User manual

(October 2008, v1.2)

CARPET

(Collection of Automated Routine Programs for Easy Tiling)

A web-based package for the analysis of ChIP-chip and expression tiling data

Matteo Cesaroni[1]*, Davide Cittaro[2], Alessandro Brozzi[1], Pier Giuseppe Pelicci[1] and Lucilla Luzi[3]*

[1] Department of Experimental Oncology, European Institute of Oncology, Via Ripamonti 435, 20141 Milano, ITALY

[2] Cogentech, Consortium for Genomic Technologies, Via Adamello 16, 20139 Milano, ITALY

[3] IFOM, FIRC Institute of Molecular Oncology Foundation, Via Adamello 16, 20139 Milano, ITALY

*To whom correspondence should be addressed.

TABLE OF CONTENTS

1. Introduction

CARPET (Collection of Automated Routine Programs for Easy Tiling) is a set of Perl, Python and R scripts, integrated on the Galaxy2 web-based platform (Blankenberg, et al., 2007), for the analysis of ChIP-chip and expression tiling data. CARPET allows rapid experimental data entry, simple quality control, easy identification and annotation of enriched ChIP-chip regions, detection of the absolute or relative transcriptional status of genes assessed by expression tiling experiments and, more importantly, it allows the integration of ChIP-chip and expression data. Results can be visualized instantly in a genomic context within the UCSC genome browser as graph-based custom tracks through Galaxy2. All generated and uploaded data can be stored within sessions and are easily shared with other users.

The program suite can be accessed through the Galaxy mirror site of IFOM-IEO-CAMPUS at http://bio.ifom-ieo-campus.it/galaxy.

For questions and suggestions, please contact:
matteo.cesaroni@ifom-ieo-campus.it
lucilla.luzi@ifom-ieo-campus.it
davide.cittaro@ifom-ieo-campus.it

2. The Galaxy platform

Galaxy was first developed in 2005 by Giardine and coworkers (Giardine, et al., 2005) as a platform for interactive large-scale genome analysis. More recently, it has been used as a framework for the collaborative analysis of ENCODE data (Blankenberg, et al., 2007). Galaxy is not a browser. Instead, it allows users to gather and manipulate data from existing resources in a variety of ways. Moreover, Galaxy provides a user-friendly interface that facilitates interactions between experimental and computational biologists by providing a simple interface (important to the former) and a robust software integration environment (important for the latter). The simplicity of Galaxy2's tool integration protocol allows developers to easily integrate their programs and make them available to biologists.
You can find further information on Galaxy2's main site, http://g2.trac.bx.psu.edu/.
For general issues regarding Galaxy usage, please consult the detailed tutorial session at the official "Galaxy Screencasts and Demos" webpage.

3. Uploading data

The first step, when using any of the CARPET programs, is to upload the data files you wish to analyse using the *Get Data/Upload File* Galaxy tool (see **Fig.1**).

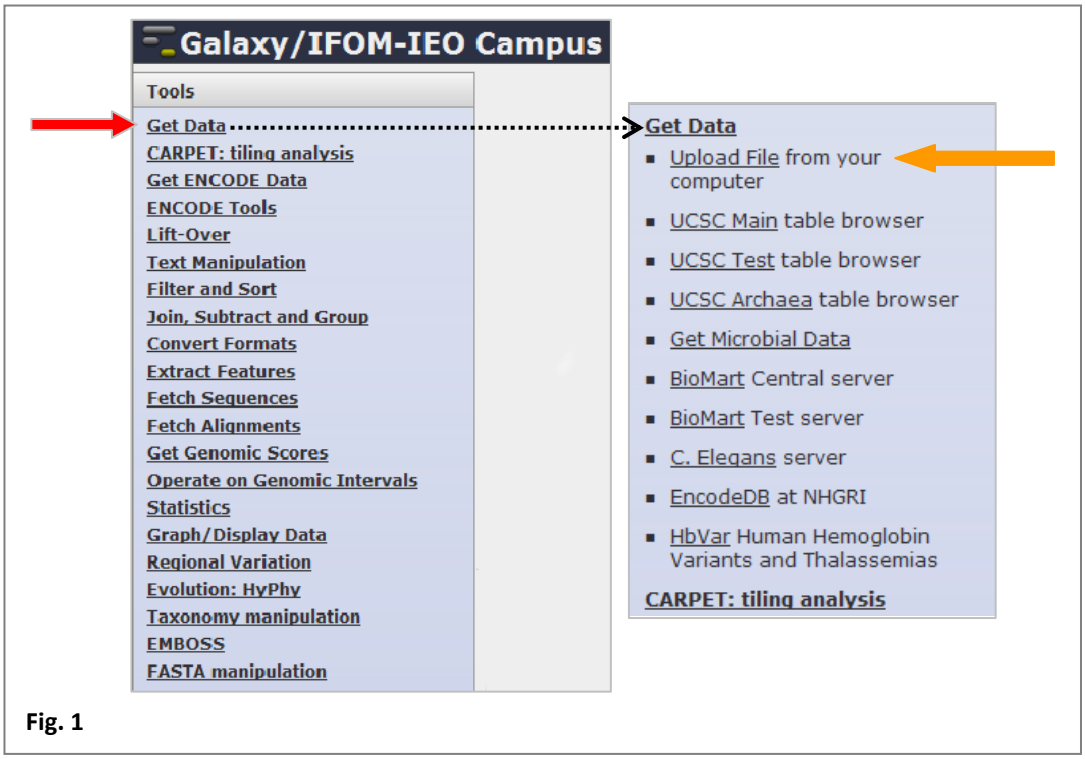


Fig. 1

Data can be uploaded by either browsing for a file on your hard disk (1 in **Fig.2**) or by cutting and pasting URL addresses or the file contents in the "URL/Text:" window space (2 in **Fig.2**).

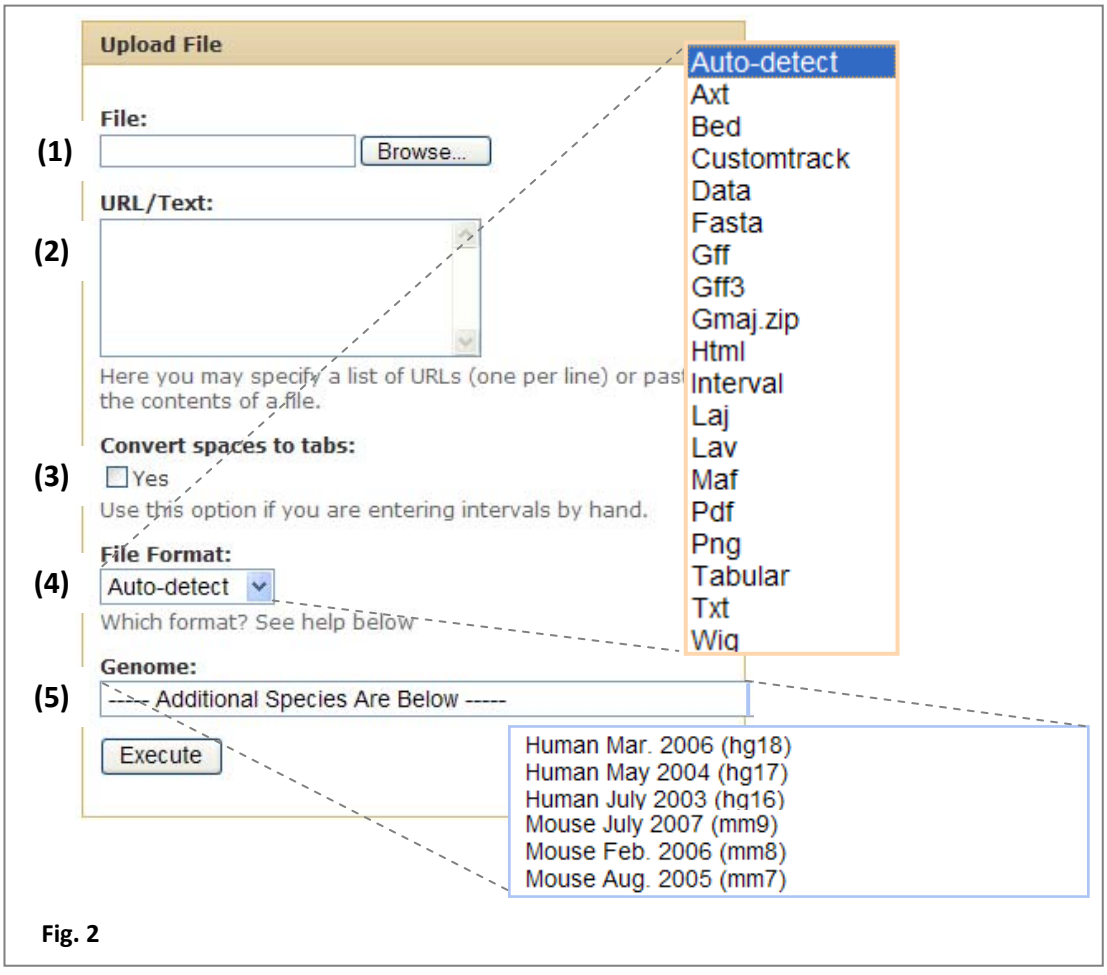


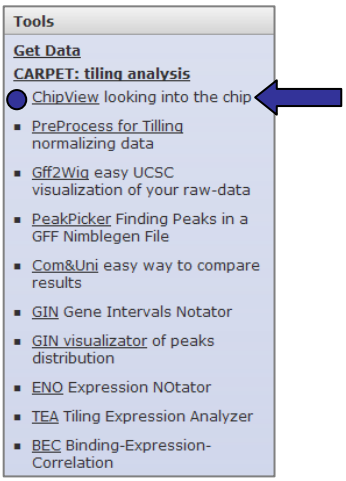
Fig. 2

Galaxy accepts many different file formats (the full list is shown on the previous screenshot; for details refer to the "Galaxy Screencasts and Demos" webpage, or to Appendix A of this manual for a detailed description of all CARPET handled file formats).
The system can auto-detect the file format or you may choose the appropriate format from the "File Format" popup menu (4, in **Fig.2**).
All "table-like" files that contain rows and (tab delimited) columns are recognized. If you are unsure, whether your file is tab delimited or not, ask the system to convert spaces to tabs (3, in **Fig.2**).

Most CARPET programs accept GFF files as the input, i.e. the standard file format provided by NimbleGen for tiling array data. For all other programs, a detailed description of the required file format is provided in the tool sections below or in Appendix A of this manual. Corresponding example files are provided on the server for downloading and are linked in this manual through digital text.

Although not mandatory, if you are uploading files that can be visualized on the UCSC Genome Browser (GFF or BED files; see Appendix A of this manual), we suggest that you associate the corresponding Organism type and Genome Assembly version with your file using the "Genome:" popup menu (5, in **Fig.2**).

4. Quality assessment by chip image visualization: ChipView



ChipView, allows you to create and visualize an image of the hybridized chip surface (see example below), a feature not normally offered by NimbleGen. ***Chipview*** works with "Pair files" data provided by NimbleGen (see Appendix A for details) or "custom raw signal-coordinates files". With ***Chipview***, the distribution of the signal over the chip can be easily inspected for the presence of artefacts or hybridization problems.
Since ***Chipview*** considers only the chip raw signals and the corresponding x and y chip coordinates when creating the chip image, a standard NimbleGen "Pair File" is not strictly required. ***ChipView*** can use all types of raw data tables, as long as the corresponding chip coordinates are also supplied. You will need to specify to the system which columns contain matching data.

For a more detailed "Pair File" format description, see Appendix A of this manual.

How to use ChipView:

- upload the Pair File (or your custom raw signal-coordinates file) using the *Get Data/Upload File* Galaxy tool, specifying a "Tabular" file format (*See Section 3*);
- (for NimbleGen Pair Files only) once the file is uploaded, control that it contains 11 columns;
- click on the ***ChipView*** link from the "CARPET: tiling analysis" tools list;
- (for NimbleGen Pair Files only) select column C10 for "PM value", C6 for "X position" and C7 for "Y position" as shown below; (for custom raw signal-coordinates files only) select "PM value", "X position" and "Y position" columns accordingly for your custom file;
- click Execute.

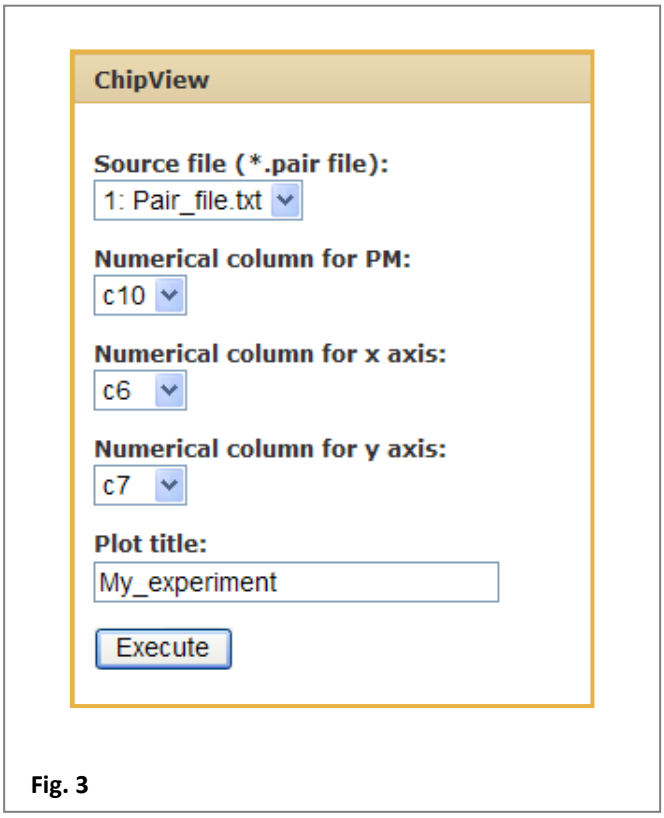


Fig. 3

ChipView produces a PDF image file, which can be downloaded and saved onto your hard disk.
An example image is shown in **Fig.4**.

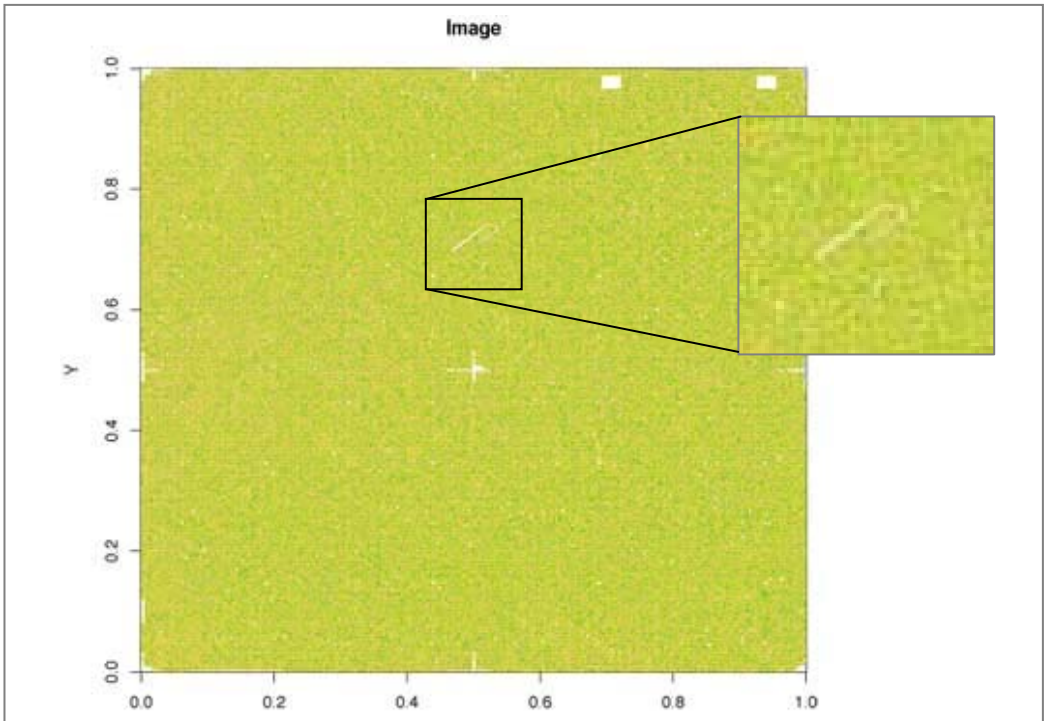
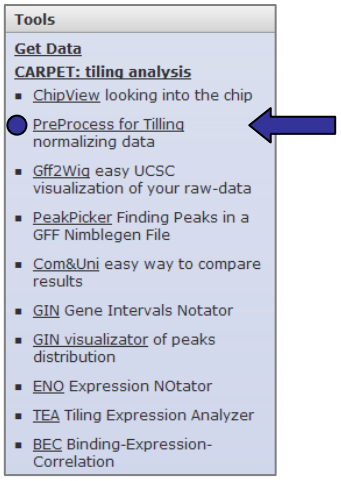


Fig. 4

5. Data normalization - PreProcess for Tiling: PPT



An important first step in chip data analysis is the normalization phase.

***NOTE:** if you are working with data created from the NimbleGen platform and you have a GFF file of your results, your data will already be scaled by a Tukey-biweight scaling procedure that centers the data around zero. This procedure is similar to the mean-normalization of each channel, therefore, you can skip this section and go directly to the peak identification tool, **PeakPicker**.*

PPT normalizes both ChIP-chip and expression tiling data (single or multi-replica experiments) using a program based, in part, on the Ringo Package (Toedling, et al., 2007). ***PPT*** also calculates and compares the correlation between replicates and finally creates a GFF file suitable for peak identification by PeakPicker or other user preferred methodologies/programs.

TEST files example *(click on the link here or go the PPT page on CARPET sever, to download 3 Test file examples - zipped compressed ~4.4Mb).*
Please NOTE: since virtually any kind of file can be used in this normalization tool, we have provided 3 example pair files from a ChIP-chip experiment randomly selected from GEO.

This tool accepts, as an input, essentially any number of replicates and any *table-like* files as long as they contain the following information in columns:

- Sequence ID, such as a chromosome number (e.g. chr1, chrX, chr2L, chrIV) or any other distinctive identifier of the sequences covered on the tiling array (e.g. ENm001, My_seq1, AnyStringX, 00000052);

HINT: *in both standard and custom single Nimblegen pair files, this information is usually placed in the third column (C3, SEQ_ID).*

- Start position of the probes, i.e. their relative position within a certain Sequence ID (e.g. 1567, 12, 1589019);

and one of the following:

- log2(ratio) of Cy5, Cy3 raw signals (e.g. -0.95, 3.28, 0.02);

or

- paired Cy5 (635nm) and Cy3 (532nm) raw signals as a single file, e.g. the "all_pair.txt" file provided by NimbleGen (e.g. 8903.44, 486.78, 10582.56). This kind of data likely corresponds to ChIP-chip records;

HINT: *If you do not have this kind of file, you can easily re-build it from single paired Cy5 and Cy3 raw files using the Galaxy "Text Manipulation" tools (see Appendix B for details).*

or

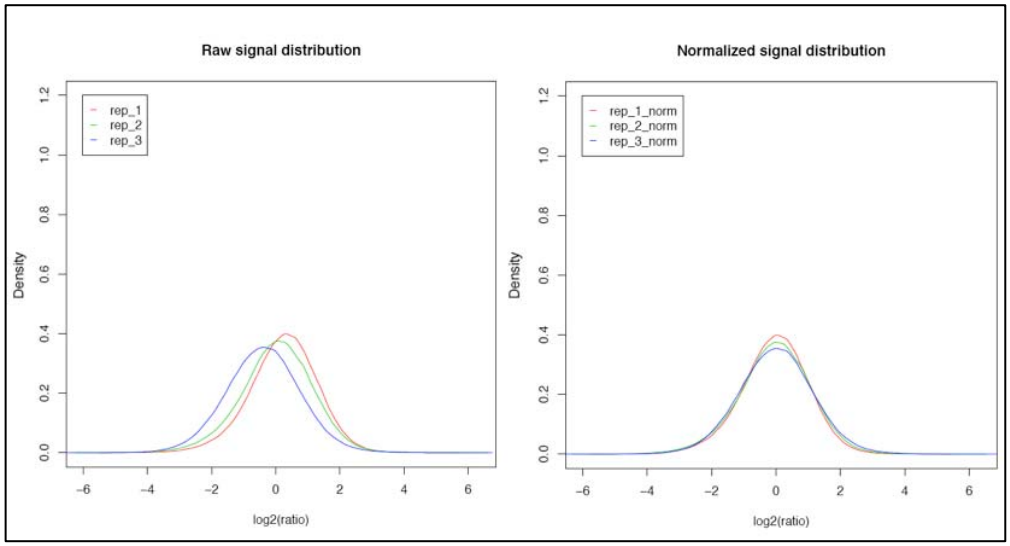
- single Cy5 (635nm) or Cy3 (532nm) raw signals (e.g. 8903.44, 486.78, 10582.56); this kind of data is typical of expression tiling records.

HINT: *if you wish, you can consider real probe lengths: if you have this information as a GFF file, you can specify the column that contains the End positions of the probes, instead of using an average probe length.*

How does ***PPT*** work?
For each chip, the log2 of Cy5 - Cy3 ratio is calculated (if not already provided). All the chips are then normalized, according to the type of normalization selected:

- bi-weight: this procedure centers the probe log2(ratio)s around zero; scaling is performed by subtracting the bi-weight mean for the log2(ratio) values from each log2(ratio) value.
- quantile: this procedure normalizes the distributions of the probe log2(ratio) of each chip with a quantile normalization.

The correlations between chip replicates can then be calculated.
The program produces two outputs: a table file of the normalized data and a pdf file of graphs showing data distribution before and after the normalization process and the correlation between replicates. An example of the graphical output is shown in **Fig. 5**.


Raw signal distribution
Normalized signal distribution
rep_1
rep_2
rep_3
rep_1_norm
rep_2_norm
rep_3_norm
Density
log2(ratio)

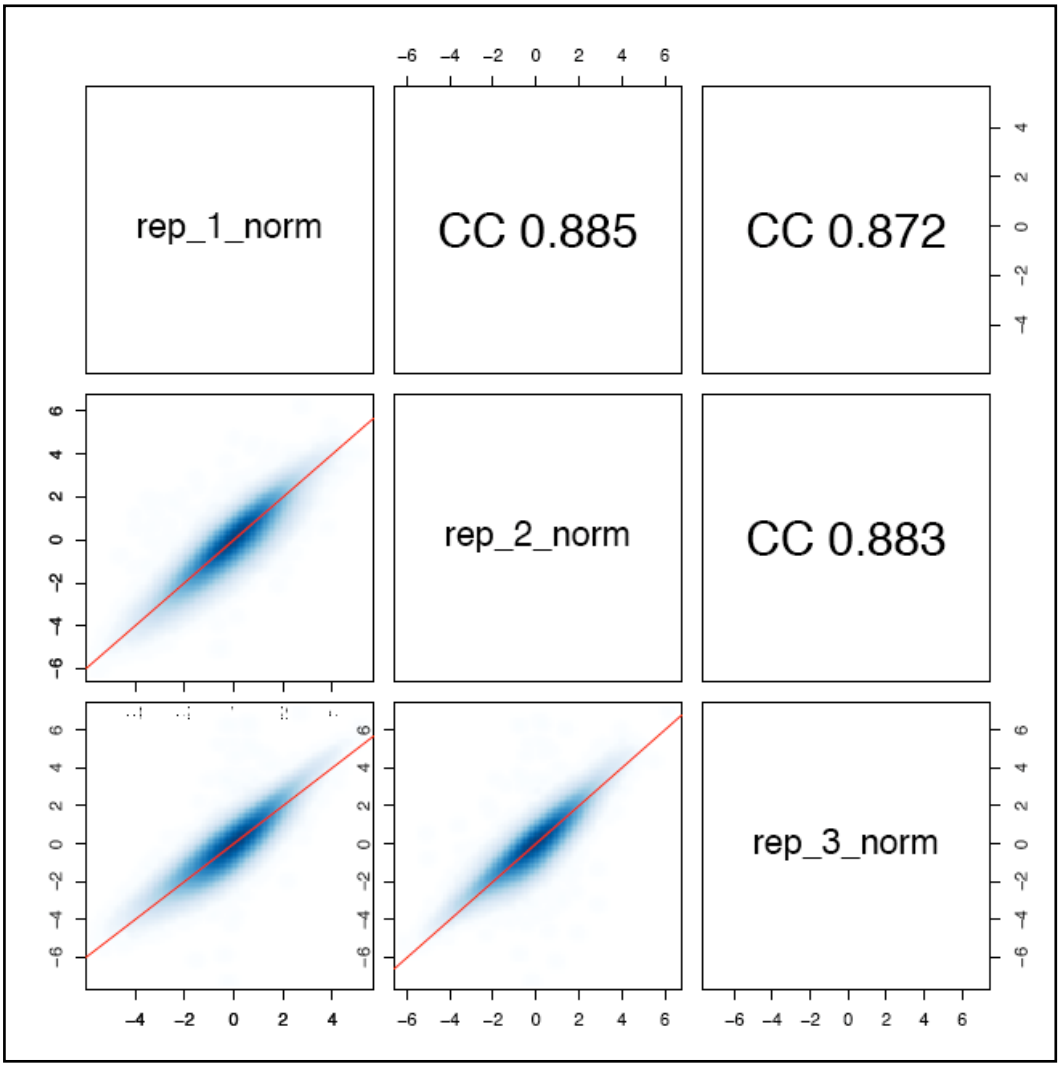

rep_1_norm
CC 0.885
CC 0.872
rep_2_norm
CC 0.883
rep_3_norm

Fig. 5

How to use PPT:

- upload your files using the *Get Data/Upload File* Galaxy tool, leaving the system to Auto-detect the file format (*See Section 3*);
- independently from the real format of your uploaded dataset, you need to convert your file to a GFF format in order for the system to interpret it correctly: click the pencil link in the History frame (**Fig. 7**);

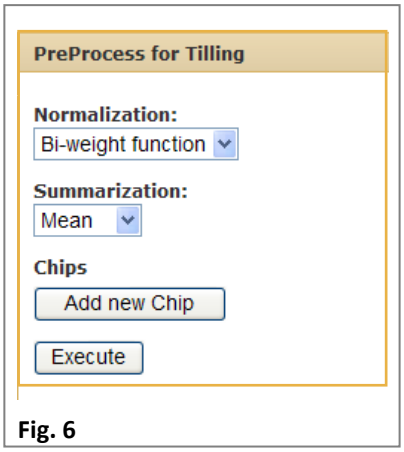


Fig. 6

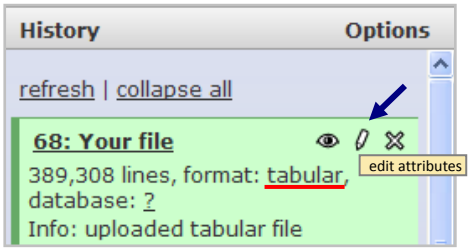


Fig. 7

- in the "Change data type" area (**Fig. 8**), select the GFF data type and Save to convert;

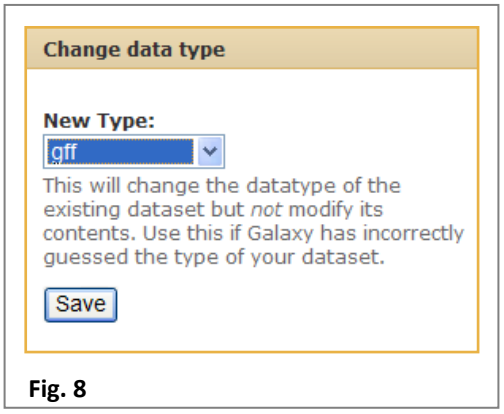


Fig. 8

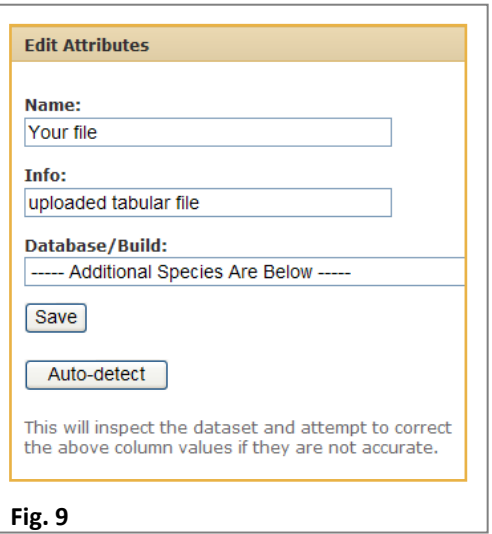


Fig. 9

- click again on the Save button in the "Edit Attributes" frame to definitely save the changes (**Fig. 9**);
- check that your file now has a GFF format in the History frame (**Fig. 10**);

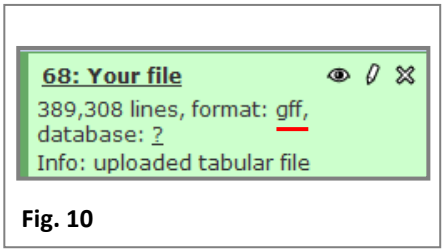
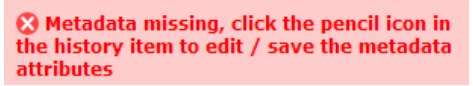


Fig. 10

WARNING: if you disregard this formatting procedure an error message, such as the one shown below, may appear during your analysis.



- repeat the procedure for all datasets/replicates you wish to analyze;

HINT: *(mainly for ChIP-chip experiments) if you want to normalize replicates, starting from single Cy3-Cy5 raw signals, you have to upload a single file containing the paired Cy3-Cy5 raw data. If you do not have this kind of file, you can easily re-build it from single paired Cy5 and Cy3 raw files, using the*

Galaxy "Text Manipulation" tools (see Appendix B for details). NimbleGen usually provides a file called "all_pair.txt": if available, you can directly upload this file.

- set the requested parameters (**Fig. 6**):
  - **Normalization**:
    - o B-weight function: centers the probe log2(ratio)s around zero;
    - o Quantile: applies quantile normalization.
    - o None: data is not normalized.
  - **Summarization**:
    - o Mean: after normalization, the program calculates, for each probe, the mean value of replicate signals;
    - o Median: after normalization, the program calculates, for each probe, the median value of replicate signals;
    - o None: the program supplies normalized values for each replicate analyzed.
- now add the **Chip** information (you have to repeat this procedure for each replicate) by clicking the "Add new chip" button (**Fig. 6**);

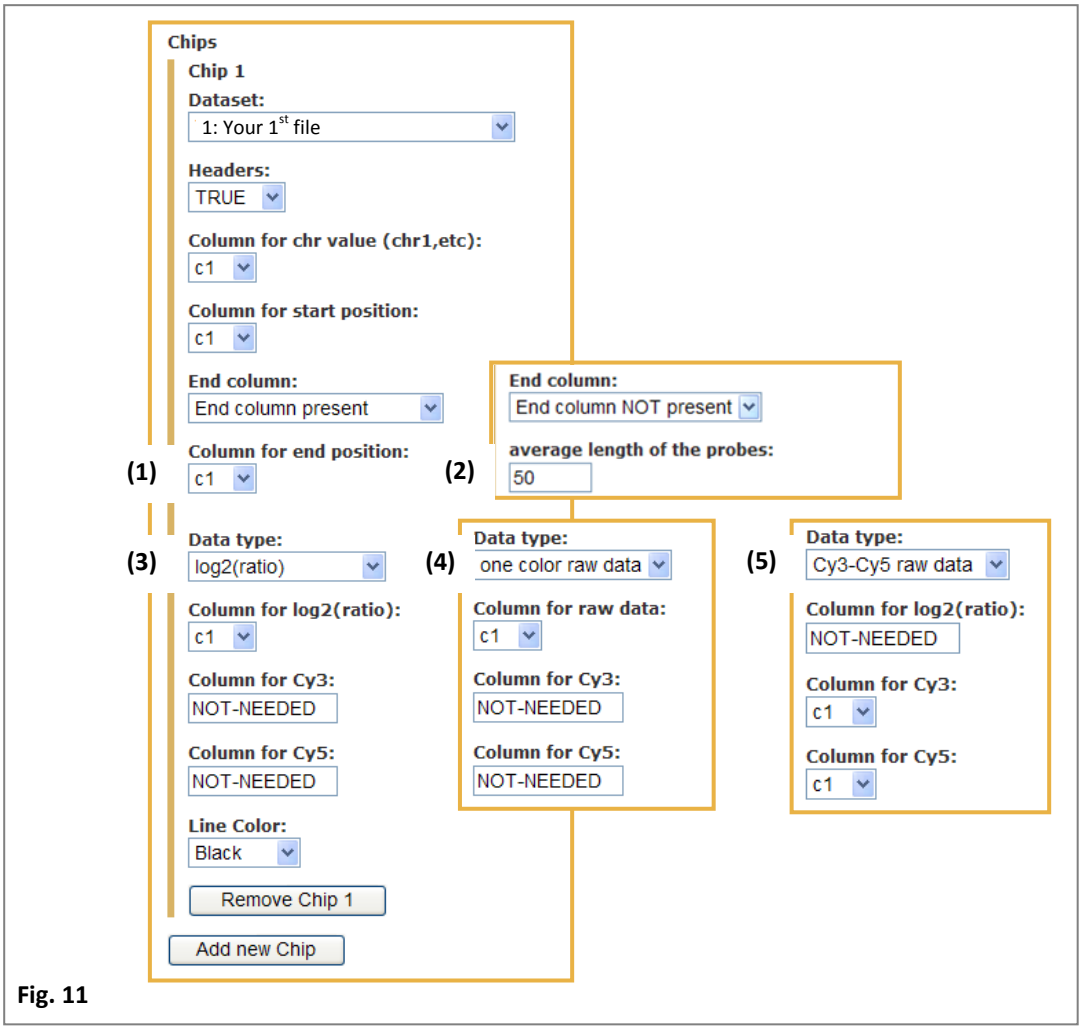


Fig. 11

- set the requested parameters:
    - select your (first) file from the Dataset list;
    - specify if your file contains a Header (TRUE) or not (FALSE);
    - select the column containing the Chr value (See the Sequence ID description above for details);
    - select the column containing the Start Position of the probes;
    - specify if your file contains (and you wish to consider it) the End Position of the probes (End column present) or not (End column NOT present);
        - if your file contains the End position, select the corresponding column (1 in **Fig. 11**);
        - if your file does NOT contain the End position (or you decide not to consider it) specify the average probe length of your tiling array (2 in **Fig. 11**);
    - select the Data type of your file from the popup menu: the page will automatically refresh and you will be presented with options associated with your data type;
        - for **log2(ratio)** (3 in **Fig. 11**):
            - select the column containing the log2(ratio) value of your experiment (normally from a ChIP-chip study);
        - for **one color raw data** (4 in **Fig. 11**):
            - select the column containing the one color raw data value of your experiment (normally from an expression tiling study);
        - for **Cy3-Cy5 raw data** (5 in **Fig. 11**):
            - select the columns containing the Cy3 and Cy5 raw data values of your experiment (See the Cy3-Cy5 raw data description above for details);
- repeat the procedure by clicking the "Add new chip" button for each replicate;
- when you have finished uploading your replicates click Execute.

The program produces different types of outputs determined by the combination of data type and options selected:

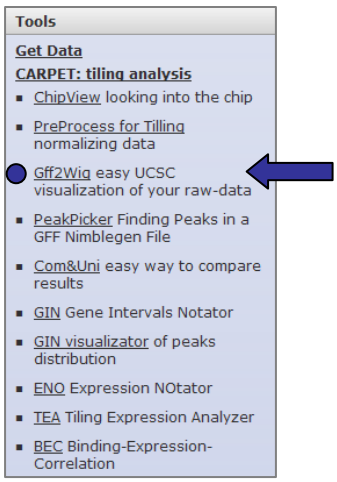
- if a summarization method was applied (or only one chip was uploaded), the program will generate a GFF file (ready to use with PeakPicker);
- otherwise a file similar to that shown below in Fig. 12 will be produced (in the example, 3 replicates were uploaded).

					rep1	rep2	rep3			
chr19	GALAXY	CARPET	11001	11051	-1.23	-1.69	-1.2	.	.	Cesaroni et al.
chr19	GALAXY	CARPET	11050	11100	0.059	-0.5	-0.539	.	.	Cesaroni et al.
chr19	GALAXY	CARPET	11099	11149	-2.89	-2.52	-2.703	.	.	Cesaroni et al.
chr19	GALAXY	CARPET	11148	11198	-2.83	-2.78	-3.198	.	.	Cesaroni et al.
chr19	GALAXY	CARPET	11197	11247	-2.19	-2.54	-2.797	.	.	Cesaroni et al.

Fig. 12

6. Visualizing chip intensity data on UCSC Genome Browser: GFF2WIG

Before performing in depth analyses of your tiling array data, you may wish to view a preliminary profile of your log2(ratio) raw signal. The intensity data for each probe (log2(ratio) of Cy5, Cy3 signals) can be visualized on the UCSC Genome Browser as a continuous histogram superimposed on the genome sequence. To do this you need to transform the GFF file of your tiling experiment to a WIG_bed format (http://genome.ifom-ieo-campus.it/goldenPath/help/wiggle.html).



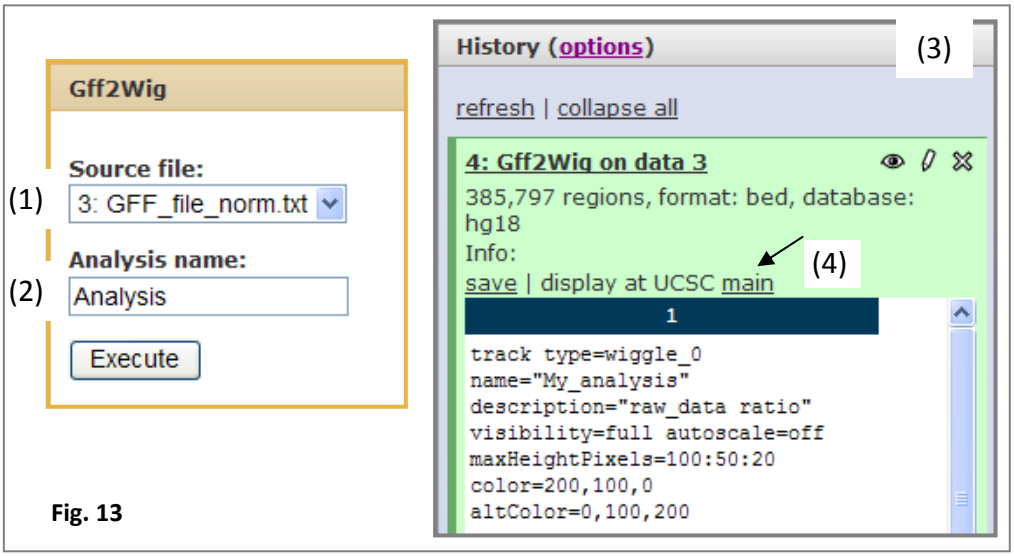


Fig. 13

GFF file example *(click on the link to download a GFF file example - zipped compressed ~4.4Mb)*

How to use GFF2WIG:

- upload your GFF File using the *Get Data/Upload File* Galaxy tool ensuring that the "GFF" format is selected (*See Section 3*);
- once your file is uploaded, click on the ***GFF2WIG*** link from the "CARPET: tiling analysis" tools list;
- select your GFF file from the "Source file:" popup menu (1 in **Fig. 13**);
- you can specify an "Analysis name" that will appear in the UCSC Genome Browser as the track name (2 in **Fig. 13)**;
- click Execute;

GFF2WIG will transform your GFF file to a WIG_bed file similar to that shown below in Fig. 14 (the first 3 rows only are shown).

```
1   track type=wiggle_0 name="Analysis name" description="raw_data ratio" visibility=full autoscale=off maxHeightPixels=100:50:20
    color=200,100,0 altColor=0,100,200
2   chr19  1000000 1000050  -1.2
3   chr19  1000100 1000150   2.9
.....
```

Fig. 14

- to visualize your chip log2(ratio) intensity data in Genome Browser, click on the UCSC link (4, in the figure above) in the history frame (2 in **Fig. 13**).

***Please note**: if you are using the GFF example file (that contains tiling array data of human chr19) and Genome Browser does not automatically display chr19, insert any chr19 coordinate (e.g. chr19:5,528,000-5,670,000) in the "position/search" window of Genome Browser. A result similar to that shown below in **Fig. 15** will be displayed.*

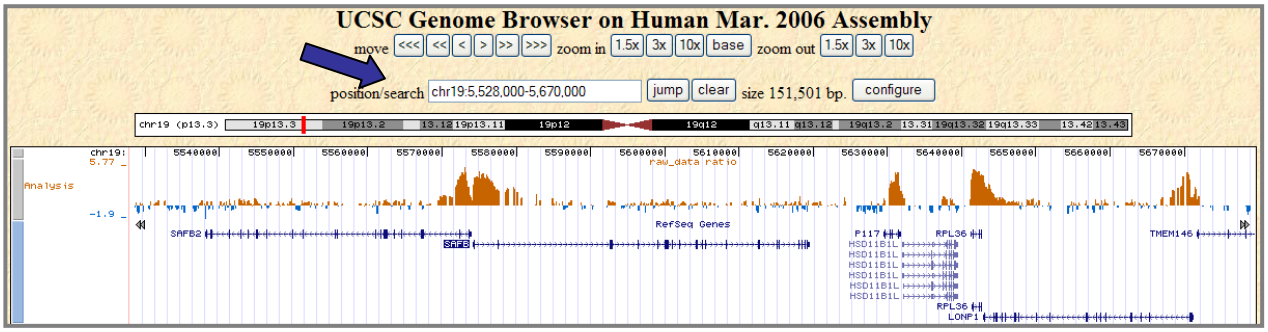


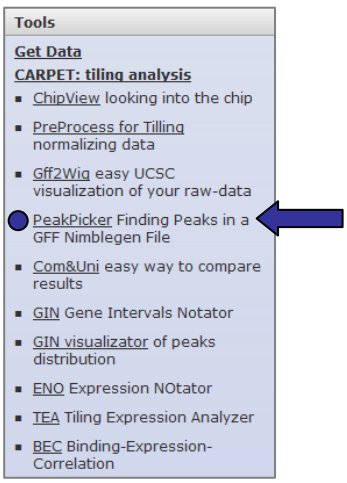
Fig. 15

Once you have viewed the chip surface with ***ChipView*** and profile of your log2(ratio) data in the Genome Browser using the ***GFF2WIG*** tool, you are now ready to start the detailed analysis of your tiling experiments. The following sections describe the analysis pipeline and the CARPET instruments we suggest for ChIP-chip (Section 7) and expression tiling data (Section 8).

7. ChIP-chip data analysis

7.1. Peak identification: PeakPicker

finding Peaks in a GFF ChIP-chip File



PeakPicker is a Perl script that is able to identify enriched regions (peaks) from a ChIP-chip experiment. The ***PeakPicker*** tool utilizes NimbleGen log2(ratio) files in GFF format (or properly reformatted GFF files obtained from other platforms) as the INPUT FILE and identifies regions of enriched signals (peaks), providing as an output a table in GFF format that contains the genomic peak coordinates and scores, alone or with statistical values.

Input GFF file example (*click on the link to download a GFF file example - zipped compressed ≈4.4Mb)*; for more details see the Appendix A of this manual.

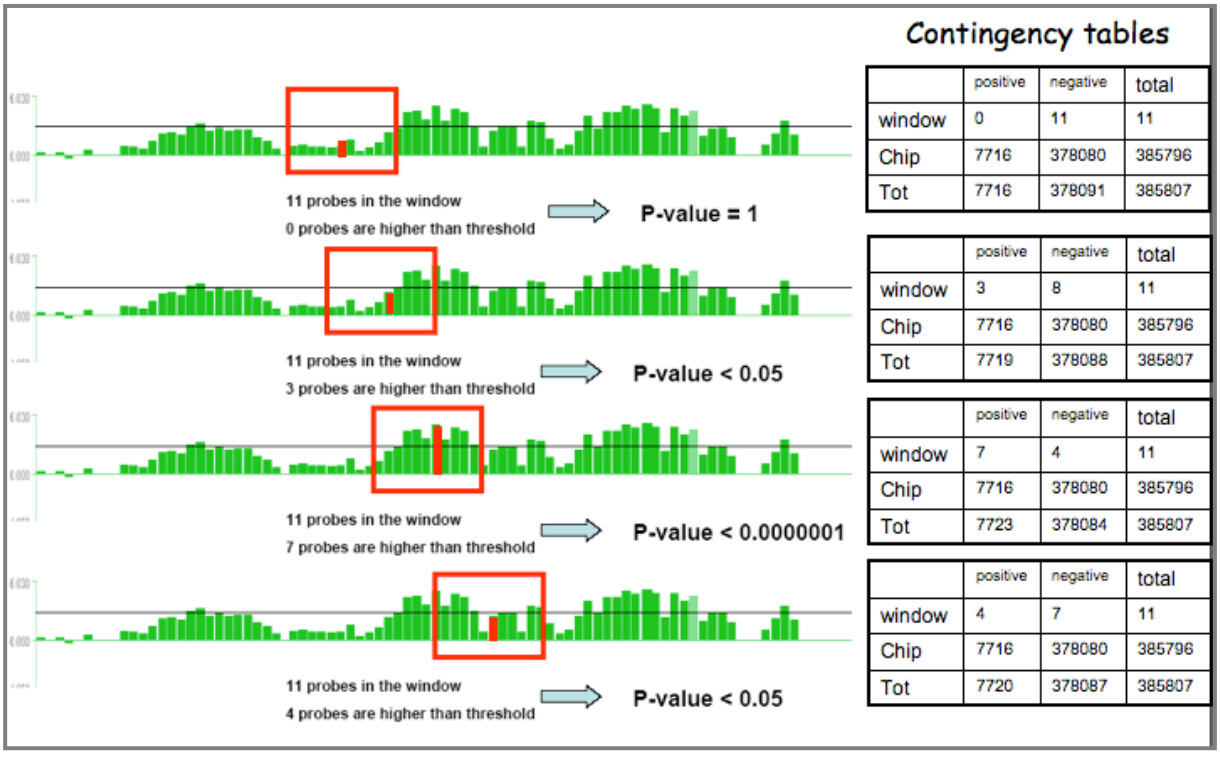
How does ***PeakPicker*** work?
A “peak” is defined as a region in the genome where multiple probes, with a log2(ratio) greater than a user-defined threshold, are located close to each another.

PeakPicker makes two assumptions: i) data are enriched for signals in the positive direction ("one-tailed"); ii) a peak is represented by multiple probes located close to each other in the genome.

PeakPicker makes use of the sliding window statistical approach, essentially as described by Scacheri *et al*. (Scacheri, et al., 2006). A window moves along the array, centering on each probe in turn; in each window Chi squared is calculated (see formula)

$$\chi^2 = \sum \frac{(f(a) - f(e))^2}{f(e)}$$

by building a contingency table for each window/probe position (see **Fig. 16**). A p-value is then assigned.



	positive	negative	total
window	0	11	11
Chip	7716	378080	385796
Tot	7716	378091	385807

	positive	negative	total
window	3	8	11
Chip	7716	378080	385796
Tot	7719	378088	385807

	positive	negative	total
window	7	4	11
Chip	7716	378080	385796
Tot	7723	378084	385807

	positive	negative	total
window	4	7	11
Chip	7716	378080	385796
Tot	7720	378087	385807

Fig. 16

Therefore, ***PeakPicker*** produces a new profile of your experiment (see example in **Fig. 17** and **18**) derived from the "-log2(p-values)" associated to each probe. Profile peak margins are finally delineated by taking into account a user-defined p-value threshold.

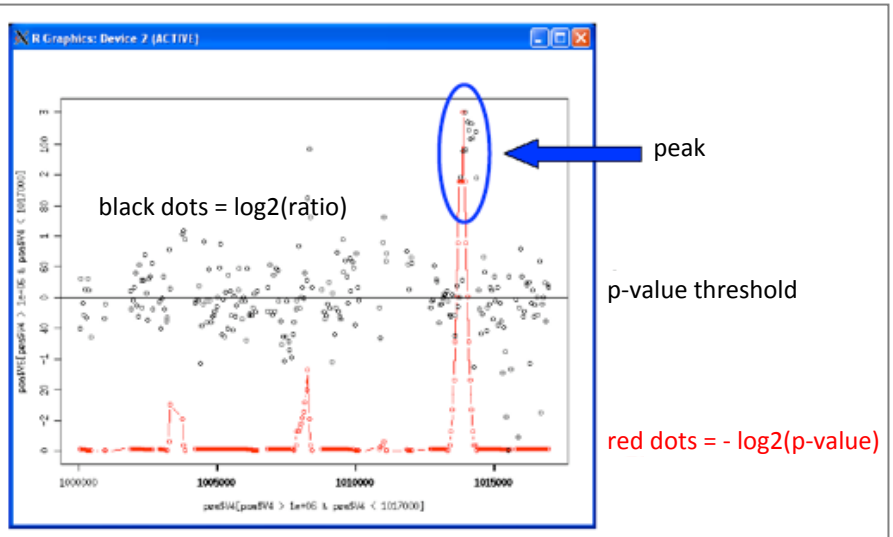


Fig. 17

"-log2(p-values)" are used, instead of plain p-values, for this procedure since they take into account the so-called “neighboring probes effect”, thereby, dramatically reducing the impact of the background signal (see **Fig. 17**).

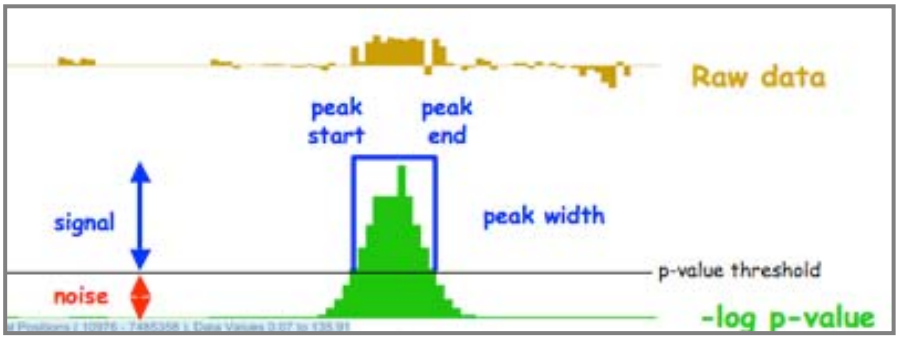
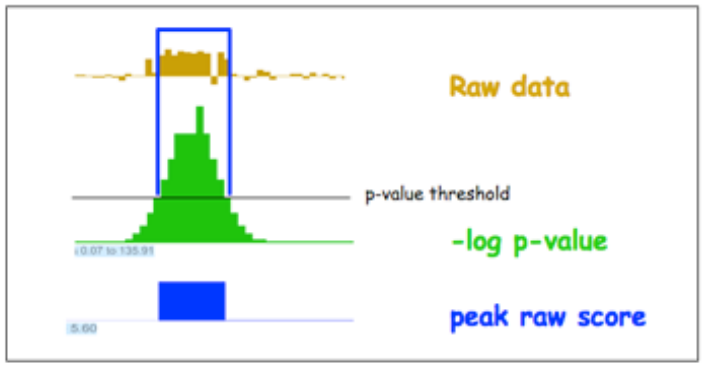


Fig. 18

Alternatively or alongside the statistical p-value calculation for each peak, ***PeakPicker*** also estimates a peak score value that takes into account the length and the intensity of the raw log2(ratio) signals under the peak using the formula reported below.



$$S = \frac{\sum_{1}^{n} \log(ratio)}{n} + \sqrt{n}$$

Average of probe raw values belonging to a peak

Height contribution

Square root of probe number belonging to a peak

Length contribution

Fig. 19

S stands for Score value, $\sum_{1}^{n} \log(ratio)$ is the sum of all log2(ratio) values of the probes under the peak, n is the number of probes under the peak and $\sqrt{n}$ is an empirically defined correction factor for the peak length contribution.

The score value is derived from two main variables: peak height, the mean log2(ratio) value of the probes below the peak; peak length, an empirical correction factor derived from the square root of the total number of probes belonging to a peak.

PeakPicker allows the user to define a number of different parameters: the minimal number of probes that must exceed the defined threshold (fix this parameter in accordance with the expected peak length), the maximum distance permitted between probes, for probes to be considered as contiguous (fix this parameter according to your chip tiling design). The stringency of your analysis can be varied by setting different p-value thresholds. Neighbor-enriched regions can also be joined together (fix this parameter according to the expected peak spread). The analysis output is in the format of a GFF file that can be visualized simultaneously with your raw log2(ratio) data on the UCSC Genome Browser, as shown in the screenshot in **Fig. 20**.

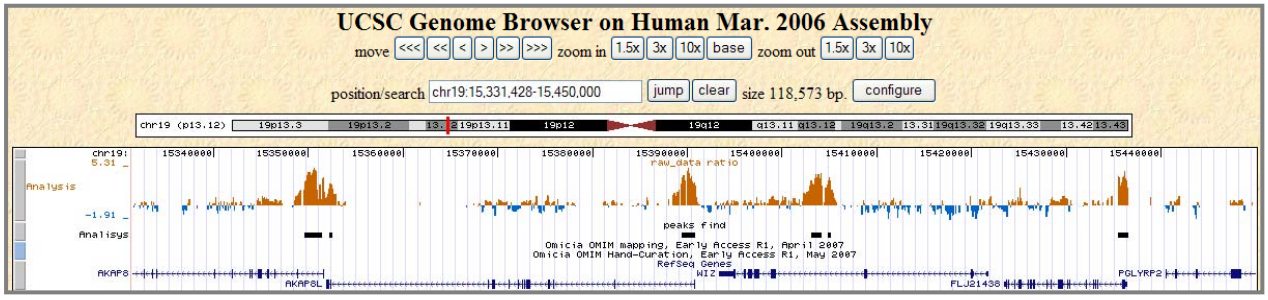


Fig. 20

How to use PeakPicker:

- upload your GFF File using the *Get Data/Upload File* Galaxy tool ensuring that the "GFF" format is selected (*See Section 3*);
- once your file is uploaded, click on the ***PeakPicker*** link from the "CARPET: tiling analysis" tools list;
- select your GFF file from the "Source file:" popup menu;
- you can specify an "Analysis name" that will be appear in the UCSC Genome Browser as the track name (1 in **Fig. 21**);
- set each of the following parameters:
  - analysis type (2 in **Fig. 21**):
    - p-value analysis performs peak determination based on p-value inference (statistical analysis);
    - score analysis performs peak determination based only on the scoring system (scoring analysis);
  - percentile value (3 in **Fig. 21**):
    - used to calculate the threshold rate, based on dataset distribution, to filter out background;
  - -log p-value cutoff (4 in **Fig. 21**):
    - (required only for p-value-based analyses) the cutoff integer used to identify a significant peak;
  - minimal number of probes (5 in **Fig. 21**):
    - minimal number of consecutive probes used to define a peak
  - max distance between two probes (6 in **Fig. 21**):
    - the greatest nucleotide distance (bp) allowed between two probes in order to consider them as adjacent;
  - min distance between two peaks (7 in **Fig. 21**):
    - the minimum nucleotide distance (bp) required in order to consider two peaks as separate entities;
  - window length (8 in **Fig. 21**):
    - length in bp of the window used for statistical analysis.

- once parameters have been set click Execute;

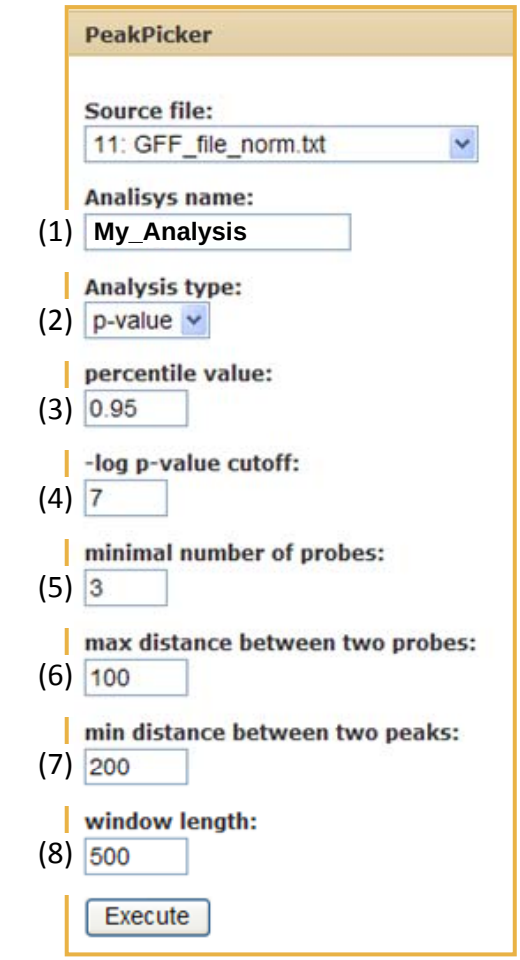


Fig. 21

If you choose to perform the **statistical** analysis, the program will produce an output file in GFF format, similar to that shown below in **Fig. 22** (the first 4 lines only are displayed): in columns 6 (C6) and 9 (C9) the maximum value of –log2(p-value) reached by the probes within a peak and the peak score are reported, respectively.

C1	C2	C3	C4	C5	C6	C7	C8	C9
#chromosome	*Source*	*Feature*	*Start*	*End*	*max -log2(p-value)*			*Score*
track name=Statical Analysis description="PeakPeaker identified peaks" visibility=2								
chr19	NimbleScan	Analysis	20724	21581	25.38	.	.	7.27
chr19	NimbleScan	Analysis	22168	22463	13.95	.	.	5.79
chr19	NimbleScan	Analysis	293061	293367	13.27	.	.	4.79

Fig. 22

If you choose to perform the **scoring** analysis the program will produce an output file in GFF format, similar to that shown below in **Fig. 23** (the first 4 lines only are displayed): in columns 6 (C6) and 9 (C9) the computed peak score is reported.

C1	C2	C3	C4	C5	C6	C7	C8	C9
#chromosome	*Source*	*Feature*	*Start*	*End*	*Score*			*Score*
track name=Scoring Analysis description="PeakPeaker identified peaks" visibility=2								
chr19	NimbleScan	Analysis	22122	22313	5.95	.	.	5.95
chr19	NimbleScan	Analysis	293015	293157	4.81	.	.	4.81
chr19	NimbleScan	Analysis	294081	295180	8.10	.	.	8.10

Fig. 23

To visualize your new results on the UCSC Genome Browser check that the output file (a bed file format) is correctly interpreted by Galaxy, by following the "Edit Attributes" hyperlink (blue arrow in **Fig. 24**, left panel) and verifying that the Database/Build (1 in **Fig. 24,** right panel), Start column (2 in **Fig. 24,** right panel), and End column (3 in **Fig. 24,** right panel) are properly set. Strand column information is not needed in this case (4 in **Fig. 24,** right panel).

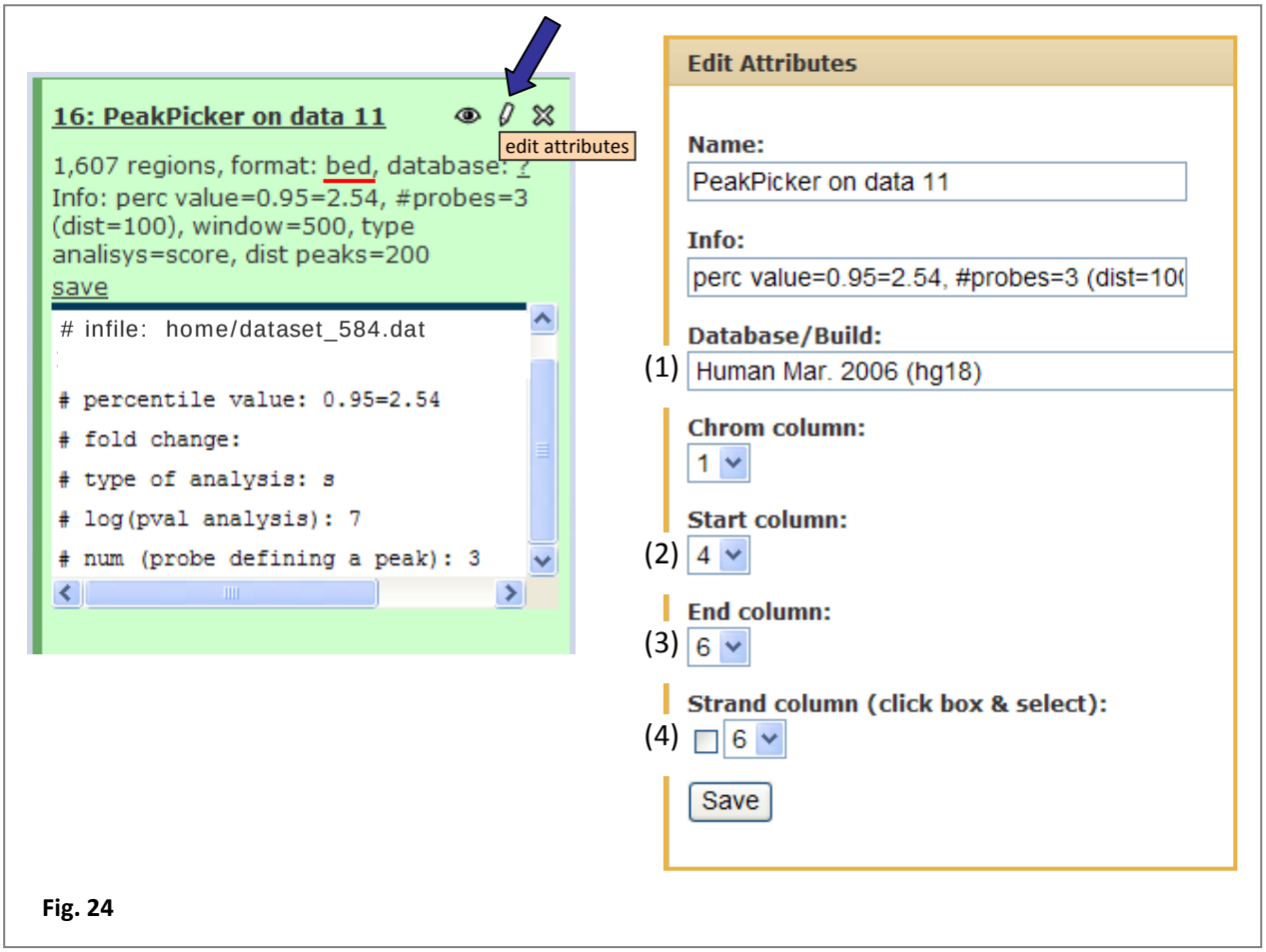
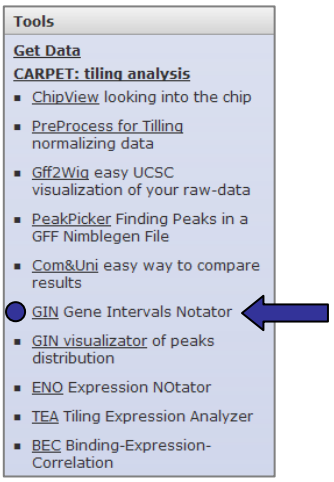


Fig. 24

7.2. Peak annotation: Genomic Interval Notator - GIN



Once you have identified and mapped enriched peaks of binding for your ChIP-chip experiment, you can now determine the relationships between your data and gene loci. The ***GIN*** (Genomic Interval Notator) tool helps you with this task by annotating peak queries using user-defined annotation tables (e.g. RefSeq, UCSC genes, Ensembl Genes) and calculating the relative positions of peaks with respect to transcript associated features (e.g. promoter, exon, intron, intergenic).

How does ***GIN*** work? It uses two files: a GFF file with genomic intervals (i.e. the output file of ***PeakPicker***) and any user-preferred transcript annotation table (e.g. Ref-Seq, UCSC genes) that can be easily downloaded from the UCSC Genome Browser database (see Appendix A of this manual for more information).

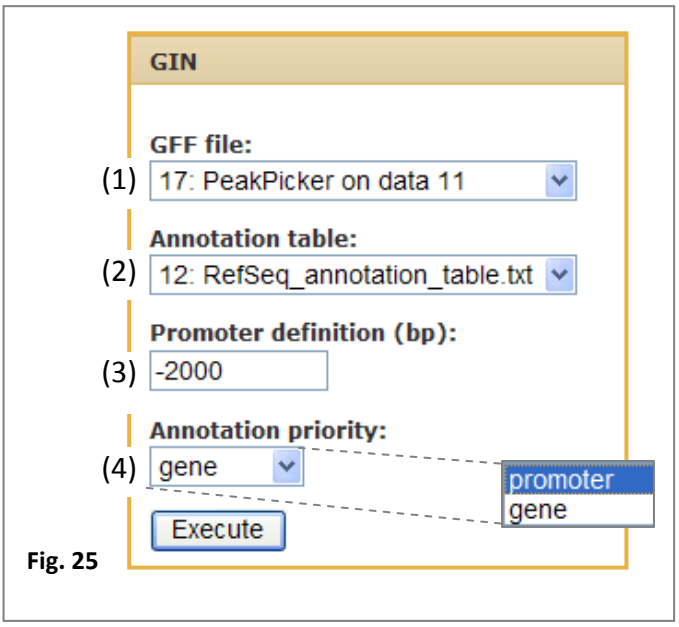


Fig. 25

GIN associates genomic interval queries (i.e. your peaks) with the matching interrogated transcripts. The output for each interval includes the name and absolute chromosome coordinates of the assigned transcriptional units, as well as a call describing its relative position with respect to the transcribed unit (e.g. first exon, fourth intron, promoter) and the relative distance from the putative TSS (Transcription Start Site). Intervals that do not intersect any gene loci are annotated as "intergenic".

The user can arbitrarily define the length (in bps) of the putative promoter regions upstream of each TSS by setting the "Promoter definition" option (3, in **Fig. 25**).

GIN has been programmed to produce a not-redundant annotation table, meaning that each peak will have a unique annotation. Since genes or gene features are often overlapping in genome sequences (e.g. antisense transcripts, bidirectional putative promoters) leading to an ambiguous annotation, the user can give priority to the annotation of genes or of (putative) promoter regions using the "Annotation priority" option (4 in **Fig. 25**). If the "promoter" option is chosen, ***GIN*** first tries to locate a peak within a promoter region. If more than one promoter is found, the peak is associated to the promoter of the closest

transcriptional unit. If the "gene" option is selected, ***GIN*** first tries to locate a peak within an exon.
A flowchart summarizing how ***GIN*** annotates peaks, depending on the priority specified, is shown in **Fig. 26**.

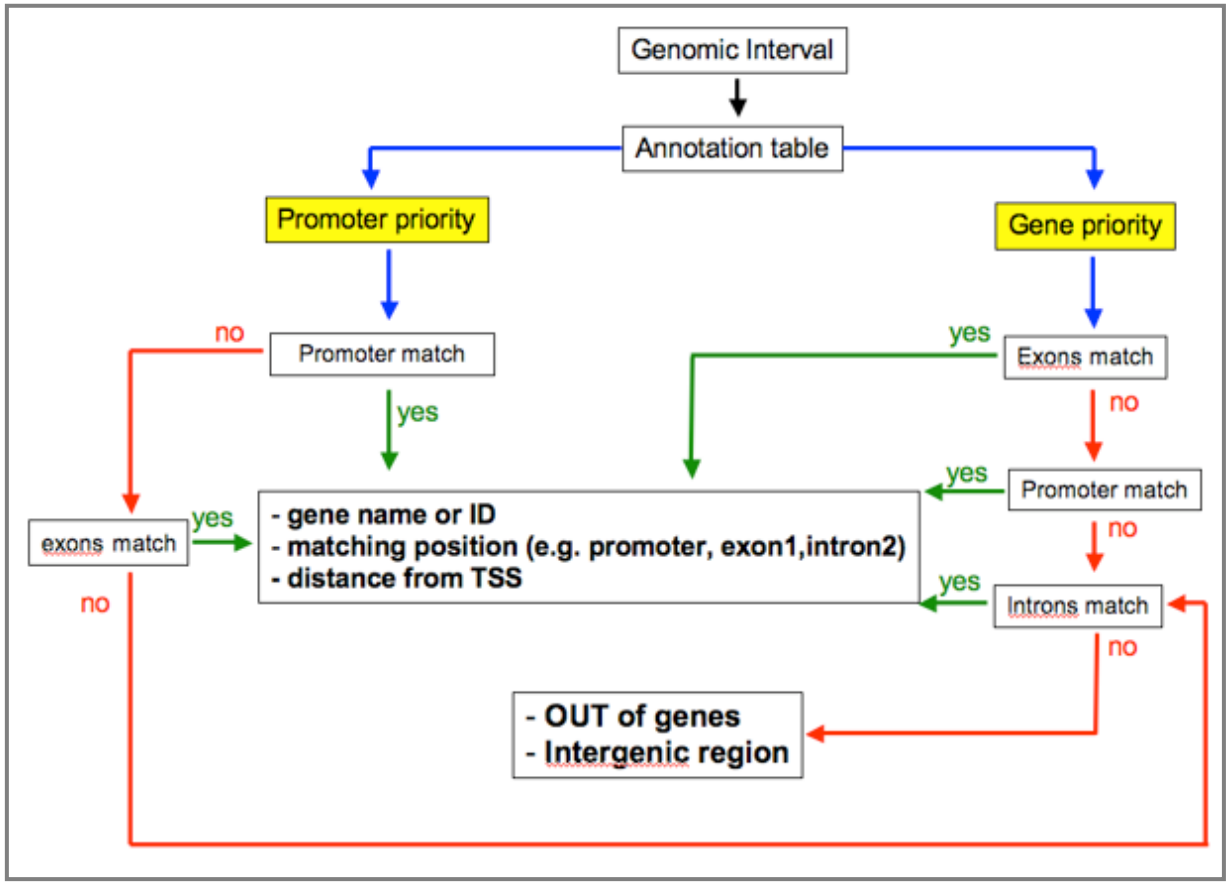


Fig. 26

GIN requires both a peak file and an appropriate "transcript annotation table" in order to correctly annotate peaks (see Appendix A of this manual for more information).

How to use GIN:

- you can either upload your peak GFF File using the *Get Data/Upload File* Galaxy tool, selecting the "GFF" format *(See Section 3)* or use the output file of ***PeakPicker*** directly *(in this case there is no need to upload as the file is already in your History frame)*;
- upload the Transcript Annotation Table you want to use for annotating your peaks (look at Appendix A for details);
- click on the ***GIN*** link from the "CARPET: tiling analysis" tools list;
- select the GFF file and Annotation Table (1 and 2, respectively, in **Fig. 25**) from the corresponding popup menu;
- set the requested parameters:
  - promoter definition (bp) (3 in **Fig. 25**):
    - defines the sequence length upstream of the TSS that you want to consider as the putative promoter region;
  - annotation priority (4 in **Fig. 25)**:
    - promoter – ***GIN*** tries to locate peaks within promoter regions first;
    - gene - ***GIN*** tries to locate peaks within exons first;
- click Execute.

GIN generates an output file containing an annotation table similar to that shown below in **Fig. 27**.

*C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
chr	peakStart	peakEnd	peakScore	GeneID	GeneName	txStart	txEnd	Strand	# exons	relativeAnnotation	Distance TSS
chr19	22122	22313	5.95	AK311358	AK311358	21651	22626	-	2	exon 1	408
chr19	293015	293157	4.8120508	BC028203	BC028203	256574	295791	-	14	intron 2	2705
chr19	457488	457777	5.248068	BC009520	BC009520	458610	470653	+	2	promoter	-977
chr19	458174	458648	5.6043227	AK126170	AK126170	458496	471516	+	4	intronexon 1	-85
chr19	458947	459530	6.9007683	AK125401	AK125401	458499	461372	+	1	exon last	739
chr19	483767	484301	6.3392777	AK024373	AK024373	389420	2034745	-	19	intron 15	1550711

Fig. 27

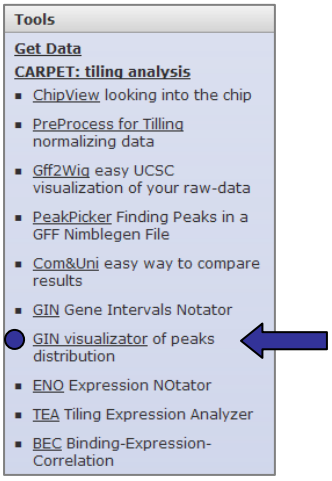
The annotation table contains the following columns:
C1: chr = chromosome name (e.g. chr1, chrY);
C2: peakStart = the start position of the peak defined by the first absolute genomic coordinate mapped on the chromosome;
C3: peakEnd = the end position of the peak defined by the last absolute genomic coordinate mapped on the chromosome;
C4: peakScore = the score or p-value derived from your peak query file;
C5: GeneID = transcript ID;
C6: GeneName = transcript name;
C7: txStart = the start position of the transcript, expressed as the first genomic coordinate;
C8: txEnd = the end position of the transcript, expressed as the last genomic coordinate;
C9: Strand = strand direction of the transcript;
C10: # exons = the number of exons of the transcript;
C11: relativeAnnotation = the position of the peak relative to the transcript;

C12: Distance TSS = the relative distance (in bp) of the center of the peak from the transcript TSS.

Please NOTE: the annotation table does not contain headers, therefore, please refer to the above list to identify columns.

After analyzing your data with ***GIN*** you will have associated your ChIP-chip enriched regions of binding to gene transcripts and know their positions relative to the associated transcript.

7.3. Peak characterization: GIN visualizator



Now that you have an annotated list of the positions of your ChIP-chiped binding peaks relative to gene transcripts, you can visualize this information using the ***GIN visualizator*** tool. For this purpose, the GIN output file can be directly submitted to ***GIN visualizator*** to portray the distribution of peak-intervals around the TSS, as shown in figure below. Knowledge of the location of binding elements with respect to the TSS of the associated transcriptional unit is often helpful when characterizing regulatory DNA binding proteins tested by ChIP.

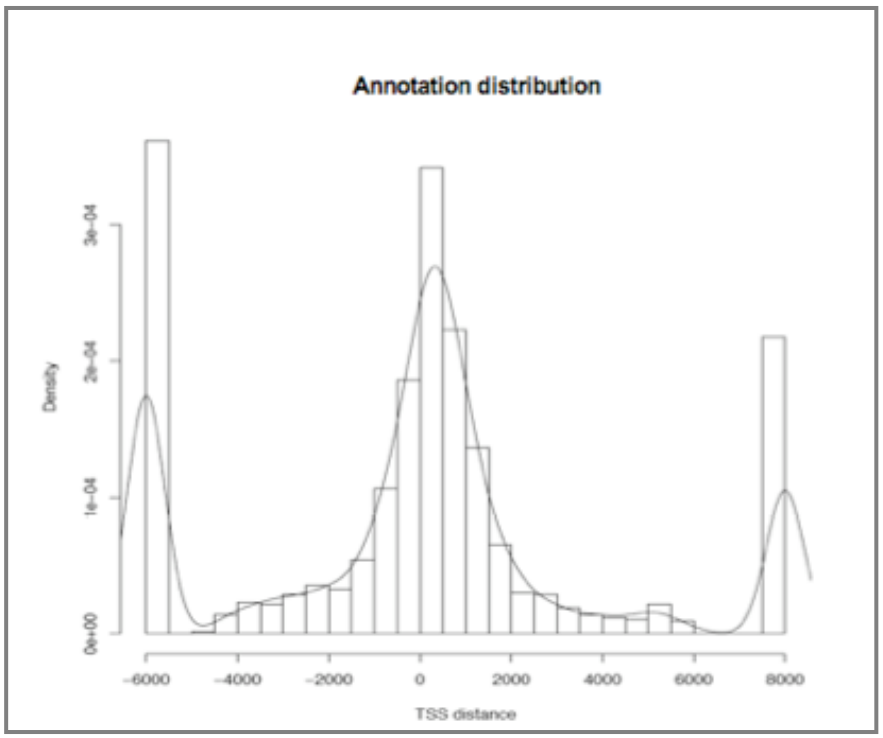


Fig. 28

How to use GIN visualizator:

- click on the ***GIN visualizator*** link from the "CARPET: tiling analysis" tools list;
- select your annotated table (***GIN*** output file) obtained in the previous step from the popup menu (1, in **Fig. 29**);
- set the requested parameters:
  - numerical column for x axis (2, in **Fig. 29**):
    - o indicate the column containing the relative "distance from the TSS" (in bp), i.e. C12 in the GIN output file;
  - number of breaks/bar (3, in **Fig. 29**):
    - o define the number of breakpoints between histogram cells: if a value of '0' is given, breakpoints will automatically be determined;
  - plot title (4, in **Fig. 29**)
    - o input the histogram title;
  - zoom visualization (5, in **Fig. 29**):
    - o define the range around the TSS to be plotted on the X axis (i.e. upper and lower limits): peaks falling beyond these limits will be plotted in the outer histogram bars;
  - include smoothed density (6, in **Fig. 29**):
    - o check the box if you want results to be presented as a continous line graph superimposed on the histogram plot;
- click Execute;

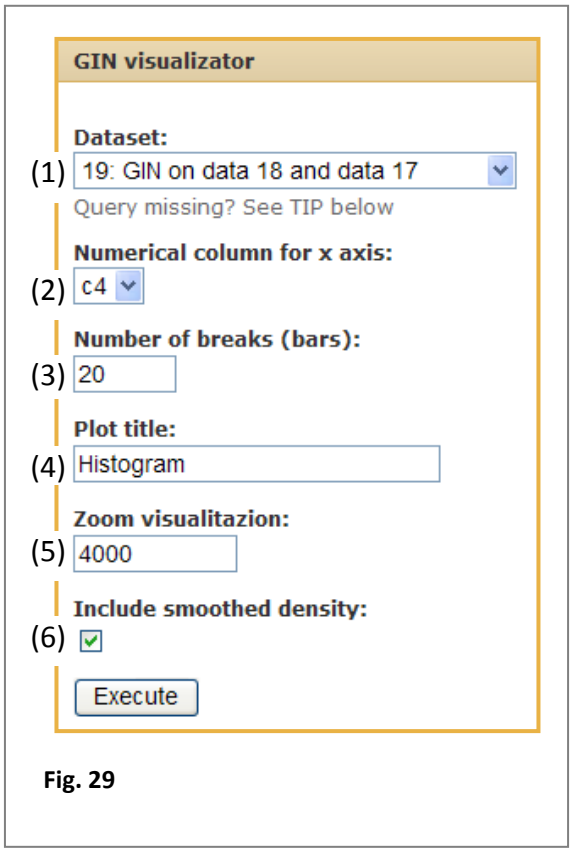
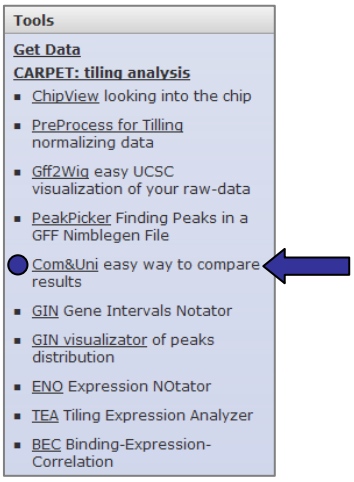


Fig. 29

7.4. Peak comparison: Common & Unique (Com&Uni)

Several binding factors or histone modifications are often ChIPed within independent, but analogous, experiments, making cross-comparison of experiments essential for interpreting results correctly. The ***Com&Uni*** tool allows the user to compare two PeakPicker GFF output files, corresponding to two independent ChIP-chip experiments, in order to identify common and unique features. The program also permits the user to analyse peak flanking regions.



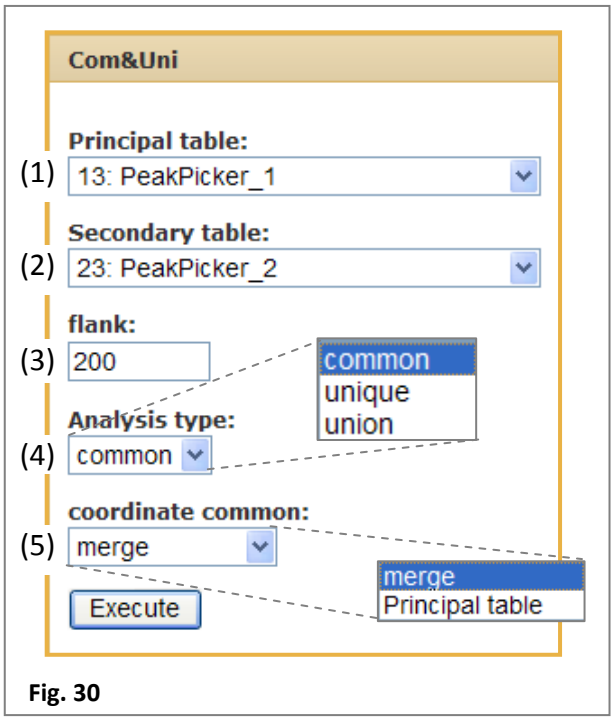


Fig. 30

How to use Com&Uni:

- Click on the ***Com&Uni*** link from the "CARPET: tiling analysis" tools list;
- select your first peak interval file (most likely a PeakPicker GFF output file) from the "Principal table" popup menu (1, in **Fig. 30**);
- select your second peak interval file (most likely a PeakPicker GFF output file) from the "Secondary table" popup menu (2, in **Fig. 30**);
- set the length of the peak flanking regions you wish to analyze (3, **Fig. 30**); type "0" if you want to consider the real peak coordinates only;
- choose the type of analysis you want to perform (combining 4 & 5, in **Fig. 30**);
    - for the **common** analysis (4) you need to choose between the following options (5, in **Fig. 30**):

- the merge option combines information from the two files to determine the outer coordinates of overlapping peaks (see "common/merge" in **Fig. 31**);
- the principal table option uses the first query file as a reference and will give the peak coordinates from the first file that overlap with peak coordinates from the second file (see "common/principal table" in **Fig. 31**);

Please NOTE: *the "common/merge" analysis is "symmetrical" with respect to the two peak query files, meaning that if you invert the files the results will be the same. The same is NOT true for the "common/principal table" analysis, since this type of analysis will give back coordinates corresponding to peaks from the first file only (see the scheme below for clarification).*

- the **unique** analysis (4) will provide the coordinates of peaks from the first file that do not overlap with peaks from the second file ("unique/principal table" scheme in **Fig. 31**);
- the **union** analysis (4) will provide both the common/merge and the unique peak coordinates for the two peak query files (see "union" scheme in **Fig. 31**);

- click Execute.

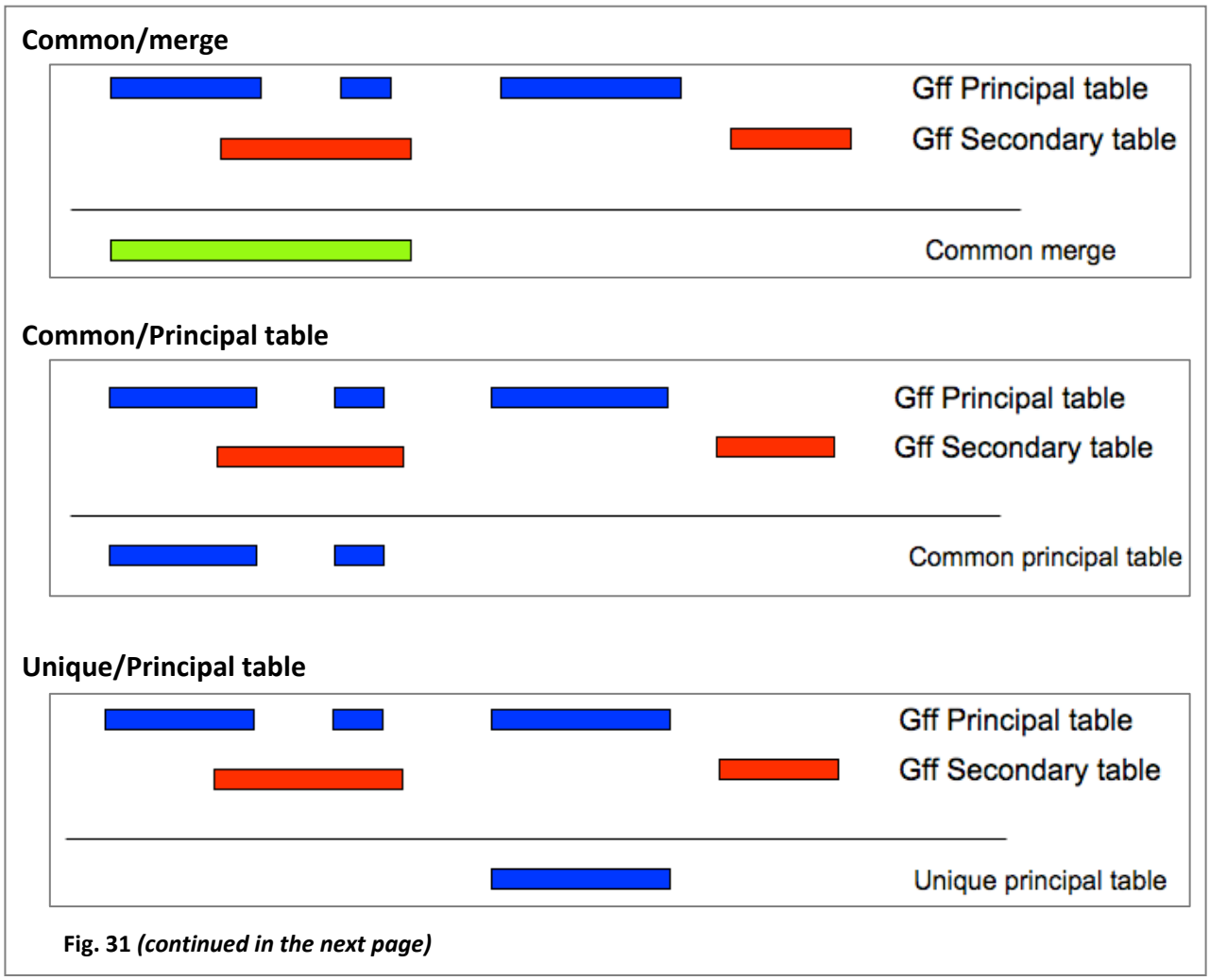


Fig. 31 ***(continued in the next page)***

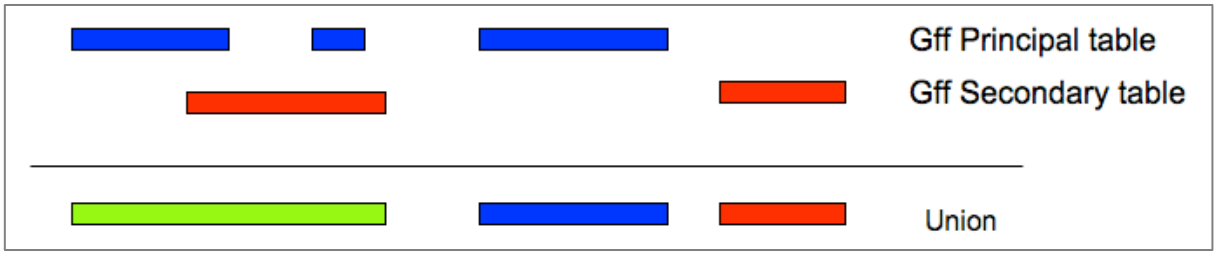
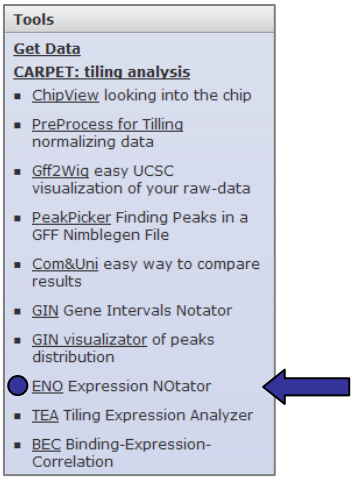


Fig. 31 ***(continued from the previous page)***

8. Expression tiling data analysis

8.1. Expression chip annotation: Expression Notator (ENO)



The first step in analyzing an expression tiling experiment (namely a cDNA hybridation on a tiling platform), is to assign chip probes to their corresponding gene, in particular, to the corresponding exon. The ***Expression Notator (ENO)*** tool annotates each probe on the chip using a user-defined transcript annotation table (e.g. RefSeq, UCSC genes) downloaded from the UCSC Genome Browser database. Since this annotation step relies on chip design, it is necessary to perform this step only once for each transcript annotation table that you want to use. If a probe matches with more than one transcript (e.g. two overlapping antisense transcripts or two different splicing isoforms) the program will take into account the same probe for both transcripts. In **Fig. 32,** a scheme of different annotation features is reported.

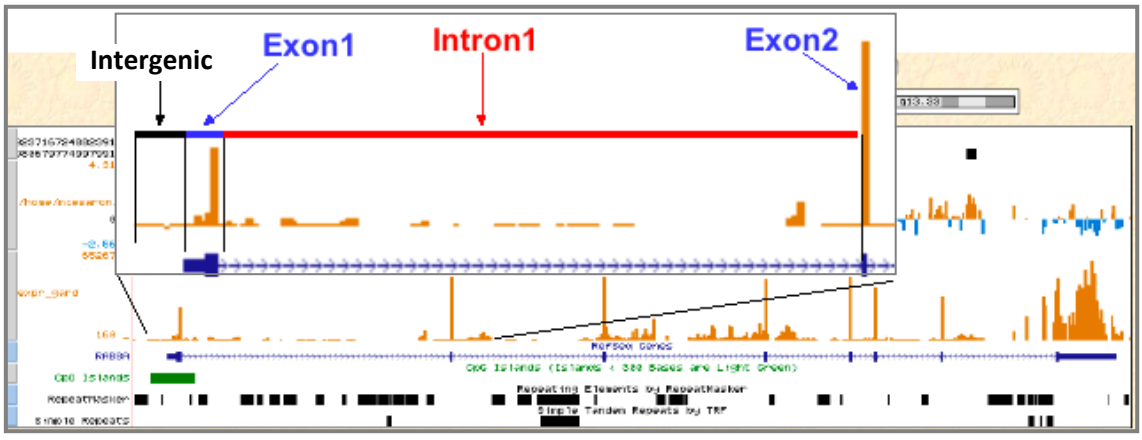


Fig. 32

***How to use ENO**:*

- upload your expression tiling GFF File in the "GFF" format using the *Get Data/Upload File* Galaxy tool (*See Section 3*);
- upload the Transcript Annotation Table you want to use for annotating your chip (see Appendix A for details); this is the same file as that requested for the GIN tool;
- click on the ***ENO*** link from the "CARPET: tiling analysis" tools list;
- select your expression file from the corresponding popup menu (1, in **Fig. 33**);
- select your Annotation Table (2, in **Fig. 33**) from the corresponding popup menu;
- click Execute;

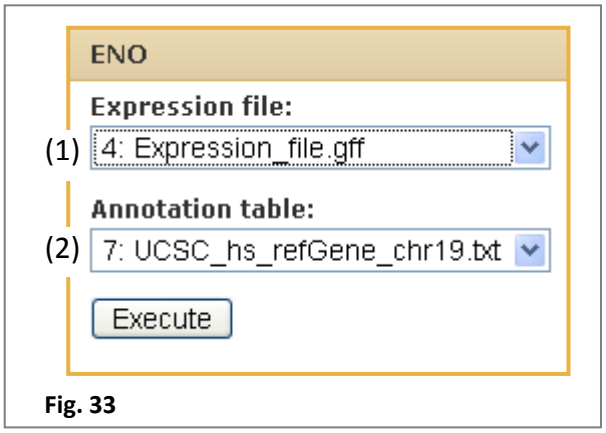
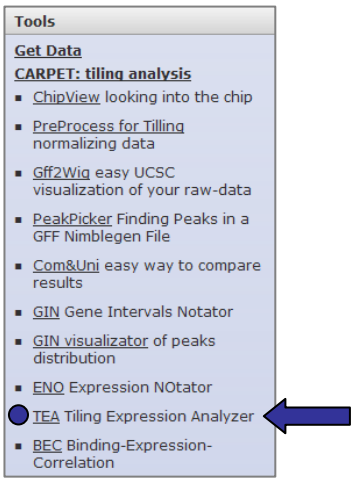


Fig. 33

ENO will generate an output table to be used in the next step of the analysis.

8.2. Analysis of Tiling expression data: Tiling Expression Analyzer (TEA)



The ***TEA*** tool (Tiling Expression Analyzer) performs two different tasks, depending on the number of experiments uploaded. For simple expression estimation, ***TEA*** (starting from your ***ENO*** annotation file) calculates an expression value based on the mean and/or the median of the probe signals associated with the exons of a particular transcriptional unit. In comparison experiments, ***TEA*** analyzes the signal distribution for each gene under different conditions (e.g. untreated vs. treated) and calculates the fold-change and statistical p-values. The user may also choose to operate a False Discovery Rate (FDR) correction (Benjamini and Hochberg, 1995). Many filters can also be applied to the data, e.g. on the raw signals, fold-change and p-values.

TEA utilizes NimbleGen expression files in GFF format and the annotated table produced by ENO as INPUT FILES to generate a table of the expression value for the transcripts studied. When comparing two different conditions, ***TEA*** computes, from the two

distinct expression level evaluations, the fold-change in expression between the two conditions and a statistical p-value.
In this analysis, the expression tiling GFF file should contain raw signals (NOT the log2(ratio)), as usually provided by NimbleGen; alternatively the transformed log2 of the raw signal, obtained after normalizing with the "PreProcess for Tilling" tool, can be use as the input.

How does ***TEA*** work?
For each gene, the program builds the signal distribution of the probes that match the exons.
In a simple expression experiment, the mean or the median signal distribution for each gene is reported in the ***TEA*** output.

In a comparison experiment the signal distribution over the gene exons is compared between the two conditions (1 and 2) and a t-test is performed (see **Fig. 34** below); in addition the user can introduce a FDR correction (Benjamini and Hochberg, 1995).

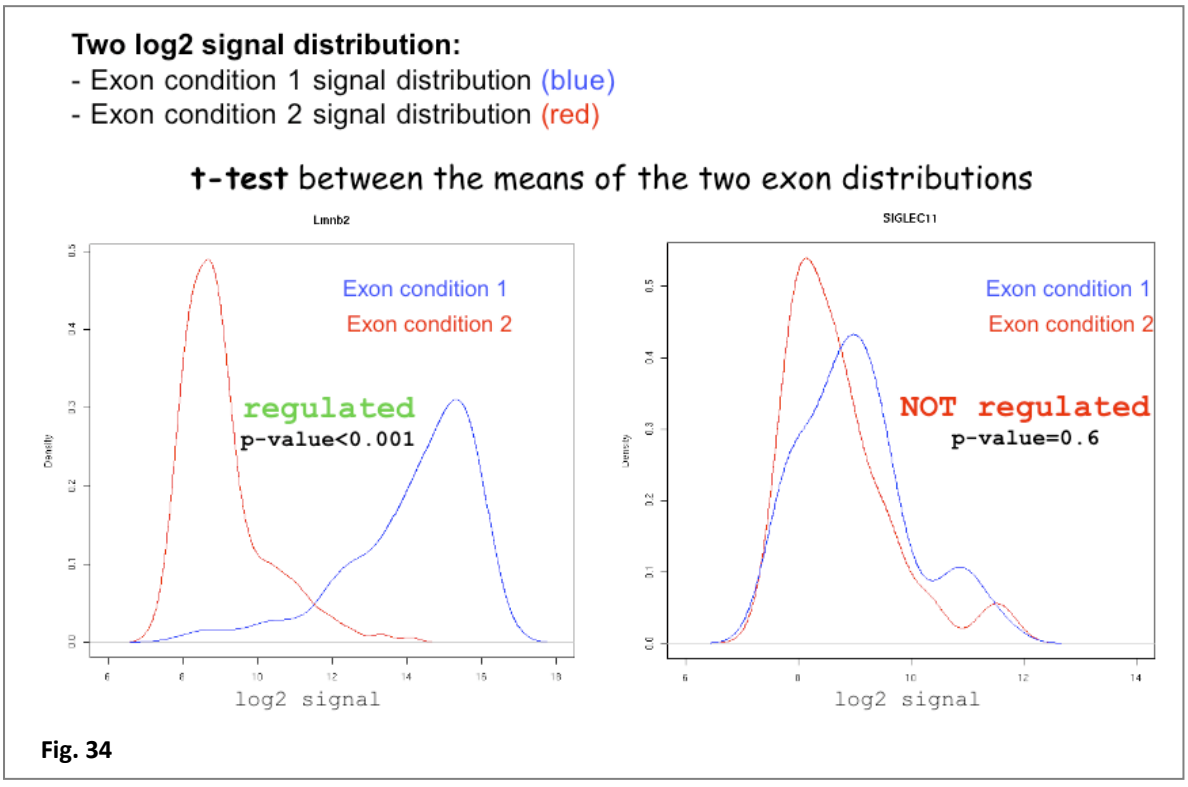


Fig. 34

How to use TEA:

- upload your expression tiling GFF files in the "GFF" format using the *Get Data/Upload File* Galaxy tool (*See Section 3*);
- once the file is uploaded, click the ***TEA*** link from the "CARPET: tiling analysis" tools list;
- choose the type of analysis you wish to perform (1, in **Fig. 35**):
    - o comparison analysis calculates an expression value for each transcript in both conditions analyzed, then

calculates a fold-change for each transcript and a p-value based on a t-Test;
    - o expression analysis calculates an expression value for each transcript derived from the mean or the median of all matching probes;
- select the ENO annotation table, produced in the previous step, from the popup menu (2, in **Fig. 35**);
- select your tiling expression file/s from the corresponding popup menus (3 and 4, in **Fig. 35**): only one file for an expression analysis; two files for a comparison analysis;
- set the requested parameters:
    - o normalization: for a comparison analysis you can choose to apply a quantile normalization between the two chips (not necessary for an expression analysis) (5, in **Fig. 35**);
    - o probe selection (6, in **Fig. 35**:
        - internal exons - only probes representing sequences completely contained within exons are used to calculate the expression value;
        - all exons - probes completely contained within exons together with probes at the boundaries of introns/exons are used to calculate the expression value (probes in intron-exon position usually have a lower signal);
        - last exons - only probes lying within the last exon of each transcript are used to calculate the expression value; this analysis may be preferred for cDNA generated by oligo-dT RT, since the 3' of transcripts are generally better represented;
    - o summary method (7, in **Fig. 35**):
        - mean – the expression level of each gene is calculated based on the mean probe signal value; for the comparison analysis, fold-change will be derived from mean values;
        - median – the expression level of each gene is calculated based on the median probe signal value; for the comparison analysis, fold-change will be derived from median values;
        - both - both methods are used;
    - o fold change cutoff: only transcripts with a fold-change higher than the specified cutoff will be retained (8, in **Fig. 35**);
    - o raw value cutoff (log2): only transcripts with raw values higher than the specified cutoff, in at least one experiment, will be retained (9, in **Fig. 35**).
    - o FDR: you can choose to apply a False Discovery Rate (FDR) multiple test correction (Storey, 2003); if FDR is used a q-value is calculated (10, in the figure below);

- p-value cutoff: only transcripts with p-values less than the specified cutoff will be reported (11, in **Fig. 35**);

- click Execute.

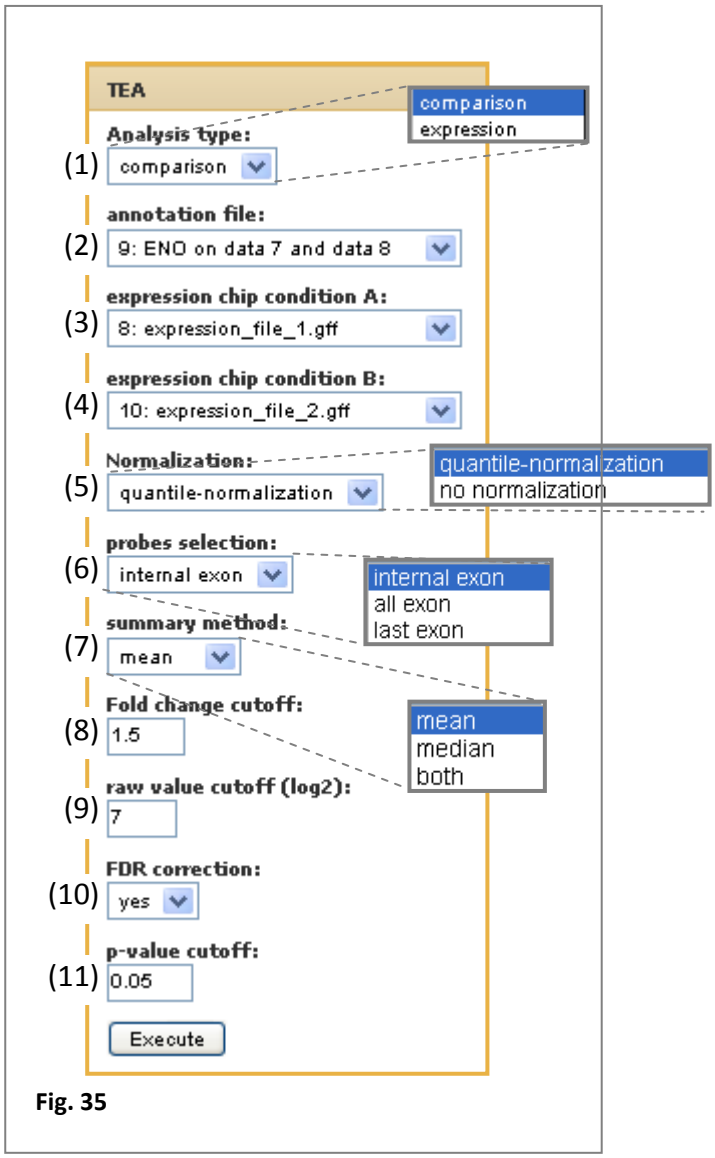


Fig. 35

TEA generates output files similar to those shown below in **Fig. 36**.

for the simple expression analysis

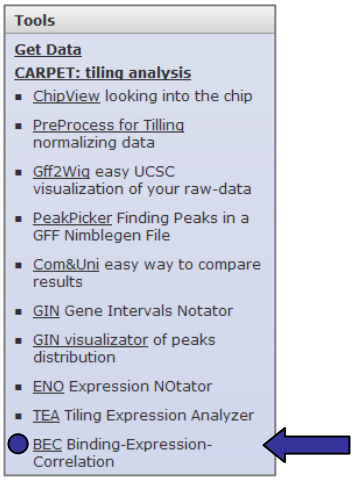
Gene Name	Gene ID	CHR	txStart	txEnd	strand	Mean	Median	# probes
EID-3	NM_152361	chr19	44713469	44715334	-	9.55	9.40	19
PRAM1	NM_032152	chr19	8460940	8473495	-	13.38	12.16	33

for the comparison analysis

Gene Name	Gene ID	CHR	txStart	txEnd	strand	Mean cond A	Mean cond B	FC mean	Median cond A	Median cond B	FC median	p-value	# probes
EID-3	NM_152361	chr19	44713469	44715334	-	9.55	9.45	-1.07	9.40	9.23	-1.12	0.59832	19
PRAM1	NM_032152	chr19	8460940	8473495	-	13.38	10.76	-6.15	12.16	9.64	-5.76	5.04E-14	33

Fig. 36

9. Comparing ChIP-chip and expression tiling data: Binding-Expression Correlation (BEC)



Merging expression and ChIP-chip results can be helpful when formulating hypotheses regarding, for example, the mechanistic implications of the binding of a transcription factor (TF) near to, or within, putative target gene loci. Outputs of ChIP-chip results from PeakPicker and expression data from TEA can be rapidly compared using ***BEC***. For each gene, ***BEC*** gives the number of peaks that match the strict transcriptional unit or the user-defined putative promoter region around the TSS.
Correlations between gene regulation or expression and TF binding can, therefore, immediately be evaluated.

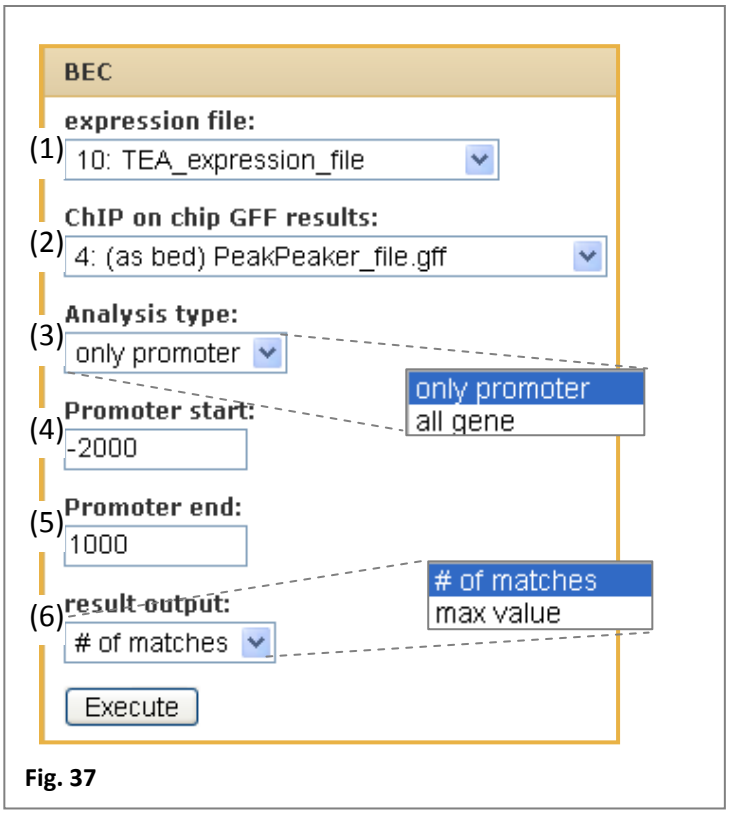


Fig. 37

BEC integrates the results of expression analyses and ChIP-chip analyses. For each transcript, the number of peaks within the promoter region and/or the gene body is calculated.

How to use BEC:

- *(it is presumed that you have already performed the peak identification analysis of your ChIP-chip data using PeakPicker and the analysis of your expression tiling data using TEA)* click on the ***BEC*** link from the "CARPET: tiling analysis" tools list;

- select the TEA result file of the expression tiling analysis and the PeakPicker result file of your ChIP-chip data from the corresponding popup menus (1 and 2, in **Fig. 37**);
- select the type of analysis you wish to perform (3, in **Fig. 37** above):
    - only promoter: only peaks matching the promoter region (defined in length and position by the user) are associated with the transcript;
    - all gene**:** peaks matching the promoter region and those within the gene body are associated with the transcript;
- set the requested parameters:
    - promoter start/promoter end: set the limits around the TSS that you wish to consider as the putative promoter region (4 & 5, in **Fig. 37**);
    - result output (6, in **Fig. 37**):
        - # of matches – if selected, the number of matching peaks will be reported
        - max value - if selected, the highest score of the matching peaks is reported;
- click Execute.

BEC will generate an output file similar to that shown below in **Fig. 38**:

Gene Name	Gene ID	CHR	txStart	txEnd	strand	Mean cond A	Mean cond B	FC mean	Median cond A	Median cond B	FC median	p-value	# probes	# of matches
EID-3	NM_152361	chr19	44713469	44715334	-	9.55	9.45	-1.07	9.40	9.23	-1.12	0.59832	19	0
PRAM1	NM_032152	chr19	8460940	8473495	-	13.38	10.76	-6.15	12.16	9.64	-5.76	5.04E-14	33	2

Fig. 38

References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing, *J Roy Stat Soc B Meth*, **57**, 289-300.

Blankenberg, D., Taylor, J., Schenck, I., He, J., Zhang, Y., Ghent, M., Veeraraghavan, N., Albert, I., Miller, W., Makova, K.D., Hardison, R.C. and Nekrutenko, A. (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly, *Genome Res*, **17**, 960-964.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J. and Nekrutenko, A. (2005) Galaxy: a platform for interactive large-scale genome analysis, *Genome Res*, **15**, 1451-1455.

Scacheri, P.C., Crawford, G.E. and Davis, S. (2006) Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen arrays, *Methods Enzymol*, **411**, 270-282.

Storey, J.D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value, *Ann. Statist.* , **32**, 2013-2035.

Toedling, J., Skylar, O., Krueger, T., Fischer, J.J., Sperling, S. and Huber, W. (2007) Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts, *BMC Bioinformatics*, **8**, 221.

APPENDIX A: File Format and Tables

BED format

The BED format is a conventional file type used to report genomic interval coordinates and their related annotational data; it is one of the formats accepted by the UCSC Genome Browser platform for uploading custom tracks. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is shown in the example below: lower-numbered fields must always be populated if higher-numbered fields are used. For more information, see this site:
http://genome.ucsc.edu/goldenPath/help/customTrack.html#BEDhttp://genome.ucsc.edu/goldenPath/help/customTrack.html#BED.

GFF format

This GFF file format is important in Galaxy and CARPET because it is one of the most commonly used file formats and, in addition, it is used by Genome Browser to define its tracks.
GFF (General Feature Format) files have nine required fields that must be tab-separated. If the fields are separated by spaces instead of tabs, the track will not display correctly in Genome Browser.

A brief description of the standard GFF fields is given below:

Seqname	(or Seq_ID) The name of the sequence. This could be a chromosome, a scaffold or just a string.
Source	The program that generated this feature.
Feature	The name of the type of feature. Some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
Start	The start position of the feature in the sequence. The first base is numbered 1.
End	The end position of the feature (inclusive).
Score	A score between 0 and 1000. If the track line useScore attribute is set to 1 for this annotation data set, the score value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). If there is no score value, enter ".".
Strand	Valid entries include '+', '-', or '.' (for do not know/do not care).
Frame	If the feature is a coding exon, the frame should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
Group	All lines with the same group are linked together into a single item.

Example:
An example of a GFF-based track is shown below.

C1*	C2	C3	C4	C5	C6	C7	C8	C9
#Seqname	*Source*	*Feature*	*Start*	*End*	*Score*	*Strand*	*Frame*	*Group*
chr19	my_chip	feature_1	10000000	10001000	500	+	.	TG1
chr19	my_chip	feature_2	10010000	10010100	900	+	.	TG1
chr19	my_chip	feature_2	10020000	10025000	800	-	.	TG2

* C1, C2, C3, ... C9 represent columns 1, 2, 3, … 9 and their relative position/number; a row beginning with "#" symbol is considered a comment. In the present example, the table is read as a matrix with 3 rows and 9 columns.

For further information on the GFF file format, go to http://www.sanger.ac.uk/Software/formats/GFF.

PAIR FILE Format
NimbleGen "Pair files" (http://www.nimblegen.com/products/methylation/data_guide.html) contain signal intensity data extracted from the scanned images of each array using NimbleScan™, the proprietary NimbleGen's data extraction software. Signal intensities for each probe are saved in Pair files (.txt), which contain eleven columns (C1, C2, C3, ... C11), but essentially the program considers only 3 of them: C10, which contains the chip raw signal, C6 and C7, which contain corresponding coordinate positions on x and y axes of the chip, respectively.

Example:
An example of a PAIR FILE format is shown below.

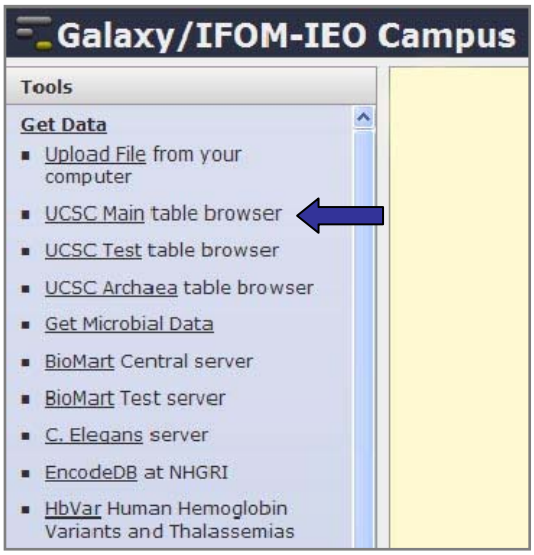
C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
IMAGE_ID	GENE_EXPR_OPTION	SEQ_ID	PROBE_ID	POSITION	**X**	**Y**	MATCH_INDEX	SEQ_URL	**PM**	MM
1251702_635	FORWARD	CHR19	CHR1900P000011001	11001	565	381	64160375	<blank>	5459.89	0
1251702_635	FORWARD	CHR19	CHR1900P000011050	11050	610	656	64160376	<blank>	865.75	0

An example of a NimbleGen "Pair File" is available for downloading *(zipped compressed, ~6.0Mb)*: the file contains tiling data from a ChIP-chip experiment on human chr19 (Human Mar.2006, hg18).

Transcript Annotation Tables
Transcript Annotation Tables used by the ***GIN*** tool to annotate user's GFF peak files can be:

- directly downloaded from the UCSC Genome Browser;
- derived from custom mapping information.

- From UCSC Genome Browser
Transcript Annotation Tables can be directly downloaded from the UCSC Genome Browser by clicking on *Get Data/UCSC Main table browser* in the Galaxy tools frame (see the figure below). For more information on how to perform this task consult the detailed tutorial session at the official "Galaxy Screencasts and Demos" webpage: in particular, see the chapter "1. Interface, Screencast 1.1 - Introduction to Galaxy interface".



When you finally download the transcript annotation table file from the UCSC site, ensure that the "all field from selected table" output format is chosen and that the "send output to Galaxy" option is checked.

It is possible to download many different annotation tables coming from different organisms and databases (e.g. RefSeq, UCSC gene, FlyBase, EST).

An example of Transcript Annotation Table is available for downloading here *(.txt file, ~400.0Kb)*: the file contains the Annotation Table of the human RefSeq of chromosome 19, relating to the Human Mar.2006, hg18, genome assembly.

- From custom mapping information

To derive a Transcript Annotation Table from custom mapping information, you need to respect the following constraints.

The custom Annotation Table must be formatted as follows. **N.B, fields names must be exactly as listed below**.

chrom - Chromosome name (e.g. chr1, chrY).
chromStart - Start position of the annotated feature in the chromosome. (The first base in a chromosome is numbered 0.)
chromEnd - End position of the annotated feature in the chromosome, plus 1 (i.e. a half-open interval).
name - The name you want to give to your annotation in the BED line/track.
strand - Defines the strand, either + or - .
blockCount - The number of blocks (exons) in the BED line/track.
blockSizes - A comma-separated list of the block sizes. The number of items in this list should correspond to blockCount.
blockStarts - A comma-separated list of block starts (e.g. start of each exon in a transcript). All of the blockStart positions should be calculated relative to chromStart. The number of items in this list should correspond to blockCount.

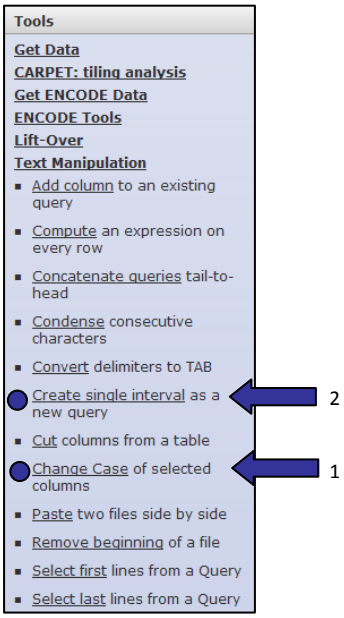
The Annotation Table must always have headers, as shown in the example below.

*C1	C2	C3	C4	C5	C6	C7	C8
chrom	chromStart	chromEnd	name	strand	blockCount	blockSizes	blockStarts
chr19	61678	62596	transcript_1	+	1	918	62596,
chr19	232043	242435	transcript_2	-	6	494,177,58,278,152,151,	232537,233310,233809,238751,239171,242435,
chr19	414359	425983	transcript_3	+	4	1005,114,108,363,	414359,418648,423393,425620,

* C1, C2, C3, ... C8 represent columns and their relative position/number, 1, 2, 3, … 8.

APPENDIX B: Editing (big) files

Joining columns derived from different (big) files in Galaxy



If you want to produce a file that contains two or more columns derived from two different files (e.g. two different replicates, Cy3 and Cy5 signals) you can make use of a combination of the Galaxy "Text Manipulation/Paste" and "Text Manipulation/Cut" tools.

For example, if you want to create a new file that contains both paired Cy3 and Cy5 signals from the two original pair files (see examples below) you can:

- upload your two original source files through the "Get Data" Galaxy section;
- open the "Text Manipulation/Paste" Galaxy tool frame;
- choose your files in the popup menu;
- execute the tools.

Cy3_532

IMAGE_ID	GENE_EXPR_OPTION	SEQ_ID	PROBE_ID	POSITION	X	Y	MATCH_INDEX	SEQ_URL	PM	MM
example	FORWARD	CHR19	CHR1900P000011001	11001	565	381	64160375		1727	0
example	FORWARD	CHR19	CHR1900P000011050	11050	610	656	64160376		3511.78	0
example	FORWARD	CHR19	CHR1900P000011099	11099	452	720	64160377		1562.89	0
example	FORWARD	CHR19	CHR1900P000011148	11148	123	843	64160378		1986.44	0
example	FORWARD	CHR19	CHR1900P000011197	11197	382	268	64160379		2040.33	0

Cy5_635

IMAGE_ID	GENE_EXPR_OPTION	SEQ_ID	PROBE_ID	POSITION	X	Y	MATCH_INDEX	SEQ_URL	PM	MM
example	FORWARD	CHR19	CHR1900P000011001	11001	565	381	64160375		1687	0
example	FORWARD	CHR19	CHR1900P000011050	11050	610	656	64160376		14101	0
example	FORWARD	CHR19	CHR1900P000011099	11099	452	720	64160377		1620.78	0
example	FORWARD	CHR19	CHR1900P000011148	11148	123	843	64160378		715.56	0
example	FORWARD	CHR19	CHR1900P000011197	11197	382	268	64160379		1068.78	0

The program will simply merge the two datasets side by side, as in the examples below.

Cy3- Cy5 side by side

IMAGE_ID	GENE_EXPR_OPTION	SEQ_ID	PROBE_ID	POSITION	X	Y	MATCH_INDEX	SEQ_URL	PM	MM	IMAGE_ID	GENE_EXPR_OPTION	SEQ_ID	PROBE_ID	POSITION	X	Y	MATCH_INDEX	SEQ_URL	PM	MM
example	FORWARD	CHR19	CHR1900P000011001	11001	565	381	64160375		1727	0	example	FORWARD	CHR19	CHR1900P000011001	11001	565	381	64160375		1687	0
example	FORWARD	CHR19	CHR1900P000011050	11050	610	656	64160376		3511.78	0	example	FORWARD	CHR19	CHR1900P000011050	11050	610	656	64160376		14101	0
example	FORWARD	CHR19	CHR1900P000011099	11099	452	720	64160377		1562.89	0	example	FORWARD	CHR19	CHR1900P000011099	11099	452	720	64160377		1620.78	0
example	FORWARD	CHR19	CHR1900P000011148	11148	123	843	64160378		1986.44	0	example	FORWARD	CHR19	CHR1900P000011148	11148	123	843	64160378		715.56	0
example	FORWARD	CHR19	CHR1900P000011197	11197	382	268	64160379		2040.33	0	example	FORWARD	CHR19	CHR1900P000011197	11197	382	268	64160379		1068.78	0

- then, open the “Text Manipulation/Cut” Galaxy tool frame;
- select the file obtained in the previous step;
- select the columns you want to keep in your final output
- execute the script.

Important! Please NOTE: you can use this tool only with files that include columns containing exactly the same features (as in the examples above), and in the same order, otherwise you **must** use other instruments, such as the Galaxy “Join, Subtract and Group” tools set.