## Background

While red and white wines are often made from very similar grape varieties, a key difference is the skin color of the grape that is used (and whether or not the skin is left on during the wine-making process). As you might expect, **red wine** is made from **red grapes**, while **white wine** is typically made from **white grapes.**

Even before the wine grape genome (*Vitis vinifera*) was sequenced, researchers and breeders were aware of a single major effect locus on chromosome 2 that controlled berry color, but they did not know what kind of mutation was occurring at this locus, or what genes were involved until many years later when they were able to obtain multiple whole genome assemblies for different varieties.

In this exercise, you'll use MUMmer and LASTZ to perform whole chromosome alignments for the **2nd chromosome** of 3 wine grape varietes:

Pinot Noir (a red wine, and also the variety used to construct the reference genome).

Cabernet Sauvignon (another red wine)

Chardonnay (a white wine).

## Questions

### Part 1: Mummer Pairwise Alignments [2 pts total]

Download the 3 fasta files that contain the chromosome 2 sequences for each of the 3 varieties: [Pinot Noir](https://uncc.instructure.com/courses/192992/files/20378494?wrap=1)

[Download Pinot Noir](https://uncc.instructure.com/courses/192992/files/20378494/download?download_frd=1)

, [Cabernet](https://uncc.instructure.com/courses/192992/files/20378499?wrap=1)

[Download Cabernet](https://uncc.instructure.com/courses/192992/files/20378499/download?download_frd=1)

, and [Chardonnay](https://uncc.instructure.com/courses/192992/files/20378503?wrap=1)

[Download Chardonnay](https://uncc.instructure.com/courses/192992/files/20378503/download?download_frd=1)

. Once you have all 3 files downloaded, upload them to the cluster (which is where you'll be running the alignments).

Your first step will be to use Mummer to perform **2 different pairwise comparisons:**

Comparison 1 (red wine vs. red wine): Query=Cabernet Sauvignon vs. Reference=Pinot Noir

Comparison 2: (white wine vs. red wine): Query=Chardonnay vs. Reference=Pinot Noir

You will need to create a bash slurm script in order to submit your job(s) to the cluster. I have installed Mummer4 in our /projects/class folder. To run nucmer (the nucleotide alignment version of Mummer), your command line will need to look like this:

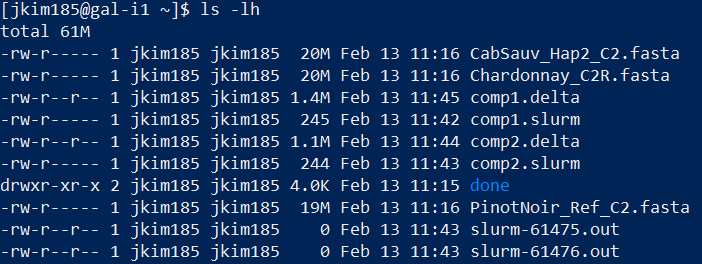
/projects/class/binf6201\_001/bin/nucmer -t CPUS --mum -p YOUR-OUTPUT-NAME REFERENCE.fasta QUERY.fasta

Mummer4 is able to use multi-threading, so replace "CPUS" with the number of threads you want to use. The more threads you add, the faster the program will go. Be sure to edit you #SBATCH resource request lines to match up with however many CPUs (or tasks per node) you need!

You will also need to edit the YOUR-OUTPUT-NAME, REFERENCE, and QUERY filenames to match the filenames that you are using.

To figure out how long the program takes, you can refer back to Varnika's notes on how to write slurm scripts that will report resource usagem, or you simply use ls -lh to see the time stamp on your output file after the program is finished (just make sure to note when you started it!)

#### Question 1: How long did Mummer take to run on each of your alignments? Be sure to note the number of threads used! (*2 pts*)



Comparison 1 (comp1) took 2 minutes and Comparison 2 (comp2) took 1 minute, each taking 4 threads.

### Part 2: Run a LASTZ alignment for comparison [2 pts total]

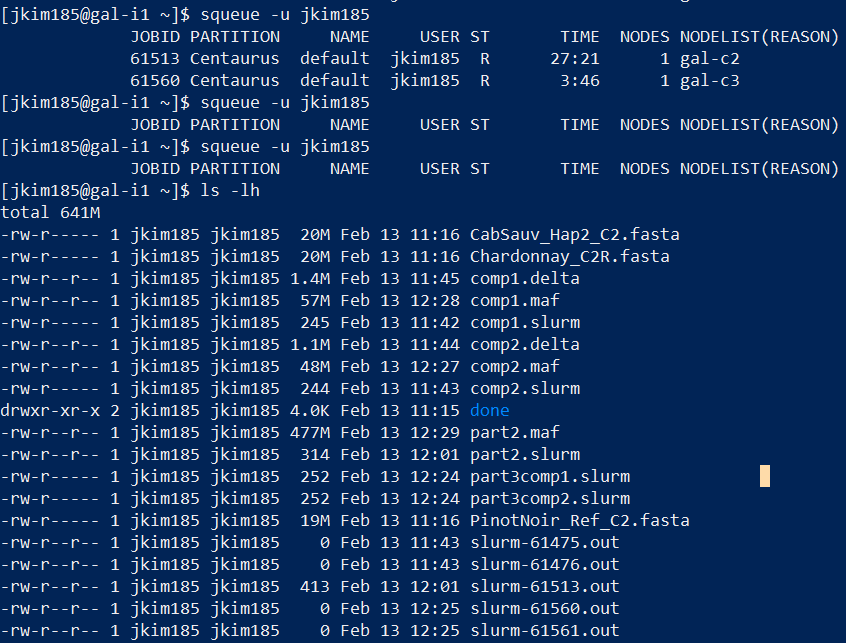
To see how long LASTZ takes to do the same alignment compared to Mummer, just pick one of your comparisons from above (you don't need to do both), and then write a slurm script to run LASTZ. We'll set the parameters of LASTZ to look for exact matches with no mutations in order to make it as comparable to Mummer as possible. Here is the general code you will use for LASTZ:

module load lastz

lastz REFERENCE.fasta QUERY.fasta --exact=20 --maxwordcount=1 --notransition --step=5 --output=YOUR\_OUT\_NAME.maf --format=maf

Again, you will need to change the file names to the ones you want to use. Note also that LASTZ is not able to multi-thread, so you only need to ask for 1 CPU (or 1 task per node). I recommend allowing at least 30 minutes of walltime.

#### Question 2: How much longer did your LASTZ run take compared to your Mummer run? (*2 pts*)



It took 29 minutes, 27 and 28 more minutes than the MUMmer runs, respectively.

### Part 3: Plot Alignments to Find Berry Color Mutation in White Wine. [6 pts total]

In order to compare our 2 alignments, we will use dot-plots. There is a web tool called [D-GENIES](https://dgenies.toulouse.inra.fr/)

that we can use for this, but first we need to get our Mummer output into the maf format.

To help you with this, I have installed another program (Mugsy) into the /projects/class folder, which contains as part of the package a tool called delta2maf (note that the Mummer output is in the "delta" format by default). You will need to create another slurm script to run this tool:

/projects/class/binf6201\_001/mugsy\_x86-64-v1r2.2/MUMmer3.20/delta2maf YOUR-MUMMER-OUTPUT.delta >YOUR-NEW-NAME.maf

Again, be sure to change the file names as appropriate. Once this program finishes, you can download the resulting .maf files for each alignment.

Go to the [D-GENIES](https://dgenies.toulouse.inra.fr/)

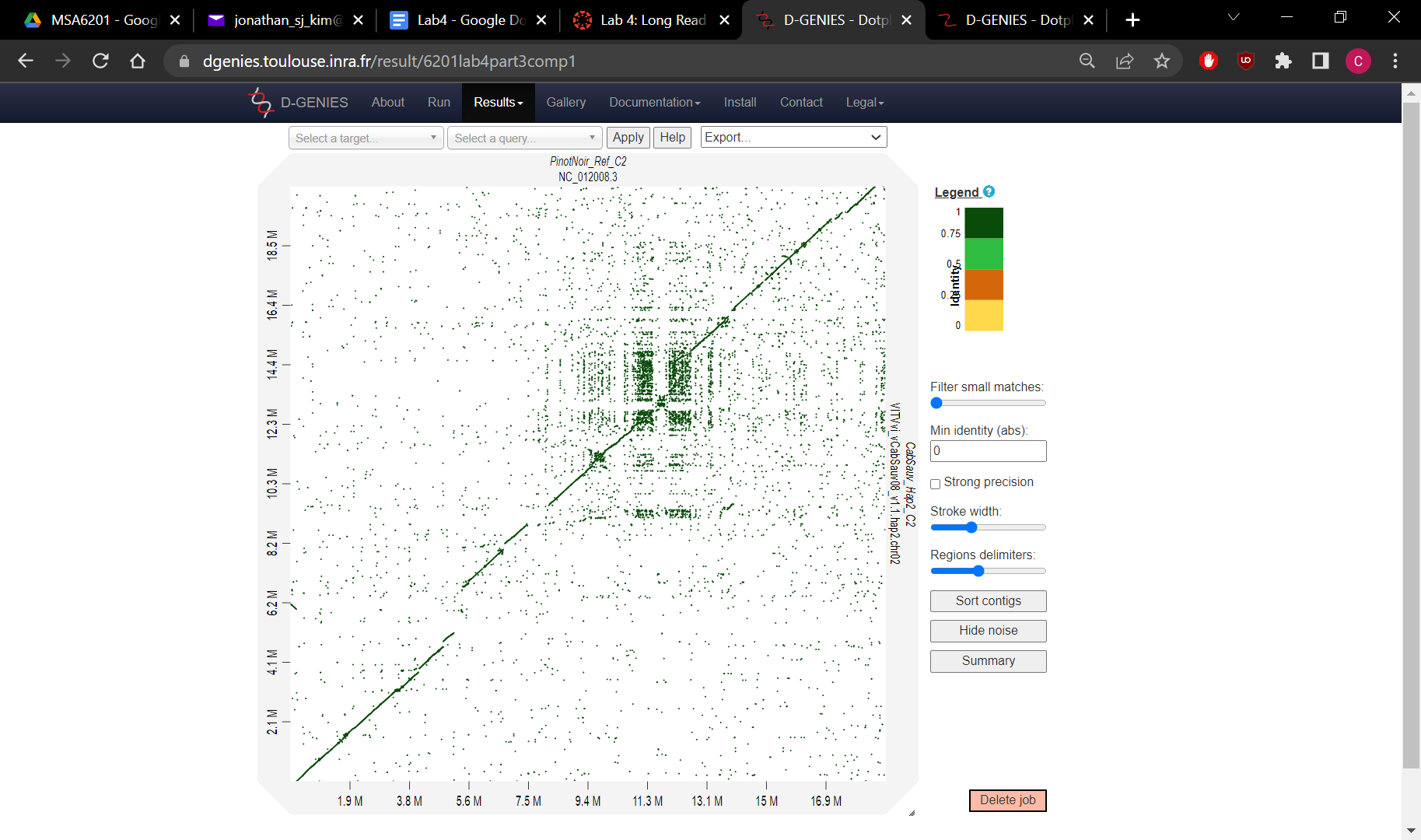
website, and click the "Run" tab in the top menu bar:

Next, click the "Plot Alignment" tab on the Run page:

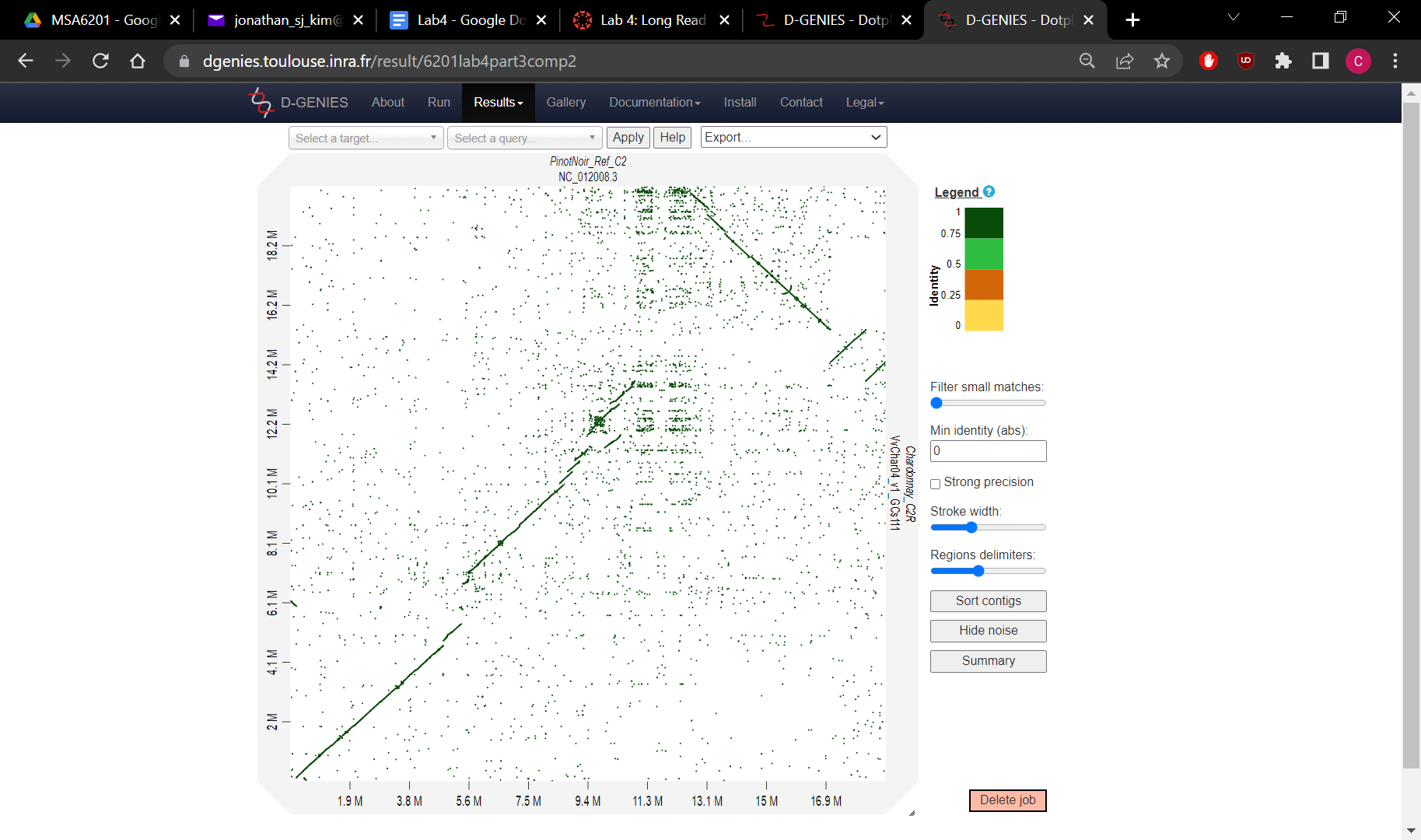
Enter a name for your job followed by your email address. Then upload the appropriate files corresponding to your Alignment (maf) file, Target (Reference) sequence, and Query sequence. Hit "Submit", and then wait for the progress bar to show that the job is complete (or wait to receive an email) before clicking to view your results.

Note that on the side bar of the page with your plot, you have a sliding filter for removing small matches, or a button that will automatically hide what it thinks are noisy matches. Play with these options until you get a plot that you like and is easy for you to identify what might be going on. Then click "Export" and select "PNG" to download a picture of your plot.

#### Question 3: Provide the Dot-Plot for your comparison of the 2 Red Wines *(2 pts*)



#### Question 4: Provide the Dot-Plot for your comparison of the White wine and the Red wine *(2 pts)*



#### Question 5: Based on your plots, what is the biggest difference you see in the white vs. red comparison (that is not apparent when you compare the two red varieties)? *(2 pts)*

*Be sure to note what kind of mutation you think there is, and the rough location of it on the chromosome.*

M = megabase

There seems to be an inversion from 11.3 M to 16.9 M (Pinot Noir reference) in the white vs. red comparison whereas the red vs. red seems pretty linear to each other.

### Part 4: Search for Genes within the Mutated Region [5 pts total]

Now that you can see the mutation in the Chardonnay genome, try using some of the NCBI database searching skills we learned earlier to find what genes might be in the mutated region. Note that the Chardonnay genome is not in NCBI, but the Pinot Noir reference genome is (species: *Vitis vinifera*). Also, the chromosomes we have been aligning are **chromosome 2.**

#### Question 6: Approximately how many genes are located in the mutation region that could be involved in the berry color difference? How did you get this number? *(3 pts)*

You do not have to be exact here, but be sure to note what positions or regions you searched, and what parameters or filters you may have applied.

#### *Approximately 118 genes. We searched Genes through the NIH/NCBI database with the following criteria:*

("Vitis vinifera"[Organism] AND NC\_012008[nucl\_accn] AND 000011300000[CHRPOS] : 000016900000[CHRPOS]) AND ("genetype protein coding"[Properties] AND alive[prop])

#### Question 7: MYB genes are a family of transcription factors previously associated with color in other plants. How many MYB genes are found within this mutated region? How did you get this number? *(2 pts)*

Again, you do not have to be exact here, but you should explain how you did your search or filtering (even if it involved downloading a gene list and filtering in Excel).

8 MYB genes were found:

("Vitis vinifera"[Organism] AND NC\_012008[nucl\_accn] AND 000011300000[CHRPOS] : 000016900000[CHRPOS]) AND ("genetype protein coding"[Properties] AND alive[prop]) AND MYB[All Fields]