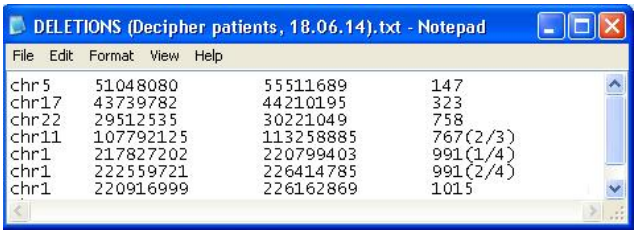
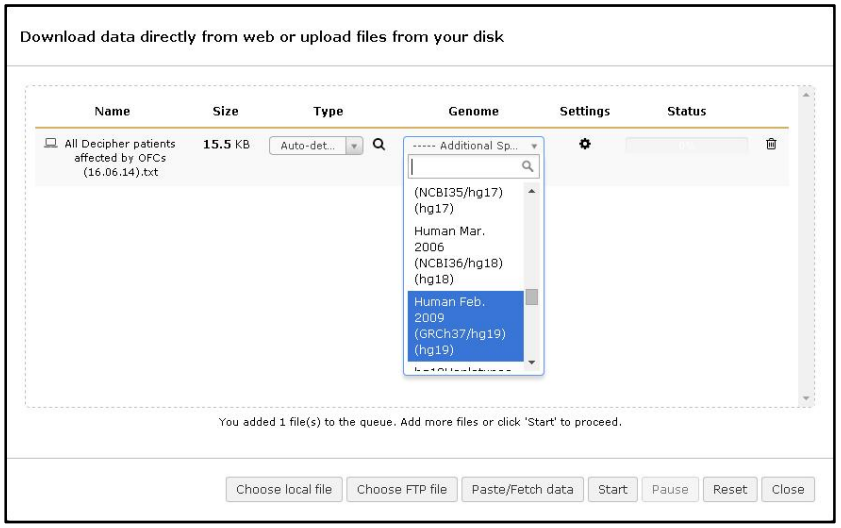
Tutorial

How to identify the regions with the highest number of overlaps using Galaxy

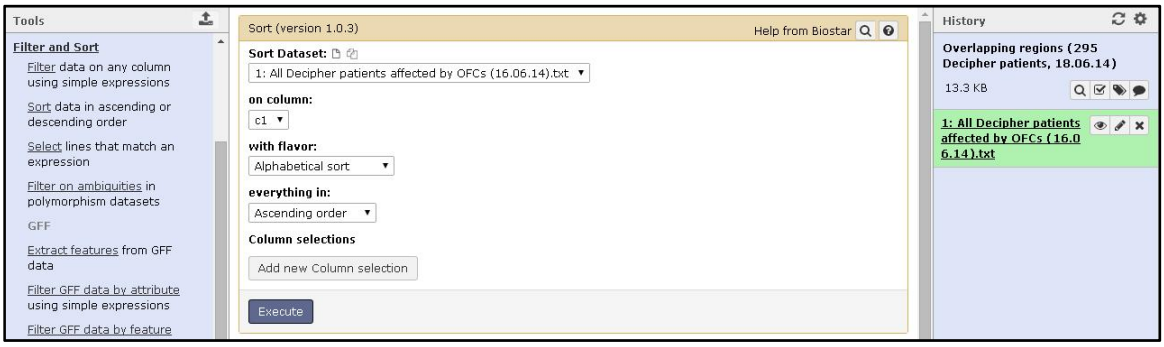
1) First of all, it's necessary to write a BED format file containing all the regions and organized as follows: *chr start end region ID*
(The region ID comes from the patient ID but, in the patients who have more than two CNVs, the patient ID is followed by two numbers in round brackets which indicate the numbers of CNV in that patient. Example: patient 991 → this patient has 4 CNVs → the first is a deletion so in "DELETIONS" BEDfile, you should write this CNV using the patient ID followed by the total number of alterations: *991(1/4)* = first of four CNVs which belong to patient 991).
(NOTE: Create different BED files dividing the regions according to the type of chromosomal abnormality).



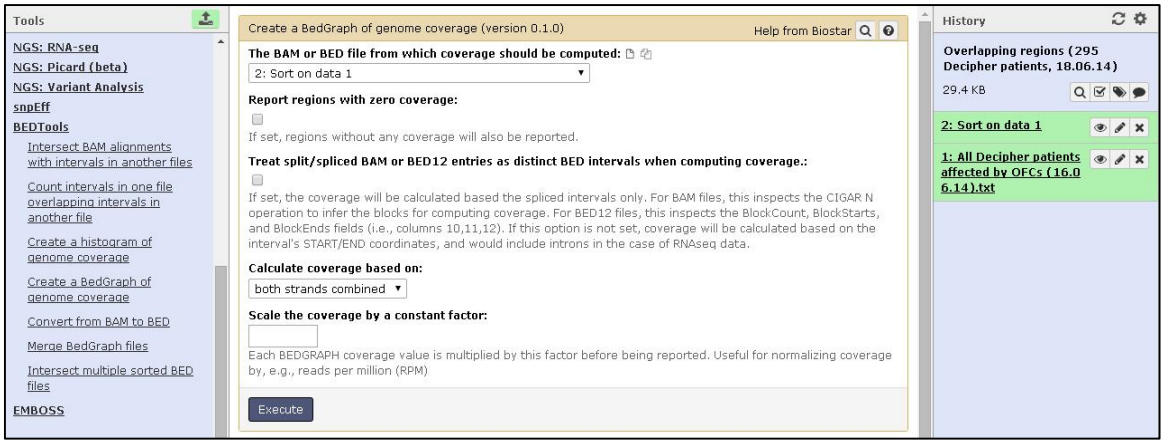
2) Open Galaxy and create a new History (or choose a saved one). Then you have to load up the BEDfile containing the regions: click on "Load your own data" in the "History" column on the right. In the windows appeared, click on "Choose local file" and in the next window (named "Open"), select the file from your computer/external memory and then click on "Open" button.
Now the selected file is displayed in "Name" textbox: in the same row, specify the type of file (in this case select "bed" from the menu in "Type" column) and then specify the type of genome (in this case select the assembly "Human Feb. 2009 (GRCh37/hg19)(hg19)" from the menu in "Genome" column).
Finally, click on "Start" to load your file definitively and wait until the "Status" bar becomes green. When the file is completely and successfully loaded, all the row is highlighted in green: click on "Close" button to close this window.
The loaded file appears as the first job in the "History" (the job title is highlighted in green).



3) A this point, the regions contained in the dataset have to be sorted according to the chromosome number. To do this, in the blue column named “Tools” on the left, search for “Filter and Sort” and click on it; then click on “Sort data in ascending or descending order”.
In “Sort dataset” textbox, select the name of the dataset you have just loaded up (step 2). Then, in “on column” textbox, select the number of “Chr” column of your first file (usually it is the first column: “c1”).
After that, in “with flavor” textbox, choose “Alphabetical sort” option.
Finally, in “everything in” textbox select “Ascending order” and then press “Execute” button.
(NOTE: this sorting is necessary for step 5).

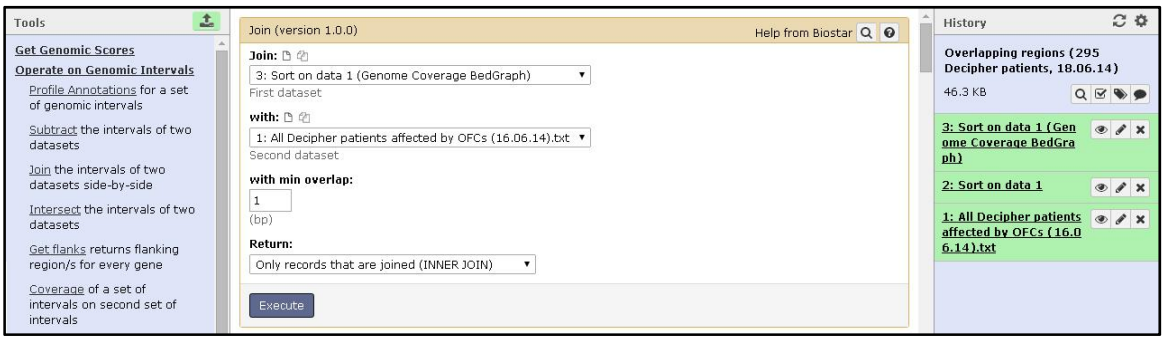


4) A new job appears in the “History” column, automatically named “Sort on data 1” (it is possible to change its name by clicking on the *pencil icon* on the right, near the job title, in the “History” column).
5) After that, in the “Tools” column, click on “BEDTools” and then on “Create a BedGraph of genome coverage”. In this page, select the file obtained from last step, named “Sort on data X”, and take off the check symbol from the option “Report regions with zero coverage”.
Finally, in “Calculate coverage based on” textbox, set the options “both strands combined”.
Don’t modify the other options presents in this page but click directly on “Execute” button.
(NOTE: in the output file there are four columns, referred to the overlapping regions: chr [c1], start [c2], end [c3], number of overlaps [c4] but not the patient IDs!).



6) When the previous job is completed, in the “Tools” menu, select “Join the intervals of two datasets side-by-side” and join dataset 3 [“Sort on data 1 (Genome Coverage BedGraph)”] with dataset 1 [file uploaded in step 1], setting ‘1’ in “with min overlap” textbox and in “Return” select “Only records that are joined (INNER JOIN)”.

(NOTE: in the output file there are eight columns: chr [c1], start [c2], end [c3] of the overlapping region; number of overlaps on each overlapping region [c4]; chr [c5], start [c6], end [c7] of the patient region; patient ID [c8]).



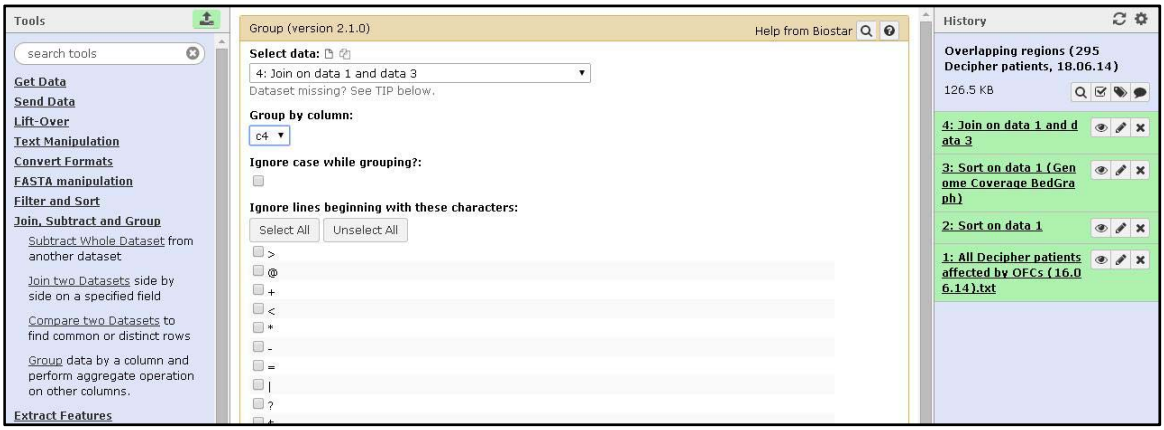
Those regions that overlap, are split in different rows (see the orange arrows), so you need to group the data in order to have all the patients whose regions overlap in the same row. Example:

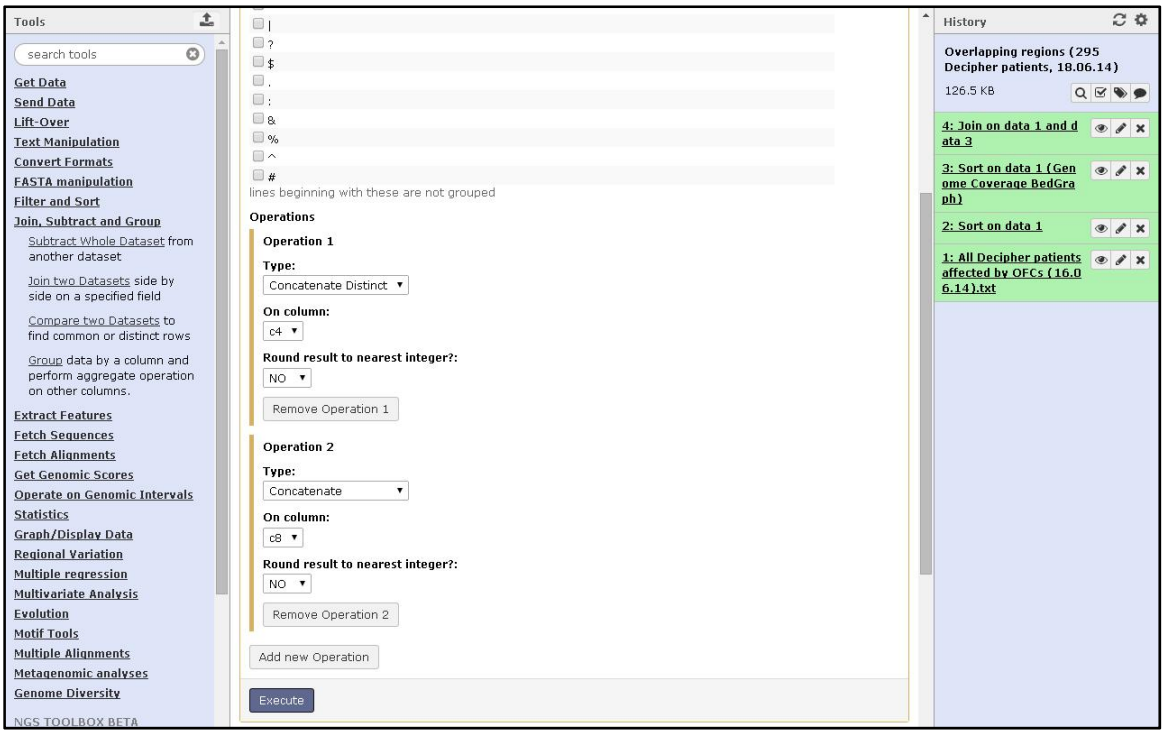
chr1	833831	4061509	1	chr1	833831	4061509	256833
chr1	4795388	5967499	1	chr1	4795388	17364849	2483
chr1	5967499	6023558	2	chr1	4795388	17364849	2483
chr1	5967499	6023558	2	chr1	5967499	6023558	288118
chr1	6023558	17364849	1	chr1	4795388	17364849	2483

7) In the “Tools” menu, click on “Join, Subtract and Group” and then select “Group data by a column and perform aggregate operation on other columns”.
In the new page, select dataset 4 [“Join on data 1 and data 3”], then in the “Group by column” textbox choose “c2”. After that, in the lower part of this page, click on “Add new Operation” button and then insert the first operation (*Operation 1*):
   - Type: Concatenate Distinct
   - On column: c4 [num. of overlaps]
   - Round result to nearest integer?: NO

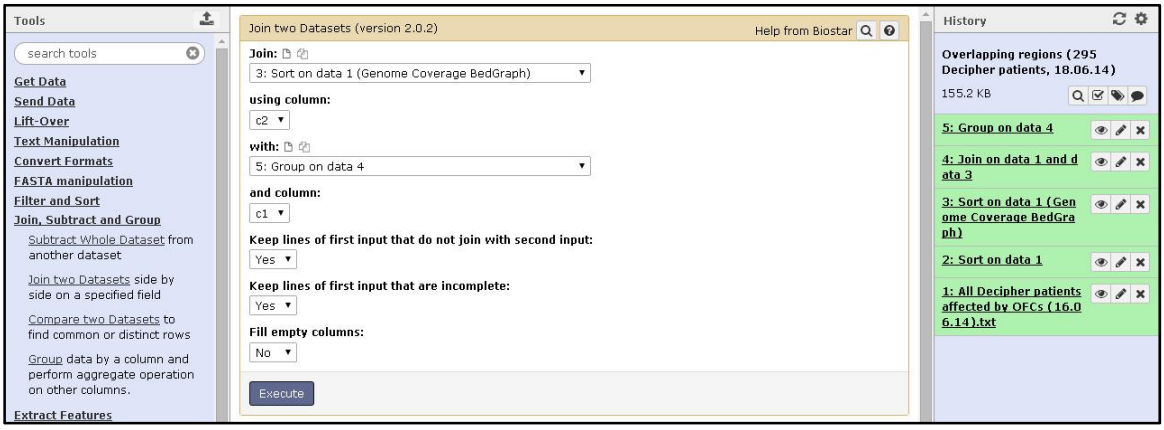
   Then, click again on “Add new Operation” button and set the conditions for the second operation (*Operation 2*):
   - Type: Concatenate
   - On column: c8 [patient IDs]
   - Round result to nearest integer?: NO

   At the end, press “Execute” button to start running this job.





8) Afterward, in the “Tools” menu, click on “Join, Subtract and Group” and then select “Join two Datasets side by side on a specified field”.
At this point, select the dataset 3 [“Sort on data 1 (Genome Coverage BedGraph)”] in “Join” textbox and ‘c2’ in “using column” textbox; then, select the dataset 5 [“Group on data 4”] in “with” textbox and ‘c1’ in “and column” textbox.

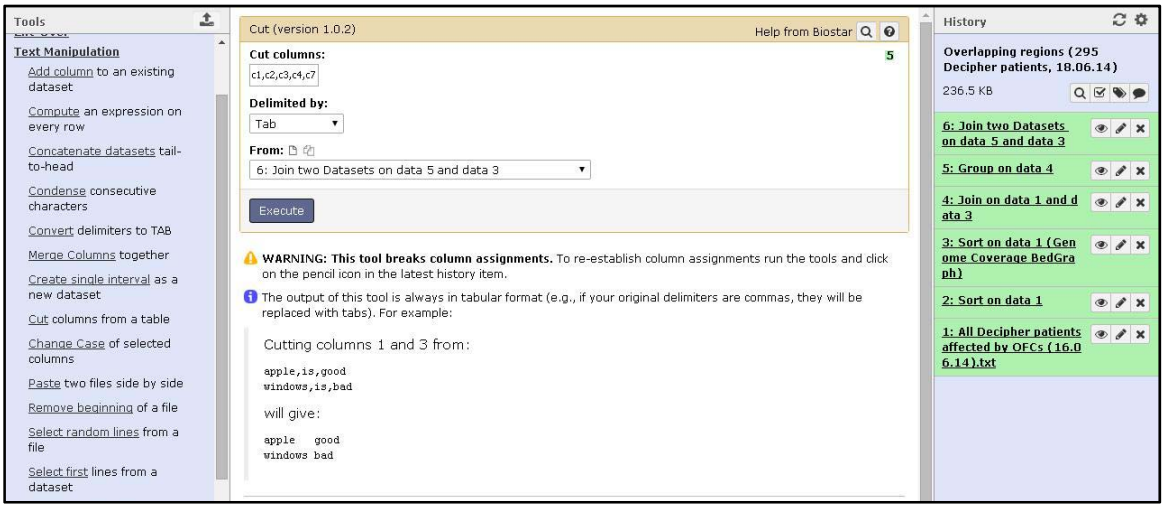


In the output file there are two columns repeated (those columns you used before for grouping, c2=c5 and c4=c6), so you need to delete them.

chr1	833831	4061509	1	833831	1	256833
chr1	4795388	5967499	1	4795388	1	2483
chr1	5967499	6023558	2	5967499	2	2483,288118
chr1	6023558	17364849	1	6023558	1	2483

9) In the “Tools” menu, select “Text Manipulation” and then click on “Cut columns from a table”, to eliminate columns c5 and c6.
Cut columns: - c1 (chr)
- c2 (start)
- c3 (end)
- c4 (num. of overlaps)
- c7 (IDs of patients whose CNV regions overlap)

(NOTE: After that, you will get the right output file you want, composed by five columns: chr [c1], start [c2], end [c3] of the overlapping region; number of overlaps on each overlapping region [c4]; IDs of the patients whose CNV regions overlap [c5]).



10) The output file you obtained from the previous step is exactly the file you are looking for, but their entries are still sorted according to the chromosome number.
In order to get a ranking based on the number of overlaps of each overlapping regions, you can sort again the output file by searching in the “Tools” menu the entry “Filter and Sort” and click on “Sort data in ascending or descending order”.
In “Sort dataset” textbox, select the name of the dataset 7 [“Cut on data 6”], and then, in “on column” textbox, choose the number of the column which contains the num. of overlaps (usually it is the fourth column: “c4”).
After that, in “with flavor” textbox, select “Numerical sort”.
Finally, in “everything in” textbox select “Descending order” and then press “Execute” button.
In the outcome file you can find the ranking of your overlapping regions, based on the num. of overlaps, with their location (chr, start, end) and the patients that overlap on them.

chr1	105300320	105601794	1	289515
chr1	113377785	113676630	2	248354,288279(1/2)
chr1	148820279	149041013	1	276232
chr3	71156	283756	2	289754(1/2),249434
chr4	1075598	1497017	2	2599,259061
chr22	49974766	50715515	3	2384(2/2),1993(2/2),282262

BEFORE

chr22	49974766	50715515	3	2384(2/2),1993(2/2),282262
chr1	113377785	113676630	2	248354,288279(1/2)
chr3	71156	283756	2	289754(1/2),249434
chr4	1075598	1497017	2	2599,259061
chr1	105300320	105601794	1	289515
chr1	148820279	149041013	1	276232

AFTER