Post-GWAS using CropGalaxy

QTL of Interest: qDTF2.2

See the workflow here:
http://cropgalaxy.excellenceinbreeding.org/u/vjuanillas/h/postgwas-hands-on

---

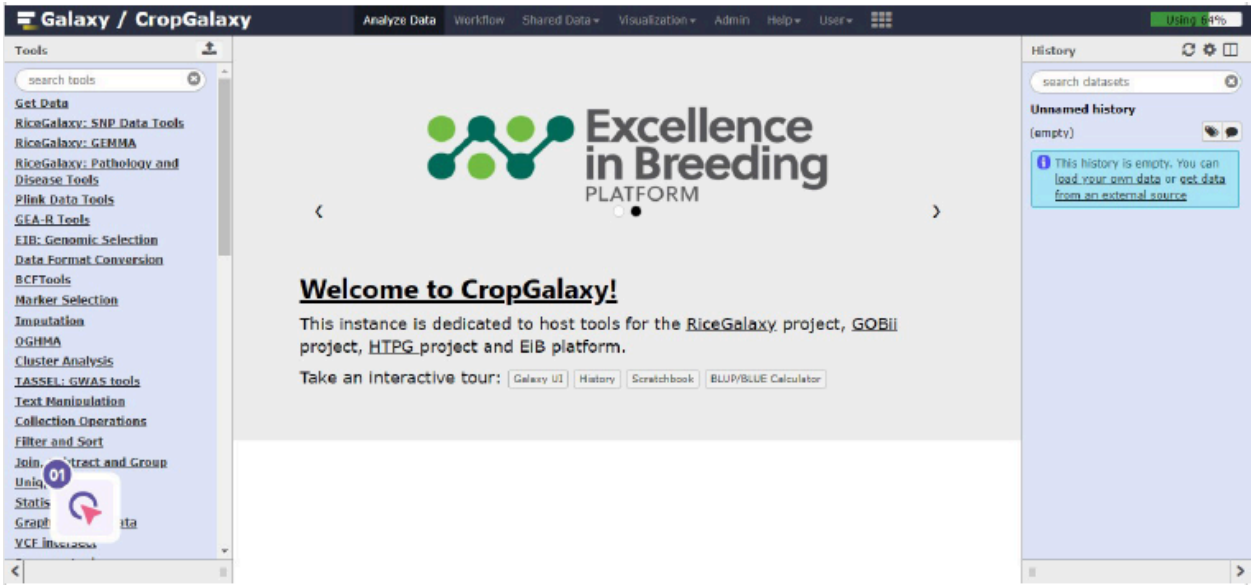
Part 1: Getting the candidate genes from a region of interest

1.) Go to http://cropgalaxy.excellenceinbreeding.org

On your browser, type "***http://cropgalaxy.excellenceinbreeding.org***". This will bring you to the CropGalaxy landing page.

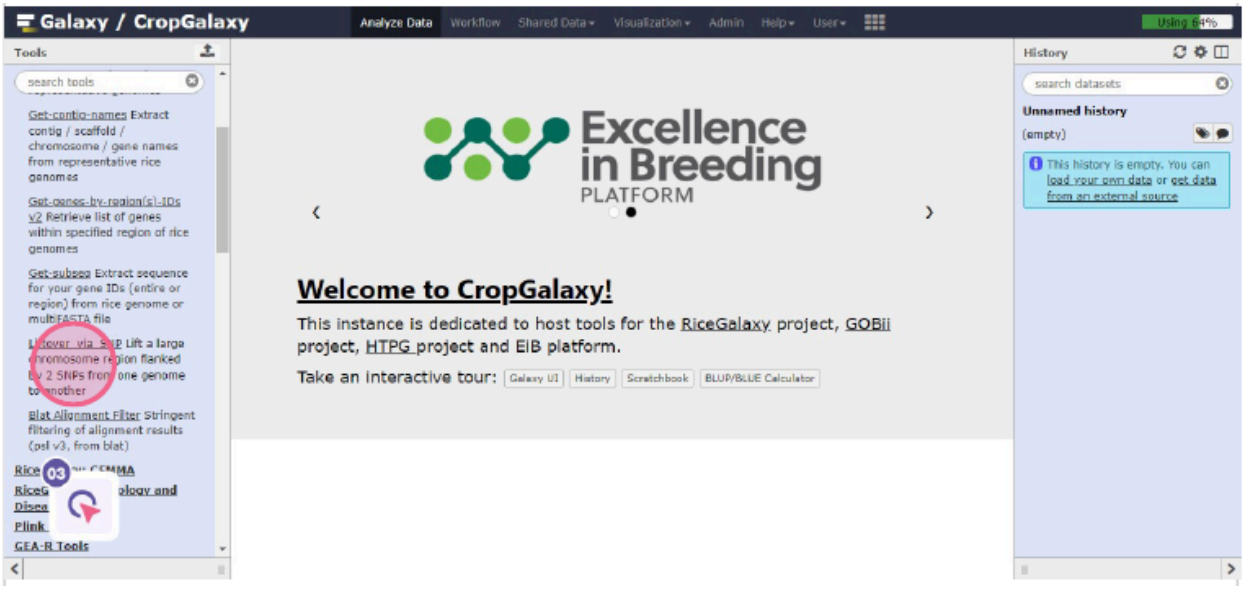
You will be greeted by the CropGalaxy landing page where you will see three main panels:

- ☐ The leftmost panel contains all the tools available for you to use.
- ☐ The rightmost panel displays the current history of the analysis. A history consists of datasets derived from each time a tool is run.
- ☐ The middle panel serves as an analysis panel which displays the tool parameters, as well as the contents of a dataset when the user selects a dataset to view.



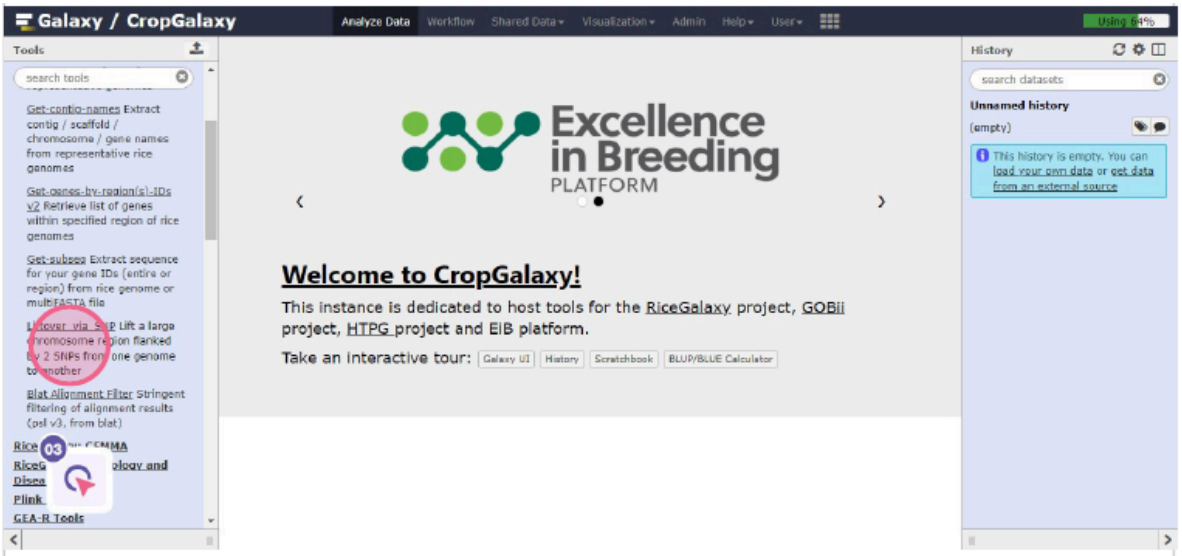
2.) SNP Data tool suite

On the tool panel, select **RiceGalaxy: SNP Data tool** suite. This set of tools contains the specific tool we are after. In the case where you do not know where to find the tool, you can always type the tool name on the "**Search**" box.



3.) Liftover_via_SNP

Liftover via SNP is a specialized alignment tool to find the location of large genome regions of interest from a source genome to another genome. It uses blat software as the alignment engine, and the output is blat **PSL** format (For more information about BLAT, please read the blat manual). The default parameters of blat should work fine for rice genomes.



4.) Set-up Lift-Over parameters

We now start our analysis by running lift-over from Nipponbare to N22 variety (aus), which, in this example, is more related to the donor variety than Nipponbare (temp japonica).

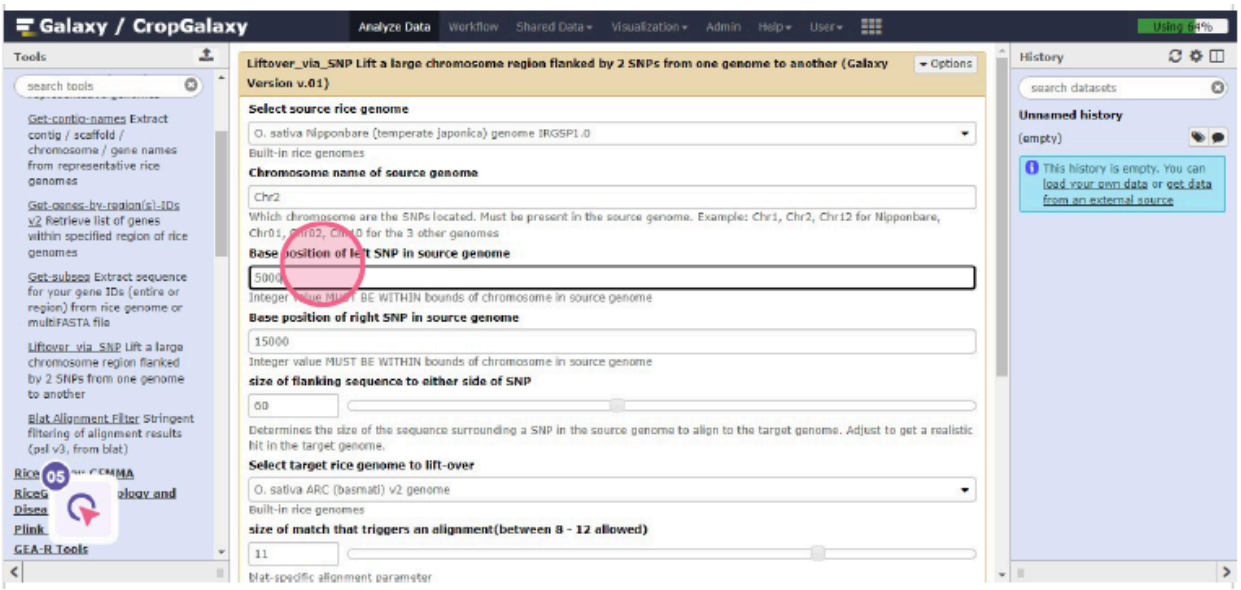
a. **Select Nipponbare as the source genome.**

b. **Type in the coordinates of the QTL: chromosome** (in this case, **Chr2**), **left SNP position** (**22001414**) and **right SNP position** (**22831782**) coordinates in the tool's form **EXACTLY** as it appears here.

Note that the Chromosome name must be correctly specified, otherwise, the tool will produce an error.

c. **Select N22 as the target genome.**

d. **Leave the other settings to default for now.**

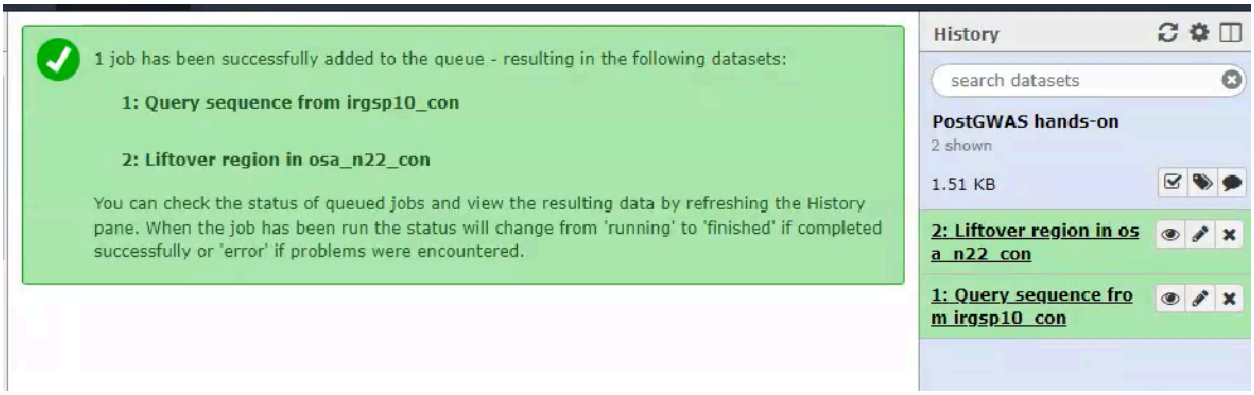


5.) Name your history: PostGWAS hands-on

Before we proceed to running the tool, let us first annotate this analysis by naming our history as "**PostGWAS hands-on**".

6.) Execute Lift-Over tool

Now we are ready. Just click on the "**Execute**" button to run the tool.

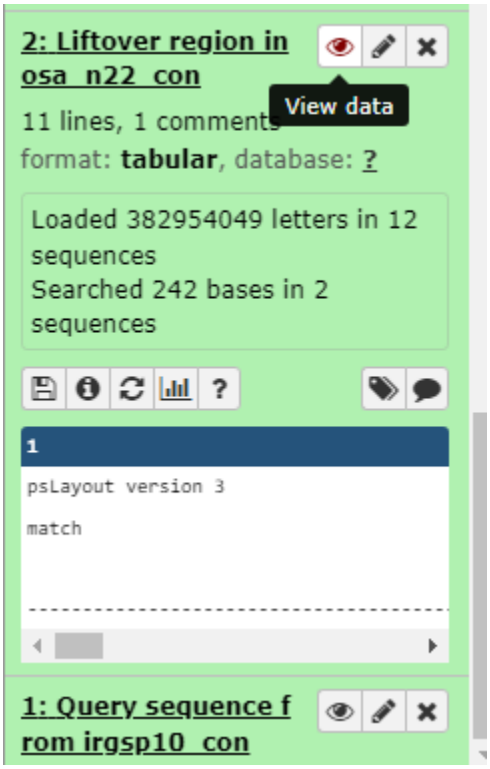


7.) Lift-over tool output files

The Lift Over tool outputs two files: the **query sequences from the source genome**, Nipponbare, and t**he blat alignment output** where you can find the coordinates where these sequences are aligned in the target genome, N22.

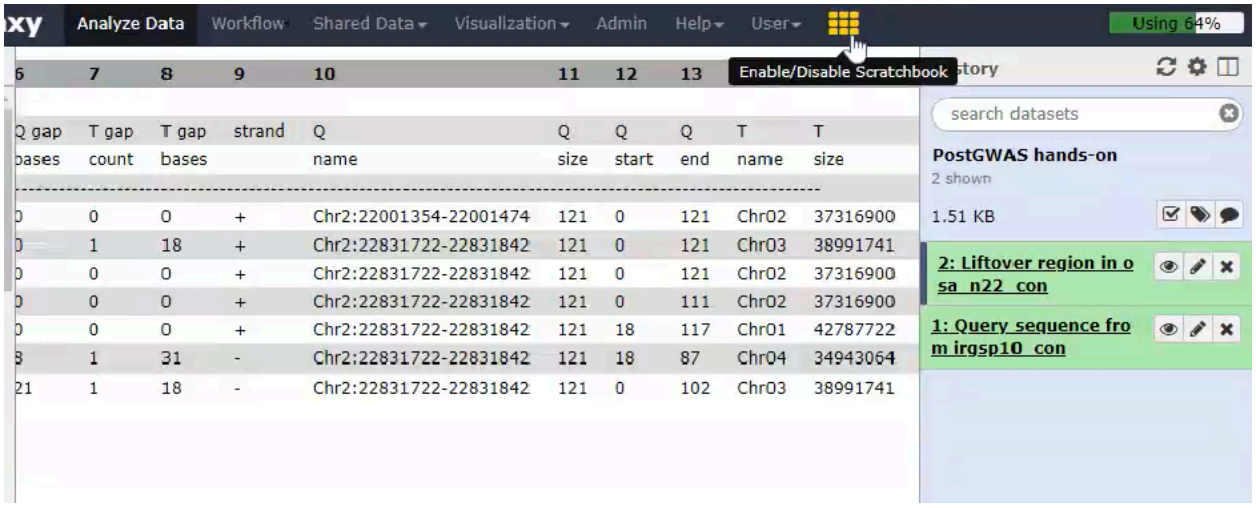
8.)Viewing a dataset

To view a dataset in the history, click the "**View data**" (eye) icon.

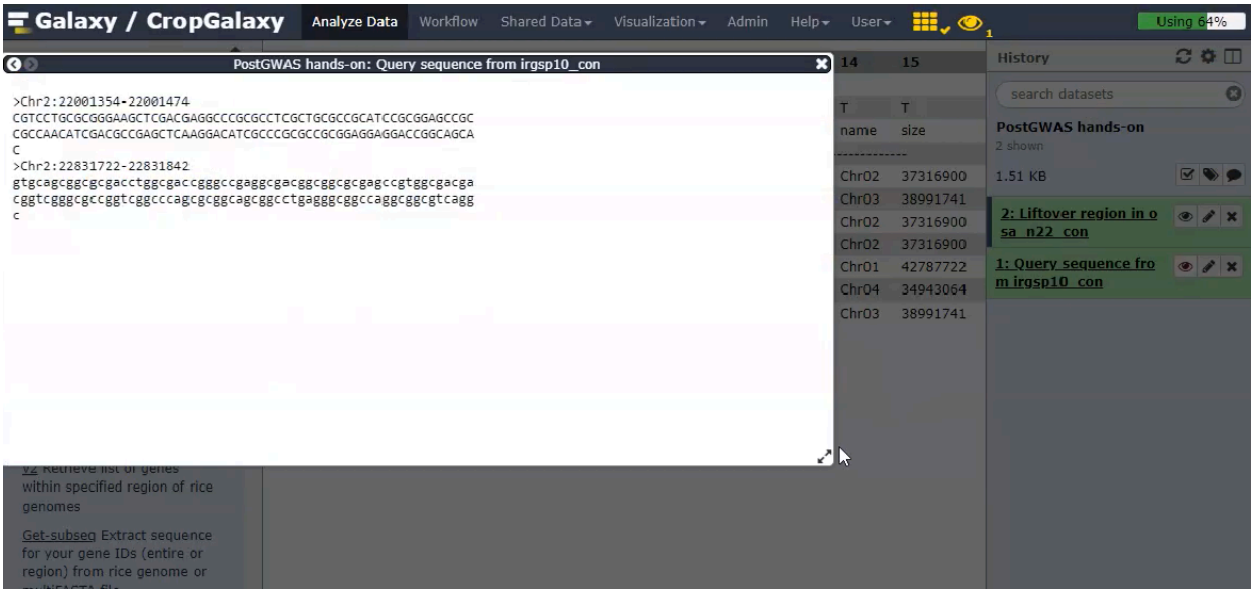


9.) Displaying dataset contents

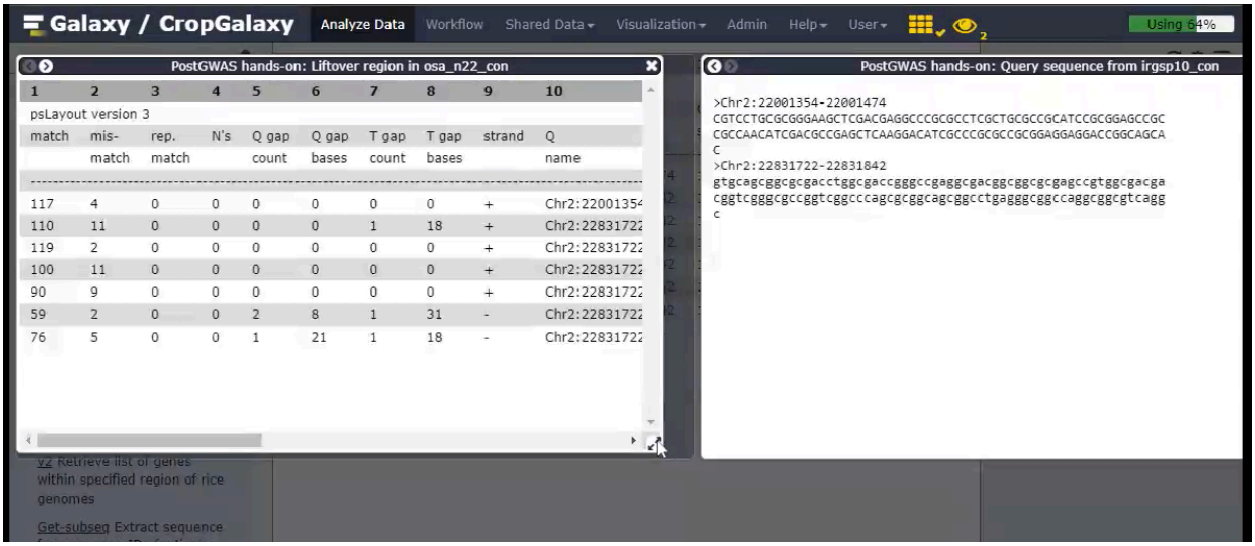
You will see the contents of the dataset displayed in the analysis panel in the middle. You may also use the **Scratchpad** feature of CropGalaxy to view multiple datasets at once.



Click on the first dataset. You can also resize the panel to adjust the size of the floating result window.



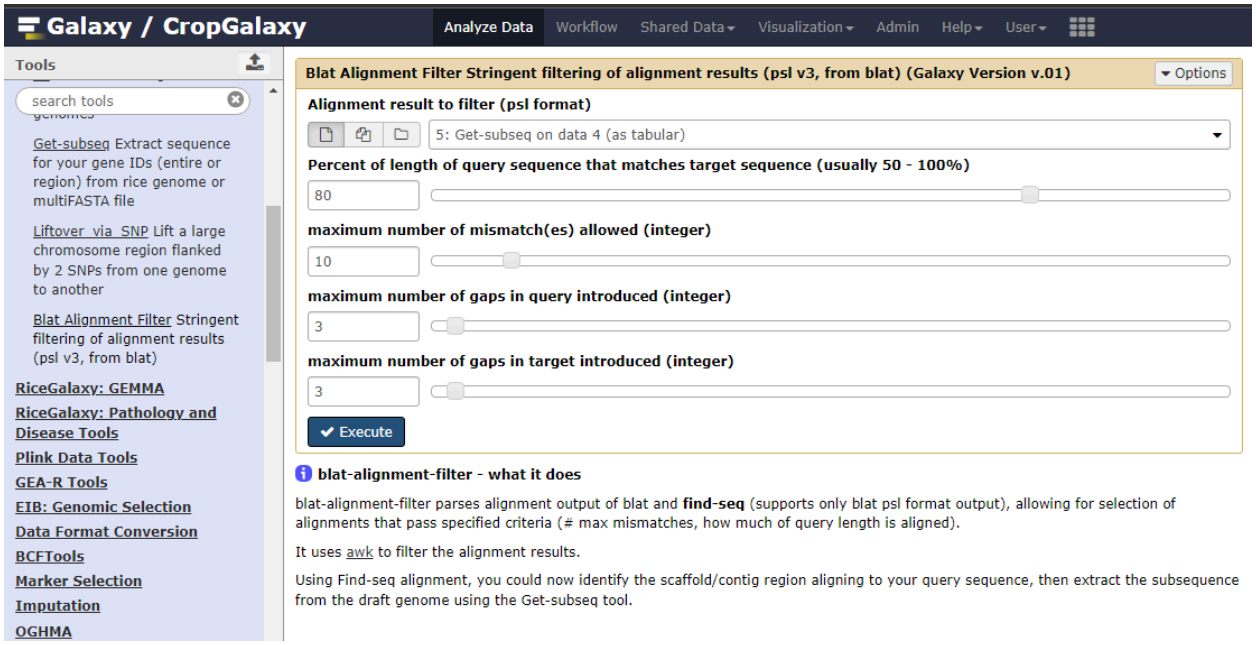
Click on the second dataset. Now you can see two datasets side-by-side.



To exit Scratchpad, just click on any gray space outside the floating window of any dataset displayed.

10.) Blat Alignment Filter

Examine the alignment result. Notice that it shows multiple hits on the probable location of the lift-over coordinates. So we need to filter this using the tool: **Blat Alignment Filter**



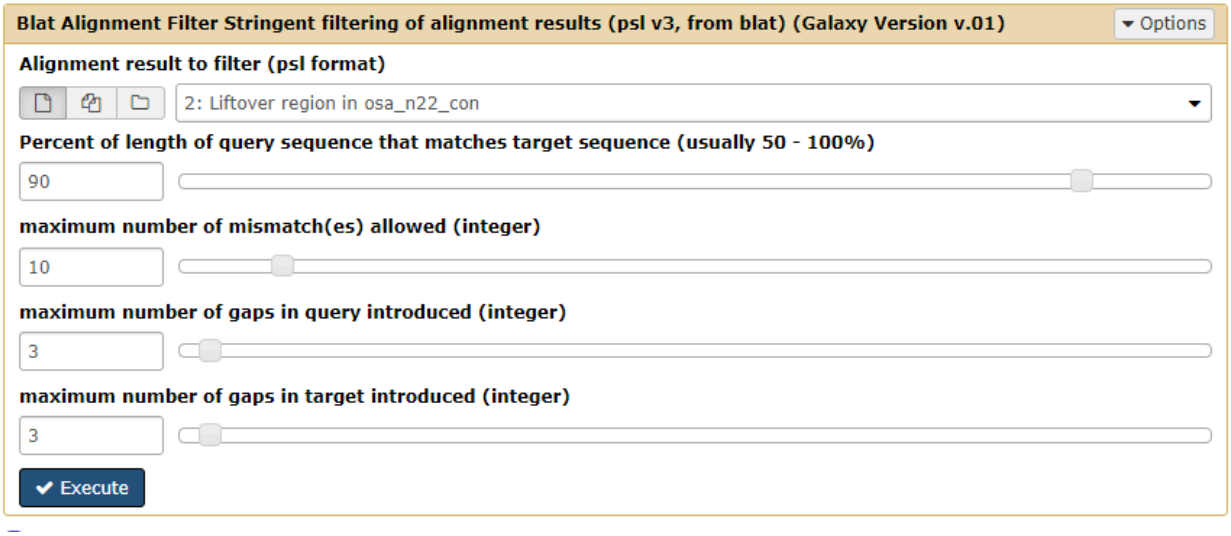
The blat-alignment-filter parses alignment output of blat and find-seq (supports only blat psl format output), allowing for selection of alignments that pass specified criteria (# max mismatches, how much of query length is aligned).

It uses "**awk**" Linux utility in the background to filter the alignment results.

11.) Filtering parameters

In this exercise, we need to set a more stringent filtering criteria to keep only the sequences closest to the target genome. So we set the value "**90**" in Percent of length of query sequence that matches target sequence (usually 50 - 100%). We will use the default values of the other parameters.

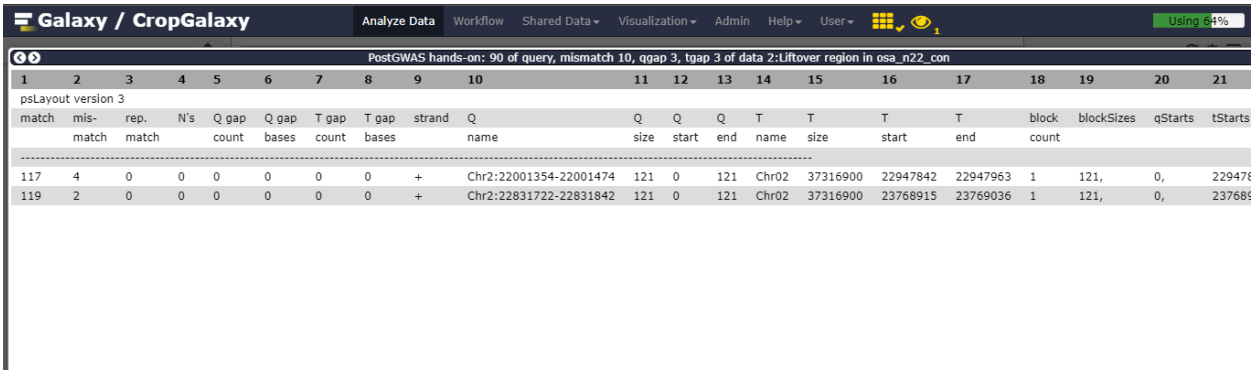
Again, click on the "**Execute**" button to run the tool.



12.) Filtering output

You will see the dataset: **"90 of query, mismatch 10, qgap 3, tgap 3 of data 2:Liftover region in osa_n22_con"** added in your history.

Click on the "**View data**" button to see the result file. You will see that the resulting file now only has two entries.



13.) Lifted-over coordinates on N22

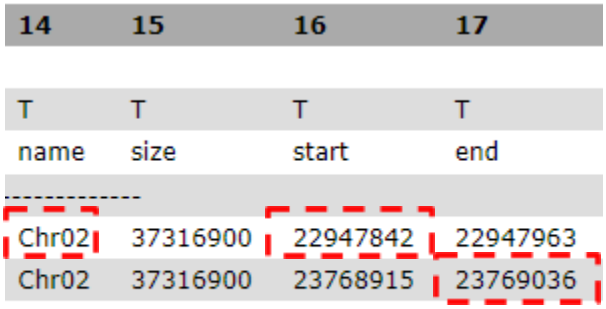
You now have the new coordinates in N22 genome.

Copy the chromosome "**Chr02**".

Copy the value of the first entry in "T start" (Column 16). This will become your lower bound coordinate: "**22947842**"

Copy the value of the second entry in "T end" (Column 17). This will become your upper bound coordinate: "**23769036**"

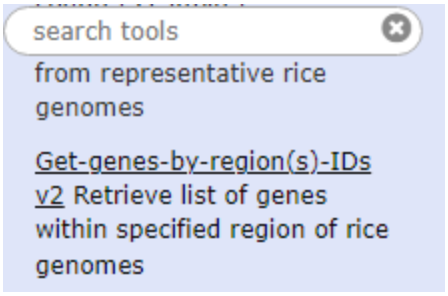
The N22 genome coordinate of the lifted-over QTL is: **Chr02:22947842-23769036**



14.) Get-genes-by-region(s) IDs v2

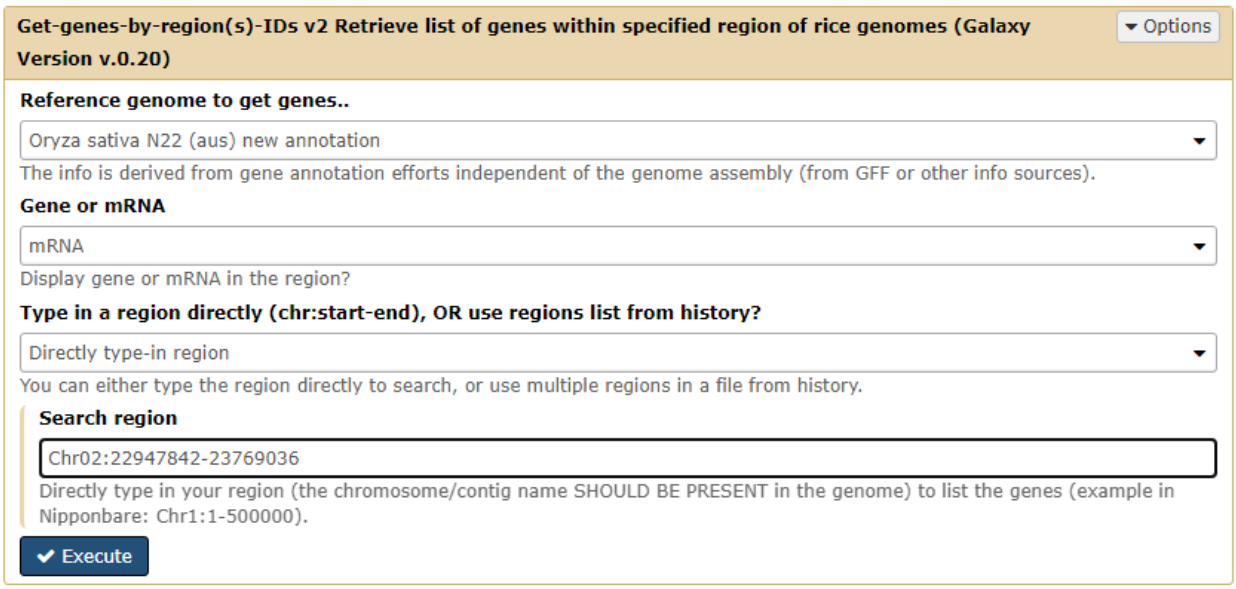
Now, we want to look for candidate genes that lie in this region. To do that, we will use the tool "**Get Genes from Region IDs v2**".

Go to **RiceGalaxy: SNP Data tools** suite again and select "**Get Genes from Region IDs v2".**



15.) Get Genes parameter set-up

a. Set the Reference genomes to get genes to : **"Oryza sativa N22 (aus) new annotation"**
b. Set Gene or mRNA to: **"mRNA"**
c. Set Type in a region directly (chr:start-end), OR use regions list from history? to: **"Directly type-in region"**
d. To extract the genes in the QTL from N22 mRNA; use the lower & upper bound of the lift-over results. We will use here the one we got from the previous step: **"Chr02:22947842-23769036"**.
e. Once the parameters are set, click "**Execute**"



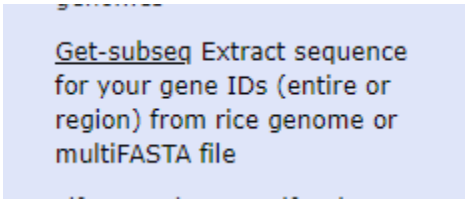
16.) Get Genes output

Click on "**View data**" to see the list of genes.

You will see that there are 149 genes (mRNA) in the lift-over region from N22 reference genome. The next thing we want to do is get the sequences of these genes. To do that, we will use the "Get-subseq" tool.

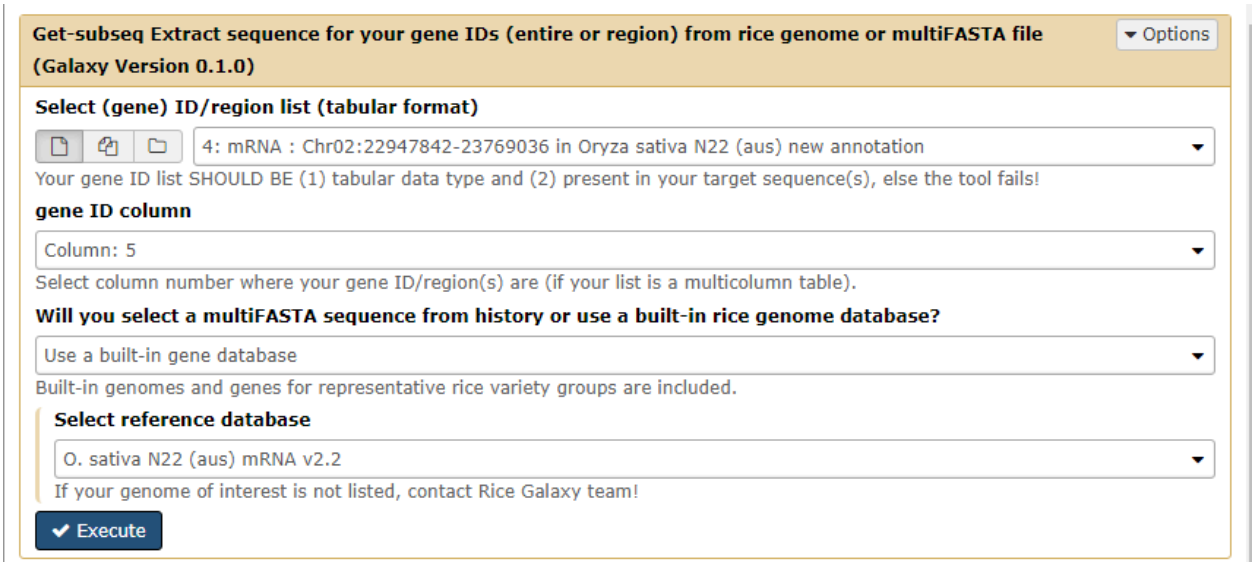
17.) Get-subset tool

This tool extracts a subsequence of interest from important rice reference genomes.



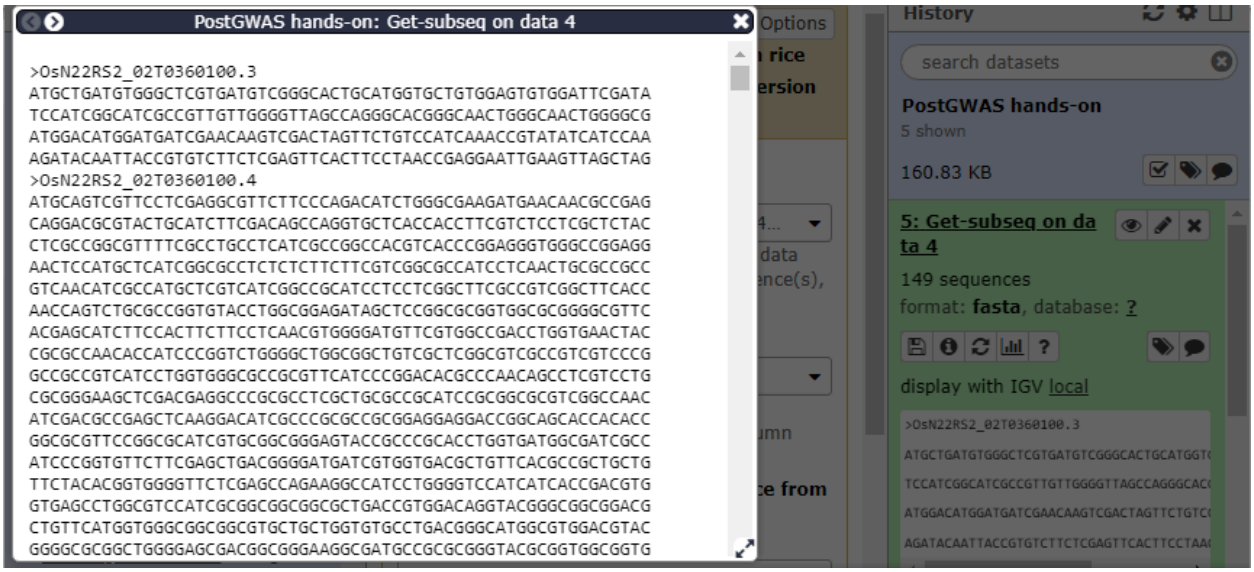
18.) Get-subset parameter set-up

a. Set the output of the previous tool as input.
b. Set gene ID column to : "**Column 5**"
c. Set "Will you select a multiFASTA sequence from history or use a built-in rice genome database?" to: "**Use a built -in gene database**"
d. Set Select reference database to: "O**.sativa N22 aus mRNA v2.2**"
e. Click "**Execute**" afterwards.



19.) Get-subset output

Click "View data" to display output.



You now have a multi-fasta file which contains the sequences of the genes within the lifted-over QTL in the N22 reference genome. You may use these sequences to look for further annotations (putative functions) from other data sources such as NCBI database, Gramene, or Rice SNP Seek.

---End of part 1 ---

Part 2: Using Nippobare annotation to learn more about the putative functions of these genes in N22

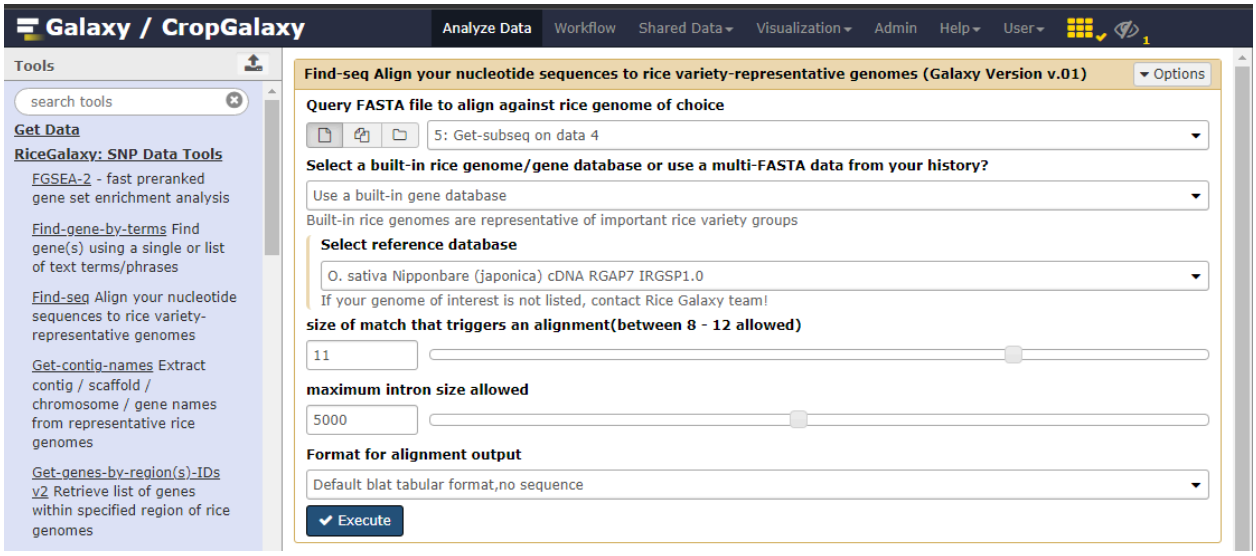
Recall: We generated a gene list and their sequences anchored on N22, an aus-type genome, for the QTL, qDTF2.2. We did this because our donor is closely related to N22.

Our goal now is to use available annotations to learn about the putative functions of the genes in our list. We will now begin another series of analysis.

1.) Find-seq

Use the sequences derived from Part 1 step 19 to align back to Nipponbare.

Set-up the parameters as shown in the image below:



After running the tool, you will get an alignment output.

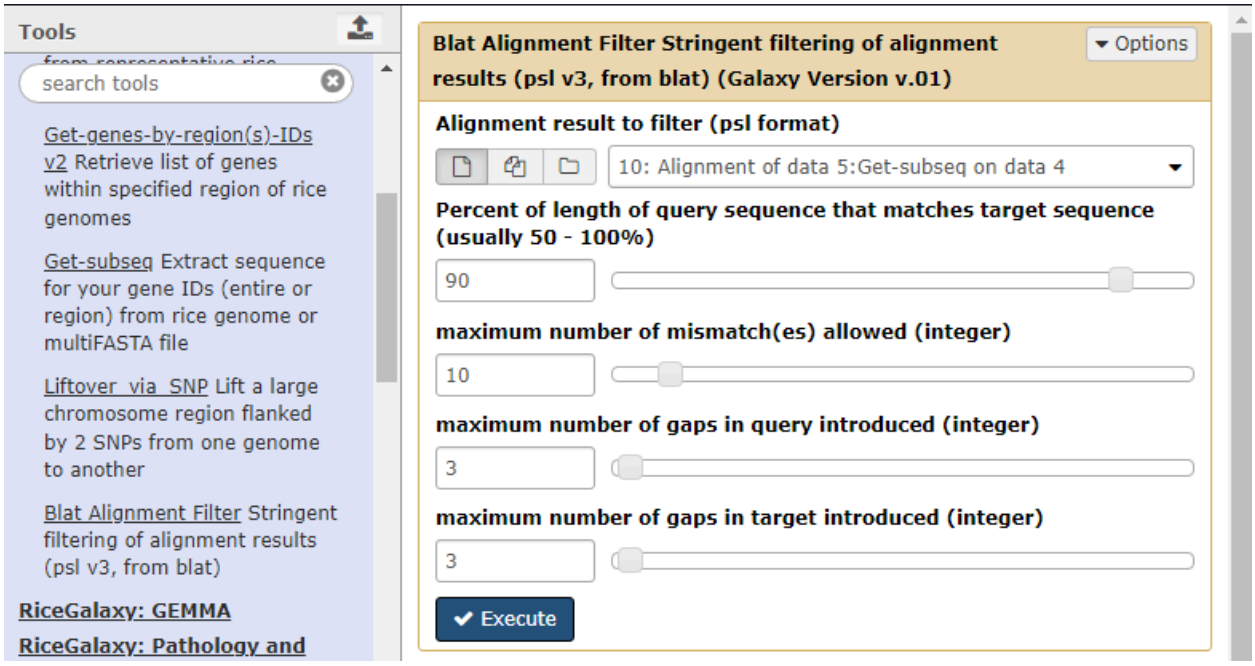
PostGWAS hands-on: Alignment of data 5:Get-subseq on data 4

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
psLayout version 3																		
match	mis-match	rep. match	N's	Q gap count	Q gap bases	T gap count	T gap bases	strand	Q name	Q size	Q start	Q end	T name	T size	T start	T end	block count	blockSizes
33	2	0	0	0	0	0	0	+	OsN22RS2_02T0360100.3	240	171	206	LOC_Os02g36450.2	1753	0	35	1	35,
33	2	0	0	0	0	0	0	+	OsN22RS2_02T0360100.3	240	171	206	LOC_Os02g36450.1	1866	0	35	1	35,
33	2	0	0	0	0	0	0	+	OsN22RS2_02T0360100.3	240	171	206	LOC_Os02g36450	1866	0	35	1	35,
158	3	0	0	2	5	3	40	+	OsN22RS2_02T0360100.3	240	0	166	LOC_Os02g36414.1	2308	1860	2061	6	7,26,57,17,46,8,
158	3	0	0	2	5	3	40	+	OsN22RS2_02T0360100.3	240	0	166	LOC_Os02g36414	2308	1860	2061	6	7,26,57,17,46,8,
45	4	0	0	0	0	0	0	+	OsN22RS2_02T0360100.4	1422	1084	1133	LOC_Os07g01560.2	2120	833	882	1	49,
1404	9	0	0	3	9	0	0	+	OsN22RS2_02T0360100.4	1422	0	1422	LOC_Os02g36414.1	2308	393	1806	4	1351,30,17,15,
1404	9	0	0	3	9	0	0	+	OsN22RS2_02T0360100.4	1422	0	1422	LOC_Os02g36414	2308	393	1806	4	1351,30,17,15,
45	4	0	0	0	0	0	0	-	OsN22RS2_02T0360100.4	1422	1084	1133	LOC_Os07g01550.1	4010	3619	3668	1	49,
45	4	0	0	0	0	0	0	-	OsN22RS2_02T0360100.4	1422	1084	1133	LOC_Os07g01550	4010	3619	3668	1	49,
32	1	0	0	0	0	0	0	+	OsN22RS2_02T0360100.1	1620	1428	1461	LOC_Os08g08070.1	2060	1485	1518	1	33,
32	1	0	0	0	0	0	0	+	OsN22RS2_02T0360100.1	1620	1428	1461	LOC_Os08g08070	2060	1485	1518	1	33,
39	2	0	0	0	0	1	24	+	OsN22RS2_02T0360100.1	1620	1254	1295	LOC_Os04g37970.1	1563	1272	1337	2	31,10,
39	2	0	0	0	0	1	24	+	OsN22RS2_02T0360100.1	1620	1254	1295	LOC_Os04g37970	1563	1272	1337	2	31,10,
1278	71	0	0	4	156	5	54	+	OsN22RS2_02T0360100.1	1620	0	1505	LOC_Os02g36450.2	1753	66	1469	6	46,9,574,268,12,440,
62	1	0	0	0	0	0	0	+	OsN22RS2_02T0360100.1	1620	1557	1620	LOC_Os02g36450.2	1753	1	64	1	63,
1383	76	0	0	3	46	5	55	+	OsN22RS2_02T0360100.1	1620	0	1505	LOC_Os02g36450.1	1866	66	1580	6	38,12,689,268,12,440,

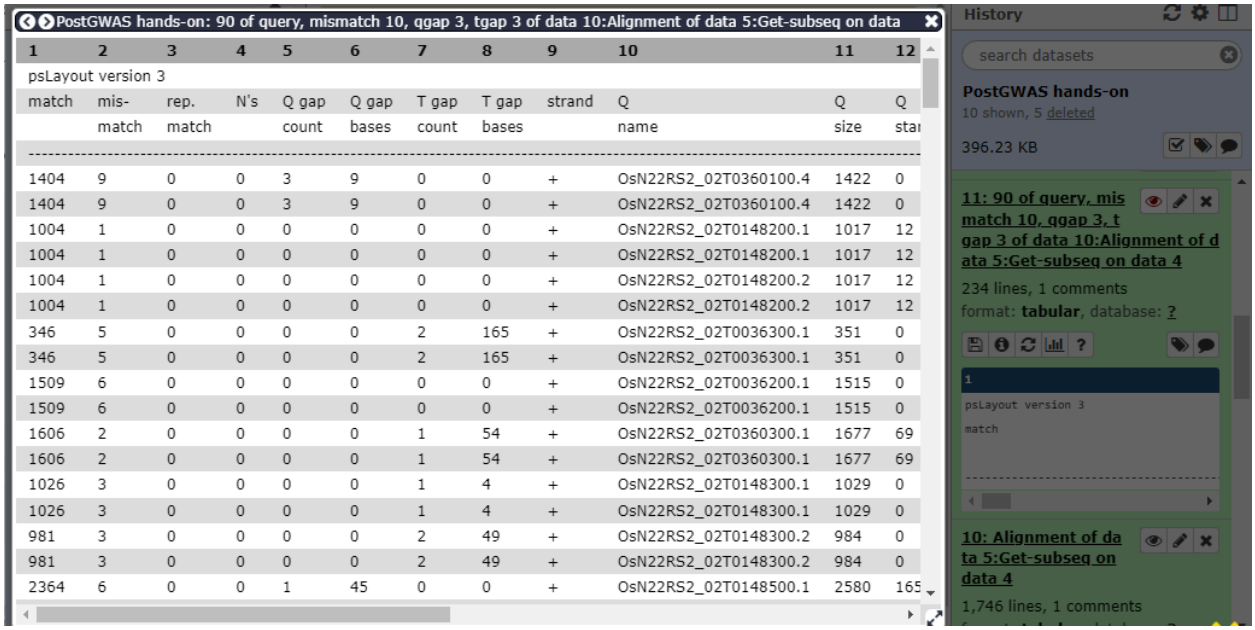
Upon closer look on the output, you will notice that there are a lot of gene duplications and genome rearrangements given the chromosome location of the Nipponbare genes. We will need to filter this using "**Blat Alignment Filter**".

2.) Blat_alignment_filter

We will filter out PSL table to get only those whose sequence matches “90%” with the Nipponbare gene sequence.



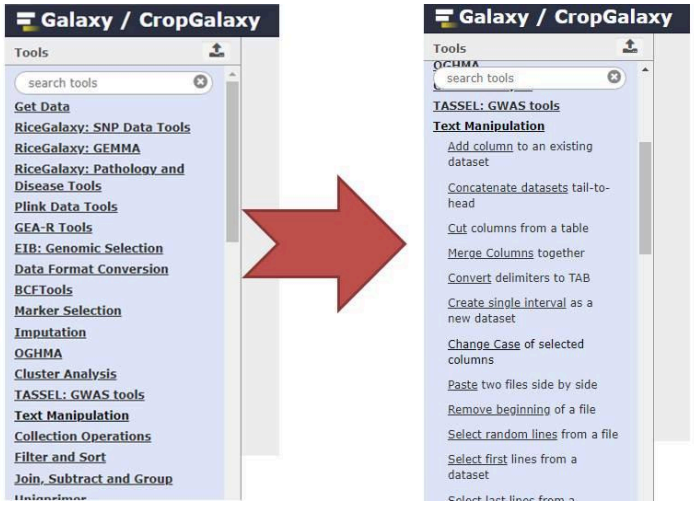
After running the tool, you will see this output. Notice the reduction of alignment results from 1742 records to 230.



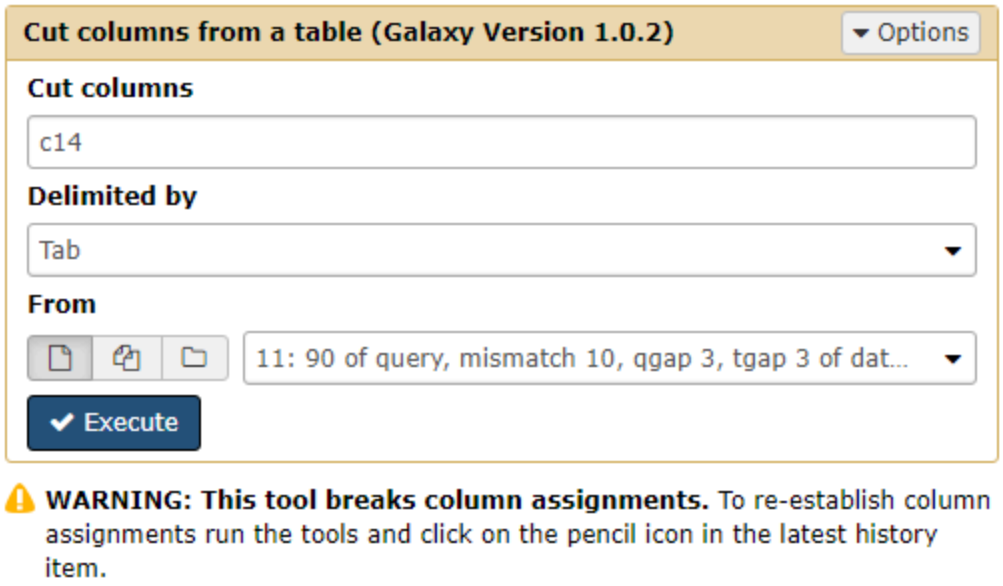
1	2	3	4	5	6	7	8	9	10	11	12
psLayout version 3											
match	mis- match	rep. match	N's	Q gap count	Q gap bases	T gap count	T gap bases	strand	Q name	Q size	Q sta
1404	9	0	0	3	9	0	0	+	OsN22RS2_02T0360100.4	1422	0
1404	9	0	0	3	9	0	0	+	OsN22RS2_02T0360100.4	1422	0
1004	1	0	0	0	0	0	0	+	OsN22RS2_02T0148200.1	1017	12
1004	1	0	0	0	0	0	0	+	OsN22RS2_02T0148200.1	1017	12
1004	1	0	0	0	0	0	0	+	OsN22RS2_02T0148200.2	1017	12
1004	1	0	0	0	0	0	0	+	OsN22RS2_02T0148200.2	1017	12
346	5	0	0	0	0	2	165	+	OsN22RS2_02T0036300.1	351	0
346	5	0	0	0	0	2	165	+	OsN22RS2_02T0036300.1	351	0
1509	6	0	0	0	0	0	0	+	OsN22RS2_02T0036200.1	1515	0
1509	6	0	0	0	0	0	0	+	OsN22RS2_02T0036200.1	1515	0
1606	2	0	0	0	0	1	54	+	OsN22RS2_02T0360300.1	1677	69
1606	2	0	0	0	0	1	54	+	OsN22RS2_02T0360300.1	1677	69
1026	3	0	0	0	0	1	4	+	OsN22RS2_02T0148300.1	1029	0
1026	3	0	0	0	0	1	4	+	OsN22RS2_02T0148300.1	1029	0
981	3	0	0	0	0	2	49	+	OsN22RS2_02T0148300.2	984	0
981	3	0	0	0	0	2	49	+	OsN22RS2_02T0148300.2	984	0
2364	6	0	0	1	45	0	0	+	OsN22RS2_02T0148500.1	2580	165

3.) Post-process the result

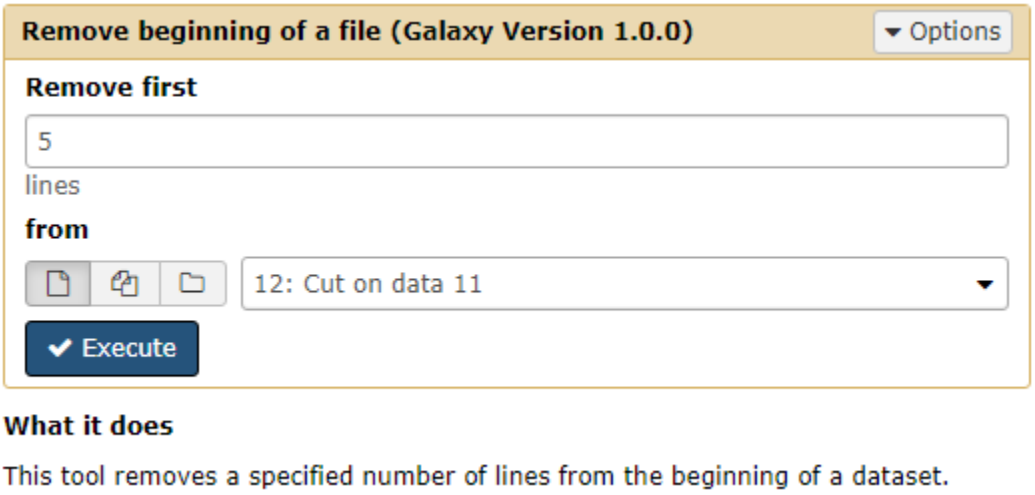
We will now extract the genes in column 14 from the output file of the blat filtering process. For this, we will use the **Text Manipulation** tools.



a. Take out column 14 (c14) from the table using "**Cut columns from a table**" tool.



b. Remove the headers and comments using "**Remove beginning of a file**"



After these two steps, you will be left with a one-column tabular file, which consists of 230 lines of genes.

1
LOC_Os02g36414.1
LOC_Os02g36414
LOC_Os02g36500.1
LOC_Os02g36500
LOC_Os02g36500.1
LOC_Os02g36500
LOC_Os02g36505.1
LOC_Os02g36505
LOC_Os02g36510.1
LOC_Os02g36510
LOC_Os02g36520.1
LOC_Os02g36520
LOC_Os02g36550.1
LOC_Os02g36550
LOC_Os02g36550.1
LOC_Os02g36550
LOC_Os02g36570.1
LOC_Os02g36570
LOC_Os02g36570.1
LOC_Os02g36570
LOC_Os02g36595.1
LOC_Os02g36595
LOC_Os02g36590.3
LOC_Os02g36590.2
LOC_Os02g36590.1

4.) Find genes by term (using the gene list)

We will now get putative functions information on these genes using an annotation from Nipponbare.

To do that, use the "**Find-genes-by-terms**" tool.

Find-gene-by-terms Find gene(s) using a single or list of text terms/phrases (Galaxy Version v.0.3.0) ▾ Options

Which genome do you wish to search?

O. sativa Nipponbare cDNA RGAP7 annotation

The info is obtained from gene annotations for several rice genomes (from GFF or other info sources).

Type in term(s) directly, OR use terms list from history?

Use a list of terms/phrases from the history

You can either type the term(s) directly to search, or use multiple terms in a file from history. Search includes partial matches.

Select list of terms to search (search is OR operation)

13: Remove beginning on data 12

UPLOAD a list file with the terms to search, 1 search term per line only, NO BLANKS!

Case-sensitive search or not?

Case-insensitive

Upper/lower case distinction enforced or not?

Whole word or partial match search?

Partial match

Match only whole words enforced or not?

✔ Execute

Find-gene rice - what it does

Retrieves list of genes that match the search terms (either directly typed in, or from a list), using the annotation description for the selected rice genomes:

After running the tool, you should be able to see on **column 7** the **annotations regarding the putative function of the genes.**

1	2	3	4	5	6	7
Chr1	34425818	34427049	+	LOC_Os01g59540	LOC_Os01g59540.1	GRF zinc finger family protein
Chr1	36317205	36318022	+	LOC_Os01g62720	LOC_Os01g62720.1	GRF zinc finger family protein
Chr1	42830103	42830920	-	LOC_Os01g73920	LOC_Os01g73920.1	hypothetical protein
Chr2	956628	960792	-	LOC_Os02g02610	LOC_Os02g02610.1	transposon protein, putative, unclassified, expresse
Chr2	21995376	22002735	+	LOC_Os02g36414	LOC_Os02g36414.1	transporter family protein, putative, expressed
Chr2	22043694	22046900	-	LOC_Os02g36500	LOC_Os02g36500.1	expressed protein
Chr2	22053938	22054860	-	LOC_Os02g36505	LOC_Os02g36505.1	expressed protein
Chr2	22056928	22058611	+	LOC_Os02g36510	LOC_Os02g36510.1	ethylene-insensitive 3, putative, expressed
Chr2	22060800	22064176	+	LOC_Os02g36520	LOC_Os02g36520.1	OsFBK9 - F-box domain and kelch repeat containing
Chr2	22068014	22070683	-	LOC_Os02g36550	LOC_Os02g36550.1	phosphopantothenate--cysteine ligase, putative, ex
Chr2	22077489	22085283	+	LOC_Os02g36570	LOC_Os02g36570.1	ABC1 family domain containing protein, putative, e
Chr2	22096006	22100897	+	LOC_Os02g36590	LOC_Os02g36590.1	CPuORF19 - conserved peptide uORF-containing tra
Chr2	22098509	22100476	+	LOC_Os02g36595	LOC_Os02g36595.1	expressed protein
Chr2	22102488	22105124	+	LOC_Os02g36600	LOC_Os02g36600.1	aldose 1-epimerase, putative, expressed
Chr2	22110121	22117660	+	LOC_Os02g36619	LOC_Os02g36619.1	exo70 exocyst complex subunit, putative, expresse
Chr2	22135516	22136944	-	LOC_Os02g36670	LOC_Os02g36670.1	expressed protein
Chr2	22140602	22141996	-	LOC_Os02g36680	LOC_Os02g36680.1	expressed protein
Chr2	22144257	22144998	-	LOC_Os02g36690	LOC_Os02g36690.1	expressed protein
Chr2	22146420	22150147	-	LOC_Os02g36700	LOC_Os02g36700.1	sucrose transporter BoSUT1, putative, expressed
Chr2	22151432	22156042	-	LOC_Os02g36710	LOC_Os02g36710.1	SET, putative, expressed
Chr2	22163299	22167383	-	LOC_Os02g36740	LOC_Os02g36740.1	zinc finger, C3HC4 type, putative, expressed
Chr2	22215057	22217951	+	LOC_Os02g36830	LOC_Os02g36830.1	cytokinin-O-glucosyltransferase 2, putative, expres
Chr2	22229412	22231714	+	LOC_Os02g36840	LOC_Os02g36840.1	cytokinin-O-glucosyltransferase 2, putative, expres
Chr2	22234374	22235445	-	LOC_Os02g36850	LOC_Os02g36850.1	oxygen evolving enhancer protein 3, identical, puta

You may also use other resources such as RicePilaf, Rice SNPSeek, Rice Gene Index to gain more insight on the function of the genes as well as gene networks.

---End of part 2 ---

Hands-on prepared by:

Ramil Mauleon

Venice Margarette Juanillas