1. **Introduction**

There is a need for quality food research more than ever. The pandemic that started in 2019 has made it clear that food security is one of the vital and foremost task people need to take care of (feedingamerica.org). Scientist community needs innovative and creative strategy to secure food for a long run. Carrot, a model plant for my project, is the 7th most consumed crop in the US. Annual crop value of carrots is $864 million (US), and US is the 4th largest producer of carrots i.e., about ~1.4 million tons/year. Carrots are rich in nutrition like carbs, fiber, vitamins and minerals (vitamin A. K1, B6, biotin, potassium), and rich in useful plant compounds like beta-carotene, alpha-carotene, lutein, lycopene, polyacetylenes and anthocyanins. As a plant molecular biologist, one of my projects specifically targets in increasing the shelf -life of carrots by making them disease resistant and healthy using tissue culture technique that increases polyacetylenes content in carrot. I use carrot stable transformation, a tissue culture technique which introduces a gene/s of interest into the plant’s genome permanently. In this method, plant is our host while agrobacteria carrying our gene/s of interest is the carrier. As carrots are a huge use source of nation’s revenue, and are also an important daily food for people, my ultimate goal here is to develop a new variety of carrot with high disease resistance, thus prolonging the shelf-life of carrots.

1. **Problem**

Carrot crop losses amounts to ~ $375 million dollars a year due to plant pathogens. The most common plant pathogen that leads to this big loss is *Sclerotinia sclerotiorum*, a necrotrophic plant pathogenic fungus, also known as white mold, which is capable of infecting wide host range i.e., up to 600 plant species. The problem with this pathogen is that its resilience to grow even in extreme environment and thrives quickly as an opportunistic pathogen in moist and cool conditions. Chemicals like fungicides are heavily used to control the pathogens like *S. sclerotiorum*, but they are costly, and this pathogen is most likely to develop resistance against the fungicides.

Also, there are **a lot of** **problems** associated with carrot stable transformation which make the process of tissue culturing daunting. It is a very long process that takes about eight to nine months if one doesn’t come across any failure of experiment on the way. It involves a lot of different parameters (about 15 in my case) which makes it harder to determine which combination of parameters works best for the procedure. This all makes it challenging to navigate the best option for transformation as I am trying to establish this protocol in our lab for the first time.

1. **Solution**

Fortunately, carrot has its own chemical defense mechanism to fight pathogens. Carrots innately produce these chemical compounds called ‘polyacetylenes’, which are

known to show antifungal, antibacterial, and anti-cancerous properties. Specifically, falcarinol and falcarindiol-type polyacetylenes are implicated to play a role in plant pathogen defense. The acetylase (FAD-2 like) gene *DCAR-013552* is known to overexpress falcarin and falcarindiol, together as falcarins, in carrots. So, to study polyacetylenes, I use this gene along with other controls- empty vector and *PEAQ\_p19\_GFP* – and perform carrot stable transformation, which is a technique to permanently introduce the gene of your interest into your host (carrot in my case). Once the gene is introduced, it will continue to express into the later generations to come. From the stable transformation, once the transformants (carrot with our genes) are plantlets and then plants, we perform various molecular techniques like PCR (Polymerase Chain Reaction), gel electrophoresis, disease assay, GC-FID (Gas Chromatography- Flame Ionization Detector), and GC-MS (Gas Chromatography-Mass Spectrometry) to confirm the presence of gene, analyze/quantify the polyacetylene content, and its function.

Right now, **the Goal of our project** is to find the combinations of parameters involved that will give the highest percentage of polyacetylene content for our carrot stable transformation.

**Open refine (OR) and R to solve the problem + speed up the process**

Since the carrot transformation and other analysis procedure is a time-consuming process experimentally, I used OpenRefine (OR) to clean up my data and R to visualize my data to predict and interpret my data. I used various parameters to automate what I need to know or can know about my experiment. Parameters involved in this project are mentioned below:

1. carrot\_variety
2. carrot\_parts
3. agrobacterium\_type
4. gene
5. conditions
6. time\_in\_MS1D\_dark
7. time\_in\_MS1D\_dl
8. time\_in\_half\_MS1D\_dl
9. time\_in\_quarter\_MS1D\_dl
10. antibiotics\_in\_MS1D\_dl
11. antibiotics\_in\_half\_MS1D\_dl
12. antibiotics\_in\_qaurter\_MS1D\_dl
13. callus\_status
14. PCR\_Gel\_status
15. GC\_MS\_polyacetylene\_percentage

Our excel data sheet contained all these parameters above along with some spaces, numbers, characters that we did not need. We cleaned up the data using OR, added metadata sheet, which contains all the details on each parameter involved in our data, and versions sheet, which contains what changes has been made to our data. The version 2.1 sheet was converted to .csv file and used in R studio to create all the graphs or visuals necessary to predict and interpret our data. Basically, we followed **OSEMN** Data Science Pipeline i.e.

1. Obtaining data (O)
2. Scrubbing/Cleaning data (S)
3. Exploring/Visualizing data (E)
4. Modeling data (M)
5. Interpreting data (N)

After obtaining and cleaning data, we generated required visuals to model and interpret our data. For our modeling, we used linear regression prediction model in R, which is one of the most common and sometimes considered the weak model to predict and interpret the data. With our data visualization and modeling, we were able to predict which combination worked best for our transformation procedure. The combination of Carrot variety – DHL (Danvers Half Long), carrot part – root, agrobacterium type – LBA4404 and gene – acet\_552 got us the best result indicating that this combination gives us the highest polyacetylene percentage. We also got the result for our prediction of polyacetylene percentage which did not match the actual polyacetylene percentage from the data. Though we can say that this combination of parameters works the best, the discrepancy on polyacetylene % means we need more data to clearly conclude this part of our prediction.

1. **Future Direction**

For our **future direction**, we need more explants, different varieties of carrots, GC data, and different timepoints for our media used, generating more data in general. We need proper conversion of our categorical data into nominal data. Understanding and using powerful predictive models- neural network, logistic regression, KNN, random forest- would help us better interpret our results. Also, finding the correlation, and using Occam's razor to eliminate unwanted parameters for polyacetylene % prediction would help us solve the issues we are facing currently in this project.

This project and the use of these tool programming tools has been an enormous help for me as I can predict and confirm the efficiency of my experiment visually and statistically. This will help me implement the carrot stable transformation protocol that is the most effective for our lab in the near future.

**Notes:** MS1D is the media we grow our carrot parts (explants) in.

Agrobacterium (bacteria) is used to transfer our genes of interest into explants.

**Thank You! ☺**