Recap with correction - How to identify the regions with the higher number of overlaps using Galaxy

(correction after step 5, shown in green)

My goal is to get a file with five columns organized as follows:

Chr Start End N. overlaps Patient IDs whose CNV regions overlaps

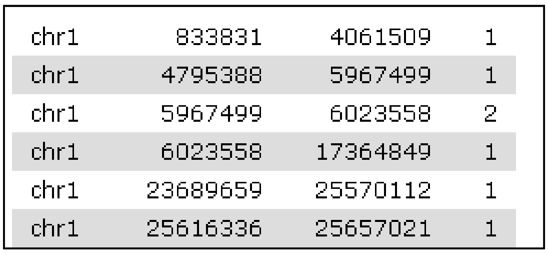
of the overlapping regions

And then sorted this table according to the n. of overlaps to get a kind of ranking (prioritization according to the number of overlaps).

1 – Load the file with the original dataset

2 – Click on "*Sort data in ascending or descending order*" and sort the original dataset according to the chromosomes (necessary for step 3).

3 – Click on "*Create a BedGraph of genome coverage*".
In the output file you can see: chr (column1), start (c2), end (c3), number of overlaps (c4) but not the patient IDs.



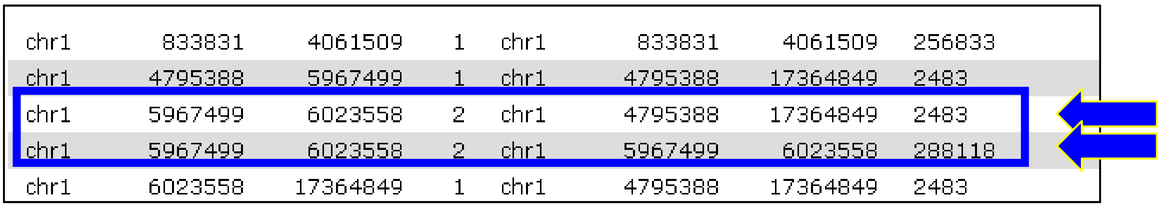
chr1	833831	4061509	1
chr1	4795388	5967499	1
chr1	5967499	6023558	2
chr1	6023558	17364849	1
chr1	23689659	25570112	1
chr1	25616336	25657021	1

Example with five regions

4 – Select *"Join the intervals of two datasets side-by-side"* and join dataset '3' and '1'.
In the output file you can see:

- chr (c1), start (c2), end (c3) of the **overlapping regions**
- number of overlaps in that regions (c4)
- chr (c5), start (c6), end (c7) of the **original regions** (I mean the regions with CNV of the patients)
- patient ID (c8)

Note: the regions that overlap are split in two rows(see blue arrows)... need to group the data!

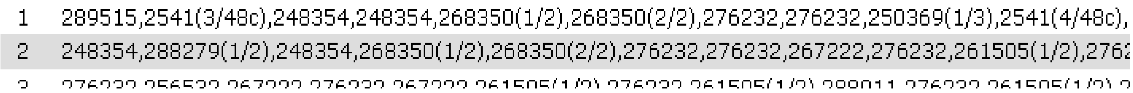


chr1	833831	4061509	1	chr1	833831	4061509	256833
chr1	4795388	5967499	1	chr1	4795388	17364849	2483
chr1	5967499	6023558	2	chr1	4795388	17364849	2483
chr1	5967499	6023558	2	chr1	5967499	6023558	288118
chr1	6023558	17364849	1	chr1	4795388	17364849	2483

Example with five regions

5 – *Group* on dataset '4' by column c4 (number of overlaps) and concatenate on column c8 (patient IDs). **WRONG!**

If I do this, Galaxy pools together all the IDs and the regions… (see the output file down below).
In this case: column1 contains the number of overlaps; column 2 contains all the IDs of the patients whose CNV regions overlap 1 time (or 2 times for the second row, etc…).

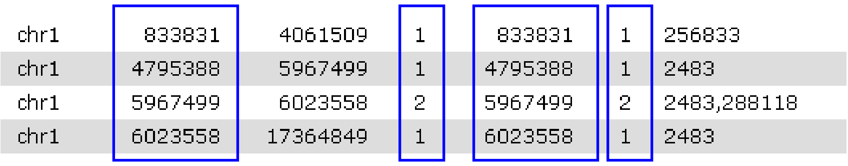


1	289515,2541(3/48c),248354,248354,268350(1/2),268350(2/2),276232,276232,250369(1/3),2541(4/48c),
2	248354,288279(1/2),248354,268350(1/2),268350(2/2),276232,276232,267222,276232,261505(1/2),2762
3	276232,256532,267222,276232,267222,261505(1/2),276232,261505(1/2),288011,276232,261505(1/2),2

POSSIBLE SOLUTION (that works!)

5 – Group on dataset '4'by column c2 (start of the overlapping regions) with two operations: concatenate distinct on c4 (num of overlaps) and concatenate on c8 (patient IDs).

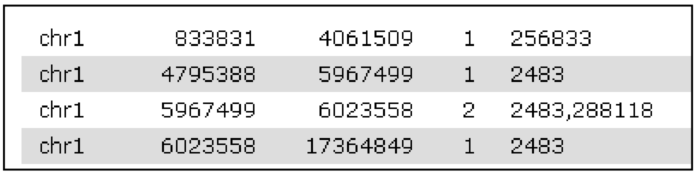
6 – Join datasets '3' (BEDGraph), using c2, and '5', using c1.
In the output file there are two columns repeated (those columns I used before for grouping).



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

7 – Click on "*Text Manipulation*" and then 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 (patient IDs whose CNV regions overlap)

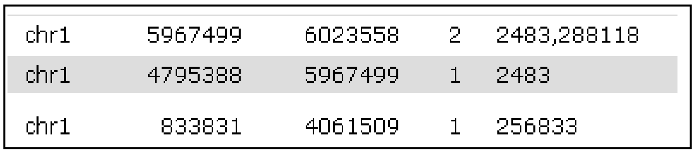
This is the correct output file:



chr	start	end	n. overlaps	IDs
chr1	833831	4061509	1	256833
chr1	4795388	5967499	1	2483
chr1	5967499	6023558	2	2483,288118
chr1	6023558	17364849	1	2483

But the output file entries are still sorted according to the chromosomes and not to the number of overlaps… so I need to sort the dataset again!

8 – Sort according to the num. of overlaps to get a ranking (a kind of prioritization), using dataset '7' on column c4 (num. o f overlaps).
The resulting output file it's exactly what I want (chr / start / end / n. of overlaps / IDs):



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