Open Galaxy and log in to your account. We will call variants using FreeBayes which you can find under the NGS: Variant Analysis. Use the BAM with marked duplicates and hg19 for the reference. We won’t limit calling to a set of regions. Choose simple diploid calling with filtering and coverage. This sets a minimum mapping quality of 30 and a minimum base quality of 20 for reads to be considered in calling a variant. If you read the description, you will see that FreeBayes is set up to only call variants that have at least 20% of reads (and at least 2 reads) supporting them. We will leave the coverage required at 0 for now.

I have also created other vcf files using GATK (chrom\_10.vcf) and SAMtools (samtool\_chrom\_10.vcf). For both tools I used a minimum mapping quality of 20 and a minimum base quality of 10. These are on the google drive in the practical folder. Download these. Then upload them into Galaxy. Set type to vcf and the Genome to Human Feb. 2009.

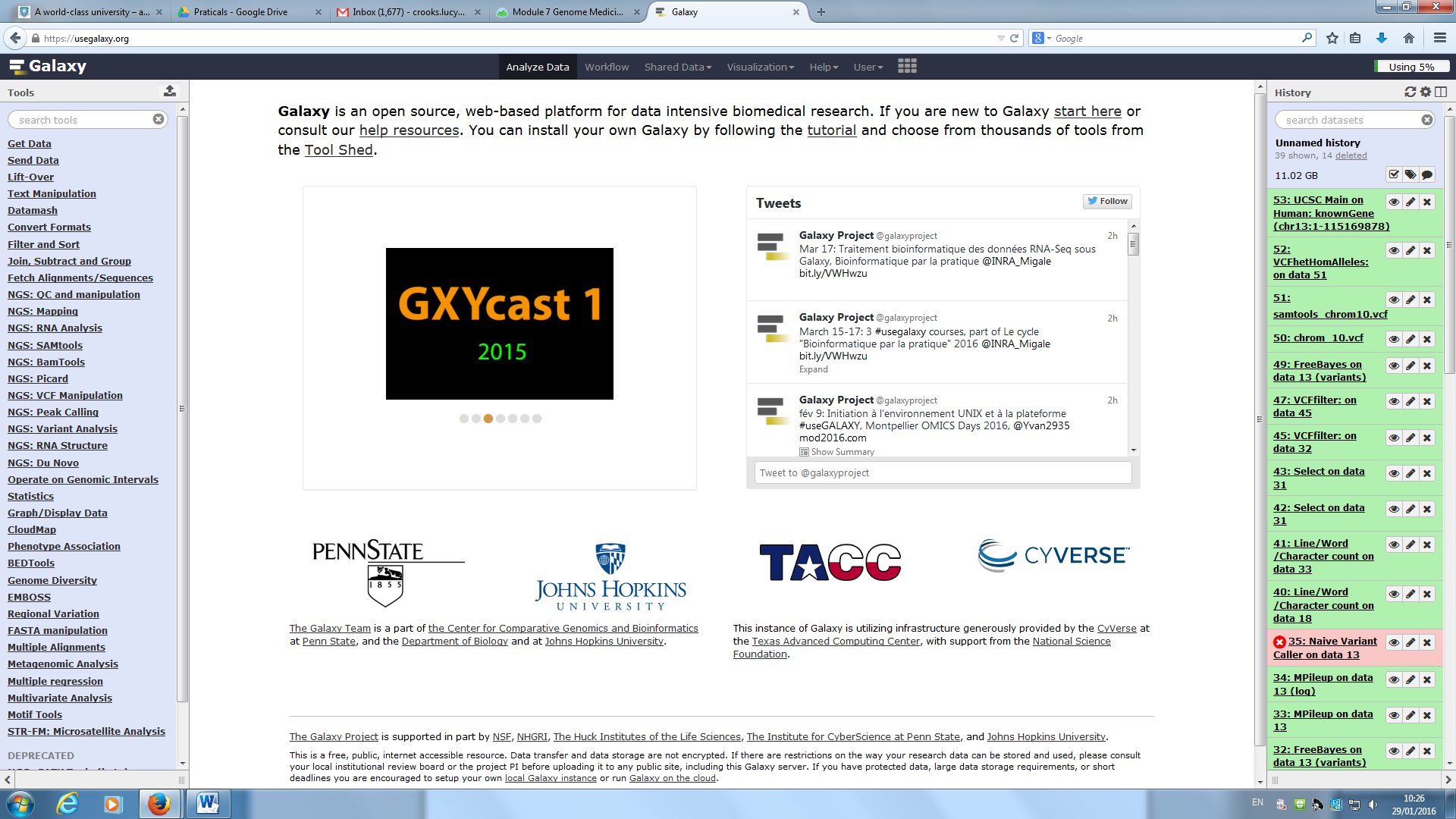
Find out how many variants were called by each caller (if you click on the dataset to get the preview it will tell you how many lines there were). Which caller called the most variants and which the least? Look inside the files and try to understand the calls. Can you see calls that are common between the different methods?

Many of the calls are not very convincing because they are based on not many reads. We are going to filter each vcf to retain only variants called with quality of at least 30 and depth of at least 6. The tool to do this is under **NGS: VCF Manipulation** and is called VCFfilter. Look at the examples and work out how to filter on quality. Do this first. Then filter the output on depth. Look at how many lines are in each of the files. Did filtering on quality or depth have the biggest effect? (be aware that many of the calls with low quality may also have had low depth so if you did the filtering the other way round you would probably have seen more reads filtered by depth than you do when you filter on depth second. However, the end result should be the same).

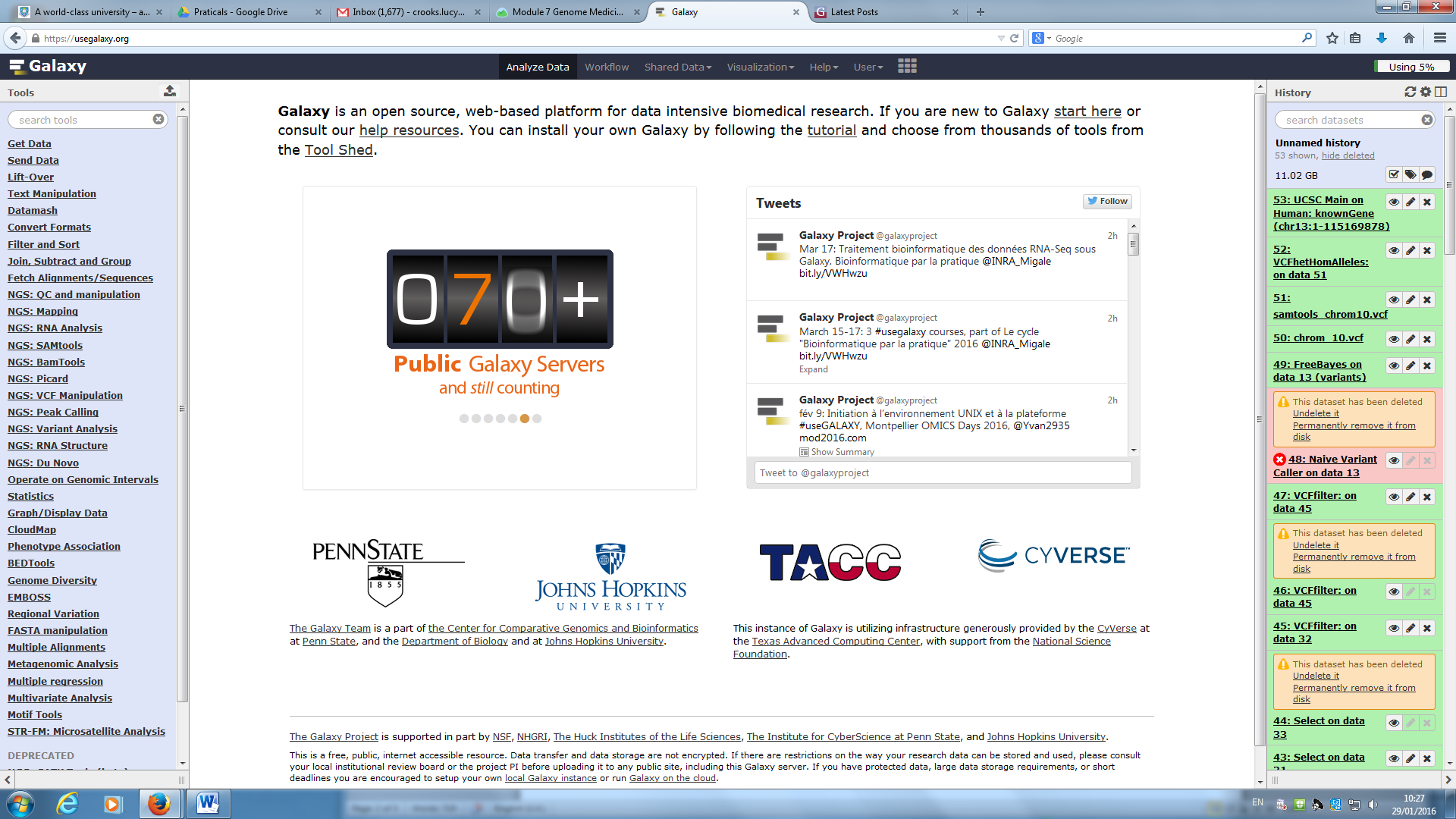
Now that we have restricted our results to calls with better quality results, we are going to look at how the calls made by the different software compared to each other. To do this, we want to find the intersection of the datasets. You can use a tool under **NGS: VCF Manipulation** called VCF-VCFintersect. This tool only lets you compare two vcfs at a time so compare samtools with GATk and with FreeBayes and then find the intersect of these results. How many variants are you left with?

These data are from a targeted sequencing experiment. We therefore expect that we will get more reliable calls in the regions that we targeted. The last step in the analysis is therefore to restrict our results to only the targeted regions. You can do this a tool under **NGS: VCF Manipulation** called VCF-BEDintersect. Have a look at your results. How many variants are left? How many of them are SNVs? Also look at the BAM file in IGV and look at some of the positions of the variants in your final filtered vcf. Do the reads show good evidence of the called variant?

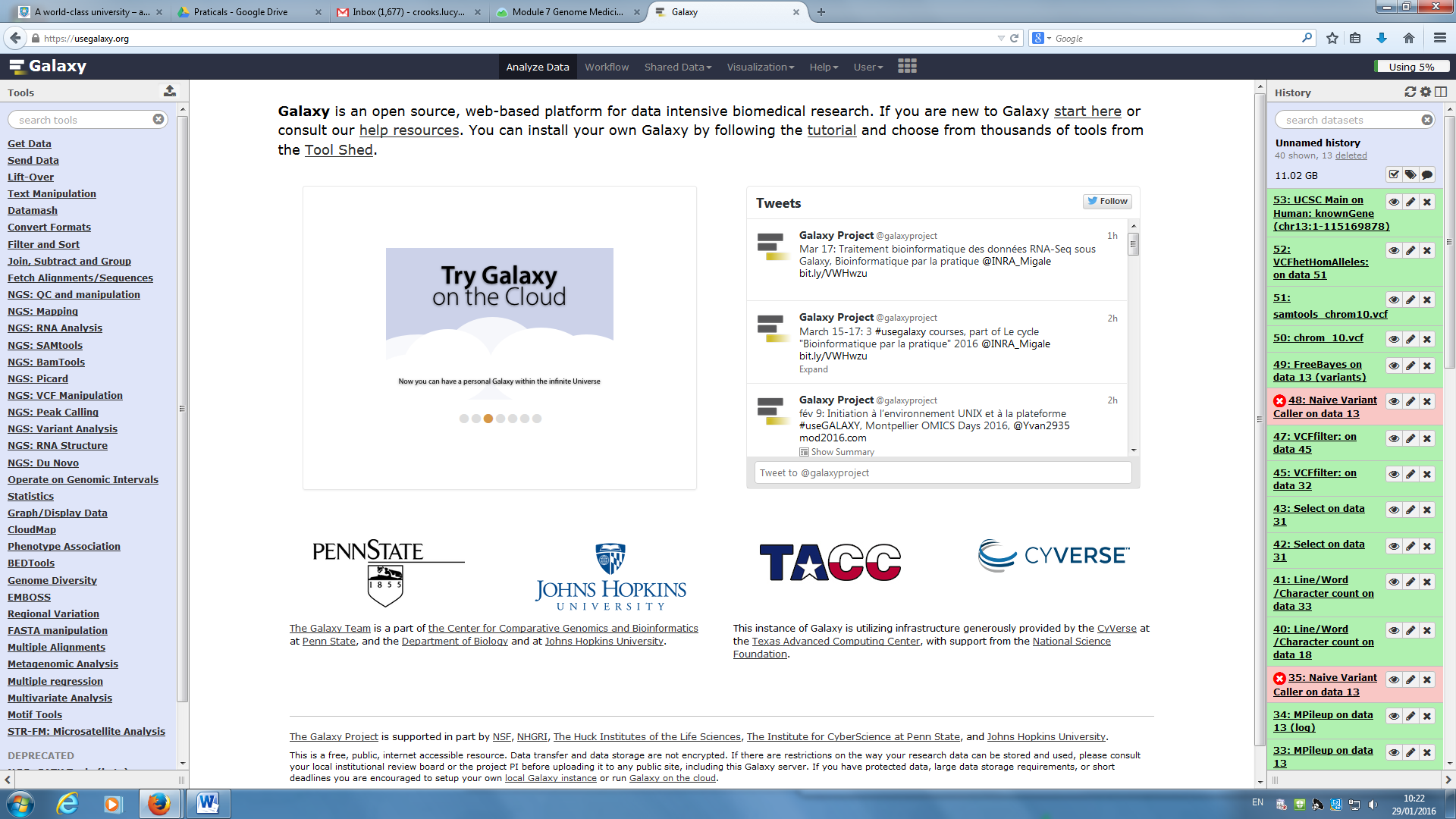
A great feature of Galaxy is the ability to make workflows that you can reuse and share with other people. You can construct a workflow from your history. We are going to make a workflow that performs the entire process of variant calling from fastqs to a quality and region filtered vcf. You probably have additional steps in your history that you do not need so I suggest that you first streamline your history to remove any unnecessary parts. I also recommend as you are going through it, that you rename the steps so that it is clear what process they are carrying out. Delete unnecessary steps by clicking the cross icon next to them. If you delete a step by accident you can retrieve it by clicking where it says deleted at the top of the workflow.



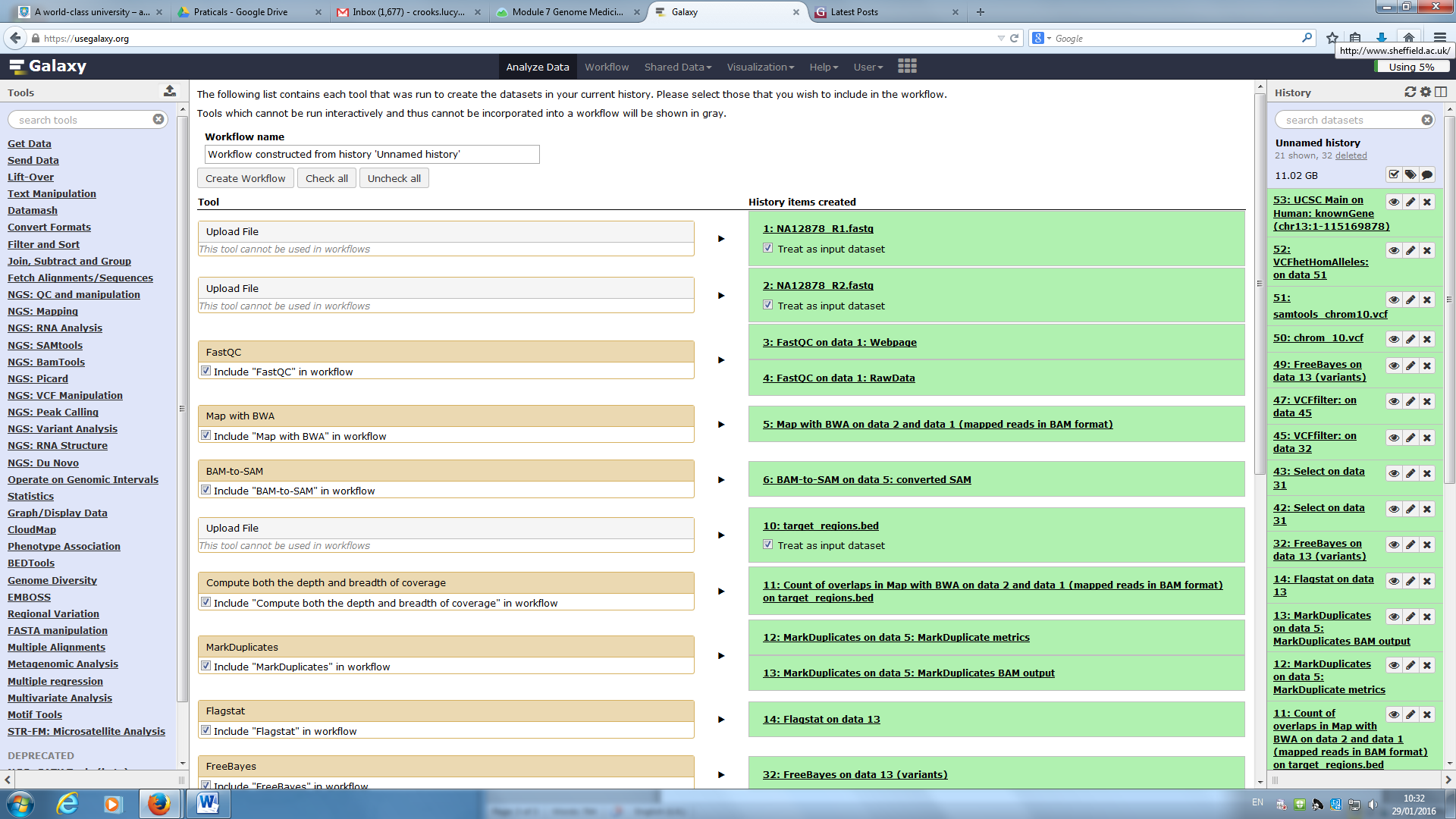
All the deleted steps will then appear in your workflow, coloured in orange and you can click on undelete it



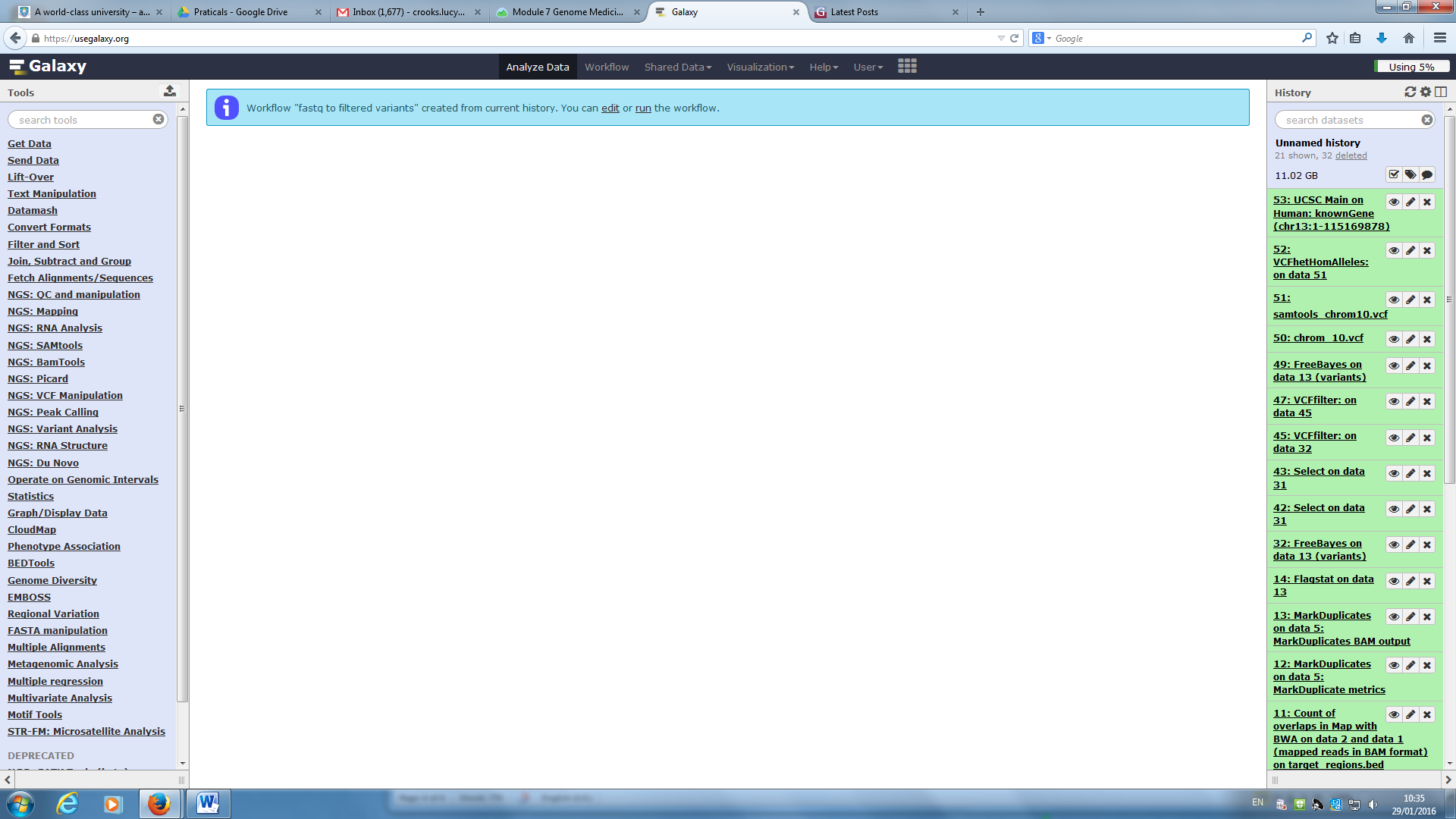
When you are happy with your history, to create the workflow, click on the cog wheel icon at the top of the history and select Extract workflow.



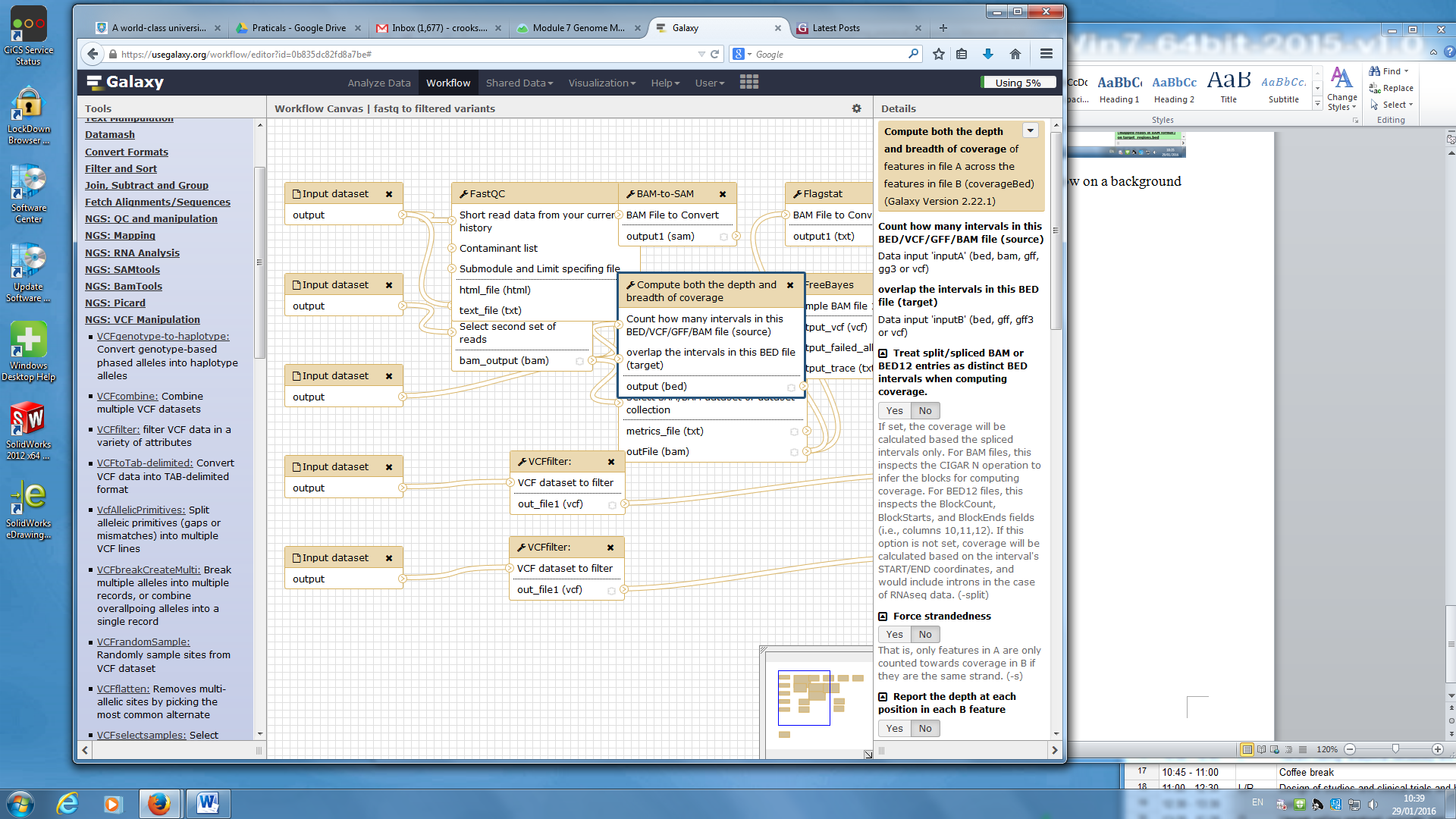
Type in a name for your workflow. There will be a series of boxes corresponding to your workflow. You can click on the tickbox to unselect any steps that you don’t want. Click on Create Workflow.



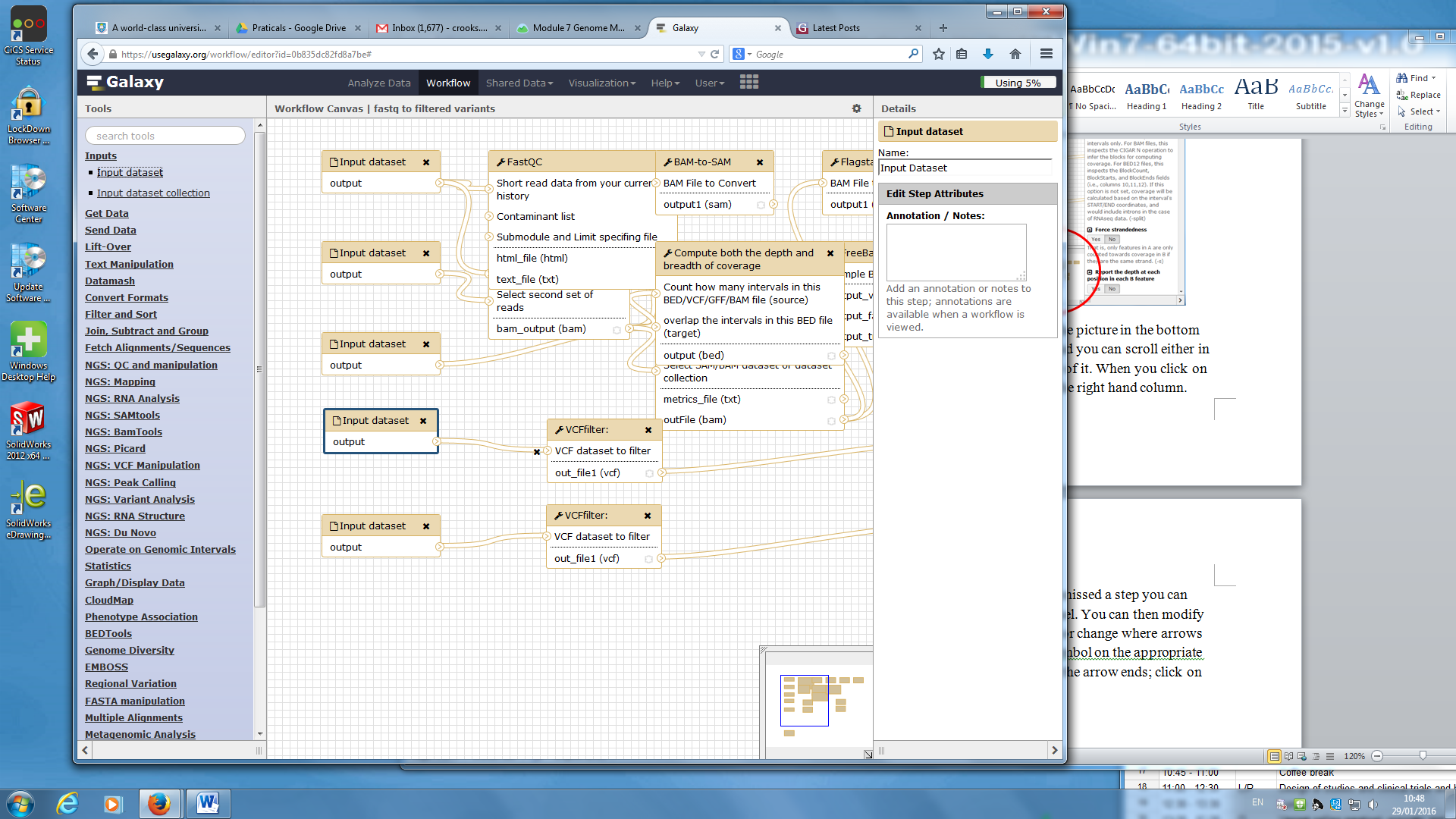
Click on edit to view the workflow.



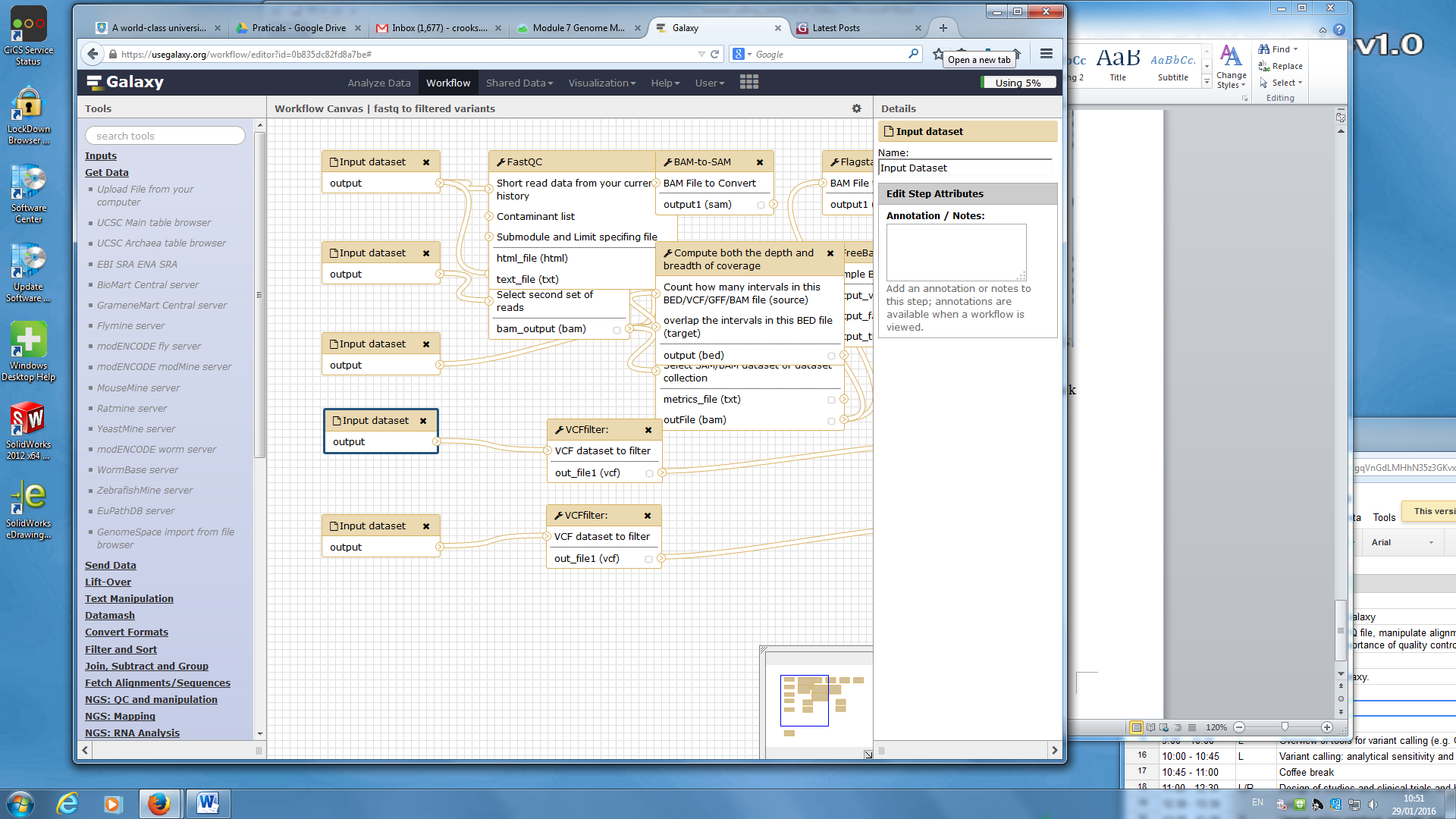
You will see the workflow editor, which shows a graphic of your workflow on a background that looks like graph paper.



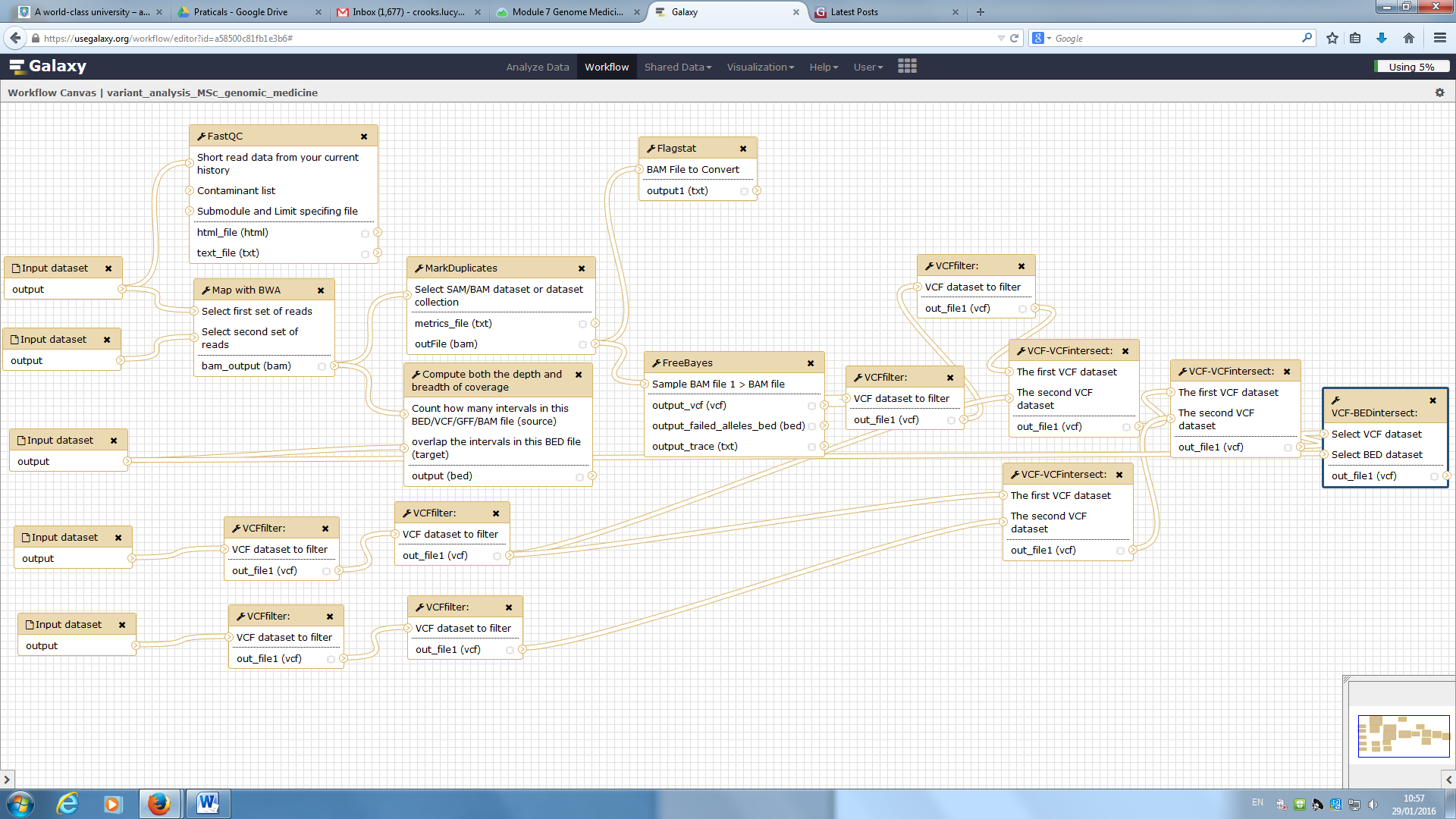
You can move the boxes around to make it more intuitive to you. The picture in the bottom right of the main panel shows you where you are in the workflow and you can scroll either in the main panel or by moving the blue box in the picture to see more of it. When you click on one of the boxes you will see details of the process it represents in the right hand column. You can also amend the options that you applied there. If you have missed a step you can click on it in the left hand column and it will appear in the main panel. You can then modify the parameters in the right hand column. You can also add or remove arrows. To add a new arrow from a step, click and drag from the > symbol on the appropriate box in the direction that you want the arrow to go. To delete an arrow, move near the > symbol on the box where the arrow ends; click on the cross that appears.



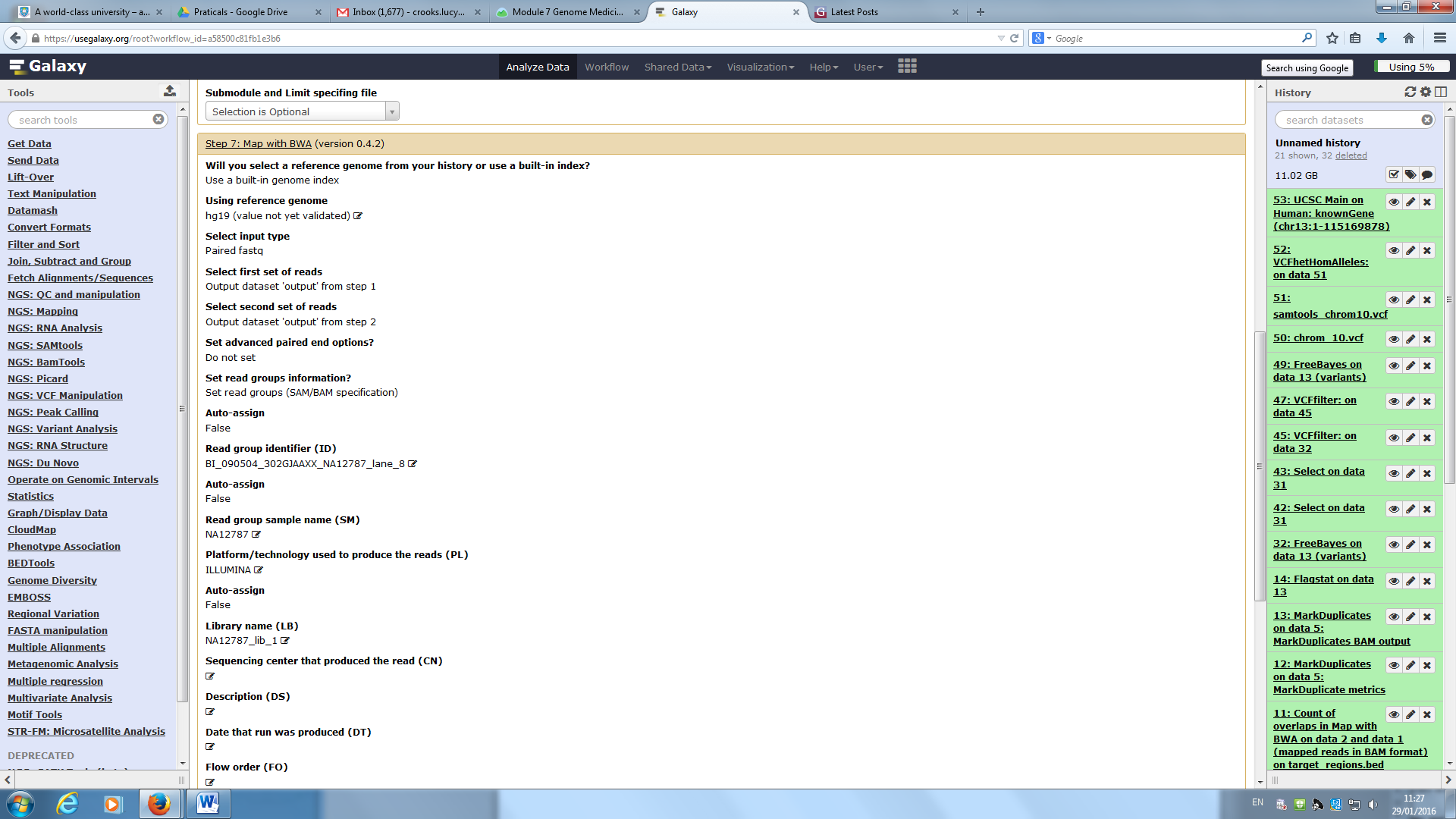
**IMPORTANT** Do not leave the workflow page without saving your work. You save by clicking on the cogwheel at the top right and selecting Save.



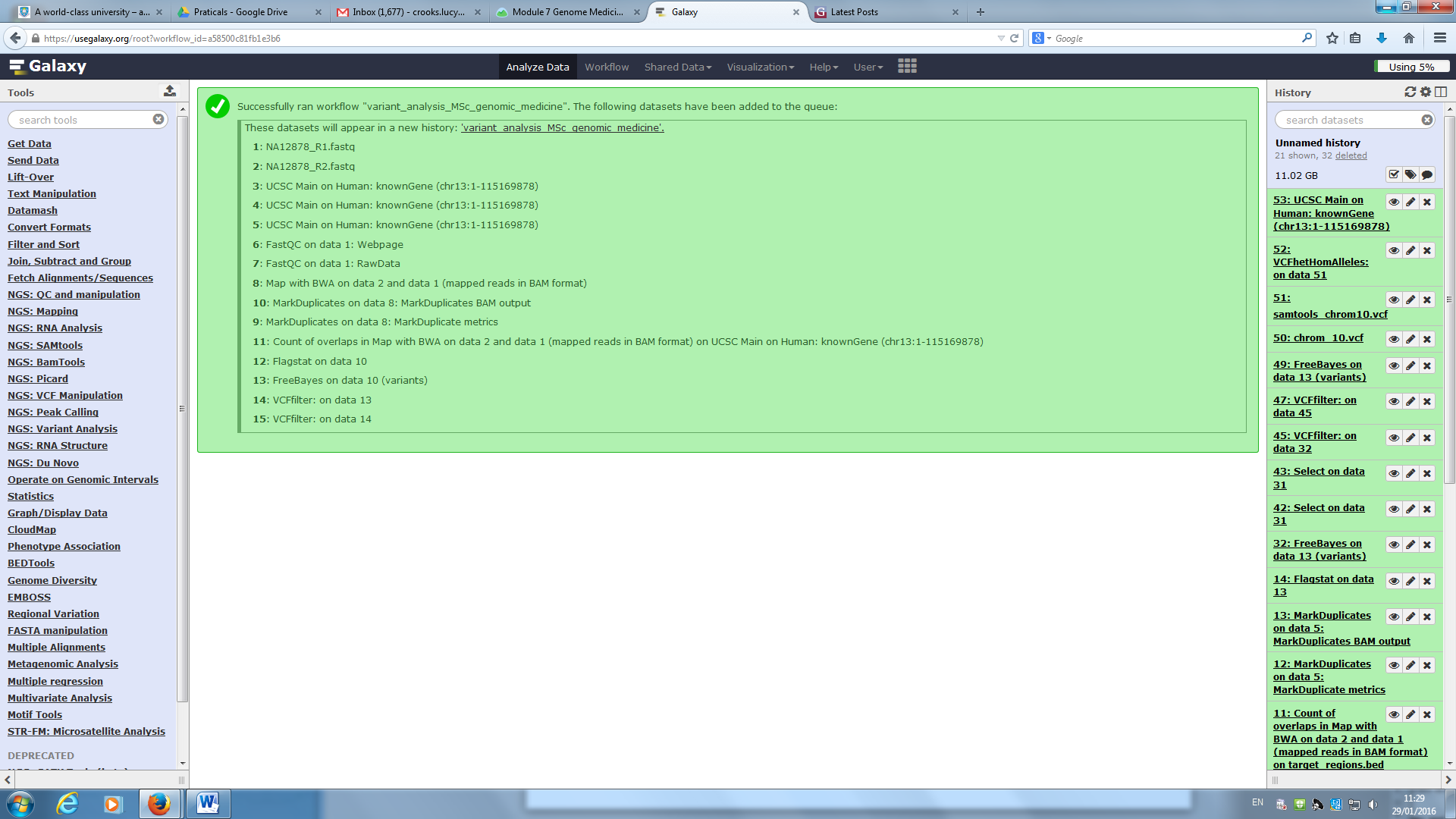
Your final workflow should resemble the picture below



Save the workflow. We will now try to run the workflow with the original data so that we can check it works. To do this, click on workflow on the top bar. Select the appropriate workflow name and chose run. You can select which input files to use. You can alter parameters to the tools by clicking on the step to get an expanded view and then clicking on the text symbol next to the parameter you want to change.



Click the box to send the results to a new history. Then click Run workflow. You should get a notification similar to that below.



To see the new history, click on the link at the top. You may find that some steps don’t work. If that is the case look at the error message to try to see what went wrong (it might be that you don’t have the correct reference genome in the attributes). You can also add the step again, just as when you performed the analysis the first time.

Finally, try to run the workflow on a new dataset. Fastq files NA12878\_run\_1\_R1.fastq, NA12878\_run\_1\_R2.fastq, target region bed file chrom\_13\_target\_regions.bed and vcf files produced by GATK, chrom\_13.vcf, and samtools, samtools\_chrom\_13.vcf, have been made available. Upload them to Galaxy. Run the pipeline and look at the results. How did the number of variants on chromosome 13 compare to the number on chromosome 10? Does this see reasonable given the length of the chromosomes? (You can see the length of the chromosomes in the reference by looking at the header lines in one of the SAM files).