Imoddus Apple mutant project

Final report

0. Aim and objectives of the "Apple mutant" project:

The apple mutant project aim is to increase the efficiency and reliability of DUS testing of apple mutants. Its final goal is to develop phenotypic, genetic and epigenetic tools (in the form of molecular markers) in order to help differentiating apple mutants.

As planned, we focused on the "Gala" variety which is one of the most planted apple varieties in the world. Furthermore, the variety exhibits a huge number of commercial mutants. Thus, the two objectives of the project were:

1. to set up, standardized, and develop new high-throughput tools and methods to phenotype apple fruit skin colour. This part has been increased in comparison to the initial plan.
2. to assess the genetic and epigenetic differences among Gala and its mutants

Initially, the project was quite fully focused on the genetic and epigenetic part but we succeeded to save some budget on sequencing; with the agreement of CPVO, we decided to dedicate more time and budget on the phenotyping part which will deliver results more applicable at short term for the DUS testing.

In this report we summarize the final results obtained from the beginning of the project until March 2020, and provide an update of the work undertaken in the period between March 2020 and October 2021.

1. Genetic, Epigenetic and Transcriptomic

Results obtained from Sept 2019 to March 2020

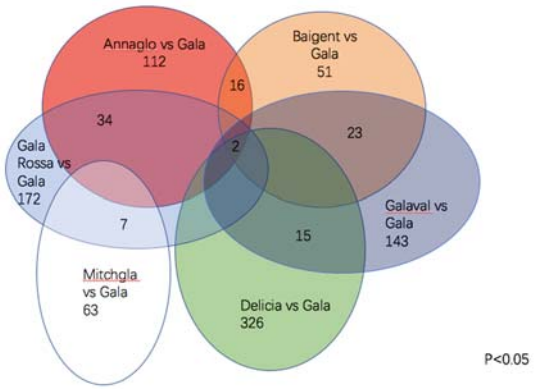
As mentioned in the intermediate report, all the tissue harvest and analyses have been performed as planned. In addition to Gala, the other varieties investigated are: Annaglo, Baigent, Delicia, Galaval, Mitchgla and Rossa.

Over the course of two consecutive years (2018 and 2019), tissues from leaves and fruit skin were harvested. The genome and epigenome were sequenced from leaves tissues (14 and 28 samples sequenced, respectively), while gene transcription level was investigated from RNA purified from fruit skin (28 samples sequenced).

<u>Transcriptome analysis of the Gala mutants:</u>

A total of 62 million reads were generated in total. All the reads were aligned to the apple reference transcriptome. Using the DeSeq bioinformatic software, differentially expressed genes among Gala and the mutants were identified (p-value < 0.05; |logFoldChange|>0.5). Among these genes several were found to be associated with the anthocyanin pathway. This pathway is of particular interest in this study as the skin colour is one of the most important characters allowing the distinction and

identification of particular mutants. Among the differentially expressed genes we identified **MD17G1272100**, which encodes a glutathione transferase and belongs to the phi class of GSTs (naming convention according to Wagner et al., 2002).



Venn diagram representing the number of differentially expressed genes identified is each comparison.

Whole genome sequencing of the Gala mutants:

A total of 120 million reads (150bp PE) were obtained for the 14 samples (2 repetitions per variety). The alignment of the reads permitted the identification of a large number of genetic modifications in the form of Single Nucleotide Polymorphism (SNP). The identification of this rather unexpected large set of SNPs did not allow us to associate the genetic and transcriptomic dataset. However, these SNPs will provide a great resource for the development of particular genetic markers allowing the differentiation among Gala mutants.

Whole Epigenome sequencing of the Gala mutants:

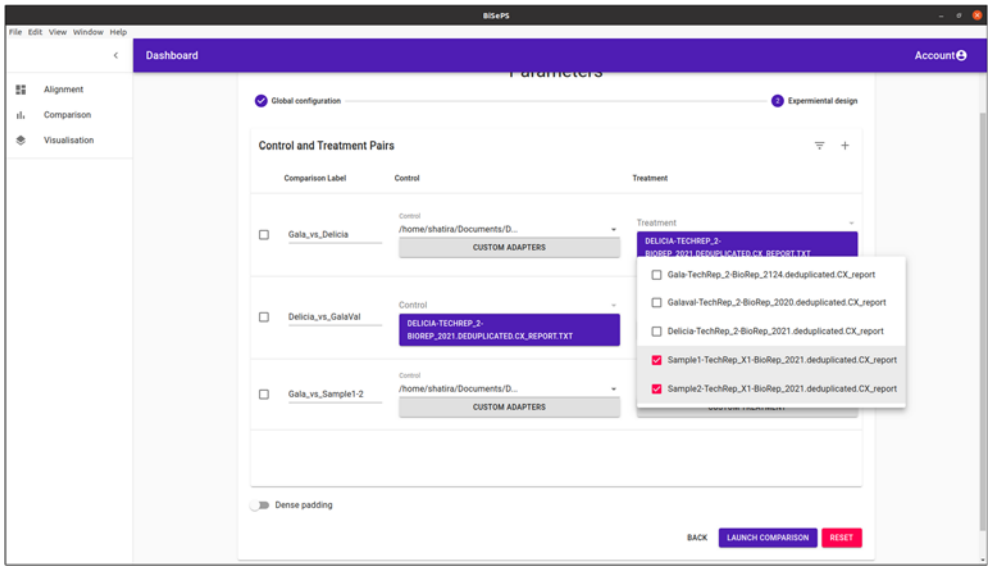
The methylome (=epigenome) sequencing was performed over two years on two trees, yielding a total of over 460 million reads (150bp PE). A primary analysis of this dataset was performed using an inhouse bioinformatic pipeline allowing the identification of Differentially Methylated Regions (DMRs). However, this analysis was not completely satisfactory: the analysis permitted the identification of Differentially Methylated Regions (DMRs) in the genomes of each variety. As such DMRs could be compared among the Gala and its mutant genomes. However, DMRs are composed of a series of Differentially Methylated Cytosines (DMCs), and the bioinformatic tools used in this first approach did not allow us to ascertain whether the DMCs were identical with the DMRs. Therefore, further development was needed to better analyze this data set and estimate the actual epigenetic differences among the varieties.

Results obtained from March 2020 to October 2021

The 18 months covered by the period was devoted to the analysis of the methylome sequencing data generated from Gala and the mutants. This work was mostly performed by bioinformatician (IE - Skander Hatira) hired by the project (cf Milestones M24-M30).

During this period, several tools and platforms were tested and compared. Following this initial step, a bioinformatic tool was selected and implemented in a pipeline (BISEPS) in order to analyze remotely

this RAM-consuming analysis of a public cluster. Much of the period was devoted to the development of this pipeline as well as to the programming of an intuitive and user-friendly interface allowing biologists to set up and test various parameters allowing (cf Milestones M30-M36).



Screenshot from the BISEPS pipeline representing the selection of datasets to be compared during the analysis (control and treatment)

Several approaches were tested to analyze such big data set. As expected from the initial analysis carried in 2020, the analysis of DMRs did not appear to be the most suited for this analysis. Thus, we decided to perform an analysis allowing the characterization of individual differentially methylated cytosines (DMCs). While this analysis generated much more data to exploit, we believe that it is more appropriate for the analysis we wish to perform as it allows the comparison of epigenetic marks located on cytosines located at identical coordinates in the genome. Altogether, this analysis allowed the identification of a total of just over 77,500 differentially methylated cytosine among Gala and the mutants. A database regrouping the various epigenetic data is implemented as an output of the bioinformatic pipeline, allowing an efficient visualization of the epigenetic results. This database, in the form of a genome browser, answers partly to one of the milestones (cf M30-M36), since SNP and Differentially Expressed Genes generated from the genetic and transcriptomic analyses still need to be incorporated.

Screenshot from the genome browser database set up as an automatic output of the BISEPS pipeline.

Among the DMCs, sets of stable epigenetic marks (identified over 2 years and on two biological replicates) were identified. An example of the selected DMCs is shown in the table below. The analysis showed that these DMCs are stable over at least two years and over two independent trees in the environmental conditions of the INRAE orchard of Angers.

selected and could be further exploited for Gala mutants identification in Angers.

	DMC1	DMC2	DMC3	DMC4	DMC5	DMC6	DMC7	DMC8	DMC9	DMC10	unique barcode
Annaglo	1	0	1	0	1	0	0	0	0		101010000-
Baigent	0	1	1	0	1	1	0			0	0110110--0
Delicia	0	1	0	1	1	1	0	0	1	0	101110010
Galaval	0	0	1	0		0		1	0	0	0010-0-100
Mitchgla	0	0			1	0	0	0	0		00--10000-
Rossa	0	1	1	0	0		1	0		1	01100-10-1

1	Gain of methylation
0	Loss of methylation
	No difference

Example of stable DMCs identified among Gala and the mutants

Perspectives

The stable DMCs identified in this analysis could be further developed in the future into molecular markers and tested over additional trees and years to validate their stability. Several of the most promising of these markers could then be used for variety identification resulting in the determination of unique barcodes derived from the molecular marker analysis, as exemplified in the above table. The setup of such a set of robust epigenetic markers will have a concrete impact on the process of variety identification. Furthermore, this set of markers could facilitate the distinction of new Gala variants from the existing ones. Finally, this approach could be expanded to other varieties.

Beyond the identification of a set of stable epigenetic marks, further statistical analysis is still required to fully exploit the generated data sets. We are currently collaborating with statisticians and bioinformaticians to perform this task. This ultimate approach will also allow us to integrate the transcriptomic and genetic data in order to complete the database/browser or encompassing all the information generated in this project.

2. Image processing to distinguish colored mutants

2.1 Materials and methods

2.1.1 Images Acquisition

The acquisition of the images of the different varieties of apples was carried out with the help of a conveyer machine allowing to move the fruits in translation while carrying out a rotation (see Fig. 1). A camera located at the top of the conveyor belt of the machine with a perpendicular viewing direction, took pictures of the apples in rotation, which allowed us to have multiple images providing almost an entire coverage of each apple. Approximately 9 to 10 views of the same apple were captured thanks to this rotation-translation process. These multiple views are important since apple may have several major colors on their skins. With the standard visual approach experts have to manually rotate the apples to have a full perception of the variation of color on a single apple. Here the machine presented in Fig. 1 can acquire a set of 30 apples in a couple of minutes. Images were acquired in burst mode with a Canon camera (10.1 megapixels resolution) controlled by a simple Raspberry-pi minicomputer. Apples were segmented automatically from background as visible in Fig. 1 and assembled in multiple view images of 30 apples as shown in Figs. 2 and 3. This machine, developed for this study, is much simpler and lower cost (approximately 10 keuros versus 100 keuros for classical apple sorting machines) than any commercial systems since it does not need to incorporate any sorting mechanism. Also, by contrast with most commercial system, access to raw image format, i.e. uncompressed format, is possible.

Figure 1. Acquisition system. Upper panel: Machine equipped with a conveyor belt, used for the acquisition of images of apples with a high surface coverage. Lower panel: view of the acquired images of apples after segmentation from the background.

2.1.2 Datasets

Currently, when experts of EO are performing distinctness, they observe directly with their own eyes boxes of 30 apples of each tested variety and reference varieties manually positioned in a same room and they decide form a pure subjective perception if these sets are distinct or not from one another. An objective of this work is to produce a step toward an automation of such examination through the use of computer vision applied on images such as Figs. 2 and 3 which are automatically produced after acquisition on the system of Fig. 1. Two datasets were produced for this study to test the proposed machine vision approach for distinctness evaluation.

- Non-Gala Mutant varieties

We first created a dataset of images of apples with highly distinct color distributions. The dataset is composed of 1293 images of apples belonging to 8 varieties (see Fig. 2) which we refer to as non-Gala Mutants. These varieties correspond to varieties identified as distinct from each other by the official examining offices. These varieties are not named, they simply have a reference number to identify them.

- Gala mutants

As a complement to the first dataset, we built a second dataset containing 4040 images of apples belonging to 9 different mutants of the variety Gala. These mutants are similar to each others in

terms of color content as shown in Fig. 3. These mutants are also considered as distinct from each other by experts of EO but they somehow reach the limit of what they consider as distinct.

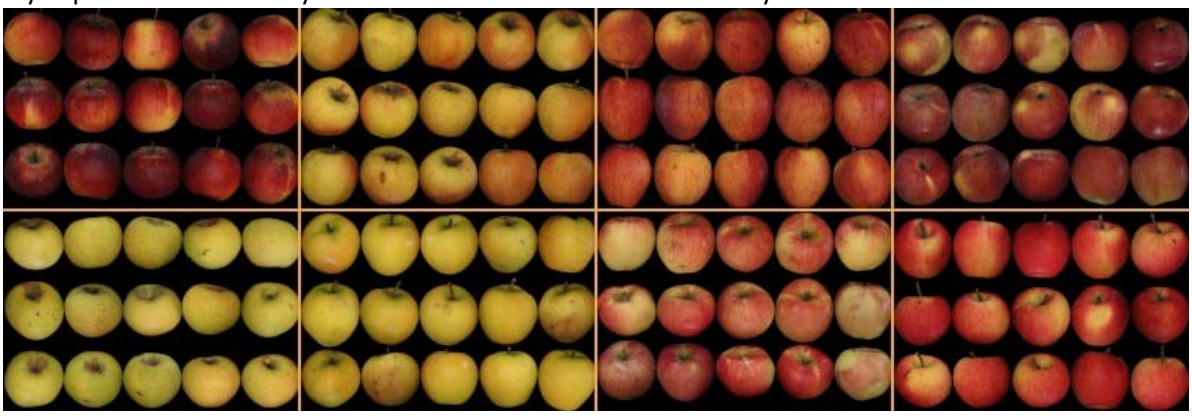


Figure 2. Images representing the 8 non-mutant Gala varieties.

2.1.3 3D RGB Histograms

With our objective being to differentiate apples mainly based on the color distribution, we extracted features from the RGB histogram of the images represented in 3 dimensions (one axis by component color). We calculated the RGB histogram of each image, and to obtain the RGB histogram of a variety, we simply calculated the sum of the RGB histograms of the images belonging to the variety. We can see the corresponding summed histogram of each of the non-Gala mutants in Fig. 4 and of the Gala mutants (except X8594) in Fig. 5. It is interesting to see that despite the loss of spatial localiation in RGB histograms a contrast between colors is clearly visible in this representation with the non-Gala mutants in Fig. 4. However, the contrast is much more difficult to perceive with RGB histograms in the case of the Gala mutants which represents a clearly challenging classification task.

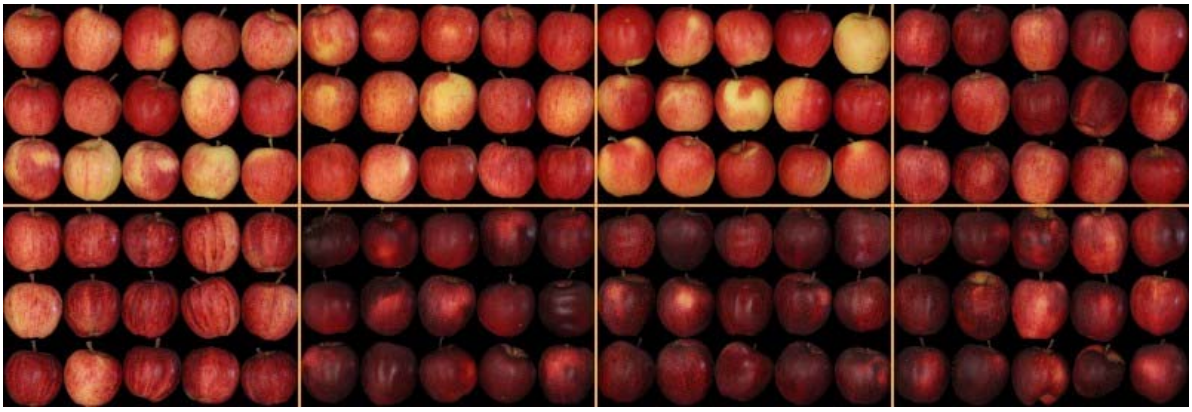


Figure 3. Images showing a subset of each Gala Mutant.

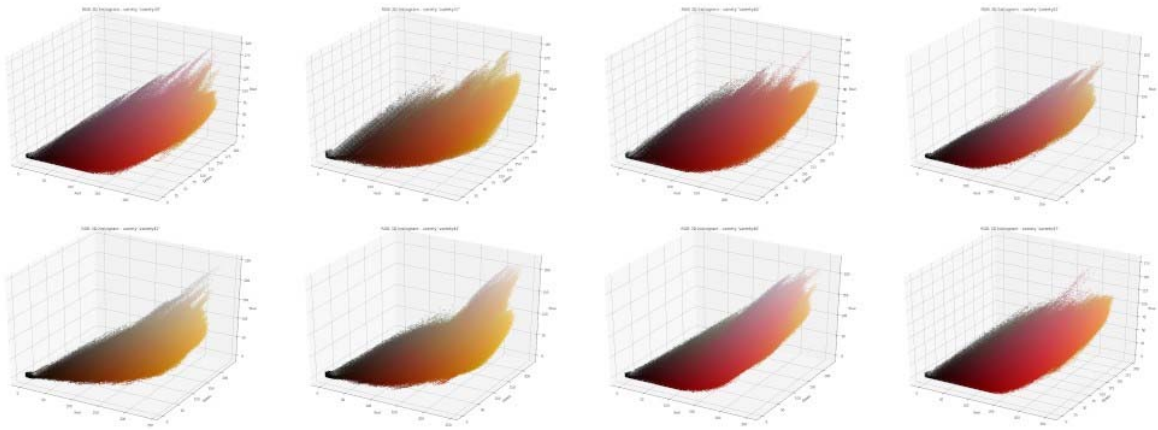


Figure 4. Images representing histograms of non-mutant Gala varieties. From left to right, by row: variety30, variety37, variety40, variety41, variety42, variety44, variety46 and variety47.

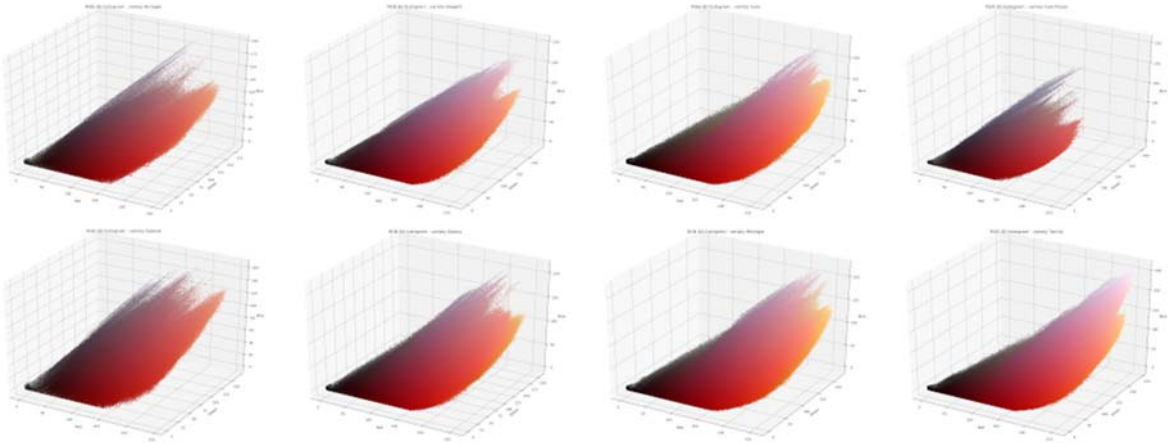


Figure 5. Images representing histograms of mutant Gala varieties. From left to right, by row: X4111, X4410, X4712, X6716, X7440, X7812, X8125, X9214.

2.1.4 Color features

Once the histogram of each image and each variety is built, we extract features allowing to characterize several aspects of the histograms. Same features were used for both datasets. We present the features used for this study in the rest of this section.

Average and variance of colors: The first two descriptors are the mean and the variance of the colors. It seems quite intuitive to use them since they give us respectively the average color of the variety of apples and the contrast of yellow and red regions are captured via the variance. These two values are easily calculated.

Fractional Anisotropy: Fractional anisotropy is a number in the interval [0, 1] which reflects the degree of anisotropy of the shape of the point cloud formed by the 3-dimensional histogram. This scalar gives a measure of the stretch of the point cloud in various directions. If its value is 1, it means that the points would all be distributed along a perfectly linear axis. If its value is zero, it means that the points are distributed homogeneously in all directions. Thus, a sphere has a zero fractional anisotropy, an ellipse has a fractional anisotropy between 0 and 1 and a straight line has a fractional anisotropy equal to 1. After obtaining the eigenvalues λ_1, λ_2 and λ_3 of the PCA (Principle Components Analysis) on the 3-dimensional histogram of an image, we can easily calculate the fractional anisotropy, denoted as *FA.*

Fractal box counting dimension: The fractal box counting method [2] subdividing the 3D color cube $[0, 255]^3$ into '*box*' of the edge length *r* counts the number of boxes *N* (*r*) necessary to cover each color cell occupied by the point cloud making up the 3D histogram. This number of boxes *N* (*r*) has been found to follow a law of the form r^{-D} where *D* is the fractal dimension of the histogram [6, 7]. This number ranging between 2 and 3 for natural images provides a description of the structure and density of the point cloud. A smaller value of fractal dimension indicates that although the histogram is distributed throughout the color space, there remain empty regions.

Mutual entropy: Mutual entropy [8] allows us to compute the information common to 2 histograms. We include this mutual entropy as the measure of the color similarity between an image and the target variety.

Cost of optimal transport: As last feature we propose to include optimal transport [21] which provides a way of transporting a set of points to another set in the least expensive way possible. In our case minimizing the total distance between the two sets of points fits with the capability of optimal transport. Since we work on 3-dimensional histograms of our images, we can measure the cost in terms of the distance between the histogram of an image and the average histogram of a target variety. If we assume we have *k* varieties (here *k* = 8 or 9 depending on the dataset used), we get *k* values representing the cost of moving our image to the *k* varieties. These *k* values are treated as color features each representing a measure of the probability of the image to belong to the corresponding variety. In practice, we calculated this cost using the python package named POT (Python Optimal Transport, https://pythonot.github.io/). The cost is computed using the method based on earth mover's distance, from 3D histograms. This algorithm has 2 advantages : histograms do not need to be normalized and they do not need to be of the same size [3].

All in all the feature space is composed of features of various dimensions. The optimal transport feature is a vector for which each component is the value of the norm of the cost between two varieties. Therefore if the dataset is composed of *k* varieties, the optimal transport is a vector with *k* components. The other features can be scalars as fractal dimension or fractional anisotropy, 3 components vectors as RGB means and RGB variance, or a vector of the same dimension as the optimal transport for mutual entropy.

2.1.5 Classification setups

In this part, we detail the machine learning classification setups tested to assess distinctness with both apple datasets presented in the previous section.

Multi-class classification A first setup is simply to perform a multi-class classification between the varieties allowing to assess if the varieties are distinguishable between them. This is a "one versus one" approach where the tested variety is tested against all the existing ones individually. For this, we simply separated an initial dataset of images to create 2 sub-datasets: a test sub-dataset containing $\frac{1}{3}$ of the images, and a training sub-dataset containing the rest. These 2 sub-sets were used respectively to train the supervised classifiers and to test their efficiency to distinguish the different varieties. For this classification, two sets of features were used, a set containing all features and the other set containing only the most relevant features among all the tested features.

The second classification setup was used to test if our model was able to differentiate the two apple datasets. This is a "one versus all" approach where the one tested corresponds to the variety compared with all the existing registered varieties at once. We gathered the 2 datasets presented previously, thus constituting a dataset of 5333 images, with 4040 images of Gala

mutants and 1293 non Gala mutants which are our 2 classes. We separated the dataset into test and training sub-sets with a 50-50 ratio. To mimic the procedure experts currently follows for apple variety testing, the algorithm made an individual prediction on each apples and a majority voting over subsets of 30 apples.

2.2 Classification results

2.2.1 Multi class classification between Gala mutant varieties

We first performed the multi-class classification between Gala mutant apples only, in order to verify that it was indeed possible to distinguish these 9 registered varieties between them. We first separated the data into test and training sets with a ratio of for the training set, then we used 3 different supervised classifiers: Support vector machine (SVM), Random Forest and Linear Discriminant Analysis (LDA).

The classification results show that the Gala mutant varieties are distinguishable in terms of color. These results are of the same order of magnitude for the three tested classifiers. However, SVM gives an F1-score over 97%, and perform slightly better than others.

A forward analysis, testing the performance of each individual type of features, identified that the best features for the classification happened to those from optimal transport. These features alone do not allow us to obtain a fully satisfactory classification, however they are relatively efficient since they yield a classification accuracy of about 50%. Since our dataset has 9 distinct classes, a random classification of the data would give 11%. accuracy. The relative superiority of optimal transport toward the other features can be explained since all histograms share the same elongated shape centered on red-yellow barycenter.

2.2.2 Multi class classification between non Gala mutant varieties

In a second step, we performed the same classification method as in the previous section, this time using the dataset composed of the non-Gala mutant varieties.

For this dataset, the results are also very satisfactory, with F1-scores close to 90%. Once again, the SVM with polynomial kernel gives the best results with a precision score of 93.76%. For the Non-Gala mutants, these results show that the varieties composing this dataset are clearly distinguishable as also confirmed by the experts from EO since these are registered as official varieties. The precision score is logically found a bit lower than for the non-Gala mutant datasets since the contrast in color between varieties is lower.

Again we performed a forward analysis which established optimal transport as providing the most significant features. The classification only based on these optimal transport features reached their best results with SVM with polynomial kernel of degree 3, which gives a precision score of 71.25%. Globally the result observed with the well contrasted dataset non-Gala mutant are robustly conserved when the method is transposed to less contrasted apples such a the one of the Gala mutants. This demonstrates the high potential of a machine learning framework equipped with color features for variety testing.

2.2.3 Binary classification with the 2 collected datasets

Once we observed that both datasets were well distinguishable, we focus on the most difficult dataset and explore the potential of our framework to determine whether a set of test images corresponds to a certain Gala mutant or not. To mimic the way experts perform their scoring, we decided to focus not only on individual classification of apples but also on a group classification from the same variety. To do so, we selected images from the test data and by a random draw without replacement of apples of the same variety, to create subsets of 30 apples. This number exactly corresponds to the size of the group of apples chosen by the experts when they observe

groups of apple for distinctness. Once our model is trained on classification of individual apple images, we tested its efficiency on the subsets through majority voting.

We get 100% F1-score with all classifiers when all features are employed. Consistent with the results of the previous section optimal transport again appeared as the most important features in a forward analysis. With optimal transport only, Random Forest gives the best results in individual classification with a precision of 88.13% and an F-score of 75.67%.

2.3 Conclusion and perspectives on phenotyping

In this work, we have considered, for the first time to the best of our knowledge, a variety testing problem with a machine learning approach. We have introduced on a use-case dedicated to apples a supervised learning scheme to identify if a new candidate for variety registration could be considered as distinct or not from an existing set of varieties. Two datasets corresponding to highly contrasted varieties and varieties contrasted at the limit of what would be considered as distinct have been tested. Distinctness was found in perfect accordance with the human expert. This demonstrates the possibility to introduce more objective and higher-throughput approaches in the domain of variety testing. We found that among the tested features optimal transport was producing the most adapted features, i.e. which contributed the most in the correct decision making. It is specially important to notice that all these results were obtained based on color histogram, i.e. with a total loss of spatial information.

This first step opens various ways of further investigations. A limit of the result presented so far stands in the absence of negative data, i.e. non registered varieties in our dataset. Access and diffusion of such historical data is complex from a legal point of view when dealing with EO. A workaround approach could consist in simulating fake non registered varieties from an existing dataset. This requires to enlarge the datasets used in this article and we are currently investigating this direction. On the machine learning side, several alternatives could be considered. We selected classical shallow learning algorithms (SVM, random forest and LDA). We produced binary decisions in accordance with the essence of distinctness which is a binary trait. All the tested algorithms could also provide probabilities and confidence intervals which would provide more insights. Such output, although not currently in practice in variety testing would nonetheless be very useful specially to provide arguments to the breeders when new variety candidates are rejected by EO. The set of hand crafted features could be extended to additional color features mentioned in the related work section. Also, all the analysis were performed in the native RGB color space and other color spaces more suitable to fit with the human perception could also be tested with the approach introduced in this paper. Alternatively deep learning approaches could be considered. An obvious match would be with generative adversarial networks (GAN) where the discriminator network could serve to decide if a variety is distinct from another after the GAN would have been trained to reproduce images of already registered varieties.

Varieties are registered based on a large set of parameters. Here, we considered only color, but it would be interesting to extend the scheme to incorporate more parameters. On apple it could be color texture (stripes on apple skin) as well as shape, other DUS traits or resistance to pests and diseases. With such extended features, the apple variety would be represented as point clouds similarly to what was found here for color histogram. In this sense, the illustration provided in this report on color would actually be extendable without any effort to any kind of characteristics to be tested for automation in variety testing. The proposed approach specially with optimal transport can be adapted to higher dimensional feature spaces and thus offers a generic framework to extend the quest for automation in variety testing.

3. General conclusions and perspectives

The imoddus apple mutant project has led to important results. Initially focused on genetic and epigenetic approaches, we have reoriented it to make much more room for phenotyping, which makes it possible to envisage applications in the fairly short term to improve the distinction of colored apple mutants. But before delivering a protocol entirely based on new phenotyping tools there are still few studies to do mainly on a statistical point of view. We are already preparing these future steps.

We knew from the beginning that the genetic and epigenetic approaches would be more ambitious and would deliver lower TRL outputs. In the first part of the project, Wuqian Wang, in her PhD thesis, has produced a lot of data and obtained preliminary results but unfortunately, we lacked a bioinformatician to analyze in depth the methylome sequencing data generated. S. Hatira, hired by the project, developed the BISEPS pipeline and an intuitive and user-friendly interface allowing biologists to analyze big sets of data and test various parameters. At the very end of the project, thanks to these tools, we have been able to identify a set of stable epigenetic marks which characterize the Gala colored mutants under study. Further statistical analyses are underway to fully exploit the generated data sets and get more marks. The next steps are now performed in the frame of the INVITE project to 1) integrate the transcriptomic and genetic data and 2) test the tools on the same set of Gala mutants in various environments.