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 available 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.

Brith					6
Settings 1					
TO Selected Features Bandis					
Garrenal Settinge					
Query File (Selected Features)				Browne	
Reference Features File				Direct	
				. Contraction	
Window with:	11M	3	94b	2008. 🛩	
			No:		
Training Settings			13		
Qvalue: D.OK					
DESTATU DES MARIO	-				
Min. Number Of Genes In Outlini	1	*			
					Revf 1.0

- 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* ≥ *r* and *S* ≥ *r*). The probability of observing by chance at least *k SF* (*x* ≥ *k*) out of *r RF* in the window is the pointwise significance of the observed numbers of *SF* in the window (p-value, p):

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

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.

File Settings 1				
re seuroi r				
0 Re 8			- Re-	
Contract Constant Functional Data Arc			1000	
Security (Security Level of 104002)			1000	120000000
(D)	Cronocome	Start	End	Description
2563	dvt.	1940702	1952058	GC01P001913 GABPD germa-annobutyric wolf (GABA) A receptor; delta G0:000
5590	dv1	1971768	2106692	GC01P001939 PRKCZ protein kinese C. zeta G0.00001664roudeotide binding 88
23261	cirl.	6767970	7752350	GC01P006779 CAMTA1 salmodule binding transcription activator 1 GO:0005516#
5293	dr1	9634389	9711956	GC01P009646 P0K3CD photphonositide 3-kinase, catalytic, delta polypeptide GO
23095	- chr1	10190417	10364241	GC01P010205 KIF38 Anein fanty member 18 GO:0000166#nucleotide binding b
4870	det	11828362	11830422	GC019011040 NFPA natriaretic peptide precursor A GO:0005179#hormone activities
7133	civit.	12149645	12191063	GC01P012161 TMFRSF18 tumor necross factor receptor expertantly member 18
11330	del	15637524	15848326	GC012015510 CTPC dymotrypen C (caldechr) G0:0004252#serine-type endrp
63036	del	15655810	15671567	GC01P015528 ELA2A NLLL GC00C04053#chynotrypon activity 8/b parent = isole
61932	ctel .	15675182	15690481	GC01P019547 ELA28 MILL G0:0004252#serine-type endopeotidese activity 68
27129	det	16213109	16210676	GC01M016085 HSP87 heat shock 276De protein family: member 7 (cardiovascular)
4237	chri	17173505	17100668	GC01M017046 MPAPU microfibrillar-associated protein 2 GD:0001527#microfibril
29400	div1	17105039	17210997	GC01M017058 ATP13A2 ATPase type 13A2 GO-0000166#nubeobide binding 68
23569	dvi	17507278	17563002	GC01P017379 FAD14 peptidyl arginine deminase, type IV GC-0004668#protein-a
23065	dri	19404744	19450633	GC01/4019289 KIAA0090 KIAA0090 21 28,5 31 25,5 32,5 29,5 23,
978	drt .	20788030	20817985	GC019020660 CDA sytidre deaminase G0:0004126#sytidre deaminase activity
1043	det .	26516997	26519600	GC01F025329 CDE2 CDE2 antigen (CAMPATH-1 antigen) GO:0005824#triembnan
6540	del	27297093	27383966	GC00M027109 SLC941 solute carrier family 9 (sodium/hydrogen exchanger), memb
9064	det	27554256	27565924	GC014027305 MMP3K6 witogen-activated protein linase kinase kinase 6 -60:000
0547	chrl	27560109	27573902	GC014027379 FCN3 Fcoln (collager/Tibrinogen domain containing) 3 (Hakata antig
2268	dr1	27611397	27823160	GC01M027622 FGR Gardner-Raiheed failne sarcona viral (v-fgr) oncogene honols
7805	det.	30977902	31003254	GCBLM030874 LAPTHS Tysosonal associated incitigrarving membrane protein 5 G.,
1307	cfv1	31890434	31942236	GCIILM031786 COL16A1 collagen type IVI. alpha 1 G0:0005198#structural nol
39832	drt:	2149935	32524350	GC01F032306 LOC lymphocyte-specific protein tyrosine lanase GO:0000074#rega
3200	det	33124634	33132833	GC01F030031 HPCA Reposition G0:0003779#attribeding 56-parent + Holecula
2201	dvt.	35031185	35033933	GC01P034927 G344 gap unchan protein: apha 4, 37kDa (connexin 37) GD:0005
1441	ctel .	36704233	36771096	GCIIDM036600 CSF3R colory stimulating factor 3 receptor (granulocyte) GO:0004
2901	chr1	42391670	42394002	GC01P042289 GUCA28 guarylate cyclase activator 28 (uroguarylin) GO:0007988
2025	drl	43539250	\$3561365	GC01F043437 TIE1 Tyrotine linese with immunophilum-like and EQF-like climane 1
1580	dy1	47037304	47057672	GC01P046976 CVP4E1 cytochrome P450, family 4, subfamily 8, polypeptde 1 GO
1579	chr1	47167434	47180004	GC01M047106 CVP4A11 tytochrone P450, family 4, subfamily A, polypeptide 11
1723	chr1	:54132948	54149346	GC01P094071 0101 deadnase adothyronine. type 1 G0:00048004thyroone 5
211	eter.	AMERICA.	57356482	activitations CIA previousit concernt II with industrial Contractors
				Bird 1.0

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



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.



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.



Cename,	Blan Tables		David Berter .		DINA	Connert	PDAYS
The second second	UCSC Ge	nome Brow	ser on Hum	an Mar. 200	6 Assembly	Contraction of the second	1000
	and the Cold Cold	Contractions)		1 th I ter =	15151	3	
	puttering	1-647-02397-0210	ENKING (MR)	See her biltra	Na Company		
1000 (att.0) 278	NOT THE OWNER.		>+CROOM D		Section and the		3
and Sectors	e - 100100	'admin	00 00	1992 (M	direct .	different (1	illior .
111			1000	d heating			
and the second s							-
make A			- including of the	TTAL BUTTON AND D	tioners and		The second second
1000. 1900			+ 4	1311551000			
12 and			1.1	Charles III			e
121 0							
12			11			3-100 1-1	 1 1 1 2 3 4 4
	Chikmas Buikagead Upri	camer for details. C le optimies (altrationity) (de lege dover, cambola)	Soli as hair poster and Count table tober and prevents	lei anni pi a nad li I Charlper I Crémi ph to sho trache de	une Chikin Mit Di alerri	neu luitono Se	Line south
	Chile on a S Stack open Used Tracker of Last Postan	ramen for density (a updates) (minut heats) (the heat dense controls) di lanz of menu and Chatter	Sell as bair poster of all Connected A before and preserved Connected by be do Connected by be do Connected by be	la zona a aronal i I Consepti I Conse relativa des tracto de played as soure comp a	ere Chille Mi alerti atorit	aas kathati Ee	COM CI
10 mm Add	Chikeys Kukeyni Upri Todar w Tael Podan Jaes W	namer for Anials C or optimum (Anial Sector) (So here down or category of the law of annual of Charge 1.4 ()	Sell as hair poster e d.) Connected, letter all prevent l'acteurs dy to de Catter delated to de Catter delated	for some at an owned to I Conseque I Control yields to show that the played as more comp and a	ere cha wite Q alerti afanisi	nee katasi Ee	
	Chikers all Backgrout Tacker of Tacker of Tacker of Tacker of Tacker of Tacker of Tacker of Tacker of Tacker of Tacker Tacker of Tacker T	Charter for decail 1 c options (and other control of the down control of the of annu with Charter 14 C	202 on here proton in (4) Course tests before all protocols Contents dy to do Dutter, defaund of all Test Course Test Course	In most a send of I Compare I Control of the data tracks in probability of the sent comp of 1000000 The most fact.	ere Chilaide Dairtí afartí MarDona	nas lumas fo	
sere Aut	Chile on 1 Solidary Unit Test Parties Area on Maria and Maria anda	Canada for Annala C Canada Salaha (Canada Salaha) Canada Salaha (Canada Salaha) Raya dawa canada (Canada Raya (Canada Salaha) Canada Canada Canada Canada (Canada Salaha) Canada (Canad	Self on here proton in AD Connected before and preserved Connected by the de Context of the de Context	In some in armalie Compare I Connel Conta to also tracks in played in some comp international from I formation I formation		see Lattain for	
Contraction	Chile on all Stationard Under Texture Area on Area on	amore for density of a spectral and the set of the set of the set of secure will be set of secure will be set of secure will be set of secure set of the set of secure set of the set of secure set of the set of secure secure set of secure secure secure set of secure se	All in her present at 2 Generative labor of provide laboration of the labor laboration of the labor laboration of the labor laboration of the labor laboration of the laboration of the laborati	to anno a armado) Conserva (Conserva- rel to des socials de provision anno 1000 provision anno 1000 provisio	ere Chikabi Unigeri atanin MacDana Mats Mats Mats	nga katana Ka	
Contraction of the second seco	Chik on 41 Stationer Under Techne Anne an Anne	Annual for density of a spinore and spinor	All as har person of a 1 Conservation before and persons determined by the data of the conservation of the	Norman Connection I Connection Connection the the Section Section page 10 more ready page 10 more read	Martina Martin Martina	sae luinas fo	
Server Aut	Chik on 4 bick on 4 Une 1 Track of the Track of the Tra	Annual for density of a spinore and dense control of all dense control o	And an bair points of all constrained with bother and points of extension of the la constrained with the Constrained with the Constrain	Al more a second in Conveyor I Conveyor I physical and the second of a second second second a second second second second a second second second second second a second sec	Martina Martin Martina		
sere Aut	Cake with Stational Unit Technic Annual Annu	A series of the density of the densi	And an bair points of all Connect table below on presented to connect of the A Connect of the A C	A mean a second in Conveyor I Conveyor I Conveyor I ph p shot tracks in phone in a second second in a second second second second in a second second second second second in a second s	ner Chinakt	nee Luthon Ev	
sere aut	Chike and Stakened Trade of Trade of Tr	Annue for Annuel 1 Se opposed annue for Annuel 1 Se opposed The Annuel for Annuel for Annuel for Annuel 1 Se opposed The Annuel for An	And an bate possion of all a state possion before and protocols development of the A Control Plant of the Control	A mass a second in Chernese I Control of the data vanish in provide a new ready provide a	And Children of the Control of the C		terre sud (1)11 (1)
Contraction of the second seco	Cake and S State and Trade of Trade of	anar (a Anda) - C - andar Andar - C - andar Andar - C - andar Andar - C - andar Andar - C - andar - a	And an bate possion of all a state possion before and possion devices and all the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the s	A mass a second in Consequence (Consequence) particular and second and particular and par	And Children Left Martineite		terre sud Califa Ca
Contraction of the second seco	Cake and Stakened Used Todar or Todar or Description d	Anna Anna Anna Anna Anna Anna Anna Anna	And an base possible and a local posterior and a local posterior design of the local control	N meso a second in Compact Comment physical activity of the physical activity of the physica	And Andrew Contractor		Rome and Colific Co

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 lds 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; 162000 Number of Reference Number of Selected Fe	000 - 163600000 Features: 6 satures: 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.