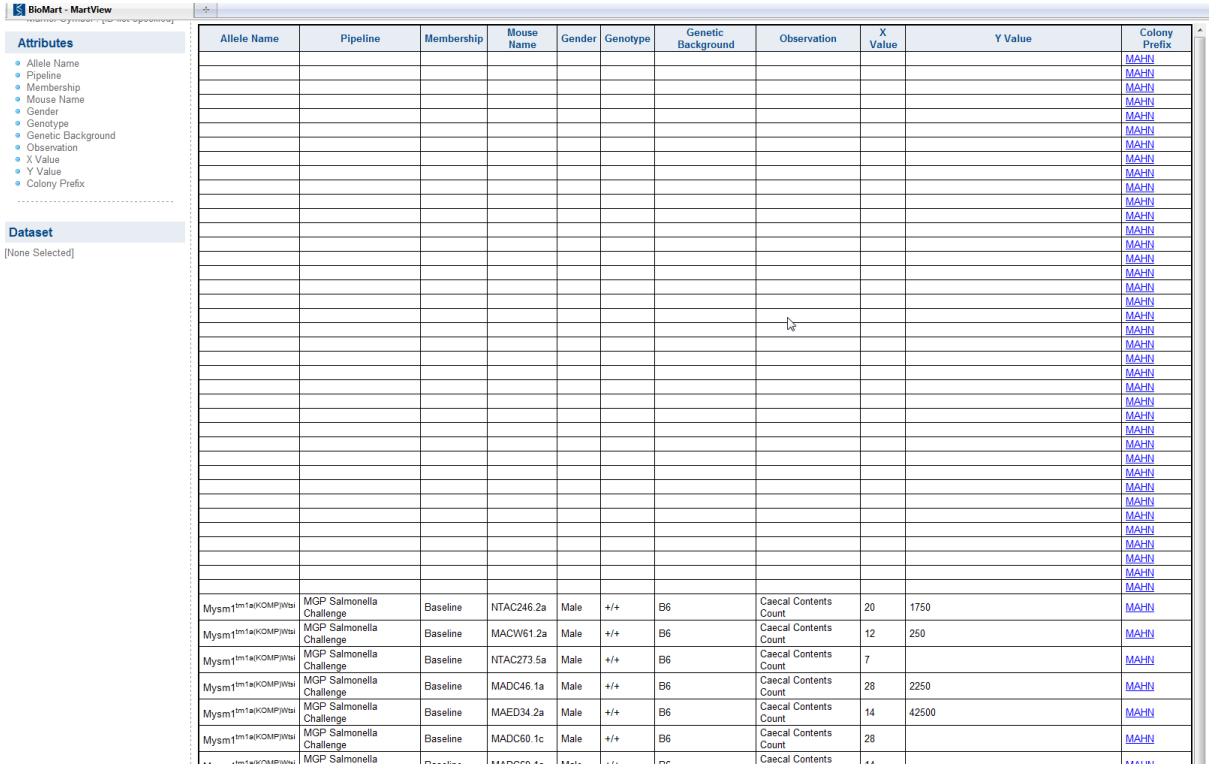


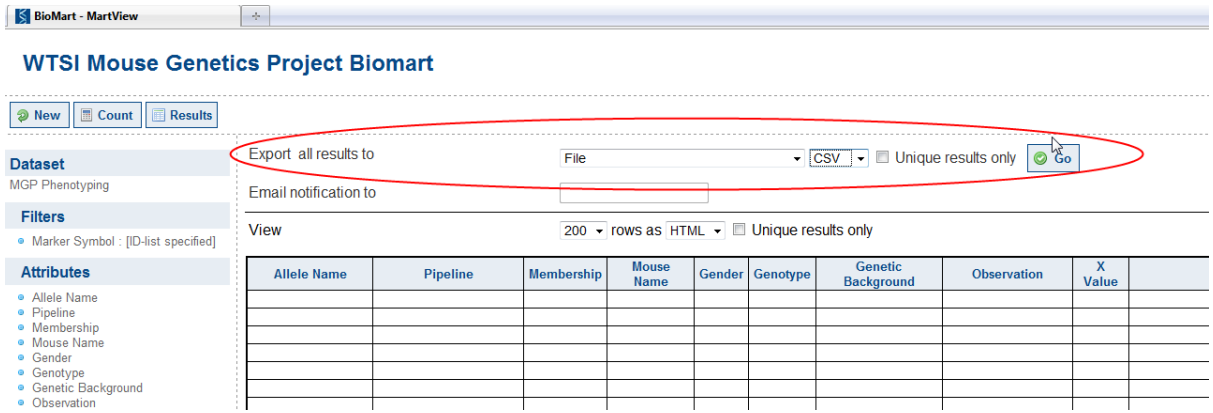
9. Edit the number of rows shown to view the data (as shown below).



10. Scroll down the dataset. Empty rows can appear at the top if entries in the database have the correct colony prefix but no information to populate the remaining columns of the table. This is a limitation of the BioMart system and, while inelegant, does not present a challenge to the analysis of the data.

You may notice large numbers of digits appearing after a decimal place in the values displayed in X or Y value columns. These are a result of displaying the numerical values at the maximum available computer precision. NB. Software, such as Excel, also internally store values in this format but truncate them before display. To allow for the significant differences in the numerical precision of different assays we have not applied truncation. Users requiring improved formatting should extract the data (see step 11) and use the number formatting capabilities within, for example, excel to choose their desired reporting precision and format.



Allele Name	Pipeline	Membership	Mouse Name	Gender	Genotype	Genetic Background	Observation	X Value	Y Value	Colony Prefix
										MAHN
										MAHN
										MAHN

11. Results can be exported using the export facility (as shown below).