REEF User Manual

Overview

REEF is aimed at identifying regions of a genome enriched in specific features, as compared with a reference landscape of features density. It takes as input a list of reference features (RF, e.g human genes) mapped on a genome sequence, a list of selected features (SF) among the RF (e.g. human genes specifically expressed in a given tissue) with their genomic positions and the number and the length of the chromosomes in the genome under consideration. It scans the genome using a sliding window approach, and calculates the statistical significance of each windows using the Hypergeometric Distribution and the False Discovery Rate (FDR). Consecutive significant windows form a cluster of regional enriched features. Results can be viewed as plots or dumped to text file for further analysis. The program also allows the user to display the results using the Custom Annotation Tracks facility from the UCSC Genome Browser.

Installation

Binary packages is avaliable for Microsoft Windows, if you choose to install from this package no other software installation is needed. Windows users can download the executable installer, run it and follow the instructions.

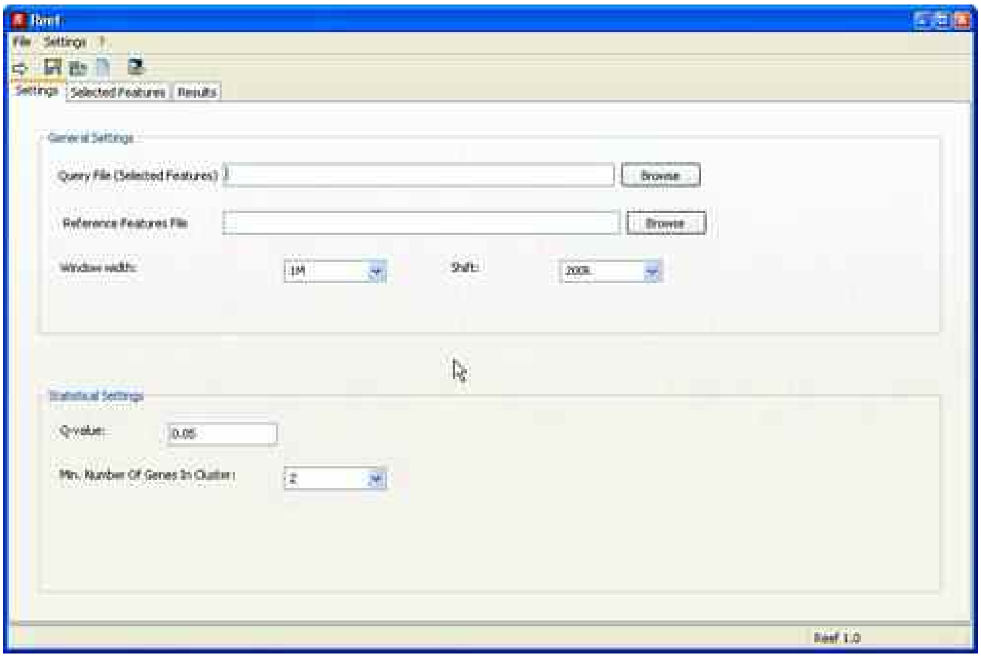
The python *source code* is multiplatform, can be run from different operating systems and it requires the following aditional packages to be installed:

- **The python interpreter** (version 2.3 or higher), download it from www.python.org
- **The wxPython gui toolkit**, you can get it form the following web site: www.wxpython.org
- **The SciPy package** tath can be downloaded from http://www.scipy.org/

Once you have installed all the dependencies and downloaded the REEF source code from the download page, extract the directory from the archive an from that directory run the main script called reef.py

Settings

The REEF settings window allows the user to choose the appropriate parameter for the analysis.



Query File (Selected Features)

Reference Features File

Window width: 1M

Shift: 200k

Q-value: 0.05

Min. Number Of Genes In Cluster: 2

- The *Query File* is the file containing the selected features (e.g. tissue specific genes). It must be a tab separated values text file containing at least four colums: the feature ID and the chromosomal localization in which the feature is found (chromosome, start and end cohordinates, e.g. NM_002291 chr7 107351498 107431040). Optional columns may contain additional description of the feature. A sample query file can be downloaded from REEF web site.
- The *Reference Features File* is the file containing the reference features. It must be a tab separated values file containing four colums: the feature ID and the chromosomal localization in which the feature is found

(chromosome, start and end cohordinates, e.g. NM_002291 chr7 107351498 107431040) *The directory of the reference features file must contain a file with the same name of the reference file but with the .chr extension containing in the first line the organism and in the following lines the name of the chromosomes and their length*; the chomosome length is used for drawing the results and the organism field is used to select the appropriate organism in the UCSC Genome Browser custom annotation view. If the organism is not present in the UCSC Genome Browser database, the visualization of REEF results in the UCSC Genome Browser is not possible. Some allowed organism names are: Human, Chimp, Dog and Cow. See genome.ucsc.edu/cgi-bin/hgGateway to check the avaliable organisms at in the UCSC Genome Browser database. A sample reference file and a sample chromosomes length file for the human genome can be downloaded from REEF web site.

- The *Window width* parameter changes the dimension of the window used to scan the genome by the sliding window approach.
- The *Shift* parameter changes the distance between the starts of adjacent windows in the sliding window algorithm. Obviously the bigger the shift, the less the number of windows considered in the analysis (N). The N parameter influences the FDR calculation: the higher N the most stringent the statistical threshold on the single window.
- The *Q-value threshold* determines the global threshol for significance. Let *S* be the total number of *SF* over the entire genome, *R* the total number of *RF* over the entire genome, and *r* the number of *RF* in a given window (with $R \geq r$ and $S \geq r$). The probability of observing by chance at least *k SF* ($x \geq k$) out of *r RF* in the window is the pointwise significance of the observed numbers of *SF* in the window (p-value, p):

$$p_{x \geq k} = \sum_{x=k}^{r} \frac{\binom{S}{x}\binom{R-S}{r-x}}{\binom{R}{r}}$$

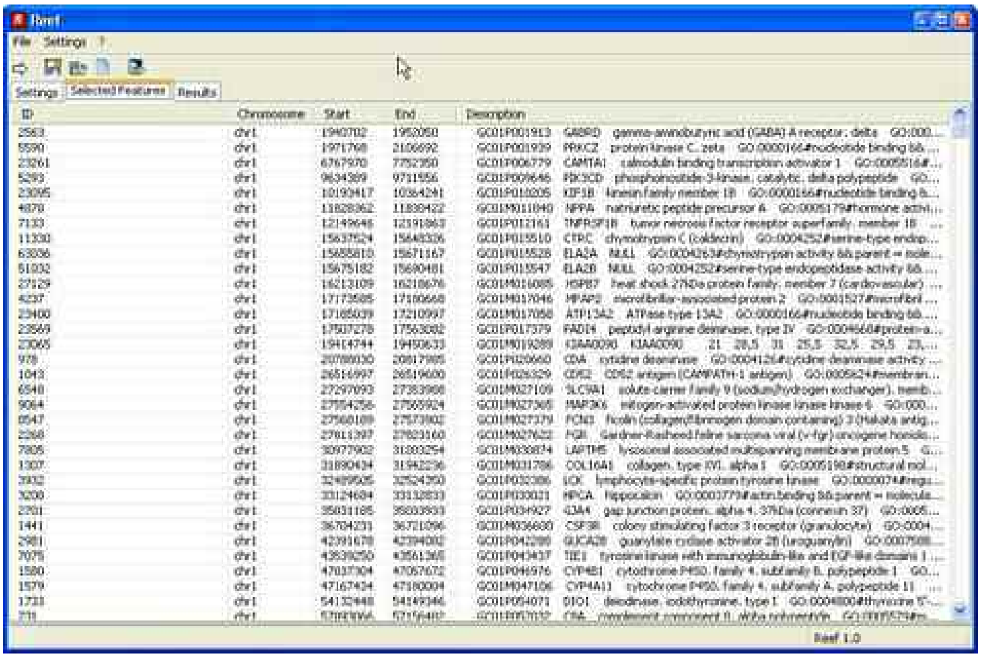
The False Discovery Rate (FDR, Storey and Tibshirani, 2003) is used to circumvent the problem of multiple testing for the genomewide calculation of statistical significance for the observed enrichment in *SF* in a given region. In particular, after sorting windows by *p-values* over the entire genome, *q-values* (FDR) were calculated. *Q* (*q-value*) for each window is defined as *Q=(p*N)/i* , where *p* is the *p-value* of the window, *N* the total number of windows considered and *i* the number of windows with a *p-value* not higher than *p*. Given a global threshold for the genome-wide FDR (e.g. 5%), the number of windows "significantly enriched in *SF*" is determined. The span of the maxi-mum number of adjacent windows showing statistical significance defines the boundaries of one cluster of *SF*.

- The *Minimun Number of Features in Cluster* parameter determines the minimun number of selected features that a window must contain in order to consider the window for further analysis. This parameter influences the FDR calculation: the higher the Minimun Number of Features in Cluster the less stringent the statistical threshold on the single window.

By clicking on the "Start Analysis" button in the toolbar, the program start to run!

Results

The "Selected Features" window contains the list of all the selected features loaded from the input query file.

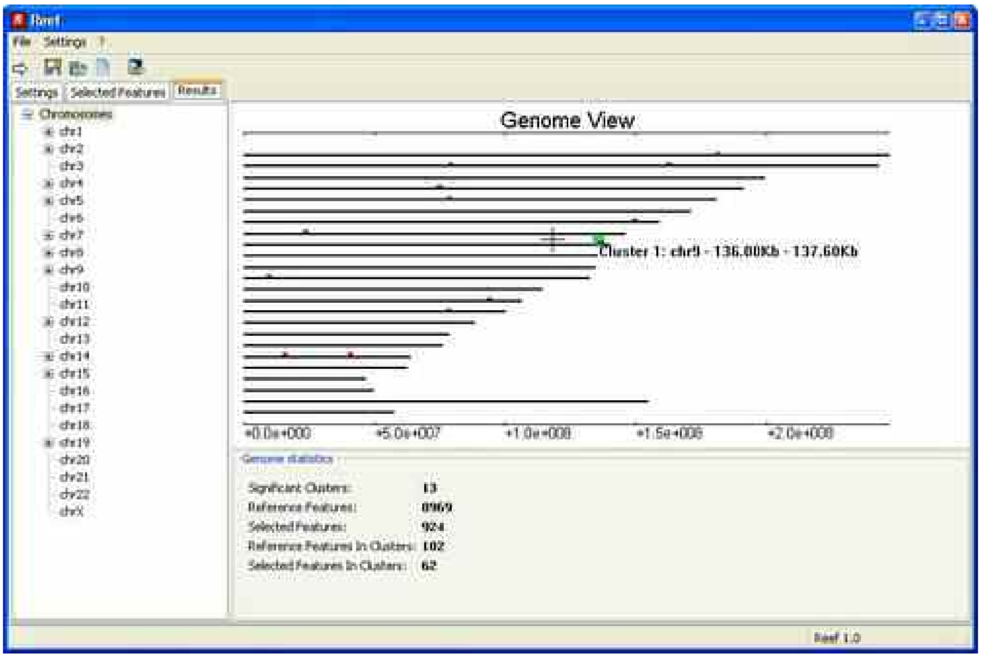


ID	Chromosome	Start	End	Description
2563	chr1	1940702	1952050	GC01P001913 GABRD gamma-aminobutyric acid (GABA) A receptor, delta GO:000...
5590	chr1	1971768	2106692	GC01P001939 PRKCZ protein kinase C, zeta GO:0000166#nucleotide binding && ...
23261	chr1	6767970	7752350	GC01P006779 CAMTA1 calmodulin binding transcription activator 1 GO:0005516#...
5293	chr1	9634389	9711556	GC01P009646 PIK3CD phosphoinositide-3-kinase, catalytic, delta polypeptide GO:...
23095	chr1	10193417	10364241	GC01P010205 KIF1B kinesin family member 1B GO:0000166#nucleotide binding &...
4878	chr1	11828362	11830422	GC01M011840 NPPA natriuretic peptide precursor A GO:0005179#hormone activi...
7133	chr1	12149646	12191863	GC01P012161 TNFRSF1B tumor necrosis factor receptor superfamily, member 1B ...
11330	chr1	15637524	15648326	GC01P015510 CTRC chymotrypsin C (caldecrin) GO:0004252#serine-type endop...
63036	chr1	15655810	15671167	GC01P015528 ELA2A NULL GO:0004263#chymotrypsin activity && parent = mole...
51032	chr1	15675182	15690481	GC01P015547 ELA2B NULL GO:0004252#serine-type endopeptidase activity && ...
27129	chr1	16213109	16218676	GC01M016085 HSPB7 heat shock 27kDa protein family, member 7 (cardiovascular) ...
4237	chr1	17173585	17180668	GC01M017046 MFAP2 microfibrillar-associated protein 2 GO:0001527#microfibril ...
23400	chr1	17185039	17210997	GC01M017058 ATP13A2 ATPase type 13A2 GO:0000166#nucleotide binding && ...
23569	chr1	17507278	17563082	GC01P017379 PADI4 peptidyl arginine deiminase, type IV GO:0004668#protein-a...
23065	chr1	19414744	19450633	GC01M019289 KIAA0090 KIAA0090 21 28,5 31 25,5 32,5 29,5 23,...
978	chr1	20788030	20817985	GC01P020660 CDA cytidine deaminase GO:0004126#cytidine deaminase activity ...
1043	chr1	26516997	26519600	GC01P026329 CD52 CD52 antigen (CAMPATH-1 antigen) GO:0005624#membran...
6548	chr1	27297093	27383988	GC01M027109 SLC9A1 solute carrier family 9 (sodium/hydrogen exchanger), memb...
9064	chr1	27554256	27565924	GC01M027365 MAP3K6 mitogen-activated protein kinase kinase kinase 6 GO:000...
8547	chr1	27560189	27573902	GC01M027379 FCN3 ficolin (collagen/fibrinogen domain containing) 3 (Hakata antig...
2268	chr1	27811397	27823160	GC01M027622 FGR Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolo...
7805	chr1	30977902	31003254	GC01M030874 LAPTM5 lysosomal associated multispanning membrane protein 5 G...
1307	chr1	31890434	31942236	GC01M031786 COL16A1 collagen, type XVI, alpha 1 GO:0005198#structural mol...
3932	chr1	32489505	32524350	GC01P032386 LCK lymphocyte-specific protein tyrosine kinase GO:0000074#regu...
3208	chr1	33124684	33132833	GC01P033021 HPCA hippocalcin GO:0003779#actin binding && parent = molecula...
2701	chr1	35031185	35033933	GC01P034927 GJA4 gap junction protein, alpha 4, 37kDa (connexin 37) GO:0005...
1441	chr1	36704231	36721096	GC01M036600 CSF3R colony stimulating factor 3 receptor (granulocyte) GO:0004...
2981	chr1	42391678	42394002	GC01P042288 GUCA2B guanylate cyclase activator 2B (uroguanylin) GO:0007588...
7075	chr1	43539250	43561365	GC01P043437 TIE1 tyrosine kinase with immunoglobulin-like and EGF-like domains 1 ...
1580	chr1	47037304	47057672	GC01P046976 CYP4B1 cytochrome P450, family 4, subfamily B, polypeptide 1 GO...
1579	chr1	47167424	47180004	GC01M047106 CYP4A11 cytochrome P450, family 4, subfamily A, polypeptide 11 ...
1733	chr1	54132448	54149346	GC01P054071 DIO1 deiodinase, iodothyronine, type I GO:0004800#thyroxine 5'-...
731	chr1	57093066	57156487	GC01P057032 C8A complement component 8, alpha polypeptide GO:0005579#m...

The "Results" window shows the results that are obtained by an analysis. On the left a tree structure allows the user to choose different kind of views. The "genome view" can be accesses by left-clicking on the root of the tree called Chromosomes, it shows all the chromosomes in the genome and the position of clusters of enriched features on the chromosomes. The clusters are represented by red squares; passing the mouse pointer over them the name and the position of the cluster is shown. The "genome statistics" on the bottom rigth shows some statistics about the clusters in the genome:

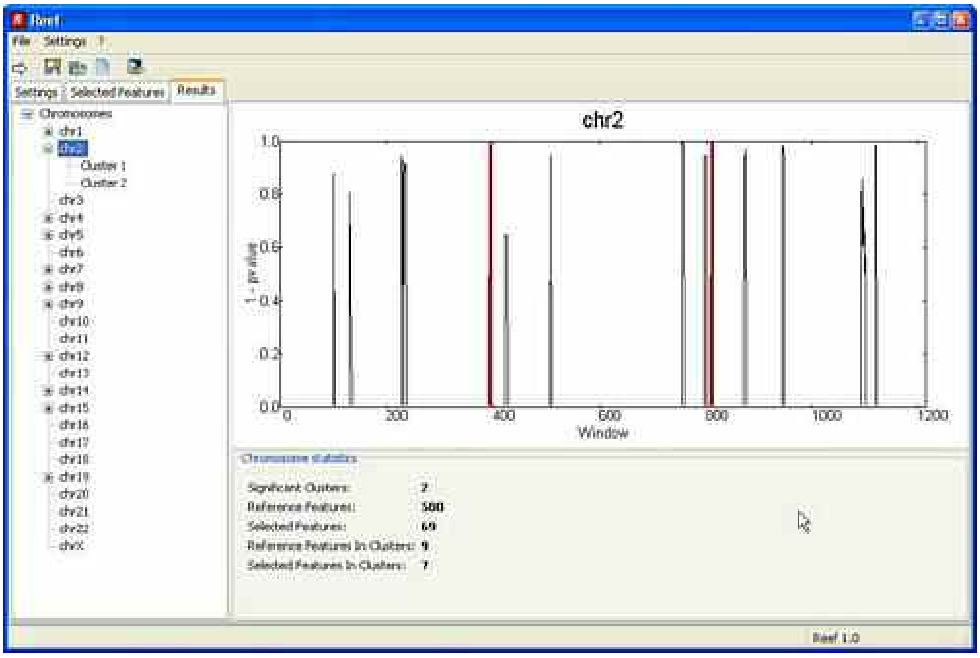
- *Significant Clusters* shows the total number of significant clusters in the genome
- *Reference Features* shows the total number of reference features in the genome
- *Selected Features* shows the total number of selected features in the genome
- *Reference Features In Clusters* shows the number of reference features contained in clusters

- *Selected Features In Clusters* shows the number of selected features contained in clusters



Significant Clusters: 13
Reference Features: 8969
Selected Features: 924
Reference Features In Clusters: 102
Selected Features In Clusters: 62

By clicking on a specific chromosome in the tree structure, a bar plot is given, showing the quantity (1 - p) of windows along the sequence-based coordinates of the chromosoem (significant values are represented by red bars). Information about the total number of significant clusters of selected features and of the total number of features in significant clusters in the chromosome are also given in the bottom panel.

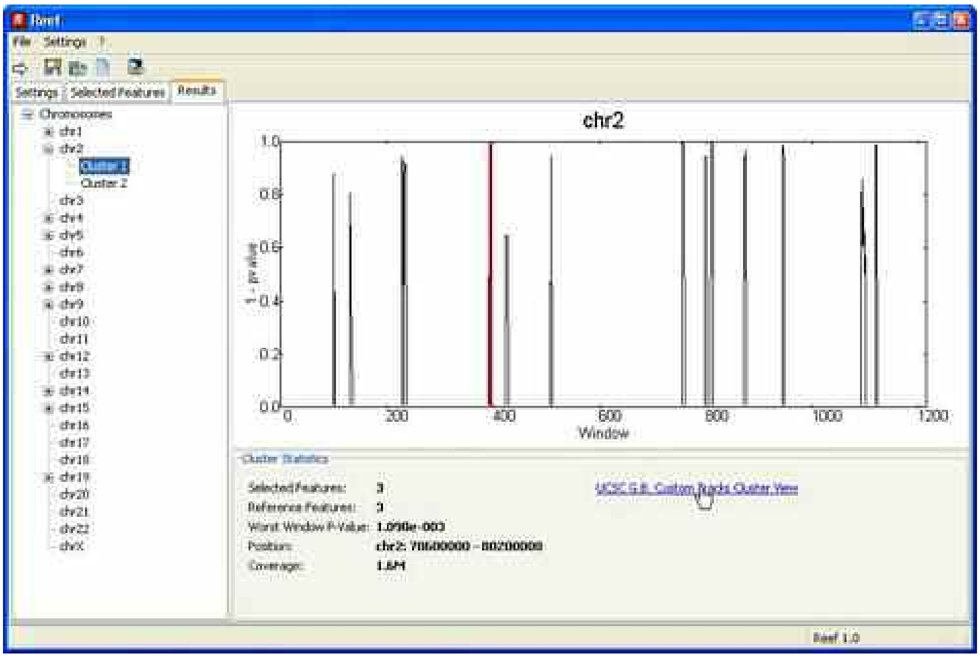


Significant Clusters: 2
Reference Features: 500
Selected Features: 69
Reference Features In Clusters: 9
Selected Features In Clusters: 7

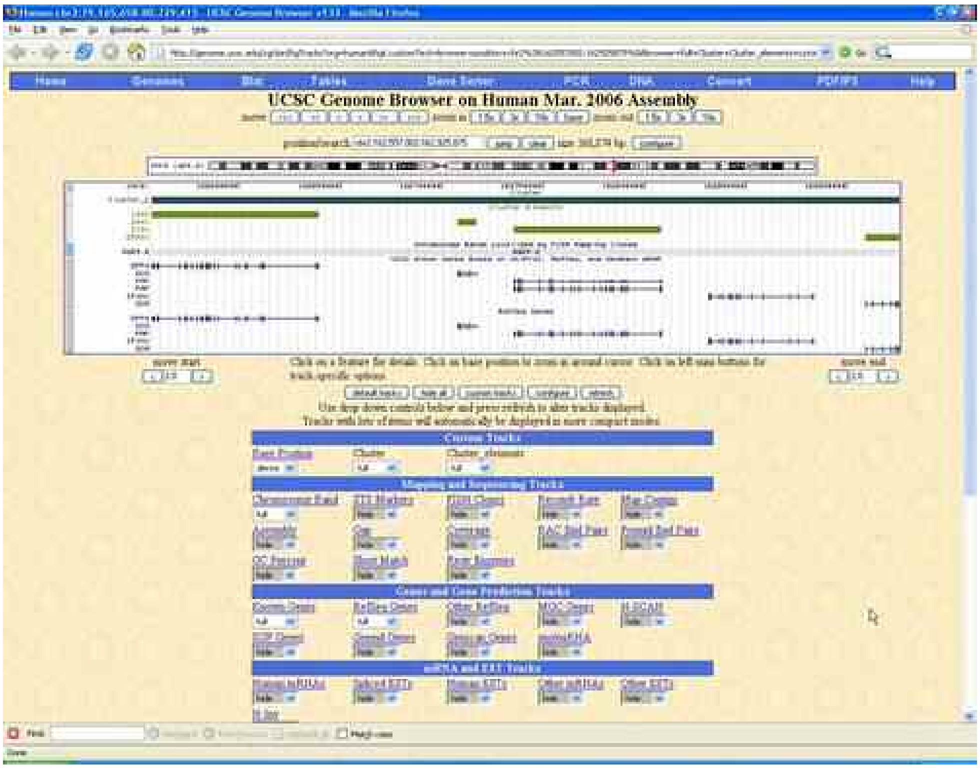
The chromosomes containing significant clusters show + symbol in the tree structure. By clicking on the + symbol the tree structure is expanded in order to show the list of clusters in the chromosome. By clicking on the cluster's name, the plot of the chromosome is shown but only the selected cluster is represented by red bars. The chromosome and the cluster plots are zoomable by left mouse buttom drag; left mouse double click resets the zoom; right mouse click zooms out centered on click location. The botton right subwindow shows the following information about the cluster:

- *Selected Features* shows the number of selected features in the cluster
- *Reference Features* shows the number of reference features in the cluster
- *Worst Window P-Value* shows the p-value of the less significative window in the cluster. Cluster are defined as consecutive significant windows
- *Position* shows the position of the cluster on the chromosome
- *Coverage* shows the extension of the cluster in Megabases
- *UCSC G.B. Custom Tracks Cluster View* is a web link that shows the cluster and it's elements on the UCSC Genome Browser. The features pertaining to the cluster are visualized as custom tracks, together with

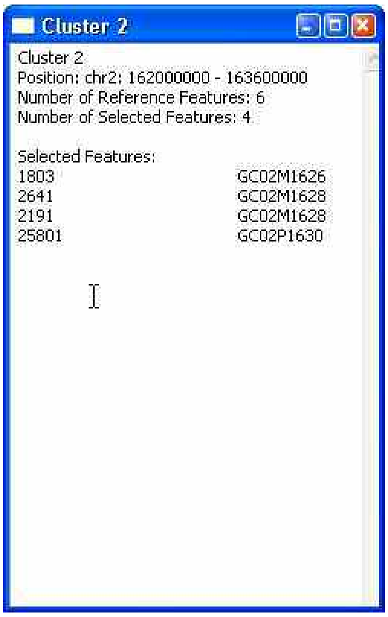
standard tracks from UCSC Genome Browser. A "cluster" track shows the chromosome position and the span of each given cluster, whereas a "cluster elements" track shows the position and the span of the different selected features in the cluster, each identified by the name/ID given by the user. In this, way, cluster information can be inspected together with the annotation information available for the considered genome. The user must set the "Cluster" and "Cluster_element" tracks to "full view" on the Genome Browser web page in order to display the custom annotation.



Selected Features: 3
Reference Features: 3
Worst Window P-Value: 1.090e-003
Position: chr2: 78600000 – 80200000
Coverage: 1.6M


UCSC Genome Browser on Human Mar. 2006 Assembly
Custom Tracks
Mapping and Sequencing Tracks
Genes and Gene Prediction Tracks
mRNA and EST Tracks

By double clicking on a cluster in the tree structure a new window containing information about the cluster is opened. The window also show the list of selected features Ids pertaining to the cluster and the first part of the feature description provided by the user in the input file.



Cluster 2

```
Cluster 2
Position: chr2: 162000000 - 163600000
Number of Reference Features: 6
Number of Selected Features: 4

Selected Features:
1803                    GC02M1626
2641                    GC02M1628
2191                    GC02M1628
25801                   GC02P1630
```

The *"Dump To Text"* button on the toolbar creates a .txt file with information about all clusters. Values are separated by the tab character in order to allow post-processing of the results with custom made scripts or spreadsheets-based programs. The text dump file shows, for every cluster, the list of all the features names/ID, chromosome, start position, end position and annotation information.