Universitat Autònoma de Barcelona

DEEP AND SHALLOW LEARNING SOLUTIONS FOR MODERN AGRICULTURE

Laura M. Zingaretti

DOCTORAL THESIS

DEEP AND SHALLOW LEARNING SOLUTIONS FOR MODERN AGRICULTURE

Author

Laura M. Zingaretti

Supervisors

Dr. Miguel Pérez-Enciso
Dra. Amparo Monfort-Vives

Tutor

Dra. Sònia Casillas Viladerrams

Universitat Autònoma de Barcelona

Departament de Genètica I de Microbiologia
Facultat de Biociències
Universitat Autònoma de Barcelona

2021

CRAG CENTRE FOR RESEARCH IN AGRICULTURAL GENOMICS

EXCELENCIA SEVERO OCHOA

DEEP AND SHALLOW LEARNING SOLUTIONS FOR MODERN AGRICULTURE

SOLUCIONES DE APRENDIZAJE PROFUNDO Y SUPERFICIAL PARA LA AGRICULTURA MODERNA

SOLUCIONS D'APRENENTATGE PROFUND I SUPERFICIAL PER A L'AGRICULTURA MODERNA

Memòria presentada per-María Laura Zingaretti Viano
per a optar al gau de Doctor en Genètica per
la Universitat Autònoma de Barcelona

María Laura Zingaretti Viano	Dr. Miguel Pérez- Enciso	Dra. Amparo Monfort-Vives.	Dra. Sònia Casillas Viladerrams
Author	Director	Directora	Tutora

Bellaterra, 24 de Març de 2021

This PhD thesis has been funded by the grant from the Ministry of Economy and Science (MINECO, Spain), by the MINECO grant RTA2013-00010-00-00 to AMV, AGL2016-78709-R to MPE and from the EU through the BFU2016-77236-P (MINECO/AEI/FEDER, EU) and the "Centro de Excelencia Severo Ochoa 2016-2019" award SEV-2015-0533.

Cover designed by Leónidas

«*La ciencia ha sido hasta ahora un proceso*
de eliminar la confusión absoluta de las cosas
mediante hipótesis que lo explican todo;
un proceso originado en la repugnancia
del intelecto por el caos»
FN

Gracias a Alexandra y a "Google", que me han permitido
desarrollar mi espíritu autodidacta

A Maruca y Alberto, mis abuelos,
de quienes aprendí
qué es el amor incondicional y
que extraño cada día

A mi hermano Leo,
al que admiro profundamente

Al "viejo" Coccioli,
porque nunca olvidaré
las incontables horas de
cafés matemáticos y literarios

TABLE OF CONTENTS

Summary

Modern agriculture relies heavily on sophisticated computational tools that involve genomics and phenomics data at a large scale. As for genomics, over the past few decades, plant and animal breeders have taken advantage of genomic selection (GS), which is the breeding strategy that consists of predicting complex traits using genomic wide genetic markers. GS has two main advantages over traditional approaches: increasing genetic gain and reducing the amount of data to be tested in the field. In parallel, the implementation of electronic, sensors, digital cameras, unmanned aerial vehicles, mass spectrometry, among others, have opened a window of opportunities in the 'phenomics' area, rapidly increasing the amount of available data. Phenotyping does not end here, as 'omics' technologies also provide a new source of information, allowing not only the characterization of the organism itself but also of its metagenome. The current challenge is to transform and combine all these heterogeneous data into valuable information that helps the breeder to make better and more effective decisions.

The present work deals with a variety of genomic prediction and phenomics problems, all with the shared objective of exploring the strengths and weaknesses of ML techniques in agriculture. The first two contributions deal with genomic prediction issues while the following two chapters are concerned with phenomics, followed by the last research on data integration.

In chapter 3, we develop a versatile forward simulation tool, called polyploid Sequence-Based Virtual Breeding (pSBVB) to evaluate genomic selection strategies in polyploids. pSBVB is an efficient gene dropping software that can simulate any number of complex phenotypes and can address many genomic selection strategies in polyploids. We use pSBVB to evaluate the potential advantage of using GP in two important polyploid species, auto-tetraploid potato, and allo-octoploid strawberry. Overall, we show that, while genomic selection is a promising breeding strategy for polyploids, the actual advantage critically depends on the underlying genetic architecture.

In Chapter 4, we compare the application of deep neural networks and traditional linear models on genomic prediction problems, using data from two important polyploid species: strawberry and blueberry. Regarding deep learning, we focus on two well-known architectures: The Multi-layer Perceptron (MLP) and Convolutional Neural Networks (CNN). Our main results indicate that there is no clear advantage of neural networks over linear methods, except when the epistasis component is important. However, using a parameterization capable of accounting for these nonlinear effects, Bayesian linear models can match or exceed the predictive accuracy of neural networks. Furthermore, we have shown that the predictive ability of neural networks is critically affected by the combination of hyperparameters, so finding the best neural network is not a trivial task and is computationally expensive.

In chapter 5, we evaluate fruit morphology by automatic analysis of digital images. We present a data analysis pipeline that segments, classifies and labels the images, extracts conformation features, including linear (area, perimeter, height, width, circularity, shape descriptor, the ratio between height and width) and multivariate statistics (Fourier Elliptical components and Generalized Procrustes). Internal color patterns are obtained using an autoencoder to smooth out the image. Besides, we develop a variational autoencoder to automatically detect the most likely number of underlying shapes in the database. We also resort to Bayesian modeling to estimate additive and dominance effects for all recorded traits. The research shows that fruit shape and color can be quickly and automatically evaluated and are moderately heritable.

In chapter 6, we revisit the problem of automatically evaluate morphological traits, focusing this time on the dairy industry. Assessing conformational cows' features accurately and rapidly is a challenge, mainly due to the difficulty of removing the background when evaluating field images. While recent developments in computer vision have greatly improved automated background removal, these methods have not been fully translated to biological studies. The work presents a composite method (DeepAPS) that combines two readily available algorithms to create a precise mask for an animal image. This method performs accurately when compared with manual classification of the proportion of coat color with an adjusted $R^2 = 0.926$. Using the output mask, we can automatically extract useful phenotypic information for fourteen additional morphological features. Using pedigree and image information from a web catalog (www.semex.com), we estimated high heritability, indicating that meaningful biological information has been extracted automatically from imaging data. This method can be applied to other datasets and requires only a minimal number of image annotations (~50) to train this partially supervised machine-learning approach. DeepAPS allows for the rapid and accurate quantification of multiple phenotypic measurements while minimizing study cost.

In Chapter 7, we address the problem of data integration. The decreasing cost of "omics" technologies facilitates the simultaneous study of gene expression, proteins, metagenomics and metabolites, and the investigation of their relationship with complex traits. However, the integration of heterogeneous data is not trivial. Here, we have developed Link-HD, an R package for integrating multiple datasets based on STATIS-ACT ('Structuration des Tableaux A Trois Indices de la Statistique -Analyse Conjointe de Tableaux'), a family of methods designed to integrate information from diverse subspaces. Our software extends the classical approach by incorporating distance matrices for numerical, categorical and compositional data, a variable selection method, a differential abundance test and a hypergeometric taxon enrichment analysis (HyTE) to analyze whether there is an enrichment of genera (families) in the selected taxa. We illustrated the methodology by integrating microbial communities (Bacteria, Archaea and Protozoa) from 65 Holstein cows from which methane (CH4y) production was measured individually, In the problem addressed, we found a common subspace composed of a mixture of the three communities, reflecting the existence of three "ruminotypes" as previously described in the literature. Additionally, the HyTe test allowed us to identify several families of bacteria and archaea associated with CH4y emission.

The results obtained here show how machine learning (ML) techniques can empower modern agriculture in multiple avenues. However, much work remains to be done and specific ML developments are required to enhance genetic gain in breeding programs.

Resumen

La agricultura moderna depende ampliamente del uso de sofisticadas herramientas informáticas para analizar datos masivos, tanto genotípicos como fenotípicos. La selección genómica (SG), que consiste en predecir rasgos complejos utilizando marcadores genéticos de amplio espectro, ha sido aprovechada por los mejoradores de plantas y animales a lo largo de las últimas décadas, para producir un considerable aumento de la ganancia genética, reduciendo el número de muestras a testear en el campo. Paralelamente, la implantación de la electrónica, los sensores, las cámaras digitales, los vehículos aéreos no tripulados y la espectrometría de masas, entre otros, han abierto una ventana de oportunidades en el ámbito de la "fenómica", aumentando rápidamente la cantidad de datos disponibles. Todas las tecnologías 'ómicas' también proporcionan nuevas fuentes de información, permitiendo no sólo la caracterización del propio organismo, sino también de su metagenoma. Por lo tanto, uno de los mayores desafíos actuales es combinar todos estos datos heterogéneos para transformarlos en información valiosa que ayude a los mejoradores a tomar decisiones eficaces.

El presente trabajo aborda una variedad de problemas de predicción genómica y fenómica, todos ellos con el objetivo común de explorar ventajas y desventajas del uso de ML en agricultura. Las dos primeras contribuciones tratan de problemas de predicción genómica, mientras que los dos capítulos siguientes se ocupan de la fenómica, y la última investigación trata de la integración de datos.

En el capítulo 3, desarrollamos una herramienta de simulación llamada Polyploid Sequence-Based Virtual Breeding (pSBVB) que se puede utilizar para la evaluación de diferentes estrategias de selección genómica en especies poliploides. Utilizando esta herramienta de simulación avanzada, valoramos la potencial ventaja del uso de predicción genómica en dos importantes especies poliploides, la patata auto-tetraploide y la fresa alo-octoploide. Nuestros resultados indican que, aunque la selección genómica es una estrategia de mejora prometedora para los poliploides, la ventaja real depende críticamente de la arquitectura genética subyacente.

En el capítulo 4, comparamos la aplicación de las redes neuronales profundas y los modelos lineales tradicionales sobre problemas de predicción genómica, utilizando datos de dos importantes especies poliploides: la fresa y el arándano. En cuanto al aprendizaje profundo, nos centramos en dos arquitecturas bien conocidas: El Perceptrón Multi-capa y las Redes Neuronales Convolucionales. Nuestros principales resultados indican que no existe una ventaja clara de las redes neuronales sobre los métodos lineales, excepto cuando el componente de epistasis es importante. Sin embargo, utilizando una parametrización capaz de tener en cuenta estos efectos no lineales, los modelos lineales bayesianos pueden igualar o superar la precisión predictiva de las redes neuronales. Además, hemos demostrado que la capacidad predictiva de las redes neuronales se ve críticamente afectada por la combinación de hiperparámetros, por lo que encontrar la mejor red neuronal no es una tarea trivial y es costosa desde el punto de vista computacional.

En el capítulo 5, presentamos un trabajo que evalúa la morfología de la fruta mediante el análisis automático de imágenes digitales. Nuestro método segmenta, clasifica y etiqueta las imágenes, extrae características de conformación, tanto lineales (área, perímetro, altura, anchura, circularidad, descriptor de forma, la relación entre la altura y la anchura), como multivariadas (componentes elípticos de Fourier y Procrustes Generalizado). También recurrimos a técnicas de aprendizaje profundo, específicamente los autocodificadores (del inglés, autoencoders) para suavizar la imagen y estimar los patrones de color interno del fruto y para determinar automáticamente el número más probable de formas subyacentes en la base de datos. Es importante resaltar que, estos métodos, permiten también la generación automática de formas. Finalmente, estimamos los parámetros genéticos de todos los rasgos identificados. La investigación

demuestra que la forma y el color de diversas frutas pueden evaluarse de forma rápida y automática y son moderadamente heredables.

En el capítulo 6, retomamos el problema de la evaluación automática de los rasgos morfológicos, centrándonos en la industria del vacuno lechero. La evaluación automática de rasgos morfológicos de vacunos a través de imágenes digitales es un desafío, dado que las imágenes tomadas en el campo tienen fondos complejos. Aunque los recientes desarrollos en visión por ordenador han mejorado enormemente la eliminación automática del fondo, su aplicación en estudios biológicos es todavía escasa. Este trabajo presenta un método (DeepAPS) que combina dos algoritmos para crear una máscara precisa de la imagen del animal y remover automáticamente el fondo. Hemos estimado la configuración del color del ganado automática y manualmente, alcanzando una correlación de $R^2 = 0{,}926$ entre ambos enfoques. Además, hemos extraído catorce características morfológicas adicionales. Utilizando la información de pedigrí e imágenes de un catálogo web (www.semex.com), estimamos los parámetros genéticos de cada uno de los rasgos de conformación y color. Este método puede aplicarse a otros conjuntos de datos y sólo requiere un número mínimo de imágenes anotadas en el entrenamiento (~50), lo que lo convierte en una herramienta interesante para cuantificar rápida y precisamente múltiples mediciones fenotípicas a bajo coste.

En el capítulo 7, abordamos el problema de la integración de datos. El descenso del coste de las tecnologías "ómicas" facilita el estudio simultáneo de la expresión génica, de las proteínas, la metagenómica y los metabolitos, y de la investigación de su relación con rasgos complejos. Sin embargo, la integración de datos heterogéneos no es trivial. Aquí, hemos desarrollado Link-HD, un paquete de R para integrar múltiples conjuntos de datos basado en STATIS-ACT ('Structuration des Tableaux A Trois Indices de la Statistique -Analyse Conjointe de Tableaux'), una familia de métodos diseñados para integrar información de diversos sub-espacios. Nuestro software amplía el enfoque clásico incorporando matrices de distancia para datos numéricos, categóricos y composicionales, un método de selección de variables, una prueba de abundancia diferencial y un análisis de enriquecimiento hipergeométrico de taxones (HyTE) para analizar si existe un enriquecimiento de géneros (familias) en los taxones seleccionados. Hemos ilustrado la metodología integrando comunidades microbianas (Bacterias, Arqueas y Protozoos) de 65 vacas Holstein de las que se midió individualmente la producción de metano (CH4y), En el problema abordado, encontramos un sub-espacio común compuesto por una mezcla de las tres comunidades, que refleja la existencia de tres "ruminotipos" como se ha descrito previamente en la literatura. Adicionalmente, la prueba HyTe nos permitió identificar varias familias de bacterias y arqueas asociadas a la emisión de CH4y.

Los resultados obtenidos aquí muestran cómo las técnicas de aprendizaje automático (ML) pueden potenciar la agricultura moderna en múltiples vías. Sin embargo, queda mucho trabajo por hacer y se requieren desarrollos específicos de ML para potenciar la ganancia genética en los programas de mejora.

Resum

L'agricultura moderna depèn àmpliament de l'ús de sofisticades eines informàtiques per analitzar dades massives, tant genotípiques com fenotípiques. La selecció genòmica (SG), que consisteix en predir característiques complexes utilitzant marcadors genètics d'ampli espectre, ha estat aprofitada pels milloradors de plantes i animals, al llarg de les últimes dècades, per produir un considerable augment del guany genètic, reduint el nombre de mostres a testar al camp. Paral·lelament, la implantació de l'electrònica, els sensors, les càmeres digitals, els vehicles aeris no tripulats i l'espectrometria de masses, entre d'altres, han obert una finestra d'oportunitats en l'àmbit de la "fenómica", augmentant ràpidament la quantitat de dades disponibles. Totes les tecnologies 'òmiques' també proporcionen noves fonts d'informació, permetent no només la caracterització del mateix organisme, sinó també seva del seu metagenoma. Per tant, un dels majors reptes actuals és combinar totes aquestes dades heterogènies per transformar-les en informació valuosa que ajudi els milloradors a prendre decisions eficaces.

El present treball aborda una varietat de problemes de predicció genòmica i fenómica, tots ells amb l'objectiu comú d'explorar avantatges i desavantatges de l'ús de ML en agricultura. Les dues primeres contribucions tracten de problemes de predicció genòmica, mentre que els dos capítols següents s'ocupen de la fenómica, i l'última investigació tracta de la integració de dades.

En el capítol 3, desenvolupem una eina de simulació anomenada Polyploid Sequence-Based Virtual Breeding (pSBVB) que es pot utilitzar per a l'avaluació de diferentes estratègies de selecció genòmica en espècies poliploids. Utilitzant aquesta eina de simulació avançada, valorem la potencial avantatge de l'ús de predicció genòmica en dos importants espècies poliploides, la patata auto-tetraploide i la maduixa alo-octoploide. Els nostres resultats indiquen que, tot i que la selecció genòmica és una estratègia de millora prometedora per als poliploides, l'avantatge real depèn críticament de l'arquitectura genètica subjacent.

En el capítol 4, comparem l'aplicació de les xarxes neuronals profundes i els models lineals tradicionals sobre problemes de predicció genòmica, utilitzant dades de dos importants especies poliploids: la maduixa i el nabiu. Pel que fa a l'aprenentatge profund, ens centrem en dos arquitectures ben conegudes: El Perceptrón Multi-capa i les Xarxes Neuronals convolucionals. Els nostres principals resultats indiquen que no hi ha un avantatge clar de les xarxes neuronals sobre els mètodes lineals, excepte quan el component de epístasi és important. No obstant això, utilitzant una parametrització capaç de tenir en compte aquests efectes no lineals, els models lineals bayesians poden igualar o superar la precisió predictiva de les xarxes neuronals. A més, hem demostrat que la capacitat predictiva de les xarxes neuronals es veu críticament afectada per la combinació de hiperparàmetres, de manera que trobar la millor xarxa neuronal no és una tasca trivial i és costosa des del punt de vista computacional.

En el capítol 5, vam presentar un treball que avalua la morfologia de la fruita mitjançant l'anàlisi automàtic d'imatges digitals. El nostre mètode segmenta, classifica i etiqueta les imatges, extreu característiques de conformació, tant lineals (àrea, perímetre, alçada, amplada, circularitat, descriptor de forma, la relació entre l'altura i l'amplada), com multivariades (components el·líptics de Fourier i Procrustes Generalitzat). També vam recórrer a tècniques d'aprenentatge profund, específicament els autocodificadors (de l'anglès, autoencoders) per suavitzar la imatge i estimar els patrons de color intern del fruit i per determinar automàticament el nombre més probable de formes subjacents a la base de dades. És important ressaltar que aquests mètodes permeten també la generació automàtica de formes. Finalment, estimem els paràmetres genètics de tots els caracters identificats. La investigació demostra que la forma i el color de diverses fruites poden ser avaluades de forma ràpida i automàtica i són moderadament heretables.

En el capítol 6, revisitem el problema de l'avaluació automàtica dels trets morfològics, centrant-nos en la indústria del boví lleter. L'avaluació automàtica de trets morfològics de bovins a través d'imatges digitals és un desafiament, donat que les imatges preses en el camp tenen fons complexos. Tot i que els recents desenvolupaments en visió per ordinador han millorat enormement l'eliminació automàtica del fons, la seva aplicació en estudis biològics és encara escassa. Aquest treball presenta un mètode (DeepAPS) que combina dos algoritmes per a crear una màscara precisa de la imatge de l'animal i remoure automàticament el fons. Hem estimat la configuració del color de la ramaderia automàtica i manualment, aconseguint una correlació de R^2 =0,926 entre tots dos enfocaments. A més, hem extret catorze característiques morfològiques addicionals. Utilitzant la informació de pedigrí i imatges d'un catàleg web (www.semex.com), estimem els paràmetres genètics de cada un dels trets de conformació i color. Aquest mètode pot aplicar-se a altres conjunts de dades i només requereix un nombre mínim d'imatges anotades en l'entrenament (~ 50), el que el converteix en una eina interessant per quantificar ràpida i precisament múltiples mesures fenotípiques a baix cost.

En el capítol 7, abordem el problema de la integració de dades, focalitzant-nos en l'anàlisi de múltiples fonts de dades del microbioma del bestiar boví. El descens de el cost de les tecnologies "òmiques" facilita l'estudi simultani de l'expressió gènica, de les proteïnes, la metagenòmica i els metabòlits, i de la investigació de la seva relació amb caràcters complexes. No obstant això, la integració de dades heterogenis no és trivial. Aquí, hem desenvolupat Link-HD, un paquet de R per integrar múltiples conjunts de dades basat en STATIS-ACT ('Structuration des Tableaux A Trois Índexs de la Statistique -Analyse Conjointe de Tableaux'), una família de mètodes dissenyats per integrar informació de diversos sub-espais. El nostre programari amplia l'enfocament clàssic incorporant matrius de distància per dades numèriques, categòriques i composicionals, un mètode de selecció de variables, una prova d'abundància diferencial i una anàlisi d'enriquiment hipergeomètric de tàxons (HyTE) per analitzar si hi ha un enriquiment de gèneres (famílies) en els taxons seleccionats. Tot i que hem il·lustrat la metodologia integrant de comunitats microbianes (bacteris, arqueobacteris i Protozous) de 65 vaques Holstein de les que es va mesurar individualment la producció de metà (CH_4y), aquesta es pot aplicar a problemes més generals. En el problema abordat, trobem un sub-espai comú compost per una barreja de les tres comunitats, que reflecteix l'existència de tres "ruminotipos" com s'ha descrit prèviament a la literatura. Addicionalment, la prova HyTe ens va permetre identificar diverses famílies de bacteris i arqueees associades a l'emissió de CH_4y.

Els resultats obtinguts aquí mostren com les tècniques d'aprenentatge automàtic (ML) poden potenciar l'agricultura moderna en múltiples vies. No obstant això, queda molta feina per fer i es requereixen desenvolupaments específics de ML per potenciar el guany genètic en els programes de millora.

Chapter 1

General Introduction

This thesis focuses on advancing machine learning techniques for a smarter agri-food system. Agriculture, sometimes dubbed as the 'Neolithic Revolution', is one of the most important human activities and has been the predominant way of food production up to the present day.

Over the last century, our farming capacity has grown exponentially, bringing with it some undesirable effects. Currently, farmers face challenges ranging from climate change to demographic pressure. The former is perhaps the utmost threat to the planet and is increasing the drought areas, causing extreme weather conditions and the extinction of species, among others. While the latter means that food production will have to increase significantly in the coming years. These major concerns could be addressed, at least in part, by taking advantage of the enormous amount of genomic and phenomics data available and advances in data analysis techniques.

The present chapter introduces the main topics of this thesis. Section 1.1 traces from domestication to current plant and animal breeding challenges and, Section 1.2 explains the main differences between plant and animal breeding. Sections 1.3 provides an overview of Statistical Inference vs. Prediction and section 1.4 discusses the use of Bayesian or Machine Learning methods for genomic prediction. Section 1.5 introduces the key concepts of Bayesian Learning and section 1.6 deals with Deep Learning modeling used in this thesis. Section 1.7 addresses generative models; Section 1.8 is concerned with Image processing and Computer Vision in agriculture and section 1.9 outlines statistical shape analysis. Section 1.10 introduces data integration techniques in biology. We finally close this chapter with the discussion about the importance of using simulations in genomic problems in section 1.11.

1.1 Plant and Animal Breeding: from phenotype to genotype and from genome to phenome

Agriculture cannot be conceived without the concept of domestication, which is probably the most important event in the last 13000 years of mankind history. There is no consensus on the meaning of domestication, but it implies a relationship between humans and the target organism, either plant or animal [1]. The anthropocentric view, especially influenced by the practices of European animal breeders during the 19th century, emphasizes the role of humans over the reproduction, movement, distribution, and protection of domesticates [2].

In a broad sense, breeding can be defined as the alteration caused in plants and animals due to human intervention, regardless of whether they were intentional or accidental. The primary goal of breeding is to improve organisms for human welfare; but there are many secondary aims and, above all, breeding tools and strategies that have shifted over the time, e.g., in the last 100 years

breeders first used the phenotypic selection, later included information on relatives, and finally resorted to molecular tools, from marker selection to the implementation of whole-genome information [3].

Phenotypic selection relies on the breeder's ability to visually identify favorable effects on a target trait, which are transmitted to the progeny either for crossing or generation advancement. This approach, besides being imprecise and time-consuming, is quite inefficient, especially in low heritable or polygenic traits [4]. So, it is not surprising that breeders were concerned about finding the best methods. Since the late 1970s, the availability of cost-effective and flexible molecular markers, coupled with the new statistical and software, has enabled the implementation of Marker Assisted Selection (MAS) in many programs.

The impact of MAS has been less than initially envisaged, which is related to the fact that it chooses individuals that have a desirable allelic effect on the target trait, assuming that the causal mutations can be well localized. If this assumption is not met, it may work even worse than traditional phenotypic selection [4–7].

Many traits are controlled by 'small-effects' genes. At first glance this may appear to be a problem, but it has a simple solution. If a dense marker map is available, multi-locus linkage disequilibrium (LD) between Quantitative Trait Loci (QTLs) and the genome markers (e.g., Single Nucleotide Polymorphism, SNPs) can be exploited. This idea, which has revolutionized breeding, is better known and Genome-enabled Selection or Genomic Selection (GS) and was formalized by Meuwissen *et al.* [8]. While MAS only uses the markers with significant associations, GS includes all available markers, irrespective of their effect on the trait and, it is currently the standard tool in many plant and animal breeding programs [5,9–12]. GS has generated a significant improvement, nearly doubling the genetic gain in many programs and considerably reducing the amount of data to be tested on the field [11,13]. Alas, GS seems to reach a plateau; neither increasing the number of molecular markers [14] nor resorting to new analysis techniques [15,16] are contributing to improving their predictive ability.

Breeding programs do not only depend on whole-genome data, but also on large-scale phenotyping and, despite the recent genomic advances, more breakthroughs are needed in their phenomics counterpart. Breeder strategies need to shift the effort from genomic to the development of high throughput phenotyping platforms, able to screening hundreds of phenotyping data in a low cost, non-invasive and fast way [17,18]. The technology is ready, which is confirmed by the available devices, including digital cameras, hyperspectral cameras, unmanned aerial vehicles, mass spectrometry, many sensor technologies, robots, among others. However, there are still many factors that cause a bottleneck in the progress of phenomics: 1) the cost of phenomics platforms is still prohibitive for many breeders; 2) Unlike the genome, the phenome is not stable, changing throughout the life of the individual, so the development of specific analysis methods that include the temporal component is a priority; 3) The phenome also has many levels of complexity, from specific molecules to dynamic metabolic networks, and can interact with pathogens or competing organisms; 4) The huge amount of data being generated requires effort in applying the FAIR (findable, accessible, interoperable and reusable) principles [19], which serve for tracing data, protocols, methods, and workflows and; 5) As phenome includes a lot of source

of information, ranging from image to multi-omics, the development of specific data - integration methodologies is urgent.

Two natural questions arise: How can high throughput-phenotyping be used to improve plant and animal breeding? and, should plant and animal breeding schemes be modified to incorporate the new strategies? Anticipate the answer to these questions is not easy. Although the phenome is being used in many programs, it is still in its infancy. In our opinion, the challenge lies not only in designing analysis methodologies and data sharing protocols but, above all, to review the breeding equation [20] and study how it can be improved through this new paradigm.

1.2 It's not all about kinship: the differences between plant and animal breeding

Plant and animal breeding main objectives consist in the genetic improvement of complex traits by maximizing the selection gain per unit of resource spent. Although plant and animal breeding disciplines have the same theoretical principles, based on Mendel's law and Charles Darwin's postulates, the breeding methods and the underlying models have diverged [21]. While animal breeding depends on sexual reproduction, the picture is quite different in plants, where breeding has an extended tradition of outcrosses, even in self-pollinating species, and sexual mechanisms have only been used in the last 250 years [20].

The discovery of Mendel's laws in the middle of the 19th Century motivated plant breeders to adopt new strategies consisting of self-pollination or vegetative propagation. In turn, animal breeding adopted a 'biometric approach', combining the concept of 'heritability' coined by Galton in 1880 and Fisher's discoveries, which have made it possible to exploit the information of relatives [20,21].

The estimation of breeding values varies between animal and plant breeding. Most plant species can produce genetically identical individuals or reproducible cultivars, so plant breeders can accurately measure phenotypic values through well-designed experiments, accounting for location and years, i.e., the GxE interaction. The situation is markedly different in animal breeding, where statistical methods and tools had to be developed to deal with large-scale unbalanced data directly collected from the production farm [20,21]. Moreover, animal breeders have to incorporate information from relatives because some traits cannot be measured in the candidates themselves (e.g., milk yield in bulls) or may only be measured late in the breeding process.

These conceptual differences have led to a distinction in the meaning of heritability, i.e., the partition of the variance of a given phenotype. In animal genetics, heritability is computed as V_A/V_P, i.e., the ratio between the additive and phenotypic variance [6]. This value expressed the extent to which the phenotype is determined by the genes transmitted from parents to the offspring in a given population. In animal breeding, genotypes and individuals are indissociable, i.e., each genotype corresponds to a single individual; but in plant breeding, many individuals share the same genotype because most plants are clones, inbred lines, or hybrid. As a consequence, 'heritability' is interpreted as the measure of the precision of a given trial. The genotypes are tested

through different environments/years and generally, the multiple observations of the same cultivar are aggregated and computed as genotype-mean [22].

In addition, the role of the individual mutations with moderate to large effects in some traits; the use of biotechnological industry, including transgenics [20]; and the characterization of hundreds of traits benefiting from high-throughput phenotyping platforms [23,24] have a fairly important tradition in plant breeding but have been used much less in animal schemes.

Another point of divergence is the ploidy level. While most animals are diploids, meaning that they have two sets of chromosomes, several plants are polyploids. The term describes the process by which some species have an extra set of chromosomes. Many crops of important commercial value are polyploids, e.g., potato, wheat, blueberry, strawberry, sugar cane, coffee, cotton, tobacco, etc. [25–27].

Polyploid species are classified into auto-polyploids, caused by one or more genome duplication events in a single species, and allopolyploids, which are the result of hybridization between closely related species. The main difference between auto and allopolyploids lies in the process of meiosis. In the first case, it is mainly described by forming either random bivalents or multivalent during the division, while in the last, the pairing is mainly preferential, exhibiting a diploid-like (or disomic) segregation [27].

Polyploidy can induce extreme phenotypes, increasing their vigor and adaptation. Most of the polyploid "phenotypes" are larger compared to their diploid ancestors. The size of the root, fruits, flowers, etc. increases with the level of ploidy, a phenomenon known as the 'gigas' effects [28]. Besides, the polyploidy induces a 'buffering' effect that protects against a single-locus deleterious mutation in inbreeding depression; and also stimulates heterozygosity, which appears to enhance vigor in some plant species, such as potato, wheat, and alfalfa [28]. Because of these potential rewards, breeders devised protocols for polyploidy induction. But it is not all advantages, as the molecular mechanisms in polyploids are quite complex and more source of variation exists, i.e., the allelic dosage of the individual locus is larger than in diploids organisms [29,30], which may introduce higher degrees of complete and partial intra-locus interactions than diploids [31,32].

Overall, the 'genomics and phenomics' era offers a great opportunity to close the gap between the theoretical bases and methods of plant and animal breeding and, while the Genomic Selection (GS) can be widely used in more plant breeding programs; animal breeding schemes can harness the biotechnological progress, the ideas behind the high-throughput phenotypic platforms and data integration.

1.3 The trade-off between statistical inference and prediction machines

Since its emergence, the concept of GS has shifted the way in animal breeding programs and - albeit to a lesser extent - in plant breeding [10,33,34]. In addition, the advent of high-throughput phenotyping platforms that can screen a large number of genotypes at relatively low cost in a non-destructive manner is driving a revolution specially in plant breeding [23,24]. All of these methods

are built upon statistical and ML concepts so, before proceeding further, we should discuss some of these key points.

Inference and **prediction** not only are at the core of statistical and machine learning but are key concepts in genetics. Both methods are supervised tasks where the objective is to find a function describing the relationships between two sets of variables, i.e., $y = f(X)$, being $\boldsymbol{y}$ the outcome and $\boldsymbol{X}$ the set of independent variables [35–38]. In a broad sense, we could think that $\boldsymbol{f}$ is a black box. However, beyond the common goal, predicting and explaining (i.e., inferring) have two different meanings, the former is a statement about the way things will happen in the future, while the last attempts to explain how the inputs determine the output [39]. One of the terms often used to differentiate them is **interpretability**. This assumption is built on the idea that **inference** models are **interpretable**, while **predictive** models are not (or not necessarily so). But what does it mean exactly?

In Tim Miller's words[35], **interpretability** is 'the degree to which a human can understand the cause of a decision'. A model is more interpretable than another if a human can better understand the decisions it has made. Mind you, although all inference models are interpretable, not all interpretable models do inference. Statistical **inference** tests some scientific hypotheses and measures the uncertainty of the estimates; its ultimate goal is to understand the population under study. The best example in the genetic context is the Genome-Wide Association Studies (GWAS), which aim to identify genetic variants affecting phenotypes.
Indeed, in a broad sense, **inference** attempts to measure causality, however, the models most commonly used in practice are association-based and rely on the idea of statistical regression [40,41]. **Inference** is founded on strong statistical assumptions (just remember the old phrase with which many theorems and papers begin "assume that the data were generated by the following model") and if they are not met, we are in trouble.

Let us explain it for the GWAS case. It tests a single association at time and is not only powerless to capture any interaction between variants, but it also needs to resort to methods that control false discoveries. The genetic individual variables identified to be associated with a given phenotype are not enough to explain all of its variance, a phenomenon known as 'missing heritability'. GWAS can and, in fact, leads to many false discoveries, even if the family-wise error or false discovery rate are applied [42] (some funny descriptions about this can be found in https://twitter.com/SbotGwa).

While statistical modeling relies on stochastic models that mainly perform **inference**, machine learning is more associated with **predictions** and is often treated as a real "black box"[38,39]. **Prediction** can be defined as the process to apply any statistical or machine learning model with the purpose of forecast new, unobserved data. A good example of prediction in genomics is Genomic Selection (GS), based on predicting future performance using molecular information from the whole genome.

In his famous article, Leo Breiman [39] suggests that the divergence between **inference** and **prediction** is associated with two Data Analysis cultures. The **data modeling** culture aims to

estimate the parameters of the following stochastic model, $y = f(X, random\ noise, parameters)$. The **algorithmic modeling** culture assumes no thought about the process that generates the data, and its only goal is to find a function able to predict unobserved $\boldsymbol{y}$ values as accurate as possible. While in the inference modeling the goal is to find $\boldsymbol{f}$ that can be assumed as a "path", i.e., the data $\boldsymbol{X}$ and y are used to find the best function ($\boldsymbol{f}$) which, in turn, is used to test statistical hypothesis; even when **X**, y and $\boldsymbol{f}$ are in the heart of a **predictive** machine, the focus is completely different, as the three are tools to generate new accurate values of unobserved y [38].

The distinction aforementioned is not only theoretical and has certain practical implications. In our opinion, the most important one is that for prediction purposes the "wrong" model may still be the best: predictive modeling is pragmatic and may sacrifice theoretical accuracy to improve empirical precision. [38]. Table 1.1 contains a comparison between prediction and inference.

TABLE 1.1. main differences between Prediction and Inference process.

Prediction	Inference
The goal is to select the model that obtains the best predictive accuracy, even if the bias is higher.	The goal is to select the model that minimizes bias.
It is focused on predicting new outcomes.	It is focused on explaining the process that generates y through X.
The decisions are based on a loss function, i.e., it empirically minimizes the loss in a test set.	Decisions are made according to Goodness-of-fit criteria.
Lack of interpretability	It has strong interpretability.
Data Modeling Culture	Data Algorithmic Culture

The lessons we have learned from the emergency of the "Big Data", the new algorithms and computing capabilities in recent decades, can be represented using three terms, according to Breiman [39]: Rashomon, Occam, and Bellman. Rashomon is the name of a Japanese movie that tells the story of four people who are witnesses of an incident where a person dies, and another is supposedly raped. Although the four witnesses tell the same fact in court, their stories are different. The Rashomon metaphor is used to exemplify that there are many models, not just one, that might be the best to describe the relationships between two sets of variables.

The second term makes refers to Occam's razor or parsimonious principle, which owes its name to the philosopher William of Ockham. It stated *pluralitas non est ponenda sine necessitate, i.e.,* "entities are not to be multiplied beyond necessity". However, accuracy and simplicity might conflict as not always the simplest model would be the best one. While it is true that the increase in data dimensionality may affect inference modeling, e.g., linear regression or logistic regression; other models more linkage to the Algorithmic culture (e.g., Support Vector Machine (SVM), neural networks) can enhance their predictive ability exploiting the data dimensionality.

Those who defend the principle of parsimony claim that it is better to have interpretable models than black boxes, but do they question the potential dangers of a bad "interpretable" model? A misleading model leads to wrong conclusions. The higher the predictive accuracy (based on cross-validation technique), the better the model describes the underlying relationship between variables,

and we believe it is this fact that gives the predictive model an advantage over inference. A data analyst should first try to obtain high predictive performance and then look for the answer to the question Why? The third term deserves a special chapter and will be dealt with in the following subsection.

1.3.1 Is dimensionality a curse or a blessing?

Imagine you have a set of 1000 observations uniformly distributed in two dimensions, which represent any two genetical variants: the data would look exactly like Figure 1 a. But now imagine that, instead of two, you have 5 or 10 genetical variants, then the first two dimensions (out of 5/10) would look like Figure 1 b and c, respectively. The filled ball gradually transforms into a kind of a ring. In these three cases, n (the number of individuals) is still lesser than p (the number of variables). Figure 1 d,e,f represent the two first dimensions when you have 100, 1000, and 10000 variables uniformly distributed in p-dimensional space. Note that although only in the last case (Figure 1 f) p>>n, increasing the number of variables has a clear effect on the data distribution.

Figure 1.1 heuristically shows that high dimensional data become more and more sparse, increasing the distance between any two points and dropping their correlation [42]. This again has many practical implications, one of the most obvious being the question of what is an outlier? Owing to the data sparsity, the fraction of point between standard deviation (σ) from mean will decrease with increasing p. Another major implication lies in the fact that when p>n, the variables span a lower-dimensional subspace, being some dimensions redundant in the sense they can be expressed in terms of the others. It might not be a problem for agnostic models that only look for good prediction, however it would be impossible to measure the contribution of each variable in a target trait. Overall, almost all statistical theory we have learned will fail.

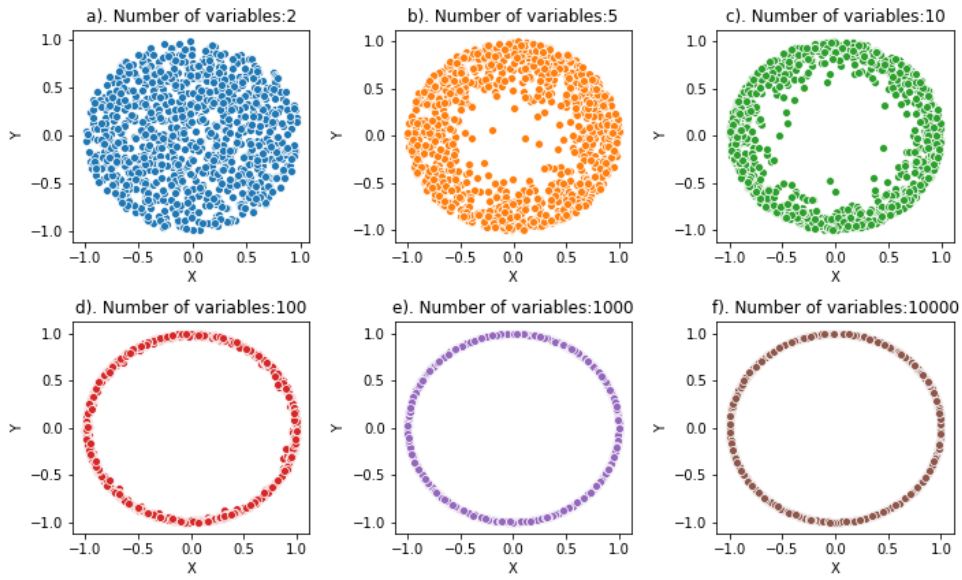


Figure 1.1: Representation of n=1000 data points uniformly distributed in a) Two dimensions, b) Five dimensions, c) Ten dimensions, d) One-hundred dimensions, e) One-thousand dimensions, f) Ten-thousand dimensions.

The issue we are discussing is called "the curse of dimensionality" and was coined by Richard Bellman [39,42–44]. The naïve intuition that stated "the more data, the better" sometimes seems to fail, or at least it can be questioned. Most of the classical statistical methods, like linear regression or logistic regression, were thought to work with more individuals than variables and they could be strongly affected by this phenomenon. However, for prediction proposes, the dimensionality could be, in fact, a blessing: we have already pointed out that the Genomic Selection (GS), a method entirely based on prediction, has taken advantage of the huge amount of genetic data available to accelerate genetic gain in many plant and animal breeding programs.

As Leo Breiman said, "the trick to being a scientist is to be open to using a wide variety of tools" and a "big data" problem requires that data scientists focused on solving problems instead of asking what model they can crate. The best solution might be a combination of the two cultures, striking a balance between prediction and inference. In this way, it may be possible to transform the curse into a blessing. There are still some practical recommendations among which we can mention the use of Bayesian models, thanks to the priors that essentially provide regularization and reduce the overfitting risk and the new methods based on Neural Networks, which are adapted to deal with high dimensional data and are better able to take into account more complex relationships between variables (e.g., epistasis).

1.4 On the prediction way: Bayesian and deep learning

Alan Turing predicted in the early 1950 that machines were going to compete with men in any knowledge area [45], and certainly, agriculture is not an exception. In a broad sense, Artificial Intelligence (AI) refers to machines able to mimic human abilities. The term is often used interchangeably with Machine Learning (ML); however, they are not the same thing. ML is a subfield of AI that needs access to a large amount of data to learn on its own, combining computer science and statistics tools [46]. Thus, it is not surprising that the development of ML tools for agriculture is a hot topic, given the enormous amount of data being generated.

Current AI is powered by ML. ML involves many concepts, including linear algebra, programming, calculus, numerical computation, information theory, and statistical skills and harnesses the computer power to learn and perform tasks without being programmed for every decision-rule. In the widest sense, we could say that ML is a Predictive Machine, intrinsically related to the Algorithmic Modeling Culture we have discussed before. Deep Learning (DL), a subfield of ML, dates from the 1940s and it only appears to be new because it was relatively unpopular for several years [47]. DL is sometimes named as Deep Artificial Neural Network since some of the earliest learning algorithms were intended to be computational models of biological learning, inspired by the workings of the brain. The modern term "deep learning" is not associated with neuroscience but is rather derived from the principle of learning multiple levels of composition.

Note ML can also be interpreted using Bayesian Learning (BL), a particular set of approaches to probabilistic ML. BL is based on Bayes' Theorem, which was first published in a post-mortem paper written by the Reverend Thomas Bayes in 1763. Although Bayes' interests were theological, he is remembered today because he proved that the probability of a cause can be deduced from

an effect[41]. BL treats the parameters of a model as random variables to be estimated from the observed data. The most important thing in BL is the need for a prior and likelihood before you can learn, i.e., you can only learn from the data based on what you already know.

Returning to the subject in hand, we are interested in Genomic Prediction (GP), i.e., those methods designed to obtain accurate predictions of genetic values of genotypes whose phenotypes are yet to be observed [8,48,49]. Here, we have used two main perspectives to handle this problem: Bayesian and Deep Learning.

1.5 Bayesian Learning for Genomic Prediction

In parametric models for GS, the phenotypes (y) are regressed on marker covariates using a linear model of the form (Eq. 1.1):

$$y = \mu + g + \varepsilon, \qquad \text{[Eq. 1.1]}$$

where $g = X\beta, \mu$ is the general intercept, $\boldsymbol{X}$ is the genotypes' matrix and β is a vector of marker effects and ε is the error term, which is assumed as $\varepsilon \sim N(0, \sigma^2 I)$. Given that the number of molecular markers (*p*) is usually higher than the number of observations (*n*), the traditional linear regression based on ordinary least squares (OLS) is not feasible and a penalization method must be used. But any statistical problem can be treated from a Bayesian perspective, where two components are needed: the likelihood and the prior density, which leads, in turn, to different models like Bayesian Lasso, Bayesian Ridge Regression, BayesC, etc. [10,50]. The general structure for GS is (Eq. 1.2):

$$\begin{aligned} & p(\mu, \beta, \sigma^2 | y, \omega) \\ & \propto p(y|\mu, \beta, \sigma^2) p(\mu, \beta, \sigma^2|\omega) \quad \text{[Eq. 1.2]} \\ & \propto \prod_{i=1}^{n} N\left(y \middle| \mu + \sum_{j=1}^{p} x_{ij}\beta_j, \sigma^2\right) \prod_{j=1}^{p} p(\beta_j|\omega)\, p(\sigma^2) \end{aligned}$$

Where $p(\mu, \beta, \sigma^2|y, \omega)$ is the posterior density of the model unknowns μ, β, σ^2 given the data (y) and the hyperparameters (ω), $p(y|\mu, \beta, \sigma^2) = \prod_{i=1}^{n} N\left(y \middle| \mu + \sum_{j=1}^{p} x_{ij}\beta_j, \sigma^2\right)$ is the conditional density of the data given the parameters, with mean $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ and variance σ^2 and $p(\mu, \beta, \sigma^2|\omega) \propto \prod_{j=1}^{p} p(\beta_j|\omega)\, p(\sigma^2)$ is the join prior density of the model unknowns, which includes the intercept (μ), to which a flat prior is generally assigned, the markers effect (β_j), to which independent and identically distributed (IID) informative priors are generally assigned and the residual (σ^2), to which a scaled-inverse chi-square prior with d.f. degree of freedom and scale parameter *S* is commonly assigned.

The key aspect in the Bayesian Framework is the choice of the prior density of the markers effects, which may be uninformative or informative. The informative priors define whether the model

induces shrinkage, variable selection, or both. The Gaussian shrinks the coefficients in the same way as Ridge Regression (RR), i.e., all of the markers are shrinking to a similar extent. The scaled-t (referred to as BayesA) and double exponential (DE) (Bayes Lasso) densities induce a type of shrinkage named "Thick tail", i.e., they have a higher mass at zero and thicker tails than the normal density. The argument behind these priors is that markers can contribute differentially to genetic variance [51]. Finally, among priors that induce both, shrinkage and variable selection, the most popular are BayesC and BayesB. The former is a mixture of a point of mass at zero and a Gaussian slab, while the latter is a mixture of a point of mass at zero and a scaled-t slab. See [10,51,52] for a more detailed discussion of Genome-Wide Bayesian methods.

Bayesian methods allow us to obtain the full posterior distribution of parameters, although they require priors' specification. In contrast, methods developed for the sole purpose of prediction such as DL can be less restrictive because its only aim is to predict new data as accurate as possible. Among the most prominent advantage of DL for genomic prediction is their ability to learn without model assumptions. This is relevant as there is no need to specify, e.g., whether the phenotype shows dominance or epistasis. Moreover, DL can model non-linear relationships since DL admits numerous non-linear activation functions. Provided enough data are available, it should be possible to find the best DL architecture, that able to learn by itself irrespective of the underlying genetic architecture.

1.6 Deep Learning Principles

A generic DL architecture is made up of a combination of several layers of 'neurons'. The neural network concept, which is the core of DL, was proposed already in the 1950s. The well-known Rosenblatt 'perceptron', inspired by brain function [53]. The DL revival during the last decade was due to the discovery of efficient algorithms that can estimate parameters in complex networks made up of several neuron layers (e.g., backpropagation [54]), and to the fact that these methods outperformed current algorithms in several automatic recognition tasks such as in image analysis [55].

1.6.1 Main Deep Learning Architectures for GP

Although all DL methods share the common principle of using stacked layers of neurons, they comprise a wide variety of architectures. The most popular ones are the Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN), and Generative Adversarial Networks (GANs). We now describe those used in genomic prediction, although the reader should be aware that many more options are available [47].

Multilayer Perceptron Network (MLP) is one of the most popular DL architectures, which consists of a series of fully connected layers, called input, hidden and output layers, respectively (Figure 1.2). In the context of genomic prediction, the first layer receives the SNP genotypes ($\mathbf{x}$) as input and the first layer output is a weighted non-linear function of each input plus the 'bias' (i.e., a constant) (Eq. 1.3).

$$z^{(1)} = b^{(0)} + \sum_{i=1}^{n_{snp}} w^{(0)} f^{(0)}(x_i) \; [Eq.\,1.3]$$

where x_i contains the i-th SNP genotypes of each individual, b is called the 'bias' and is estimated together with the rest of weights w. In successive layers, the same expression as above is used except that neuron inputs of a given layer $(a_l^{(k)})$ are the outputs from the previous layer $(z_l^{(k)})$ (Eq. 1.4):

$$z_l^{(k)} = b^{(k-1)} + \sum_{j=1}^{n_{k-1}} w_{lj}^{(k-1)} a_j^{(k-1)} \; [Eq.\,1.4]$$
$$a_l^{(k)} = f^{(k)}\left(z_l^{(k)}\right)$$

The final layer produces a vector of numbers if the target is a real-valued phenotype or an array with probabilities for each level is the target is a class (i.e., a classification problem). Although MLPs represent a powerful technique to deal with classification or regression problems, they are not the best option to manage spatial or temporal datasets. To face these constraints, other DL techniques such as Convolutional Neural Networks, Recurrent Neural Networks, or Deep Generative Networks have been proposed in recent years.

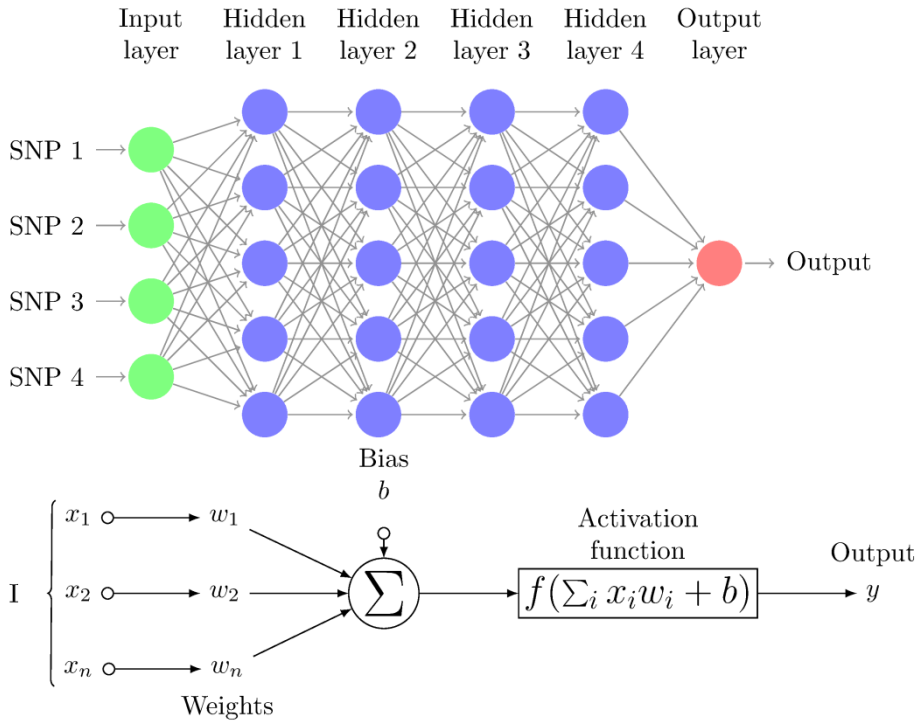


Figure 1.2. (a) Fully connected Neural network (MLP) diagram with four hidden layers and a collection of SNPs as input. **(b)** illustrates a basic 'neuron' with n inputs. One neuron is the result of applying the nonlinear transformations of linear combinations (x_i, w_i, and biases b). These figures were redrawn from tikz code in http://www.texample.net/tikz/examples/neural-network. (Figure from Pérez-Enciso and Zingaretti[56])

Convolutional Neural Networks (CNNs) were proposed to accommodate situations where input variables are distributed along a spatial pattern, say one-dimension (e.g., SNPs or text), two- or three-dimensions (e.g., images). CNNs are a special case of Neural Networks which use convolution instead of full matrix multiplication in the hidden layers [47]. A typical CNN is made up of dense, fully connected layers and 'convolutional layers' (Figure 1.3b). In each convolutional layer, a convolution operation is performed along the input of predefined width and strides. Each of these convolutional operations is called a 'kernel' or a 'filter' and is somewhat equivalent to a 'neuron' in an MLP. An activation function is applied after each convolution to produce the output. Finally, an operation called 'pooling' is usually applied to smooth out the result. It consists of merging the kernel outputs of different successive positions by taking the mean, maximum, or minimum of all values of those positions. One of the main advantages of convolution networks is their capability to reduce the number of parameters to be estimated. These networks also have sparse interactions and are equivariant to translations. An illustration of a 1D convolution with a 3-K kernel size is depicted in Figure 1.3a. Figure 1.3b shows the steps involved in a convolution network.

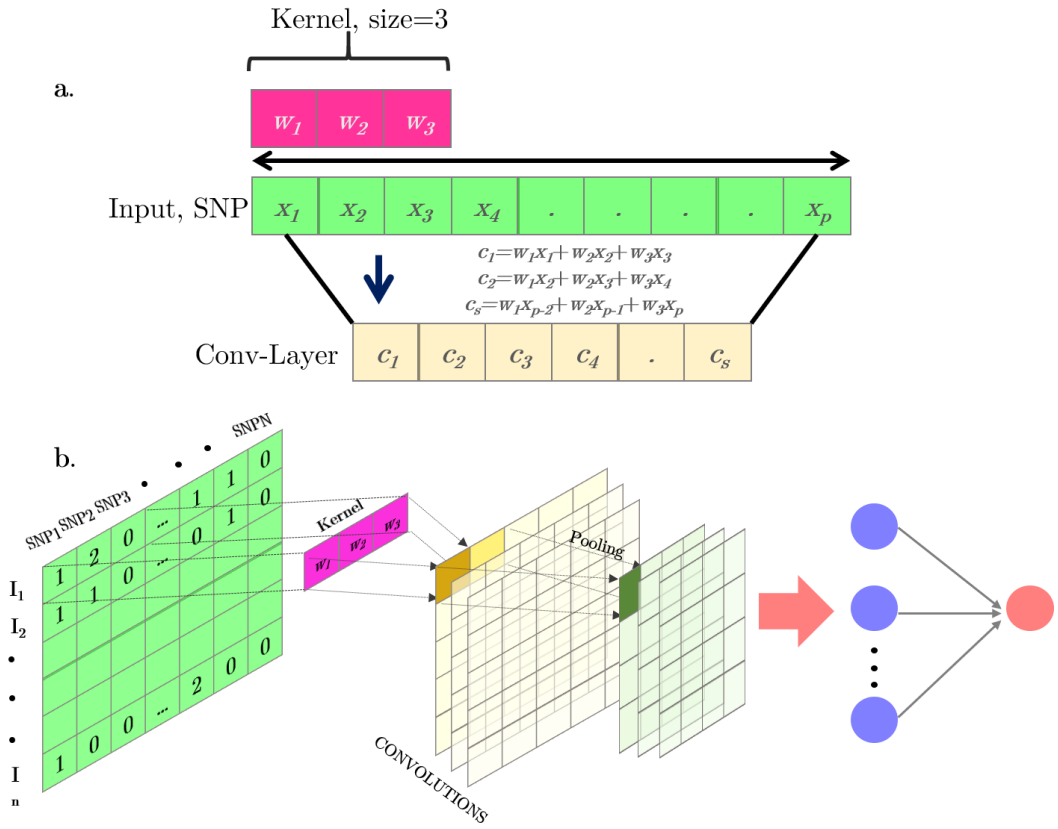


Figure 1.3. (a) Simple scheme of a 1D convolution operation. **(b)** Full representation of a 1D Convolutional Neural Network for an SNP-matrix. The convolution outputs are represented in yellow. Pooling layers after convolution operations combining the output of the previous layer at certain locations into a single neuron and are represented in green. The final output is a standard MLP. (Figure from Pérez-Enciso and Zingaretti [56])

1.6.2 Algorithms and Optimization Issues

Irrespective of the architecture chosen, all DL algorithms are based on a few principles that are used to minimize the cost function and, hopefully, maximize predictive ability. Here we describe the main concepts.

Backpropagation and Stochastic Gradient Descent are behind the modern revival of neuron-associated methods. Backpropagation [57] is a clever method that propagates the error backward at the output layer level. Then, the gradient of previous layers can be computed easily using the chain rule for derivatives, which greatly simplifies optimization in complex models. The basis of gradient descent [58] is also simple. The algorithm requires a set of initial solutions and a loss function, which usually has good mathematical behavior, i.e., it is convex or at least quasi-convex (metaphorically, this means reaching the lowest elevation point simply going downhill).

Stochastic Gradient Descent (SGD, Algorithm 1) is one of the most widely used optimizers in DL. SGD, and a plethora of related methods, randomly partition the whole dataset into subsets called 'batches' or 'minibatches' and updates the gradient using only that subset. The next batch is used in the next iteration. This intelligent strategy allows us not only to manipulate datasets of any arbitrary large size but also introduces stochastic noise that reduces the risk of converging at the local maxima. An 'epoch' is the set of iterations that comprises all samples in the dataset. For the next epoch, a different data partition is used in each batch. In addition to batch size, SGD requires initial values for all parameters and specifying the 'learning rate', i.e., the value that controls the update of the gradient (Algorithm 1).

Algorithm 1: Stochastic Gradient Descent (SGD)

Given: L the loss function,
Input: Training sample $(x_{(train)}, y_{(train)})$, regularization parameters $\Omega(\theta)$, learning rate α, λ is a parameter to controlate the penalization importance, initialize θ
Output: Model parameters $\hat{\theta} = (\mathbf{b}, \mathbf{W})$
Initialize $\theta = \{W^{(1)}, b^{(1)}, ..., W^{(L+1)}, b^{(L+1)}\}$;
repeat
  for $i \in n_{epochs}$ **do**
    Given a training set $(x_{(train)}, y_{(train)})$
    Compute: $\Delta = -\frac{\partial}{\partial\theta} l(f((x_{(train)}; \theta), y_{(train)})) - \lambda \frac{\partial}{\partial\theta} l(\Omega(\theta))$
    $\theta_i \leftarrow \theta_i + \alpha\Delta$
  end
until *stopping criterion/ convergence*;
where: $\frac{\partial}{\partial\theta} l(f((x_{(train)}; \theta), y_{(train)}))$ is the function gradient

Note that Algorithm 1 incorporates a regularization term, **Ω**. Regularization consists of adding a 'penalty' or 'constraints' to the model parameters, incorporating a restriction over the weight (w) estimations in the loss function. The two most frequent regularizations are the L1 and L2 norms,

which set restrictions on the sum of absolute values of w (L1) or of the square values (L2). There are three main variants of the GD algorithm: Batch gradient descent, stochastic gradient descent (SGD), and minibatch gradient descent. These are the most popular optimizers used nowadays. A description of all the keras implementation to DL optimization can be found in https://github.com/miguelperezenciso/DLpipeline#Optimizers.

Initial weight values are an additional factor that needs to be considered seriously and has deserved numerous contributions. In the words of Goodfellow and collaborators[47]: 'Our understanding of how the initial points affect generalization is especially primitive, offering little to no guidance for how to select the initial point' (p. 293). This is because DL algorithms are iterative and the function to minimize cost is too complex. Therefore, initial values may affect whether convergence is attained or not. In our experience, it is critical to compare prediction performance with a few different training runs with the same hyperparameter values, using either random uniform or normal values. This should indicate how reliable the initialization strategy is.

The **activation function** is the mathematical function that transforms the linear input of the neuron into its output (Figure 2.1). The most popular activation function in the past was the logistic (sigmoid) function. This function often results in numbers that are either 0's or 1's and is not flexible enough for most applications. Therefore, other functions are currently more popular, including 'relu' or 'selu'. Plots and descriptions of the most popular functions can be found (e.g., in https://en.wikipedia.org/wiki/Activation_function) and are not further described here. We recommend that the activation function should be considered as a hyperparameter to be optimized among two to four possible values, e.g., 'tanh', 'relu', 'selu', etc. In general, hidden units may need different activation functions from those of input or output layers. For instance, for classification problems, a 'softmax' activation function in the output layer is frequently used. The Table in the accompanying GitHub (https://github.com/miguelperezenciso/DLpipeline#loss) shows the most common combinations of Loss Function and Last Layer Activation for different problems.

In a genomic prediction context for a quantitative trait, the simplest activation function is 'linear', and it only can model additive effects since it sums up through the allelic frequency of the SNP (0,1,2 in a diploid organism). SNPs are not numerical, but categorical data, though DL techniques only accept numerical input. A set of encoding methods were developed to overcome these constraints [59]. One hot encoding, which is simply recoding the three SNP genotypes as three 0/1 dummy variables, is the most popular approach for genomic prediction purposes. Using this approach, non-linear relationships can be modeled by using non-linear activation functions at the first layer [60].

1.6.3 Avoiding Overfit

A model overfits when it cannot separate noise from the signal. An overfitted model results from a poor fit on the validation set, i.e., prediction of unobserved data is very poor. This is one of the most, if not the most critical problem in DL applications. Most of the time in DL optimization will be spent avoiding an overfit of the data, to improve the predictive abilities of the algorithm. Given the unknown behavior of various algorithms, no general guidance can be given, and trial-and-error is needed. Success here is very much dependent on the specific problem. There are

broadly three non-mutually exclusive techniques to minimize the risk of overfitting: early stopping, regularization, and dropout.

Early stopping is the simplest and sometimes also the most effective strategy since an excessive number of epochs tend to result in overfitting. Conversely, the number of epochs should be sufficient as too few iterations result in underfitting. As with any hyperparameter optimization, the optimum number of epochs should be chosen using only the training set. This training subset is normally partitioned in a proper training dataset and a test subset which is used for internal cross-validation.

Regularization, as mentioned, is the procedure whereby constraints are imposed upon the weights' estimates, which are incorporated in the general loss function (see SGD Algorithm 1). Regularization is a key parameter to avoid overfitting and should be very carefully optimized in your own data, perhaps using a grid or random search. In our experience with a large human dataset, the optimum regularization was surprisingly small [15]. This is likely a consequence of the algorithm failing in finding the 'regularities' in the data, which is needed for an optimum prediction of new data.

Dropout is another clever strategy that is specific to neural networks. Given an initial, completely connected network, dropout consists of setting to zero the output of a random subset of neurons. This strategy is equivalent to 'bagging' (i.e., sampling) sub-networks to produce an ensemble (i.e., joint) estimator. The usual recommended dropout rate, i.e., percentage of inactivated neurons, is 20 - 50% [55]. However, in our experience with the UK Biobank human dataset, the optimized dropout rate was very small (<0.05). Again, the optimum rate is problem-specific and should be optimized with the data at hand. In practice, it is probably unnecessary to combine regularization, either L1 or L2, and dropout, whereas early stopping is always a good practice.

1.7 Generative Models

Some ML models belong to either the **discriminative** or **generative** categories. But what does this mean? Suppose you have an image database containing two classes of fruits: strawberries and blueberries. Imagine that you need to build a model to differentiate them and then you decide to hire two data analysts (A and B) to perform the task. After a few days, the analysts have delivered two algorithms. So, it's time to see how they work by providing a new observation for classifying (e.g., a strawberry). Model A) just outputs "it's a strawberry" relying on the properties it has already learned. Model B) makes new images of the dataset and decides that it is a strawberry based on the degree of agreement between this observation and the images it has produced. Both models were successful, but they did not do the same. Model A) is an example of discriminative modeling, which models the decision boundary between the classes, whereas model B) is a generative modeling case, which models the actual distribution of each class [61] (Figure 1.4).

Although both models have the same goal, i.e., to estimate $p(fruit|features)$, the probabilities they learn are different. A discriminative model learns the conditional probability $p(fruit|features)$, estimating the parameters from the training data. A generative model instead

uses the training data to derive parameters of $p(fruit)$, and $p(features|fruits)$. The final output $p(fruit|features)$ can only be derived indirectly from the Bayes' rule.

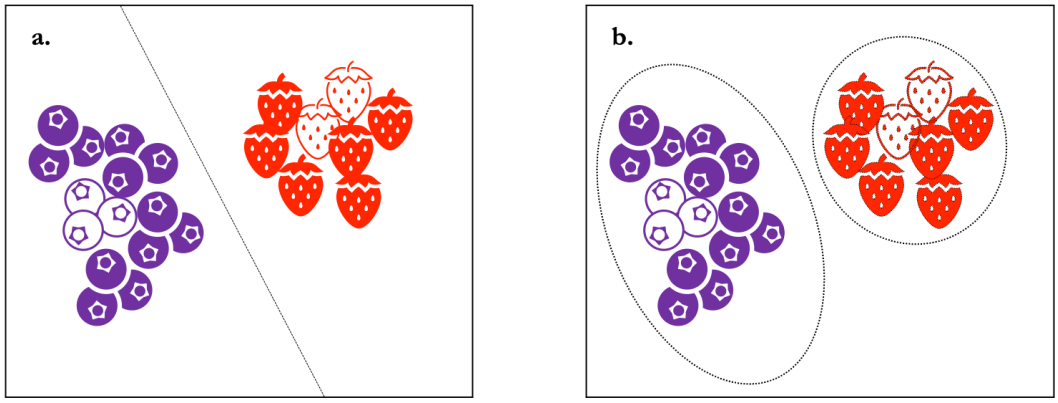


Figure 1.4. (a) Representation of a discriminative model, which learns $p(y|x)$ and (b) a generative model, which learns $p(y,x)$.

Discriminative models are computationally cheap compared to generative models and have achieved astounding successes [47,62,63]. The ML literature is plenty of examples of successful application of the discriminative model to learn classes from high-dimensional datasets, e.g., Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Decision Trees, Artificial Neural Networks (ANN). On the other hand, Naïve Bayes, Mixture of Gaussians, Bayesian Networks, Boltzmann Machines, etc. are examples of classical generative models[61].

1.7.1 Generative adversarial Networks

Until the emergence of the Goodfellow et al. [62], Rezende et al.[64] and Kingma and Welling [65] papers, most of the Deep Generative Models were based on intractable likelihood functions that therefore require numerous approximations to their gradient; the most famous of all is Boltzmann Machines (BM) proposed in 1983 by Hinton and Sejnowski [66]. Deep Generative Networks are particularly striking because they can deal with unsupervised problems, so the developments in the field and the attempts to make them simpler, are not surprising. In 2014 emerged the Generative Stochastic Networks (GSN): inspired in the idea behind the Denoising Autoencoders, these networks can be trained with simple backpropagation by transforming an unsupervised task (to estimate $p(x)$) into a supervised learning framework, consisting in parametrizing the transition operator of a Markov Chain [63].

Generative Adversarial Networks (GANs) (Figure 1.5) [62] extend GSN but do not need to use Markov chains. They are based on a simple but powerful idea: train two networks simultaneously, the Generator (G), which defines a probability distribution based on the information from the samples, and the Discriminator (D), which distinguishes data produced by G from the real ones. These networks contain two key ideas. The first one assumes that the data we are trying to generate can be well described using a probability function, i.e., if you aim to generate strawberries, you are supposing that there is an underlying probability distribution of strawberries. Once trained, a generative network uses the inverse transform method to take points from a simple distribution (e.g., uniform) and outputs points from the target distribution (e.g., strawberries distribution).

The trick relies on the **adversarial** notion (the second key point). When training a generative network, the target (e.g., strawberries samples) and the generated distribution (from random noise) are used to train a discriminative network, whose purpose is the same as any discriminative model: to classify the samples as accurate as possible. However, the goal of the generator is to learn to mimic the target distribution to fool the discriminator and it does this better and better with successive training steps. The scheme can be formalized as a minimax game theory, where the process ends when the discriminator is unable to distinguish true from fake observations.

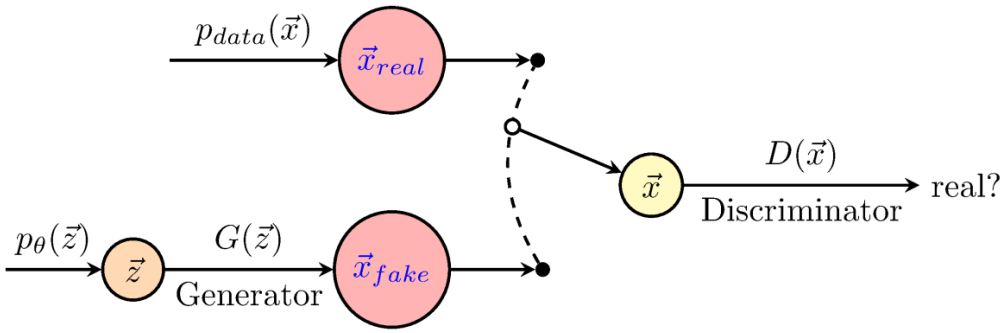


Figure 1.5. Scheme of Generative Adversarial Networks (GANs): The Generator (G), defines a probability distribution based on the information from the samples, whereas the Discriminator (D) distinguishes data produced by G from the real data. The figure was redrawn using code from http://www.texample.net/tikz/examples/neural-network.

1.7.2 Variational Autoencoders

Although perhaps the most popular, **GANs** are not the only models capable to generate new observations simply. **Variational Autoencoder (VAE)** [64,65] is a model that can be trained with gradient-based methods. First, we will briefly explain what a vanilla autoencoder is because we need to get an idea about that to understand VAE. A vanilla autoencoder (AE) is a non-linear generalization of Principal Component Analysis (PCA)[54]. PCA can be thought of as an optimization problem aiming to derive the best linear approximation in terms of L_2 norm (Eq. 1.5) in the subspace where q<p:

$$\underset{A \in R^{p\times q}, A'A = I_q}{minimize} \sum_{i=1}^{n} \|x_i - \boldsymbol{A}\boldsymbol{A}'x_i\|_2^2 \quad [Eq.\,1.5]$$

By analogy, a single layer autoencoder solves the following optimization problem:

$$\underset{W \in R^{p\times q}}{minimize} \sum_{i=1}^{n} \|x_i - \boldsymbol{W}'g(\boldsymbol{W}x_i)\|_2^2 \quad [Eq.\,1.6]$$

For some non-linear g. Eq. 1.6 shows the problem behind an autoencoder with one hidden layer, but it is not difficult to extend it to deeper hidden layers. While AE learns a compressed representation of the original data, VAE learns a probability distribution representing the data ($\boldsymbol{x}$),

which makes them suitable to generate new unseen observations. In the encoder step, VAE realizes what are the vector containing μ and σ, the parameters of a normal distribution, i.e., it learns $p(z|x)$. The big idea is that the decoder learns that in the latent space, there is a set of points (a continuum) that refers to a given data point, allowing smooth interpolation and the construction of new samples. The decoder step (or input reconstruction), $d(z)$ is obtained after sampling $z{\sim}p(z|x)$ from the latent representation. But how does VAE do it?

The answer lies in the loss function. VAE loss function includes a "reconstruction" and a "regularization" term. VAE is trained by maximizing the variational lower bound ($\mathcal{L}$) associated with the data point (x_i), which is expressed in Eq. 1.7.

$$\mathcal{L}(q) = \mathbb{E}_{z\sim q_\theta(z|x_i)}\big[log p_\phi(x_i|z)\big] - D_{kL}\big(q_\theta(z|x_i)||p(z)\big)\ [Eq.\ 1.7]$$

As $\mathcal{L}(q) \leq \log\left(p(x_i)\right)$ (Eq. 1.7), maximizing $\mathcal{L}(q)$, maximizes the logarithm of our data by proxy, transforming a typically intractable likelihood in a simple problem. The first term of the Eq. 7 is the same as in a classical AE, representing the log-likelihood of the reconstructed data output at the decoder. The second term of Eq 7., i.e., "the regularized" is the entropy corresponding to the Kulback-Leibler (K-L) divergence [47] between the latent distribution $N(\mu_x, \sigma_x)$ and the standard normal distribution $N(0,1)$. This term forces the latent distribution to be close to the standard normal, generating a continuous low variance space centered at the origin, suitable for clustering and data generation.

In other words, the core idea here is the simplification of the problem by using the evidence lower bound (ELBO), $\mathcal{L}(q)$ which is nothing but a combination of the cross-entropy (the reconstruction term) and the K-L divergence (the regularization term), a measure of the dissimilarity between the approximate and the real posterior distribution, representing the information difference between both distributions.

1.8 Image Processing and Computer vision in agriculture

First, let us define and differentiate the two key concepts of this section: Image Processing and Computer Vision. Both go hand in hand, they share many techniques and the line of differentiation between them is blurred, yet they are not the same.

Image processing comprises all the models and algorithms for manipulating images, including digitization, histogram manipulation, sharpening, blurring, smoothing, stretching, edge detection, morphological operation, segmentation, etc. Computer vision, in turn, involves all the ML algorithms that enable computers to process and understand images or videos as humans would. Image processing techniques may be used without computer vision, but not the opposite since image processing is a previous step to apply any Computer Vision algorithm.

There are many automatic computer vision machines capable of performing numerous tasks. It is a hot topic of research in medical image analysis, satellite image analysis, self-driving vehicles, financial industry, inventory management, etc. [67]. It is also revolutionizing agriculture in multiple avenues, such as disease detection in plants, crop monitoring using an autonomous flying object (e.g., drones), the automatic phenotyping of many plants and animals, the application of precision agriculture, the harvesting of fruit, etc. For example, semantic segmentation algorithms allow the identification of animals in livestock production, discriminate and classify good crops from bad ones (determining product quality and shipping options), automatically picks rock, and monitor soil moisture [68–72].

Overall, image processing and computer vision in agriculture enable accurate monitoring of plants and animals on a large scale. The goal of plant/animal imaging is to measure physiological, growth, development, shape among other phenotypes. These measurements can be made for different purposes (not just breeding), in different environments, and using a wide variety of devices (digital camera, tomography, magnetic resonance, fluorescence, etc.)[67,73–75]. There are still several challenges that are preventing computer vision in agriculture from reaching its maximum potential. Most of the Computer Vision problems can only be tackled using supervised learning, which requires a large number of annotated images. Also, people often assign ML too difficult tasks, but machines are not as wise as we think they are only smart at solving very simple tasks [46].

Here, we will mainly focus on morphological traits phenotyping, which are known to exhibit high heritabilities and therefore respond quickly to selection. It is not surprising that the origin of many breeds is associated with mutations affecting general appearances such as coat color, where humans could rapidly reproduce animals or plants with a novel and attractive phenotype. Even today, animal breeders' associations can spend much time defining the 'racial standard'. Ornamental plants are appreciated by tolerance to biotic and abiotic stress, development potentialities, and aesthetical factors [76], consumer preferences of fruit and vegetables are determined by their apparency [77]. Morphological traits are not only aesthetic features but convey essential information on animal welfare and fruit quality.

1.8.1 Image segmentation

Before defining image segmentation, we must briefly explain what a digital image is. A grayscale digital image is a two-dimensional function $f(x, y)$ where x and y are the spatial coordinates, and the amplitude of $f(x, y)$ at any pair of coordinates is the pixel intensity. It is simply an array of dimensions $n \times m$, indicating the number of rows and columns, respectively. The pixels intensities vary in the interval $[0,255]$. A pixel with an intensity of 0 is black, whreas a pixel with an intensity of 255 is white.

On the other hand, an RGB color digital image is composed of three channels, representing the degree of the red, green, and blue color of each pixel, i.e., for each spatial coordinate in an RGB image $f(x, y) = (r, g, b)$, i.e., the output is a vector of size 3, that indicates the intensity of red, green, and blue, respectively. Note that an RGB image is an array of $n \times m \times 3$ dimensions.

Although there are more ways to represent color images, they are all based on the same principles and there is no need to go into further details.

Recognizing and differentiating objects in an image is quite easy for humans, but not trivial for machines. **Image segmentation** is the image processing task that breaking down the image into multiple segments, which are easier for machines to understand. It is a preliminary step to obtain different measurements of the objects in an image. and can be broadly divided into two categories: semantic segmentation and instance segmentation [78].

Semantic segmentation classifies the pixel belonging to a particular label without differentiating objects from the same category, e.g., if there are 5 strawberries in the image, it gives the same label to all of them. Instance segmentation gives a unique label to every instance of a particular object into the image. In the case of the 5 strawberries, an instance segmentation algorithm can individualize each of them. Note that, since instance segmentation is challenging compared to semantic segmentation, it needs to appeal to supervised deep convolutional neural networks [78]. The algorithms described below belong to the Semantic segmentation approach, which has a long tradition and is not only based on complex neural networks but can also be carried out using simple image processing techniques.

The simplest way to identify different objects in an image is **threshold segmentation**. It is founded on the idea that the background and foreground pixels have different properties (Figure 1.6), so they can be easily split using a simple threshold value, which classifies the pixels as belonging to one out of two categories. The algorithm may also be used for classifying more than two categories by defining multiple local thresholding. One of the most popular thresholding algorithms is Otsu's, a method that automatically chooses the threshold by maximizing the variance between the two groups (foreground and background) [79].

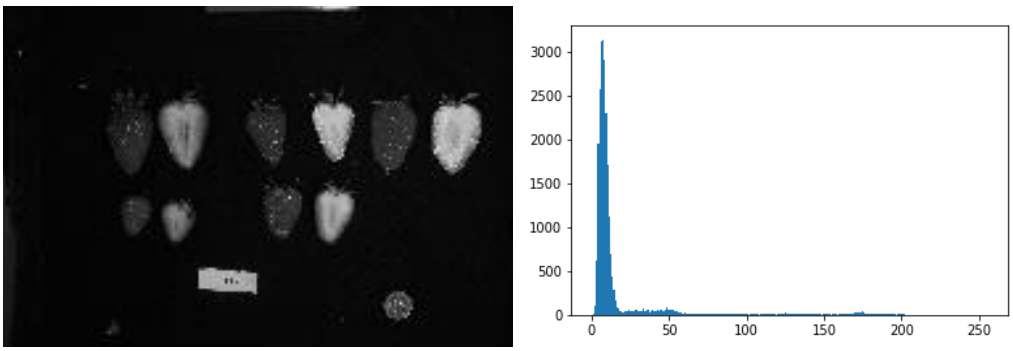


Figure 1.6. The right-side shows the histogram of pixel intensities of the left image. Note that the threshold segmentation of this image is straightforward, as background and foreground pixels are well differentiated.

The **edge-based** segmentation procedures rely on the rapid change of intensity value in an image. These techniques are founded on the differences between neighboring pixels (i.e., the derivatives). Sobel, Gaussian, and Canny are among the most popular. Sobel edge detector computes the gradient by using the discrete differences between rows and columns of a 3×3 neighborhood, weighting by 2 the central pixel of each window.

The Gaussian (also named the Laplacian of Gaussian) uses a gaussian function ($G(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}}$), being σ the standard deviation determined by the degree of blurring. The method convolves the image using the Laplacian of $G(x,y)$, i.e., $\nabla^2 G(x,y) = \left[\frac{x^2+y^2-2\sigma^2}{\sigma^4}\right] e^{-\frac{x^2+y^2}{2\sigma^2}}$. Applying this filtering has two effects: smooth the image and yields a double edge image.

The Canny Detector is the most powerful of these filters and involves a series of steps. The first one consists of finding the edges by looking for the local maxima of the gradient of the function, which is computed using after smoothing out the image using gaussian filtering. Then, after computing the local gradient, an edge point is defined as a point whose strength is locally maximum in the direction of the gradient. Finally, the algorithm can discriminate between strong, weak, and non-relevant edges, by applying a double threshold. Threshold and edge-based segmentation are one channel-based segmentation, i.e., you must convert the color image to grayscale before applying the algorithms or you can use one of the color channels instead.

The **clustering-based** treated the thresholding problem as a general classification problem, either hierarchical or partitional. These techniques find clusters of pixels where each pixel is represented as a point in space, whose axes can be color components, local texture or gradient, etc. Hierarchical clustering is "heuristic" and "agnostic", i.e., does not need to make assumptions about the data distribution and it classifies the pixels based on the distance among them. Partitional clustering, e.g., *k-means* assumes that a mean value (or median for the *k-medoids* approach) is a proper descriptor for a given class. It is an iterative process, which starts by selecting a set of k seeds and assigning each observation to the nearest seed. The initial seeds are then replaced by the cluster means (median) and the points are reassigned according to the distance to the new centers. The process continues until no further changes occur in the clusters. Mean-shift clustering [80] is essentially the same as k-means but does not require specifying the cluster number initially.

Simple Linear Iterative Clustering (SLIC) is an iterative process [81] that is built on 'superpixels', a set of pixels sharing perceptual and semantic, and proximity information. It generally works with the CIELAB color space, starting with k centers regularly spaced. The algorithm moves the k clusters to seed locations corresponding to the lowest position of the gradient in a 3×3 neighborhood, assigning each pixel to the nearest cluster center whose search area overlaps this pixel. The centers are then recalculated by averaging all pixels belonging to that superpixel.

The power of **deep learning** can be also used to enhance image segmentation and the literature is plenty of papers where deep learning has been applied with astonishing results [78,82–85]. DL powered image segmentation taking advantage of the Convolutional Operations (see section 1.4.2.1), which can produce accurate feature maps of the image, extracting all the meaningful information, i.e., it can identify the feature maps that make an object unique, in a similar way that we humans do. However, the main shortcoming of deep learning-based segmentation is that most of them are supervised and require a large number of annotated images, an expensive and time-consuming process. Table 1.2 summarizes the pros and cons of each of these approaches.

TABLE 1.2 Pros and Cons of the four image segmentation approaches.

Approach	Pros	Cons
Threshold-based	-Simplicity -Speed -It works almost perfectly if the background is homogeneous, and objects and background have a high contrast	-It is not suitable if the objects and the background do not have a high contrast (i.e., if the histogram doesn't have noticeable peaks).
Edge Detection- based	-Based on the recognition of discontinuity -Find the correct places of edges -Works pretty well if the objects have high contrast and the difference between regions is thin.	-Sometimes are complex, time-consuming. -These methods don't work well if images have many edges or are not well defined. -Sensible to noise.
Clustering-based	-Are based on the shape and spatial information. - Can use several channels of a given image.	-Time-consuming - The choice of the number of clusters.
Deep Learning	-These methods are accurate and work very well, any with complex images.	-Images have to be annotated -Require a large training dataset

1.9 Statistical shape analysis: a general overview

The analysis of shapes has a long history in Evolution, which has fostered most of the analysis tools available today [86–88]. Traditional morphometrics is based on the analysis of summary statistics such as length, width, ratios, and areas [73]. However, the shape is highly dimensional, and we restrict the list of potential candidate genes by focusing on single univariate statistics. Also, these summary statistics do not allow reconstructing the original shape.

One of the problems addressed by the statistical analysis of shape is the extraction of meaningful information from the (segmented) objects into an image [89]. A shape includes all the geometrical information that remains invariant in a given object after removal of location, scale, and rotation effects [90]. In other words, a shape is invariant to Euclidean transformations. Hence, it can be well described by locating a finite number of points on the outline, a concept better known as 'landmarks' [87]. A landmark is an anatomical position that can be identified in all samples, e.g., the tip of the nose in cattle; while the "pseudo-marks" are points sampled along the contour of all samples with only geometrical meaning. In landmark-based geometric morphometrics, the spatial information is contained in the data, which are precisely landmark coordinates. Shapes can be compared once a common reference scale is found; this can be done via Generalized 'Procrustes' Analysis (**GPA**) [91], which consists of finding an optimal superimposition of several shapes such that distances between them are minimized.

GPA is a generalization of the Procrustes transformation for two configurations (i.e., two shapes). Suppose that $X_1$and X_2 are the two centered configurations, Procrustes involves the least-squares matching of them, which can be expressed as in Eq. 1.8:

$$D_P^2(X_1, X_2) = \|X_2 - \beta X_1 \Gamma - 1_k \gamma^T\|^2 \text{[Eq. 1.8]}$$

where $\beta \; and \; \Gamma$ are the scale parameter and the rotation matrix, respectively, and γ is the location vector. The problem has an analytic solution [90], where $\hat{\gamma} = 0$, $\hat{\Gamma} = UV^T$, being U and V the eigenvectors matrix of $X_2^T X_1 X_1^T X_2$, $X_1^T X_2 X_2^T X_1$ and $\hat{\beta} = \frac{trace(X_2^T X_1 \hat{\Gamma})}{trace(X_1^T X_1)}$.

GPA is the generalization of Procrustes for $X_1, X_2, \ldots, X_n$ configurations, which involves translating, rescaling, and rotating the configurations relative to each other, considering the pairwise differences:

$$G(X_1, X_2, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1+i}^{n} \left\| (\beta_i X_i \, \Gamma_i + 1_k \gamma_i^T) - (\beta_j X_j \, \Gamma_j + 1_k \gamma_j^T) \right\|^2 \text{ [Eq. 1.9]},$$

Subject to the constraint of the constraint on the size of the average, i.e., $S(\bar{X}) = 1$ (the centroid of the average configuration, which is simply the mean of all the configurations in Eq. 1.9). Although the GPA has not an analytical solution, the algorithm to approximate it is quite straightforward and can be summarized in the following 4 steps:

1. Translations, consisting in center the configurations to remove locations. Initially, $X_i^P = X_i, i = 1, \ldots, n.$
2. Rotations. For the ith configuration let $\bar{X}_{(i)} = \frac{1}{n-1} \sum_{j \neq i} X_j^P$. The new X_i^P is the Procrustes superimposition (Eq. 7) between $\bar{X}_{(i)}$ and the old X_i^P. The n configurations are rotated until the Procrustes sum of squares of Eq. 9 cannot be reduced further.
3. Scaling. Let Φ the $n \times n$ correlation matrix of the $vec(X_i^P)$ with eigenvector $\Phi = (\Phi_1, \Phi_2, \ldots, \Phi_n)^T$ corresponding to the largest eigenvalue, for all i take $\hat{\beta}_i = \left(\frac{\sum_{k=1}^{n} \|X_k^P\|^2}{\|X_i^P\|^2} \right)^{1/2} \Phi_i$.
4. 2 and 3 are repeated until convergence criteria (Eq.1.9 cannot be reduced further).

Figure 1.7 shows an example of Procrustes superimposition on strawberries landmarks.

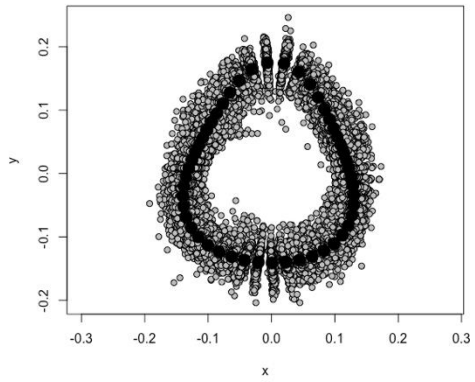


Figure 1.7: Generalized Procrustes Superimposition on strawberries landmarks.

The overall consensus configuration (Black points in Fig. 1.7) is perfectly symmetrical allowing to measure the symmetry/asymmetry of any individual configuration. Besides, principal component analysis (PCA) on the overlapping points permits obtaining latent features that are good descriptors of regions of highest shape variation. The main feature of the global shape analysis is that it allows for simple statistical modeling, providing more information than some linear descriptors that compare, for example, only the ratio between height and width.

1.10 Methods for exploration and integration of heterogeneous biological data: a multi-way view

Recent advances in next-generation technologies have revolutionized biological research providing enormous amounts of heterogeneous data. High-throughput sequencing technologies enable quantitative analysis of multiple-omics data, facilitating the study of gene expression, proteins, metagenomics, and metabolites, and the assessment of their relationship with complex traits. The omics era has directly changed the paradigm of studying the organism from a single data type to an integrated systems biology approach [92]. Although several statistical methods have been developed [93–95], the integration of these heterogeneous datasets remains a challenge.

There are different ways to integrate biological data. One of the most popular and widely used is Gene Set Enrichment Analysis, which combines gene expression with biological information [96–98]. An alternative method is the use of co-expression networks, which estimate similarities between different structures, associating gene expression with functional annotation or identifying transcription factors [99–101]. A third approach is relying on multivariate statistical methodologies. Partial Least Squares (PLS) have been successfully applied to integrate the information from two datasets [102], while a multiway extension of Factor Analysis, termed Multiple Factor Analysis (MFA) and the Multiple Co-inertia (MCOA) have been used to integrate information from several omics data [95,103,104]. The main drawback of these approaches is that most of them are only suitable for continuous data. We strongly believe that more efforts are needed to deliver multivariate approaches capable of integrating different data sources and supporting different types of data as input.

STATIS-ACT [105,106] belongs to the family of methods to address the problem of analyzing multiple datasets. It is a multiway extension of PCA that can deal with multiple datasets of variables collected on the same individuals or multiple individuals collected on the same variables [106]. Suppose you wish to analyze $X_1, X_2, \dots, X_k$ microbial communities' or microarray matrices, all of them measured on the same samples, i.e., their sizes are $n \times p_1, n \times p_2, \dots, n \times p_k$. The k arrays are transformed into cross-product (i.e., for all $j = 1..k$, $W_j = X_{[j]} M X_{[j]}^T$, being M a diagonal matrix with weights for each variable in $X_{[j]}$). The method consists of applying the following three steps:

Inter-structure: It measures the similarity between all the W_j computing the normalized version of the Hilbert-Smith product ($\langle W_{[j]}, W_{[l]} \rangle = tr\left(D W_{[j]} D W_{[l]}\right)$ if $j \neq l$ and $\langle W_{[j]}, W_{[j]} \rangle = \left\| W_j \right\|^2$). Note that $tr(.)$ indicates the trace operator, $\|.\|^2$ is the L_2 norm and D is a matrix of weights for the observations in each configuration. This step computes the correlation vectorial coefficient (RV) for each pair of tables, which is nothing but $\rho_{jl} = \frac{\langle W_{[j]}, W_{[l]} \rangle}{\langle W_{[j]}, W_{[j]} \rangle \langle W_{[l]}, W_{[l]} \rangle}$. $\mathrm{P} = \{\rho_{ij}\}$ is the similarities matrix between all the configurations and is therefore generally referred to as a generalization of the correlation coefficient for arrays.

Compromise: This step consists of finding the optimal weights (α_i) for building the consensus configuration ($\boldsymbol{W} = \sum_{i=1}^{K} \alpha_i W_i$). This is nothing but an optimization problem, which can be viewed either as a minimization problem or as its dual. The former refers to minimize the distance $D = \sum_{i=1}^{K} \left\| W_{[i]} - \alpha_i W \right\|$, being $\alpha' \alpha = 1$. With a little of algebra, the equation can be developed and simplified as follows: $D = \sum_{i=1}^{k} \|W_k\|^2 - \alpha^T \mathrm{P} \alpha$, which yields the results that α is the first eigenvector of P matrix, given that D will be minimal when $\alpha^T \mathrm{P} \alpha$ is maximum. The problem can be alternatively defined as a maximization of the vectorial correlation between the $\boldsymbol{W}$ and each $W_{[i]}$, i.e., $C = \sum_{i=1}^{k} \langle W_{[i]}, \boldsymbol{W} \rangle^2$. Both approaches lead to the same solution, since it can be shown that $C = \alpha^T \mathrm{P}^2 \alpha$, which, as already mentioned is maximum when α is the first eigenvector of P.

Intra-structure: This step consists of applying a PCA analysis to the common configuration $\boldsymbol{W}$.

STATIS has some shortcomings, as it can only be used for continuous data. An additional drawback is that the subspace generated by STATIS only allows ascertaining relationships between common elements in all datasets, i.e., observations (samples) or variables (genes/transcripts/proteins/microbial communities). It is also not possible to establish relationships between observations and variables, nor does it provide a variable selection strategy.

In the last chapter of this thesis, we show how we have enhanced the original methodology by generalizing a multidimensional scaling approach that allows for many types of data, from continuous to compositional. Exploiting the old idea of the regression biplot, we have developed a way to project variables into the common configuration [107]. Finally, based on the Set Enriched Analysis we have developed hypergeometrical taxa set enrichment approach that can integrate and analyze multiple microbial data to a different level of resolution.

1.11 Genomic Prediction in animal and plant breeding schemes: Why is simulation worthwhile?

Modern genetic evaluation methods are sometimes too complex to be evaluated analytically. Further, although the forces that govern changes in patterns of variability in genomes are known (e.g., mutation, drift, selection, etc.), their joint action is also difficult to predict. For these reasons, geneticists and breeders, in particular, have heavily employed computer simulation tools throughout the years. Today, computer simulation is critical for optimizing breeding schemes and evaluating the robustness of genomic prediction strategies in the face of uncertainty regarding the genetic architecture of complex traits.

For the non-expert, the fact that computers can reproduce stochasticity may seem a contradiction in terms. The mystery is solved, of course, once we learn that there exist algorithms capable of generating lists of numbers that are uncorrelated between them. Numeric transformations can then be applied so that random values sampled from any distribution can be mimicked.

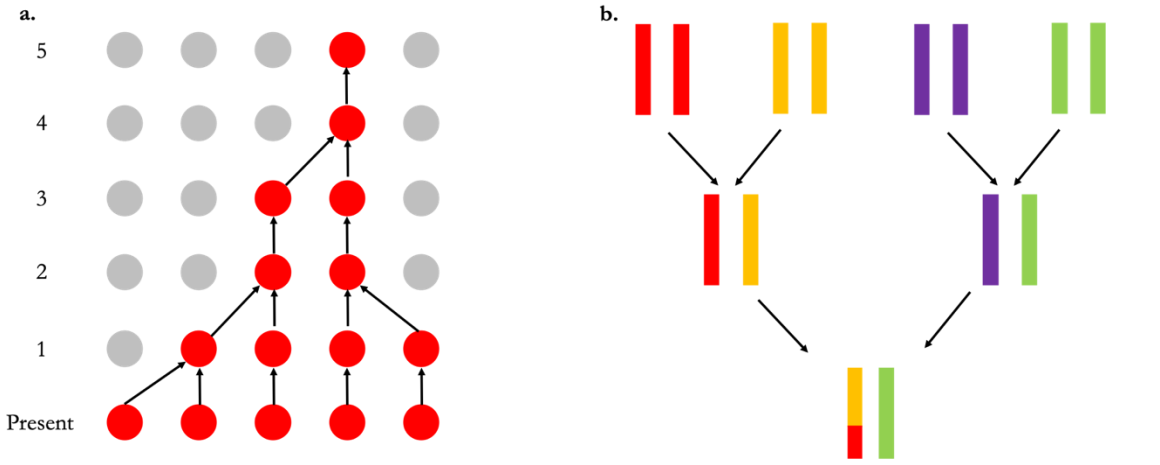


Figure 1.8. (a) Illustration of a coalescent process. It shows a simplified scheme of DNA evolution. Each row is a single generation, and the circles denote a DNA sequence. (b) An illustration of a forward-in-time process. This simplified scheme illustrates an F2 population.

Mutation, drift, meiosis, mating, or even selection are stochastic processes that can be reproduced with specialized software. Two broad approaches exist: forward and backward (i.e., coalescence) simulations. The forward simulation consists of generating offspring from parent genomes (Fig. 1.8b). Phenotypes are then simulated conditional on individuals' genotypes and predetermined genetic architecture. Forward simulation is the standard approach in breeding and is the one utilized in pSBVB (Chapter 3).

Coalescence simulations follow the opposite path (Fig 1.8a). Given current genotypes, ancestors are backward simulated up to the 'most recent common ancestor', i.e., the hypothetical founder of all observed genomes in our sample. Coalescence simulations have been especially popular to model DNA variability patterns in population genetics, but not so much in breeding. The reason is that modeling artificial selection in the coalescence framework is not easily tractable. Forward simulation, in turn, is much more computationally demanding than the coalescence. Some authors

have then proposed mixed approaches whereby the coalescence is used for generating the base population and forward simulation for the selection process [14]. Figure 1.8 illustrates both approaches.

1.12 References

[1] M.A. Zeder, Central questions in the domestication of plants and animals, Evol. Anthropol. 15 (2006) 105–117. https://doi.org/10.1002/evan.20101.

[2] F.B. Marshall, K. Dobney, T. Denham, J.M. Capriles, Evaluating the roles of directed breeding and gene flow in animal domestication, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 6153–6158. https://doi.org/10.1073/pnas.1312984110.

[3] E. Jonas, D.-J. De Koning, Does genomic selection have a future in plant breeding?, Trends Biotechnol. 31 (2013) 497–504. https://doi.org/10.1016/j.tibtech.2013.06.

[4] J.A. Bhat, S. Ali, R.K. Salgotra, Z.A. Mir, S. Dutta, V. Jadon, A. Tyagi, M. Mushtaq, N. Jain, P.K. Singh, G.P. Singh, K. V. Prabhu, Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding, Front. Genet. 7 (2016) 221. https://doi.org/10.3389/fgene.2016.00221.

[5] R. Bernardo, Molecular markers and selection for complex traits in plants: Learning from the last 20 years, Crop Sci. 48 (2008) 1649–1664. https://doi.org/10.2135/cropsci2008.03.0131.

[6] D. Falconer, T. Mackay, Introduction to quantitative genetics, 4a ed., Essex: Benjamin Cummings, 1996.

[7] G. de los Campos, J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, M.P.L. Calus, Whole-genome regression and prediction methods applied to plant and animal breeding, Genetics. 193 (2013) 327–345. https://doi.org/10.1534/genetics.112.143313.

[8] T.H. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, Genetics. 157 (2001) 1819–1829.

[9] J.M. González-Camacho, G. de los Campos, P. Pérez, D. Gianola, J.E. Cairns, G. Mahuku, R. Babu, J. Crossa, Genome-enabled prediction of genetic values using radial basis function neural networks, Theor. Appl. Genet. 125 (2012) 759–771. https://doi.org/10.1007/s00122-012-1868-9.

[10] G. de los Campos, J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, M.P.L. Calus, Whole-genome regression and prediction methods applied to plant and animal breeding, Genetics. 193 (2013) 327–345.

[11] T. Meuwissen, B. Hayes, M. Gddard, Accelerating Improvement of Livestock with Genomic Selection, Annu. Rev. Anim. Biosci. 1 (2013) 221–237. https://doi.org/10.1146/annurev-animal-031412-103705.

[12] P.M. VanRaden, Efficient Methods to Compute Genomic Predictions, J. Dairy Sci. 91 (2008) 4414–4423. https://doi.org/10.3168/jds.2007-0980.

[13] D. Boichard, M. Brochard, New phenotypes for new breeding goals in dairy cattle, Animal. 6 (2012). https://doi.org/10.1017/S1751731112000018.

[14] M. Pérez-Enciso, J.C. Rincón, A. Legarra, Sequence- vs. chip-assisted genomic selection: accurate biological information is advised., Genet. Sel. Evol. 47 (2015) 43. https://doi.org/10.1186/s12711-015-0117-5.

[15] P. Bellot, G. de Los Campos, M. Pérez-Enciso, Can Deep Learning Improve Genomic Prediction of Complex Human Traits?, Genetics. 210 (2018) 809–819. https://doi.org/10.1534/genetics.118.301298.

[16] O.A. Montesinos-López, J. Martín-Vallejo, J. Crossa, D. Gianola, C.M. Hernández-Suárez, A. Montesinos-López, P. Juliana, R. Singh, A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding, G3 Genes, Genomes, Genet. 9 (2019) 601–618.

https://doi.org/10.1534/g3.118.200998.
[17] A.M. Bolger, H. Poorter, K. Dumschott, M.E. Bolger, D. Arend, S. Osorio, H. Gundlach, K.F.X. Mayer, M. Lange, U. Scholz, B. Usadel, Computational aspects underlying genome to phenome analysis in plants, Plant J. 97 (2019) 182–198. https://doi.org/10.1111/tpj.14179.
[18] R.R. Mir, M. Reynolds, F. Pinto, M.A. Khan, M.A. Bhat, High-throughput phenotyping for crop improvement in the genomics era, Plant Sci. 282 (2019) 60–72. https://doi.org/10.1016/j.plantsci.2019.01.007.
[19] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. Van Der Lei, E. Van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, Comment: The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data. 3 (2016) 1–9. https://doi.org/10.1038/sdata.2016.18.
[20] J.M. Hickey, T. Chiurugwi, I. Mackay, W. Powell, Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery, Nat. Genet. 49 (2017) 1297–1303. https://doi.org/10.1038/ng.3920.
[21] C.C. Schön, H. Simianer, Resemblance between two relatives - animal and plant breeding, J. Anim. Breed. Genet. 132 (2015) 1–2. https://doi.org/10.1111/jbg.12137.
[22] P. Schmidt, J. Hartung, J. Bennewitz, P. Hans-Peter, Heritability in plant breeding on a genotype-difference basis, Genetics. 212 (2019) 991–1008. https://doi.org/10.1534/genetics.119.302134.
[23] L. Araus, J.E. Cairns, Field high-throughput phenotyping : the new crop breeding frontier, 19 (2014). https://doi.org/10.1016/j.tplants.2013.09.008.
[24] B. Convergef, L. Araus, L. Araus, High-throughput Phenotyping and Genomic Selection : The Frontiers of Crop High-throughput Phenotyping and Genomic Selection : The Frontiers of Crop Breeding Converge, (2012). https://doi.org/10.1111/j.1744-7909.2012.01116.x.
[25] P.S. Soltis, D.B. Marchant, Y. Van de Peer, D.E. Soltis, Polyploidy and genome evolution in plants, Curr. Opin. Genet. Dev. 35 (2015) 119–125. https://doi.org/10.1016/j.gde.2015.11.003.
[26] F. Dufresne, M. Stift, R. Vergilino, B.K. Mable, Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools, Mol. Ecol. 23 (2014) 40–69. https://doi.org/10.1111/mec.12581.
[27] G.L. Stebbins, Types of Polyploids: Their Classification and Significance, Adv. Genet. 1 (1947) 403–429. https://doi.org/10.1016/S0065-2660(08)60490-3.
[28] M.C. Sattler, C.R. Carvalho, W.R. Clarindo, The polyploidy and its key role in plant breeding, Planta. 243 (2016) 281–296. https://doi.org/10.1007/s00425-015-2450-x.
[29] P.M. Bourke, R.E. Voorrips, R.G.F. Visser, C. Maliepaard, Tools for genetic studies in experimental populations of polyploids, Front. Plant Sci. 9 (2018). https://doi.org/10.3389/fpls.2018.00513.
[30] D. Gerard, L.F.V. Ferrão, A.A.F. Garcia, M. Stephens, Genotyping polyploids from messy sequencing data, Genetics. 210 (2018) 789–807.
[31] A. Gallais, Quantitative genetics and breeding methods in autopolyploids plants, 2003.
[32] L.F. V. Ferrão, J. Benevenuto, I. de B. Oliveira, C. Cellon, J. Olmstead, M. Kirst, M.F.R. Resende, P. Munoz, Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using

Diploid and Polyploid Models in a GWAS Context, Front. Ecol. Evol. 6 (2018). https://doi.org/10.3389/fevo.2018.00107.
[33] T. Meuwissen, B. Hayes, M. Goddard, Accelerating Improvement of Livestock with Genomic Selection, Annu. Rev. Anim. Biosci. 1 (2013) 221–237. https://doi.org/10.1146/annurev-animal-031412-103705.
[34] T.H.E. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, Genetics. 157 (2001) 1819–1829. https://doi.org/11290733.
[35] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38. https://doi.org/10.1016/j.artint.2018.07.007.
[36] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Interpretable machine learning: definitions, methods, and applications, n.d. https://arxiv.org/pdf/1901.04592.pdf (accessed May 14, 2019).
[37] Z.C. Lipton, The Mythos of Model Interpretability, ArXiv Prepr. ArXiv1606.03490. (2016). https://doi.org/10.1145/3233231.
[38] G. Shmueli, To Explain or to Predict?, Stat. Sci. 25 (2010) 289–310. https://doi.org/10.1214/10-STS330.
[39] L. Breiman, Statistical modeling: The two cultures, Stat. Sci. 16 (2001) 199–215. https://doi.org/10.1214/ss/1009213726.
[40] B. Chen, J. Pearl, Graphical tools for linear structural equation modeling, Psychom. Forthcom. (2015) 1–23. ftp://ftp.cs.ucla.edu/pub/stat_ser/r432.pdf.
[41] J. Pearl, E. Ed D D B B, Y.Y. Le E El L La A An N Nd D D Ge, E. Er, R. Rs, S. So O On N, N. Ne, E. Eu, CAUSALITY: MODELS, REASONING, AND INFERENCE RE E EV V VI I IE E EW W WE, 2000.
[42] N. Altman, M. Krzywinski, The curse(s) of dimensionality this-month, Nat. Methods. 15 (2018) 399–400. https://doi.org/10.1038/s41592-018-0019-x.
[43] G. de Los Campos, D. Gianola, G.J. Rosa, K.A. Weigel, J. Crossa, Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods, Genet. Res. (Camb). 92 (2010) 295–308. https://doi.org/10.1017/S0016672310000285.
[44] R. Bellman, Dynamic Programming Princeton University Press, 1957. https://press.princeton.edu/books/paperback/9780691146683/dynamic-programming (accessed October 13, 2020).
[45] A.M. Turing, COMPUTING MACHINERY AND INTELLIGENCE, 1950.
[46] J. Shane, You Look Like a Thing and I Love You, headline publishing group, 2019. https://www.janelleshane.com/book-you-look-like-a-thing (accessed March 2, 2021).
[47] A. Goodfellow, I., Bengio, Y., Courville, Deep Learning, MIT Press Cambridge, 2016.
[48] P. Pérez, G. de Los Campos, J. Crossa, D. Gianola, Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R., Plant Genome. 3 (2010) 106–116. https://doi.org/10.3835/plantgenome2010.04.0005.
[49] P. Pérez, G. De Los Campos, Genome-wide regression and prediction with the BGLR statistical package, Genetics. 198 (2014) 483–495. https://doi.org/10.1534/genetics.114.164442.
[50] P. Pérez, G. de los Campos, Genome-Wide Regression and Prediction with the, 198 (2014) 483–495. https://doi.org/10.1534/genetics.114.164442.
[51] D. Gianola, Priors in whole-genome regression: The Bayesian alphabet returns, Genetics. 194 (2013) 573–596. https://doi.org/10.1534/genetics.113.151753.
[52] P. Pérez, G. de los Campos, Genome-wide regression and prediction with the BGLR statistical package, Genetics. 198 (2014) 483–495.
[53] B.W. White, F. Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Am. J. Psychol. (1963). https://doi.org/10.2307/1419730.

[54] B. Efron, T. Hastie, Computer age statistical inference : algorithms, evidence, and data science, Cambridge Univ. Press. 5 (2016) 475 pp. https://books.google.es/books?hl=es&lr=&id=Sj1yDAAAQBAJ&oi=fnd&pg=PR15&dq=computer+era++efron&ots=KTx09owHuA&sig=Ah-3WI0fskCUtbkTdV9My7dgp6I#v=onepage&q=computer era efron&f=false (accessed March 2, 2019).

[55] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature. (2015). https://doi.org/10.1038/nature14539.

[56] Pérez-Enciso, Zingaretti, A Guide on Deep Learning for Complex Trait Genomic Prediction, Genes (Basel). 10 (2019) 553. https://doi.org/10.3390/genes10070553.

[57] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature. 323 (1986) 533–536. https://doi.org/10.1038/323533a0.

[58] A.-L. Cauchy, Methode generale pour la resolution des systemes d'equations simultanees, Compte Rendu Des Seances L'Acad'emie Des Sci. 25 (1847) 536–538.

[59] C. Pai, K. Potdar, A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers, Artic. Int. J. Comput. Appl. 175 (2017) 975–8887. https://doi.org/10.5120/ijca2017915495.

[60] P. Waldmann, Approximate Bayesian neural networks in genomic prediction, Genet. Sel. Evol. 50 (2018) 70. https://doi.org/10.1186/s12711-018-0439-1.

[61] C.M. Bishop, J. Lasserre, Generative or discriminative? getting the best of both worlds, Bayesian Stat. 8 (2007) 3–24.

[62] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. (2014) 2672–2680. https://doi.org/10.1016/B978-0-12-088571-8.01001-9.

[63] Y. Bengio, Webé. Thibodeau-Laufer Guillaume Alain, J. Yosinski, Deep Generative Stochastic Networks Trainable by Backprop, 2014.

[64] D.J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, in: 31st Int. Conf. Mach. Learn. ICML 2014, International Machine Learning Society (IMLS), 2014: pp. 3057–3070. https://arxiv.org/abs/1401.4082v3 (accessed October 13, 2020).

[65] D.P. Kingma, M. Welling, Auto-Encoding Variational Bayes, n.d.

[66] G.E. Hinton, T.J. Sejnowski, Optimal perceptual inference, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 1983: pp. 448–453. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.445.4504&rep=rep1&type=pdf (accessed March 1, 2019).

[67] N. Fahlgren, M.A. Gehan, I. Baxter, Lights, camera, action: high-throughput plant phenotyping is ready for a close-up, Curr. Opin. Plant Biol. 24 (2015) 93–99. https://doi.org/10.1016/J.PBI.2015.02.006.

[68] C. Costa, U. Schurr, F. Loreto, P. Menesatti, S. Carpentier, Plant phenotyping research trends, a science mapping approach, Front. Plant Sci. 9 (2019). https://doi.org/10.3389/fpls.2018.01933.

[69] T. Van Hertem, V. Alchanatis, A. Antler, E. Maltz, I. Halachmi, A. Schlageter-Tello, C. Lokhorst, S. Viazzi, C.E.B. Romanini, A. Pluk, C. Bahr, D. Berckmans, Comparison of segmentation algorithms for cow contour extraction from natural barn background in side view images, Comput. Electron. Agric. 91 (2013) 65–74. https://doi.org/10.1016/j.compag.2012.12.003.

[70] M.R. Chandraratne, D. Kulasiri, S. Samarasinghe, Classification of lamb carcass using machine vision: Comparison of statistical and neural network analyses, J. Food Eng. 82 (2007) 26–34. https://doi.org/10.1016/j.jfoodeng.2007.01.003.

[71] R. Pérez-Zavala, M. Torres-Torriti, F.A. Cheein, G. Troni, A pattern recognition strategy for visual grape bunch detection in vineyards, Comput. Electron. Agric. 151 (2018) 136–

149. https://doi.org/10.1016/j.compag.2018.05.019.
[72] X.L. Li, Z.H. Ma, C. Giagnocavo, F. Qin, H.G. Wang, J.A. Álvarez-Bermejo, Development of automatic counting system for urediospores of wheat stripe rust based on image processing, Int. J. Agric. Biol. Eng. 10 (2017) 134–143. https://doi.org/10.25165/j.ijabe.20171005.3084.
[73] M.T. Brewer, L. Lang, K. Fujimura, N. Dujmovic, S. Gray, E. Van Der Knaap, Development of a controlled vocabulary and software application to analyze fruit shape variation in tomato and other plant species, Plant Physiol. 141 (2006) 15–25. https://doi.org/10.1104/pp.106.077867.
[74] F. Tardieu, L. Cabrera-Bosquet, T. Pridmore, M. Bennett, Plant Phenomics, From Sensors to Knowledge, Curr. Biol. 27 (2017) R770–R783. https://doi.org/10.1016/j.cub.2017.05.055.
[75] L.L. Klein, M. Caito, C. Chapnick, C. Kitchen, R. O'Hanlon, D.H. Chitwood, A.J. Miller, Digital morphometrics of two north american grapevines (Vitis: Vitaceae) quantifies leaf variation between species, within species, and among individuals, Front. Plant Sci. 8 (2017) 1–10. https://doi.org/10.3389/fpls.2017.00373.
[76] P. Santagostini, S. Demotes-Mainard, L. Huché-Thélier, N. Leduc, J. Bertheloot, V. Guérin, J. Bourbeillon, S. Sakr, R. Boumaza, Assessment of the visual quality of ornamental plants: Comparison of three methodologies in the case of the rosebush, Sci. Hortic. (Amsterdam). 168 (2014) 17–26. https://doi.org/10.1016/j.scienta.2014.01.011.
[77] K.S. Lewers, M.J. Newell, E. Park, Y. Luo, Consumer preference and physiochemical analyses of fresh strawberries from ten cultivars, Int. J. Fruit Sci. (2020) 1–24. https://doi.org/10.1080/15538362.2020.1768617.
[78] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, n.d.
[79] R.C. Gonzalez, R.E. Woods, S.L. Eddins, Digital Image Processing Using MATLAB, 2da., 2009. https://doi.org/10.1007/s00147-003-0648-5.
[80] Y. Liu, S.Z. Li, W. Wu, R. Huang, Dynamics of a mean-shift-like algorithm and its applications on clustering, Inf. Process. Lett. 113 (2013) 8–16. https://doi.org/10.1016/j.ipl.2012.10.002.
[81] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC Superpixels, n.d.
[82] M. Liang, X. Hu, Recurrent convolutional neural network for object recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2015: pp. 3367–3375. https://doi.org/10.1109/CVPR.2015.7298958.
[83] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.
[84] A. Işin, C. Direkoğlu, M. Şah, Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods, in: Procedia Comput. Sci., Elsevier, 2016: pp. 317–324. https://doi.org/10.1016/j.procs.2016.09.407.
[85] R. Xu, C. Li, A.H. Paterson, Y. Jiang, S. Sun, J.S. Robertson, Aerial Images and Convolutional Neural Network for Cotton Bloom Detection, Front. Plant Sci. 8 (2018). https://doi.org/10.3389/fpls.2017.02235.
[86] C. Peter Klingenberg, Evolution and development of shape: integrating quantitative approaches, Nat. Publ. Gr. (2010). https://doi.org/10.1038/nrg2829.
[87] M.L. Zelditch, D.L. Swiderski, H.D. Sheets, Geometric Morphometrics for Biologists: A primer, Academic Press, 2004.
[88] J. Claude, Morphometrics with R, Springer, 2008. https://doi.org/10.1007/978-0-387-78171-6.
[89] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, COMPUTER VISION AND IMAGE UNDERSTANDING Active Shape Models-Their Training and Application, Comput. Vis. Image Underst. 61 (1995) 38–59.

[90] K. V. Dryden, Ian L., Mardia, Statistical shape analysis, Wiley series in probability and statistics, 1998.
[91] J.C. Gower, Generalized procrustes analysis, Psychometrika. 40 (1975) 33–51. https://doi.org/10.1007/BF02291478.
[92] M. Bylesjö, D. Eriksson, M. Kusano, T. Moritz, J. Trygg, Data integration in plant biology: The O2PLS method for combined modeling of transcript and metabolite data, Plant J. 52 (2007) 1181–1191. https://doi.org/10.1111/j.1365-313X.2007.03293.x.
[93] J. Mariette, N. Villa-Vialaneix, Unsupervised multiple kernel learning for heterogeneous data integration, Bioinformatics. 34 (2017) 1009–1015.
[94] F. Rohart, B. Gautier, A. Singh, K.-A. Lê Cao, mixOmics: An R package for 'omics feature selection and multiple data integration, PLOS Comput. Biol. 13 (2017) e1005752. https://doi.org/10.1371/journal.pcbi.1005752.
[95] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J.C. Marioni, F. Buettner, W. Huber, O. Stegle, Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets, Mol. Syst. Biol. 14 (2018) e8124.
[96] B. Berger, J. Peng, M. Singh, Computational solutions for omics data, Nat. Rev. Genet. 14 (2013) 333–346. https://doi.org/10.1038/nrg3433.
[97] M. Chagoyen, F. Pazos, Tools for the functional interpretation of metabolomic experiments., Brief. Bioinform. 14 (2013) 737–44. https://doi.org/10.1093/bib/bbs055.
[98] D. Stöckel, T. Kehl, P. Trampert, L. Schneider, C. Backes, N. Ludwig, A. Gerasch, M. Kaufmann, M. Gessler, N. Graf, E. Meese, A. Keller, H.-P. Lenhof, Multi-omics enrichment analysis using the GeneTrail2 web service, Bioinformatics. 32 (2016) btv770. https://doi.org/10.1093/bioinformatics/btv770.
[99] Y. Ramayo-Caldas, N. Mach, P. Lepage, F. Levenez, C. Denis, G. Lemonnier, J.-J. Leplat, Y. Billon, M. Berri, J. Doré, C. Rogel-Gaillard, J. Estellé, Phylogenetic network analysis applied to pig gut microbiota identifies an ecosystem structure linked with growth traits, ISME J. 10 (2016) 2973–2977. https://doi.org/10.1038/ismej.2016.77.
[100] Y. Ramayo-Caldas, M. Ballester, M.R.S. Fortes, A. Esteve-Codina, A. Castelló, J.L. Noguera, A.I. Fernández, M. Pérez-Enciso, A. Reverter, J.M. Folch, From SNP co-association to RNA co-expression: Novel insights into gene networks for intramuscular fatty acid composition in porcine, BMC Genomics. 15 (2014) 232. https://doi.org/10.1186/1471-2164-15-232.
[101] A. Acharjee, Systems biology and statistical data integration of~ omics data sets, 2013.
[102] W. You, Z. Yang, M. Yuan, G. Ji, TotalPLS: Local dimension reduction for multicategory microarray data, IEEE Trans. Human-Machine Syst. 44 (2014) 125–138. https://doi.org/10.1109/THMS.2013.2288777.
[103] C. Meng, B. Kuster, A.C. Culhane, A. Moghaddas Gholami, A multivariate approach to the integration of multi-omics datasets., BMC Bioinformatics. 15 (2014) 162. https://doi.org/10.1186/1471-2105-15-162.
[104] Y. Liu, V. Devescovi, S. Chen, C. Nardini, Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties., BMC Syst. Biol. 7 (2013) 14. https://doi.org/10.1186/1752-0509-7-14.
[105] H.L. des Plantes, Structuration des tableaux à trois indices de la statistique: théorie et application d'une méthode d'analyse conjointe, Université des sciences et techniques du Languedoc, 1976.
[106] C. Lavit, Y. Escoufier, R. Sabatier, P. Traissac, The ACT (STATIS method), Comput. Stat. Data Anal. 18 (1994) 97–119. https://doi.org/10.1016/0167-9473(94)90134-1.
[107] K.R. Gabriel, The Biplot Graphic Display of Matrices with Application to Principal Component Analysis, Biometrika. 58 (1971) 453. https://doi.org/10.2307/2334381.

Chapter 2

2.1 Objectives

The generic aim of this thesis is to develop statistical and machine learning solutions that address some of the new agricultural challenges. Hopefully, some of these tools will contribute to optimize resources and increase Agriculture sustainability.

The specific objectives are:

- To evaluate the impact of different genetic architectures and selection strategies on genomic selection in clonally propagated species.

- To evaluate the impact of the diverse modeling approaches in genomic prediction in polyploids.

- To develop pipelines for automatic phenotyping from digital images suitable for the analysis of plant and animal data. This includes automatic shape evaluation in animals and plants.

- To provide an integrative data analysis workflow capable of dealing with heterogeneous phenotypic data sources.

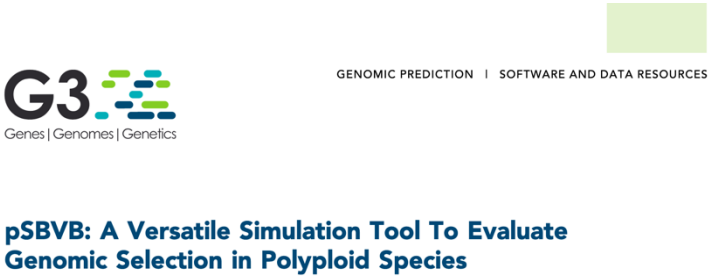
Chapter 3

pSBVB: a versatile simulation tool to evaluate genomic selection in polyploid species

Laura M. Zingaretti[1], Amparo Monfort[1,2], Miguel Pérez-Enciso[1,3]

Affiliations

1. Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Barcelona, Spain.
2. IRTA (Institut de Recerca i Tecnologia Agroalimentàries), Barcelona, Spain.
3. ICREA, Passeig de Lluís Companys 23, 08010 Barcelona, Spain.



Correspondence:

Laura M. Zingaretti (laura.zingaretti@cragenomica.es)

Miguel Pérez-Enciso (miguel.perez@uab.es)

G3 (Bethesda). 2019 Feb 7;9(2):327-334. doi: 10.1534/g3.118.200942. PMID: 30573468; PMCID: PMC6385978.

Published online 2018, Dec 20

Keywords: GenPred; Genomic Prediction; Genomic selection; Polyploids; Shared Data Resources; Simulation; pSBVB.

Abstract

Genomic Selection (GS) is the procedure whereby molecular information is used to predict complex phenotypes and it is standard in many animal and plant breeding schemes. However, only a small number of studies have been reported in horticultural crops, and in polyploid species in particular. In this paper, we have developed a versatile forward simulation tool, called polyploid Sequence Based Virtual Breeding (pSBVB), to evaluate GS strategies in polyploids; pSBVB is an efficient gene dropping software that can simulate any number of complex phenotypes, allowing a very flexible modeling of phenotypes suited to polyploids. As input, it takes genotype data from the founder population, which can vary from single nucleotide polymorphisms (SNP) chips up to sequence, a list of causal variants for every trait and their heritabilities, and the pedigree. Recombination rates between homoeologous chromosomes can be specified so that both allo- and autopolyploid species can be considered. The program outputs phenotype and genotype data for all individuals in the pedigree. Optionally, it can produce several genomic relationship matrices that consider exact or approximate genotype values. pSBVB can therefore be used to evaluate GS strategies in polyploid species (say varying SNP density, genetic architecture, or population size, among other factors), or to optimize experimental designs for association studies. We illustrate pSBVB with SNP data from tetraploid potato and partial sequence data from octoploid strawberry, and we show that GS is a promising breeding strategy for polyploid species but that the actual advantage critically depends on the underlying genetic architecture. Source code, examples, and a complete manual are freely available in GitHub https://github.com/lauzingaretti/pSBVB.

3.1 Introduction

Genomic selection (GS) [1] is the breeding strategy consisting of predicting future performance using DNA information from the whole genome, typically SNPs (single nucleotide polymorphisms). It relies on genome-wide linkage disequilibrium (LD) between markers and the causal mutations, without the need to identify them. Due to dramatic reduction in genotyping costs, GS is becoming standard in many animal and plant breeding schemes, replacing or complementing traditional methods based solely on pedigree information. So far, GS has been mainly applied to diploid species. Although polyploidy is a very common phenomenon in evolution and includes numerous species of interest (e.g., strawberry, potato, wheat), the impact of GS on polyploid breeding remains largely unexplored. Traditionally, polyploid species have been classified into autopolyploids, caused by one or more genome duplication events in a single species, and allopolyploids, the result of hybridization between closely related species [2].

In principle, the application of GS in polyploid species can have a positive impact on the rates of genetic gain through improved accuracy of predicted breeding values and/or reduction of generation intervals [3–7]. However, the complex genetic structure of polyploids has delayed the availability of genome-wide genotyping SNP arrays that are needed for GS. Polyploid SNP detection can be challenging due to the high similarity between homologous and homoeologous sequences, which generates complications to differentiate true SNPs from nuisance paralogous variants [4,8].

Further, accurate genotyping is also important but becomes more complex as the ploidy level increases. Several tools to perform genotype estimation from SNP array platforms are already available [9–12]. However, the arising of Next Generation Sequencing technologies requires new tools adapted for this type of data, which are also being developed [13–15].

Computer simulation is a fundamental tool to evaluate alternative breeding schemes since it allows the exploration of a wide range of hypotheses at no cost and can help to interpret the outcome of selection in complex situations. In this regard, numerous simulation tools have been developed such as easyPOP [16], simuPOP [17,18], forqS [19] Slim [20], PedigreeSim [21] among others. However, simulation approaches may not be straightforward to interpret owing to unknowns on the genetic architecture, among other factors. These problems are exacerbated in polyploid species and, to the best of our knowledge, only simuPOP and PedigreeSim allow polyploids organisms. simuPOP is not developed to compare breeding schemes, whereas PedigreeSim does not directly generate phenotypes nor produce genomic relationship matrices.

Here we present a flexible simulation tool for complex phenotypes adapted to polyploids and we propose several approaches to compute the molecular relationship matrix in polyploids. The software is an extension of Sequence-Based Virtual Breeding (SBVB, [22]), called pSBVB. This tool employs complete or partial genome data as input and simulates new genomes by gene dropping. We illustrate the software with data from two economically important polyploid species: potato, an autopolyploid, and strawberry, an allopolyploid.

3.2 Methods

3.2.1 Polyploid sequence based virtual breeding (pSBVB)

pSBVB is a modification of SBVB software [22] that allows simulating genotypes and phenotypes of an arbitrary genetic complexity in polyploids. Compared to SBVB designed for diploid organisms only, pSBVB enables simulating meiosis in autopolyploid or allopolyploid species (see below). It takes ploidy into account to generate the phenotypes and incorporates several options to compute the molecular relationship matrix that are pertinent to polyploids, as described below.

The source codes and the documented functions are distributed from GitHub: https://GitHub.com/lauzingaretti/pSBVB. The manual includes a full tutorial of all functions at the program and a user guide with the installation guidelines and examples to simulate polyploid organisms. The software is accompanied by R scripts [23] to generate a pedigree file, compute the numerator relationship matrix, perform GBLUP [24] or assess predictive ability (PA). Examples showing the software capabilities with alternative parameter options are also available.

3.2.2 Software algorithm

As input, pSBVB needs genotypes in vcf format (https://samtools.GitHub.io) or a text file with genotypes coded to 0 up to h (where h is the ploidy level). For diploids, the vcf genotype format is of the kind 0/0, 0/1, and 1/1 for the three possible genotypes in a biallelic SNP. The polyploid

vcf format is an extension of the type 0/0/0/0, 0/0/0/1, and so on in the case of unphased tetraploid genotypes. Phased genotypes are represented by vertical bars, (e.g., genotype 0|0|0|1 is different from 1|0|0|0) . No missing values are allowed. Phased genotypes are needed in pSBVB to identify which chromosomes are passed to offspring. A number of accurate phasing algorithms for diploids are available such as beagle [25] or minimac [26]. For polyploids, several approaches are also developed [27,28], but their accuracy has not been completely validated and seems critically dependent on ploidy level. If the phase is unknown, pSBVB randomly generates a phase configuration. Further, linkage disequilibrium can be obtained by generating an individual genome out of a random pedigree starting with the founders' genotypes. To do that, pSBVB incorporates the option 'EXPAND_ BASEPOP', which generates additional founders' by randomly crossing the available ones and random breeding for a pre-specified number of generations (see SBVB manual, https://lauzingaretti.GitHub.io/pSBVB/). A list with QTNs (Quantitative Traits Nucleotides) positions, a list of SNP positions to be used for GS, a pedigree file, and a parameter file are also necessary. The pedigree file is used to perform the gene dropping simulation, i.e., genotypes' of the descendants along the pedigree are generated following Mendelian rules and a pre-specified pairing rate between homologous and homoelogous pairs; for autopolyploids, pairing is at random. While performing gene dropping, pSBVB stores only the recombination breakpoints, which results in an efficient algorithm to recover marker genotypes and phenotypes.

pSBVB is very flexible in terms of genetic architectures; it can simulate any number of traits with their specific QTNs and allelic effects. QTNs effects can be specified in a file or sampled from gamma, normal, or uniform distributions. In contrast to SBVB, though, pSBVB does not allow for epistasis. Figure 3.2 shows a general representation of the pSBVB software, as well as screenshots.

As output, pSBVB produces phenotype and marker data of the individuals obtained from the pedigree-based gene-dropping procedure. In addition, pSBVB can also compute molecular relationship matrices $\boldsymbol{G}$ using predefined marker subsets (e.g., a genotyping array) or the whole sequence. By default, $\boldsymbol{G}$ is computed from:

$$\mathbf{G} = \frac{(\mathbf{M} - h\mathbf{p})(\mathbf{M} - h\mathbf{p})^T}{h\mathbf{p}(1-\mathbf{p})^T} \qquad \text{[Eq. 3.1]}$$

where $\boldsymbol{M}$ is a $\boldsymbol{n \times m}$ matrix with elements containing the number of copies of the alternative allele for $i\,th$ individual ($i = 1..n$) and $j\,th$ SNP ($j = 1..m$), and p is a m-dimension vector with marker allele frequencies. Note that Eq. 3.1 reduces to the standard formula in the case of diploidy ($h = 2$) [24].

Assessing the genotype for polyploids can be inferred from fluorescence intensity in SNP arrays or from read count in sequence data [13] but may not be as accurate as in diploid organisms, especially at high ploidy levels. If genotyping is not accurate, a simple alternative is to assume that only one full homozygous can be distinguished for the rest of genotypes, i.e., that a given marker allele behaves as fully dominant. To accommodate this possibility, pSBVB allows computing a modified $\boldsymbol{G^*}$ where element m_{ij} is coded as 0 if all alleles are 0 and 1 otherwise. This is specified

with the MIMIC_HAPLOID statement in the parameter file. The software also incorporates a 'MIMIC_DIPLOID' option, which assumes that only the presence or absence of the alternative allele can be ascertained for genotype values higher than 2. In summary, the software is able to generate three ***G*** matrices:

- **Default option:** The true genotype, i.e., number of copies of the alternative allele, is known without error **(GT**). In this approach, M (Eq. 3.1) has elements varying between 0 to *h*.

- **MIMIC_DIPLOID:** Only 0, 1 and 2 or more copies of a given allele can be distinguished. In this case, all genotypes with values larger than 2 are assigned a value '2', thus M (Eq. 3.1) has elements ranging between 0 and 2, and ploidy (*h*) is set to 2.

- **MIMIC_HAPLOID:** It considers that only one full homozygous can be distinguished for the rest of genotypes, then M (Eq. 3.1) has elements ranging between 0 and 1, and ploidy is set to 1.

3.2.3 Modeling meiosis in polyploids

Autopolyploid species have polysomic inheritance where homologous and homoeologous chromosomes are randomly paired during meiosis. In contrast, most allopolyploids have a disomic inheritance, resulting from a preferential pairing between homologous chromosomes. However, there is a continuum between both extreme meiotic behaviors that can be modeled by the preferential pairing factor (θ), which represents the deviations from random paring [29]. In a generic case with $h/2$ sub-genomes, where h is the ploidy level, there are $\frac{\binom{h}{2}}{\binom{(h/2)}{(h/2)-1}} = \frac{(h)(h-1)/2}{h/2} = h-1$ possible paring combinations between homologous and homoeologous chromosomes. The deviation from the random pairing scenario can be modeled as:

$$c_h = \frac{1}{h-1} + \theta_h$$
$$c_{\bar{h}} = \frac{1}{h-1} - \frac{\theta_h}{h-2} \qquad \text{[Eq. 3.2]}$$

where $0 \le \theta_h \le \frac{h-2}{h-1}$ and h $\bar{h}$ indicate whether recombination occurs between homologous or homoelogous pairs, respectively. If the recombination occurs only between homologous chromosomes then, $\theta_h = \frac{h-2}{h-1}$ and $c_h = 1$, whereas $\theta_h = 0$ means that the probability of recombination is the same between all chromosomes. pSBVB allows modeling meiotic pairing via a recombination matrix containing the c elements in (Eq. 3.2) between all chromosome pairs. In a generic case for a tetraploid, the matrix needed by pSBVB is

$$R = \begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \\ . \\ . \\ r_h \end{array} \begin{array}{c} \begin{array}{ccccc} r_1 & r_2 & r_3 & \ldots & r_h \end{array} \\ \begin{pmatrix} 0 & \frac{1}{h-1}+\theta_{12} & \frac{1}{h-1}+\theta_{13} & \ldots & \frac{1}{h-1}+\theta_{1h} \\ \frac{1}{h-1}+\theta_{12} & 0 & \frac{1}{h-1}+\theta_{23} & \ldots & \frac{1}{h-1}+\theta_{2h} \\ \frac{1}{h-1}+\theta_{13} & \frac{1}{h-1}+\theta_{23} & 0 & \ldots & \frac{1}{h-1}+\theta_{3h} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \frac{1}{h-1}+\theta_{1h} & \frac{1}{h-1}+\theta_{2h} & \frac{1}{h-1}+\theta_{3h} & \ldots. & 0 \end{pmatrix} \end{array} \quad \text{[Eq.3.3]}$$

where r_{ij} (Eq. 3.3) is the pairing probability between *i* and *j* homologous/homoelogous chromosomes. Note that, the chromosomes have to be sorted, rows and columns have to sum 1, i.e., to a *i*-- row, $\sum_{j=1}^{j=h} \theta_{ij} = 0$ and diagonal elements always are set to 0.

For example, the matrix for a strict auto-tetraploid is:

$$R = \begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \\ r_4 \end{array} \begin{array}{c} \begin{array}{cccc} r_1 & r_2 & r_3 & r_4 \end{array} \\ \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \end{bmatrix} \end{array} \quad \text{[Eq.3.4]}$$

And for a strict allopolyploid would be

$$R = \begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \\ r_4 \end{array} \begin{array}{c} \begin{array}{cccc} r_1 & r_2 & r_3 & r_4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \quad \text{[Eq. 3.5]}$$

3.2.4 Phenotype simulation

In a diploid organism, the phenotype for *i-th* individual can be simulated from (Eq. 3.6).

$$y_i = \mu + \sum_{j=1}^{Q} \gamma_{ij} a_j + \sum_{j=1}^{Q} \delta_{ij} d_j + \varepsilon_i, \qquad \text{[Eq. 3.6]}$$

where μ is the general mean, a_j is the additive effect of j-th locus, that is, half the expected difference between homozygous genotypes, and it takes values -1, 0, and 1 for homozygous, heterozygous, and alternative homozygous genotypes, respectively, d_j is the dominance effect of

j-th locus and takes value 1 if the genotype is heterozygous, 0 otherwise, and ε_i is a normal residual of the *i-* observation. For polyploids, the equivalent equation can be expressed as in the Eq. 3.7.

$$y_i = \mu + \sum_{j=1}^{Q} \eta_{ij} a_j + \sum_{j=1}^{Q} \varphi_{ij} d_j + \varepsilon_i, \qquad \text{[Eq. 3.7]}$$

where η_{ij} is the number of copies of the alternative allele (coded say as 1) minus half the ploidy ($h/2$) for *j*-th locus and *i*-th individual, and a_j is, therefore, the expected change in phenotype per copy of allele '1' in the *j*-th locus. In polyploids, as many dominance coefficients as ploidy level (h) minus two can technically be defined. However, this results in an over-parameterized model that is of no practical use. Here instead we define the φ_{ij} parameter as the minimum number of copies of allele 1 such that the expected phenotype is d. In our modeling, all genotypes with a number of copies over φ_{ij} have the same phenotype. See Figure 3.1 for a graphical representation of this modeling approach. By default, pSBVB takes $\varphi_{ij} = 1$.

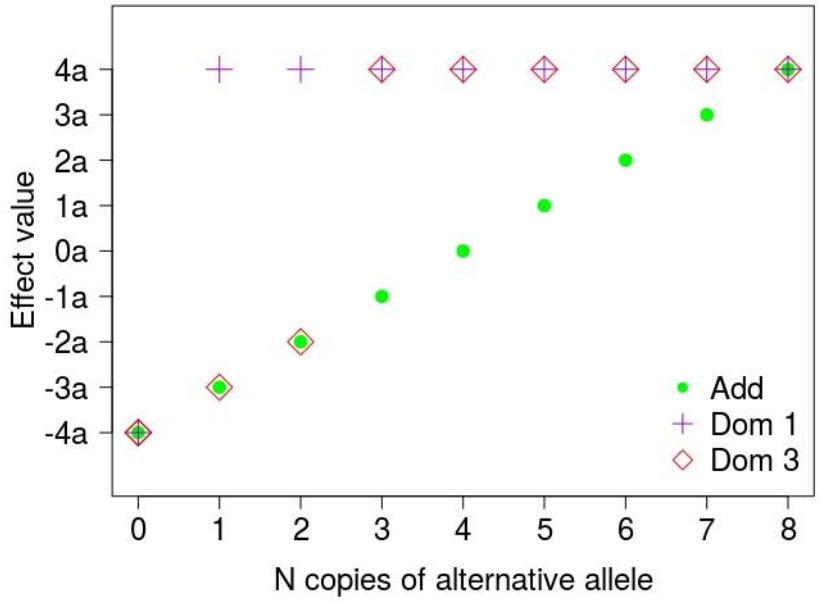


Figure. 3.1. Additive and dominance modeling in polyploids used by pSBVB. The figure represents three possible genic actions in an octoploid. Under a strict additive action (○), the phenotype is expected to increase in an '*a*' unit per copy of the alternative allele (note that *a* can be negative or positive). Under dominance action and □ = 1 (Dom 1, see Eq. 3.5), the phenotype is expected to be the same for any heterozygous genotype (+). With dominance and □ = 3 (Dom 3), the genotype is expected to be the same for all heterozygous genotypes containing more than 3 copies of the alternative allele (◊).

3.2.5 Statistical Model for Genomic Prediction

There are currently numerous statistical methods that address the large p small n problem and use genome-wide markers to predict breeding values (eg., [30,31]). pSBVB does not compute genomic breeding values but can produce genomic relationship matrices suitable to obtain GBLUP [24], as detailed above. Otherwise, pSBVB outputs genotypes of all or a subset of markers, and any desired GS algorithm can be applied. R scripts are provided in GitHub that performs GBLUP.

3.3 Results

In order to illustrate the software capabilities, we have used a dataset from two polyploids species: autopolyploid potato (*Solanum tuberosum, 2n=4x=48*) and allopolyploid strawberry (*Fragaria* x *ananassa 2n=8x=56)*.

3.3.1 Potato genotypes

The availability of an 8,300 SNP array has allowed the development of GS studies in potato, one of the most important crops worldwide (e.g., [5,7]). To illustrate our tool, here we used a subset of 407 SNPs and 150 individuals from Enciso-Rodriguez *et al.* [7]. SNP positions were obtained from Rosyara *et al.* [32]. We used these genotypes to generate a vcf file where genotypes were coded between 0 and 4 (the potato ploidy level), phases were randomly generated.

Next, to generate linkage disequilibrium in the randomly phased dataset, we included additional dummy founders using the 'EXPAND_BASEPOP' statement in the parameter file (see reference manual, HTTPS://lauzingaretti.GitHub.io/pSBVB/). With this option, new base population individuals are obtained via randomly generated pedigrees. A new base population with 100 founders was obtained. The total pedigree size was 700, including 250 founders (150 initial individuals and 100 new base population individuals) and four generations with 100, 100, 100, and 150 individuals, respectively.

Phenotypes were simulated using 140 randomly chosen QTNs and heritability (h^2) was set to 0.5. As numerous studies suggest that allele distribution is highly leptokurtic [33,34] with many near-zero effects and a few large effects, we used a gamma $\Gamma(\alpha = 0.2, \beta = 5)$ distribution to simulate additive effects as in [35]. **G** matrix was computed assuming that all markers are known without error since the potato chip ensures that the true genotype can be obtained. Finally, to illustrate GS performance, we predicted breeding values using GBLUP and quantified the predictive ability (PA), which was assessed by removing the 150 individuals from the last generation and computing the correlation between predicted and observed phenotypes of these 150 individuals. Figure 3.3 plots the observed vs. predicted phenotypes in training (400 individuals) and test (150 individuals) population. In this example, PA was reasonably high (= 0.52), and illustrates that reasonable accuracies can be obtained even with small population sizes provided linkage disequilibrium and h^2 are relatively high.

The pedigree and the numerator relationship matrix files were generated using the pedigree.R and RelationshipMatrix.R functions, respectively; breeding values were predicted with GBLUP using GBlupFunction.R script. The whole source code and scripts to run this example are available at the GitHub site.

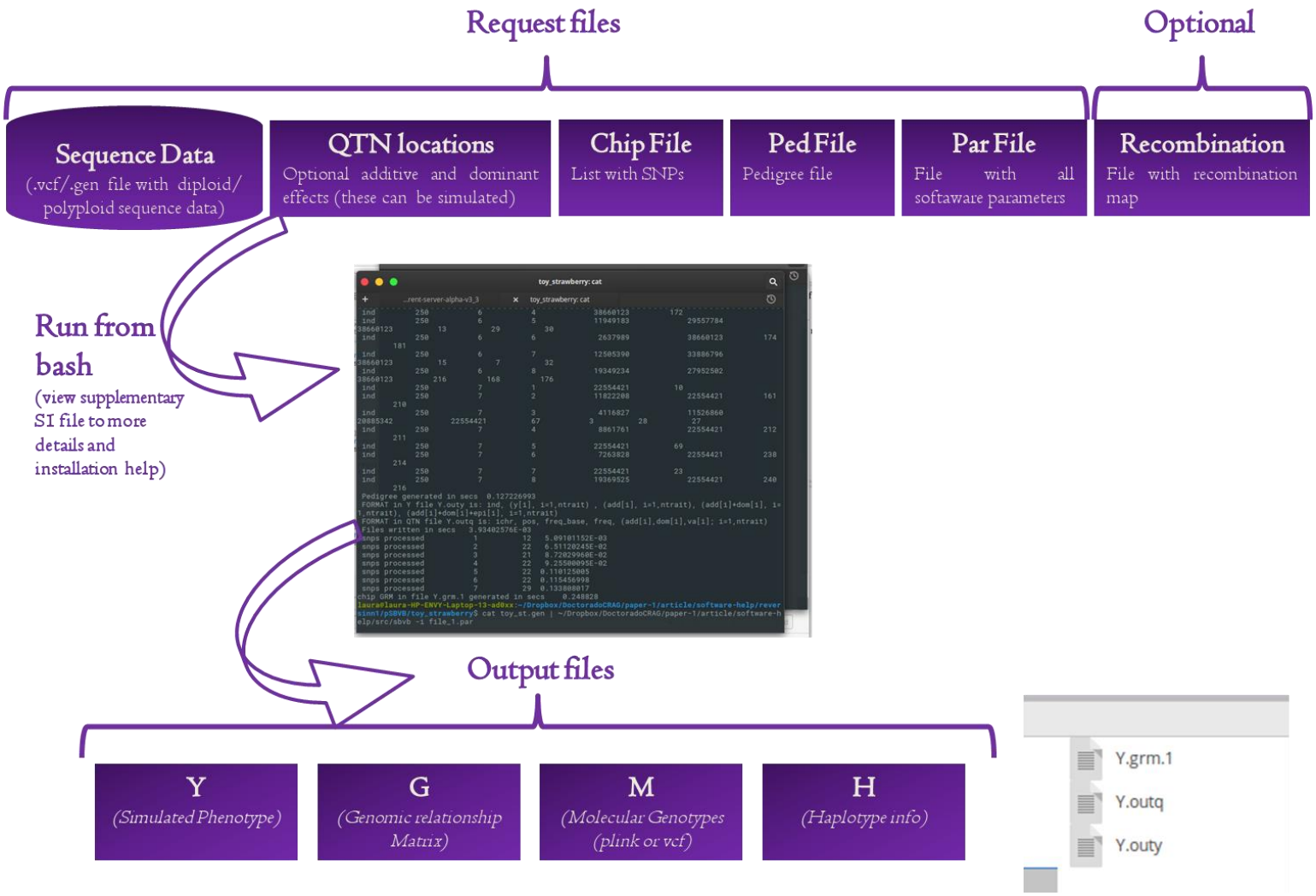


Figure 3.2: General representation of pSBVB software. As input, the software reads the vcf file containing all phased SNPs from founder haplotypes. Additional files specify the genetic architecture (it may include additive and dominant effects), the lists of SNPs (each corresponding to one genotyping array and/or complete sequence), and the recombination map for each sex and genome location (optional). pSBVB then performs gene dropping following a predetermined pedigree, generating phenotypes and true genotypes (Y) genomic relationship matrices (G), one per SNP list), and genotypes for each individual in the pedigree and for each SNP list in Plink or generic format, an optional file containing haplotype information that allows quick restart of the program, and information on QTN contribution to variance. Genomic Relationship Matrix can be computed using several options (see main text). As output, the software provides genomic relationship matrix (Y.grm.1), QTN's effects (y.outq), and simulated phenotypes (y.outy)

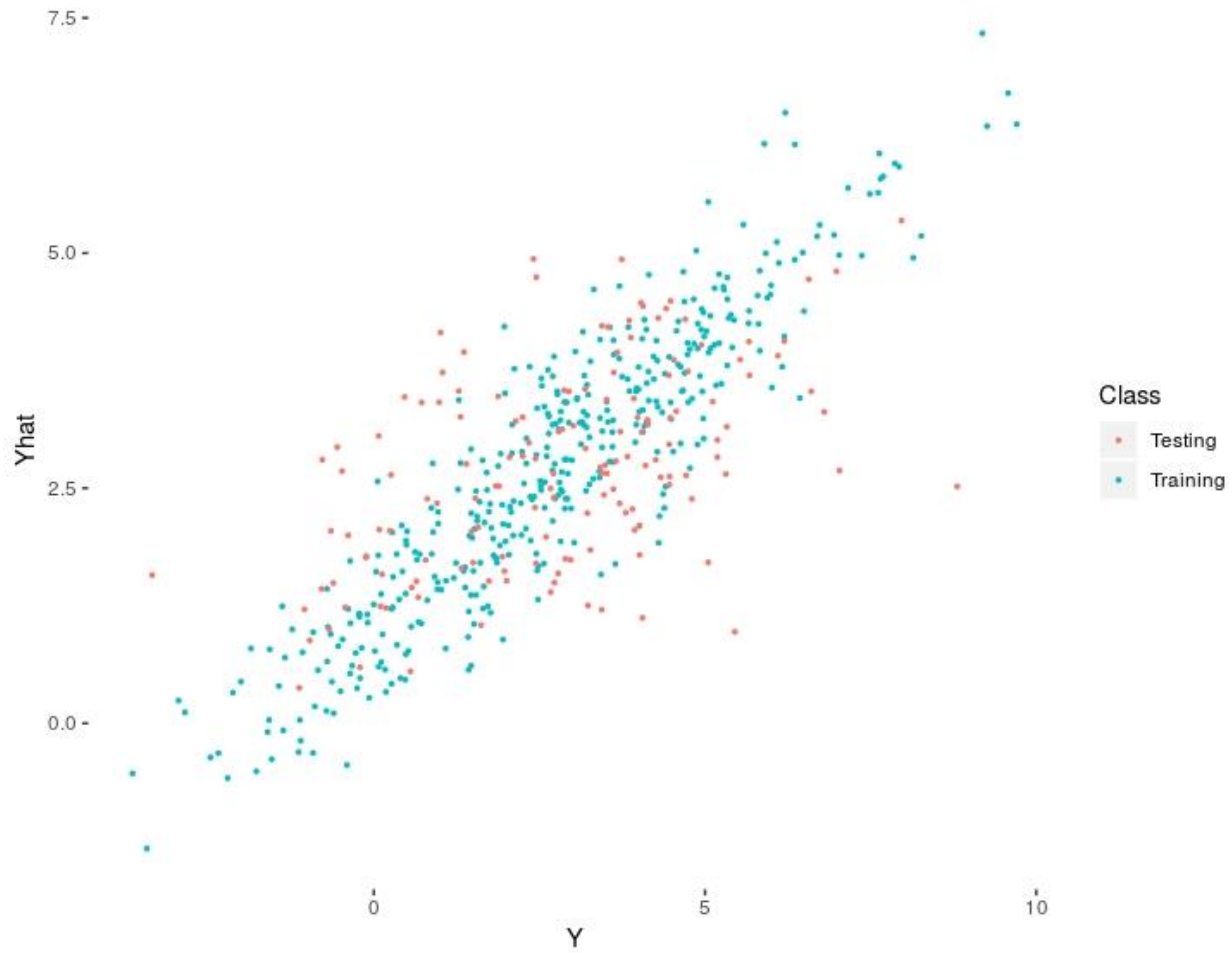


Figure 3.3 Predicted breeding values from the GS model in the simulated potato dataset. Correlations between observed and predicted values from training and testing populations were 0.91 and 0.52, respectively.

3.3.2 Application to strawberry GBS data

We also applied our program to octoploid strawberry *F. x ananassa*. In the absence of a reasonable number of strawberry sequenced genomes, we used unpublished data obtained with GBS (Genotyping by Sequencing) from 47 strawberry cultivars. Genotype-by-Sequencing libraries were prepared by Heartland Plant Innovations (http://www.heartlandinnovations.com/). Samples were multiplexed and sequenced 92 cycles on the Illumina MiSeq at the Oklahoma Medical Research Foundation. Data quality was checked by FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). To obtain reasonably realistic genotypes based on these data, we applied the following pipeline. GBS reads were aligned against F. vesca (diploid strawberry) reference genome (F. vesca-genome.v2.0.a1), bam files were filtered setting minimum base and mapping qualities to 37 and 20, respectively, and parsed with snape (https://github.com/EmanueleRaineri/snape-pooled, [36]), an SNP caller developed for pools.

This software requires as input the number of diploid individuals in the pool, which was set to four. Polymorphic positions with fewer than 20 high-quality reads were removed, as well as those where more than 60% of the cultivars were not covered. Logically, only genotype values 0, 1, to 8

are allowed in an octoploid genome SNP, whereas the number of reads per position follows a quasi-continuous distribution. To convert the number of reads to genotype score, we computed the fraction of alternative allele reads divided by the total number of reads (*f*) and inferred its genotype from the nearest possible integer to $f \times 8$. This was done for each SNP and cultivar. Missing genotypes were sampled according to the genotype frequency in the non-missing positions for that SNP. We assumed independence to perform the assignations. A total of 50,609 variant positions were obtained (5779, 7985, 7328, 6362, 8282, 9012, 5862 in linkage groups GL1, GL2, GL3, GL4, GL5, GL6, and GL7, respectively). These markers were used as genetic file input for the program. Among those SNPs, ~ 36%, 37%, 14% and 13% variants were classified as segregating in 1, 2, 3 or all sub-genomes: 2x, 4x, 6x and 8x, respectively.

Strawberry breeding programs are based on evaluating crosses between elite lines. Traditional crop breeding is expensive and time-consuming, and GS can accelerate strawberry improvement if only a subset of these crosses were fully tested in the field. To mimic this scenario, we generated a pedigree file with five generations of intercrossing starting with the 53 base population lines. Each generation was made up of 100 lines. In the last generation, 1000 crosses with unknown phenotype were generated from the 100 current parental lines. As a measure of predictive accuracy, we computed the correlation between observed and predicted phenotypes of the 1000 crosses, when the phenotypes from these 1000 crosses were removed. One hundred replicates were run per case.

To simulate the phenotypes, we considered a range of genetic architectures with a focus on sugar content:

- Random QTNs in sugar-associated pathways (RQP): 100 SNPs were randomly chosen as causal among the SNPs in the sugar pathway associated genes ± 10 kb.

- Diploid QTNs in sugar-associated pathways (DQP): 100 SNPs were randomly chosen as causal among the diploid SNPs in the sugar pathway-associated genes ± 10 kb.

- Random QTNs genome-wide chosen (RQG): 100 SNPs were randomly chosen as causal among all detected SNPs.

In the first two architectures, we aimed at mimicking a trait of economic interest such as sucrose content. The gene information was obtained from FragariaCyc (http://pathways.cgrb.oregonstate.edu, [37]). In total, there were 159 genes containing 499 SNPs associated with these pathways. Within each of the three architectures, phenotypes were simulated according to two extreme gene actions: fully additive and complete dominance ($\Phi = 1$, Figure 3.1). Heritability was set to 0.5.
For each architecture, phenotypes were simulated according to two extreme gene actions: fully additive and complete dominance. In the dominant approach, we set $G(\alpha = 0.2, \beta = 5)$ (Figure 3.1). Each phenotype was generated from its genotypic value adding an environmental effect, where was adjusted such that heritability was $h^2 = 0.5$.

Simulated PAs are in Figure 3.4. We estimated the PA using the following matrices:

- **GT:** The true genotype, i.e., number of copies of the alternative allele, was known without error and all SNPs were used. In this approach, ***M*** (Eq. 3.2) has elements varying between 0 and 8.

- **G2:** Only diploid SNPs were used, and genotypes were known without error. ***M*** (Eq. 3.2) has elements ranging between 0 and 2.

- **G2*:** All SNPs were employed but only genotypes of diploid SNPs were known without error, whereas for the remaining, although the organism was polyploid, the Genomic matrix is computed mimic diploid . ***M*** (Eq. 3.2) has elements ranging between 0 and 2.

- **Numerator Relationship Matrix (P-BLUP):** The breeding values were predicted using the pedigree relationship matrix.

Figure 3.4 shows the obtained accuracies across genetic architectures and for each evaluation method. Overall, these results indicate that the performance of GS in polyploids may critically depend on the underlying genetic architecture. Unsurprisingly, accuracy also drops when dominance exists compared to the additive scenarios. Several additional observations of interest can be drawn from Figure 3.4. First, there were no differences in the ranking of methods irrespective of whether QTN were scattered throughout the genome (RQG) or localized in given segments (RQP). This was observed for both additive and dominant architectures. Second, using the true genotype values to build **G** (**GT**) did not always outperform the rest of GBLUP methods considered. In fact, this was observed only when the architecture was fully additive and the QTNs were segregating in more than one homeolog group. In these cases, **GT-**BLUP was ~ 4 - 8% better than **G2**-BLUP or **G2***-BLUP**. G2**, which employs only diploid SNPs, should be preferred to **GT-**BLUP only if QTNs are exclusively diploid. A relevant result is that **G2*** -BLUP, which treats markers as dominant, was a quite robust strategy, in particular with complete dominance and with the exception of DQP scenario (i.e., when all QTNs were diploid).

Finally, note that the advantage of GBLUP over **P-BLUP** is not always guaranteed. At least in the breeding scenario analyzed here, **G2**-BLUP might actually perform worse than **P-BLUP** when QTNs segregate randomly (RQP and RQG) and genic action is additive. If true SNP genotypes could be known without error (**GT**), the increase in accuracy compared to **P-BLUP** would vary between ~ 7% and 18%. As for using **G2* -**BLUP, the increase in accuracy was between ~ 3% and ~ 16% across all cases examined here. The advantage, though, would diminish if genic action were additive and QTN would segregate in all homologous.

The genetic file used as input includes 1500 SNPs from the whole vcf file. More examples combining a set of different parameters (additive and dominance effects, Genetic Matrix calculation, pedigree, and Genomic Relationship Generation, among others) are available on GitHub.

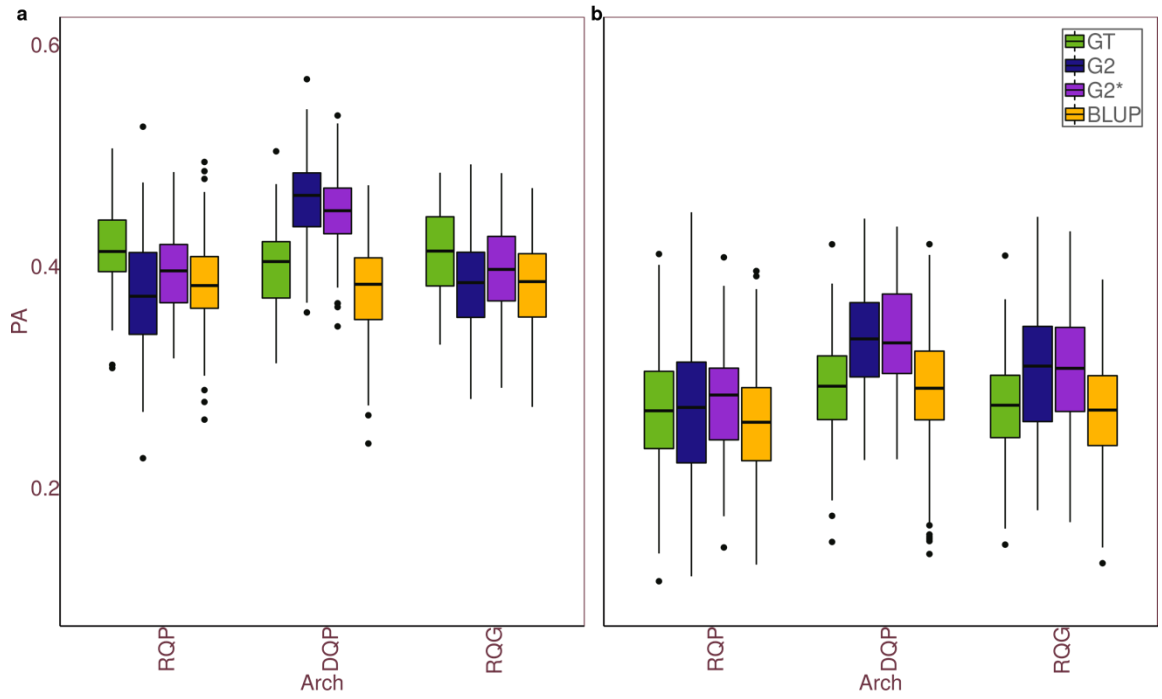


Figure 3.4: Predictive Ability $cor(y, \hat{y})$ of GBPLUP and **P-BLUP** models for each of the three genetic architectures considered in the strawberry dataset: random QTNs in sugar associated pathways (RQP), diploid QTNs in sugar associated pathways (DQP), and genome-wide chosen (RQG), and each of the three GBLUP models. Three GBLUP models were compared: In **GT**, genetic matrix **G** was computed assuming SNP allele frequencies were known without error; in **G2**, only diploid SNPs were used, and genotypes were known without error; and in **G2***, **G** Genomic relationship matrix is computed assuming than only presence or absence of the alternative allele could be known for the remaining, i.e., although the organism was polyploid, Genomic relationship matrix is computed assuming than only presence or absence of the alternative allele can be ascertained. (a) additive architecture; (b) dominant architecture.

3.4 Discussion

Certainly, polyploid sequence data will be increasingly available, which will be used to achieve a better understanding of complex trait genetics and to optimize GS strategies. To help in the latter task, here we have developed an extension of SBVB software (pSBVB) that feeds from real sequence data of polyploid organisms. It uses efficient forward algorithms and allows simulating meiosis in polyploid species, suited for both auto and allopolyploid organisms. Further, pSBVB generalizes genetic modeling in polyploids to generate phenotypes and incorporates several options to compute predefined molecular relationship matrices that are specific to polyploid organisms. Note though that, since pSBVB can print the whole SNP dataset, any custom-made **G** can be computed and any alternative GS method can be evaluated. There are some limitations though. An important one is that epistasis cannot be modeled in pSBVB - in contrast to the diploid version (SBVB) - but this limitation stems from the lack of realistic modeling on epistasis for polyploids rather than out of computational constraints.

To the best of our knowledge, there are no simulation tools that allow estimate genetic matrix in polyploid organisms with a range of options like the one described here. Among the available forward-time simulation tools, only simuPOP [17,18] and PedigreeSim [21] consider polyploids.

Compared to simuPOP, pSBVB allows simulating both auto and allo-polyploid organisms, accepting as input a recombination matrix between homeolog groups. PedigreeSim is not specifically designed for GS and is not able to simulate complex genetic architectures and compute relationship matrices as pSBVB. A further outcome of our work is the proposal of several G matrices that are robust to genotype misspecification, an important problem in polyploids [13].

To conclude, we have developed a flexible GS simulation tool capable of using real sequence data from polyploids. We show the tool capabilities using potato and strawberry real datasets. With potato genotypes, we illustrate how new base population individuals can be generated and show that accuracy can be relatively high even with modest population sizes. Among the molecular relationship matrices proposed, assuming that only diploid genotypes can be identified seems overall a good compromise in terms of performance, at least in strawberry data. Our study suggests that GS may increase response to selection compared to **P-BLUP**, but this will depend on the true genetic architecture of the trait, as also shown by Gezan *et al.* [6] with real strawberry data. We urge advancing on the quantitative and molecular dissection of complex traits in polyploids, which should provide important parameters such as prevalent genic action or number of segregating homeolog groups, in order to design optimum GS breeding schemes for these species.

Author contribution statement

MPE and AM conceived and designed research; MPE and LZ developed software and methods; LZ performed research; LZ wrote the manuscript with help from MPE and AM. All authors read and approved the manuscript.

Acknowledgments

LZ is recipient of a FPI grant to achieve the PhD research from Ministry of Economy and Science (MINECO, Spain), associated with 'Centro de Excelencia Severo Ochoa 2016-2019' award SEV-2015-0533 to CRAG. Work funded by MINECO grant RTA2013-00010-00-00 to AM and AGL2016-78709-R to MPE. We also acknowledge genotyping data provided by General Mills.

3.5 Literature Cited

[1] T.H.E. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, Genetics. 157 (2001) 1819–1829. https://doi.org/11290733.

[2] G.L. Stebbins, Types of Polyploids: Their Classification and Significance, Adv. Genet. 1 (1947) 403–429. https://doi.org/10.1016/S0065-2660(08)60490-3.

[3] A.T. Slater, N.O.I. Cogan, J.W. Forster, B.J. Hayes, H.D. Daetwyler, Improving genetic gain with genomic selection in autotetraploid potato, Plant Genome. 9 (2016).

[4] N. V Bassil, T.M. Davis, H. Zhang, S. Ficklin, M. Mittmann, T. Webster, L. Mahoney, D. Wood, E.S. Alperin, U.R. Rosyara, H. Koehorst-vanc Putten, A. Monfort, D.J. Sargent, I.

Amaya, B. Denoyes, L. Bianco, T. van Dijk, A. Pirani, A. Iezzoni, D. Main, C. Peace, Y. Yang, V. Whitaker, S. Verma, L. Bellon, F. Brew, R. Herrera, E. van de Weg, Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry Fragaria × ananassa, BMC Genomics. 16 (2015) 155. https://doi.org/10.1186/s12864-015-1310-1.

[5] E. Sverrisdóttir, S. Byrne, E.H.R. Sundmark, H.Ø. Johnsen, H.G. Kirk, T. Asp, L. Janss, K.L. Nielsen, Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing, Theor. Appl. Genet. 130 (2017) 2091–2108. https://doi.org/10.1007/s00122-017-2944-y.

[6] S.A. Gezan, L.F. Osorio, S. Verma, V.M. Whitaker, An experimental validation of genomic selection in octoploid strawberry, Hortic. Res. 4 (2017) 16070. https://doi.org/10.1038/hortres.2016.70.

[7] F. Enciso-Rodriguez, D. Douches, M. Lopez-Cruz, J. Coombs, G. de los Campos, Genomic selection for late blight and common scab resistance in tetraploid potato (Solanum tuberosum), G3 Genes, Genomes, Genet. 8 (2018) 2471–2481. https://doi.org/10.1534/g3.118.200273.

[8] J.P. Clevenger, P. Ozias-Akins, SWEEP: A Tool for Filtering High-Quality SNPs in Polyploid Crops, G3 Genes, Genomes, Genet. 5 (2015) 1797–1803. https://doi.org/10.1534/g3.115.019703.

[9] R.E. Voorrips, G. Gort, B. Vosman, Genotype calling in tetraploid species from bi-allelic marker data using mixture models., BMC Bioinformatics. 12 (2011) 172. https://doi.org/10.1186/1471-2105-12-172.

[10] R.E. Voorrips, G. Gort, fitPoly: Genotype Calling for Bi-Allelic Marker Assays, (2018). https://cran.r-project.org/package=fitPoly.

[11] C.A. Schmitz Carley, J.J. Coombs, D.S. Douches, P.C. Bethke, J.P. Palta, R.G. Novy, J.B. Endelman, Automated tetraploid genotype calling by hierarchical clustering, Theor. Appl. Genet. 130 (2017) 717–726. https://doi.org/10.1007/s00122-016-2845-5.

[12] P.D. Blischak, L.S. Kubatko, A.D. Wolfe, SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data, Bioinformatics. 1 (2017). https://doi.org/10.1093/bioinformatics/btx587.

[13] P.M. Bourke, R.E. Voorrips, R.G.F. Visser, C. Maliepaard, Tools for genetic studies in experimental populations of polyploids, Front. Plant Sci. 9 (2018) 513-.

[14] P.G. Meirmans, S. Liu, P.H. van Tienderen, The Analysis of Polyploid Genetic Data, J. Hered. 109 (2018) 283–296. https://doi.org/10.1093/jhered/esy006.

[15] D. Gerard, L.F.V. Ferrão, A.A.F. Garcia, M. Stephens, Genotyping Polyploids from Messy Sequencing Data, Genetics. (2018) genetics--301468.

[16] F. Balloux, EASYPOP (Version 1.7): A Computer Program for Population Genetics

Simulations, J. Hered. 92 (2001) 301–302. https://doi.org/10.1093/jhered/92.3.301.

[17] B. Peng, M. Kimmel, simuPOP: A forward-time population genetics simulation environment, Bioinformatics. 21 (2005) 3686–3687. https://doi.org/10.1093/bioinformatics/bti584.

[18] B. Peng, C.I. Amos, Forward-time simulations of non-random mating populations using simuPOP, Bioinformatics. 24 (2008) 1408–1409. https://doi.org/10.1093/bioinformatics/btn179.

[19] D. Kessner, J. Novembre, forqs: forward-in-time simulation of recombination, quantitative traits and selection, Bioinformatics. 30 (2014) 576–577. https://doi.org/10.1093/bioinformatics/btt712.

[20] P.W. Messer, SLiM: simulating evolution with selection and linkage., Genetics. 194 (2013) 1037–9. https://doi.org/10.1534/genetics.113.152181.

[21] R.E. Voorrips, C.A. Maliepaard, The simulation of meiosis in diploid and tetraploid organisms using various genetic models, BMC Bioinformatics. 13 (2012) 248. https://doi.org/10.1186/1471-2105-13-248.

[22] M. Pérez-Enciso, N. Forneris, G. de los Campos, A. Legarra, Evaluating Sequence-Based Genomic Prediction with an Efficient New Simulator, Genetics. 205 (2017) 939–953. https://doi.org/10.1534/genetics.116.194878.

[23] R Core Team, R: A Language and Environment for Statistical Computing, (2018). https://www.r-project.org/.

[24] P.M. VanRaden, Efficient methods to compute genomic predictions, J. Dairy Sci. 91 (2008) 4414–4423. https://doi.org/10.3168/jds.2007-0980.

[25] S.R. Browning, B.L. Browning, Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering, Am. J. Hum. Genet. 81 (2007) 1084–1097. https://doi.org/http://dx.doi.org/10.1086/521987.

[26] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, G.R. Abecasis, Fast and accurate genotype imputation in genome-wide association studies through pre-phasing, Nat. Genet. 44 (2012) 955.

[27] D. He, S. Saha, R. Finkers, L. Parida, Efficient algorithms for polyploid haplotype phasing, BMC Genomics. 19 (2018) 110.

[28] J. Shen, Z. Li, J. Chen, Z. Song, Z. Zhou, Y. Shi, SHEsisPlus, a toolset for genetic studies on polyploid species., Sci. Rep. 6 (2016) 24095. https://doi.org/10.1038/srep24095.

[29] P.M. Bourke, P. Arens, R.E. Voorrips, G.D. Esselink, C.F.S. Koning-Boucoiran, W.P.C. van't Westende, T. Santos Leonardo, P. Wissink, C. Zheng, G. van Geest, R.G.F. Visser, F.A. Krens, M.J.M. Smulders, C. Maliepaard, Partial preferential chromosome pairing is

genotype dependent in tetraploid rose, Plant J. 90 (2017) 330–343. https://doi.org/10.1111/tpj.13496.

[30] G. De Los Campos, H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, J.M. Cotes, Predicting quantitative traits with regression models for dense molecular markers and pedigree, Genetics. 182 (2009) 375–385. https://doi.org/10.1534/genetics.109.101501.

[31] M.E. Goddard, B.J. Hayes, Mapping genes for complex traits in domestic animals and their use in breeding programmes, Nat. Rev. Genet. 10 (2009) 381–391. https://doi.org/10.1038/nrg2575.

[32] U.R. Rosyara, W.S. De Jong, D.S. Douches, J.B. Endelman, Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato, 9 (n.d.). https://doi.org/10.3835/plantgenome2015.08.0073.

[33] A. García-Dorado, J.L. Monedero, C. López-Fanjul, The mutation rate and the distribution of mutational effects of viability and fitness in Drosophila melanogaster., Genetica. 102–103 (1998) 255–65. http://www.ncbi.nlm.nih.gov/pubmed/9720284 (accessed May 3, 2018).

[34] A. Eyre-Walker, P.D. Keightley, The distribution of fitness effects of new mutations, Nat. Rev. Genet. 8 (2007) 610–618. https://doi.org/10.1038/nrg2146.

[35] A. Caballero, A. Tenesa, P.D. Keightley, The nature of genetic variation for complex traits revealed by GWAS and Regional Heritability Mapping analyses, Genetics. 201 (2015) 1601–1613.

[36] E. Raineri, L. Ferretti, A. Esteve-Codina, B. Nevado, S. Heath, M. Pérez-Enciso, SNP calling by sequencing pooled samples, BMC Bioinformatics. 13 (2012) 239. https://doi.org/10.1186/1471-2105-13-239.

[37] S. Naithani, C.M. Partipilo, R. Raja, J.L. Elser, P. Jaiswal, FragariaCyc: A Metabolic Pathway Database for Woodland Strawberry Fragaria vesca, Front. Plant Sci. 7 (2016) 1–10. https://doi.org/10.3389/fpls.2016.00242.

Chapter 4

Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species

Laura M. Zingaretti[1*], Salvador Alejandro Gezan[2], Luis Felipe V. Ferrão[3], Luis F. Osorio[4], Amparo Monfort[1,5], Patricio R. Muñoz[3], Vance M. Whitaker[4] and Miguel Pérez-Enciso[1,6*]

Affiliations

1. Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, Barcelona, Spain
2. School of Forest Resources and Conservation, University of Florida, Gainesville, FL, United States
3. Blueberry Breeding and Genomics Lab, Horticultural Sciences Department, University of Florida, Gainesville, FL, United States
4. IFAS Gulf Coast Research and Education Center, University of Florida, Wimauma, FL, United States
5. Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Barcelona, Spain
6. ICREA, Passeig de Lluís Companys 23, Barcelona, Spain

Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species

Correspondence:

Laura M. Zingaretti (laura.zingaretti@cragenomica.es)

Miguel Pérez-Enciso (miguel.perez@uab.es)

Front. Plant Sci. 11:25.doi: 10.3389/fpls.2020.00025
Published 2020, Feb 06

Keywords: Genomic Prediction, Genomic Selection, Polyploid Species, Deep Learning, Epistasis, Complex Traits, Strawberry, Blueberry

Abstract

Genomic Prediction (GP) is the procedure whereby the genetic merits of untested candidates are predicted using genome wide marker information. Although numerous examples of GP exist in plants and animals, applications to polyploid organisms are still scarce, partly due to limited genome resources and the complexity of this system. Deep Learning (DL) techniques comprise a heterogeneous collection of Machine Learning algorithms that have excelled at many prediction tasks. A potential advantage of DL for GP over standard linear model methods is that DL can potentially take into account all genetic interactions, including dominance and epistasis, which are expected to be of special relevance in most polyploids. In this study, we evaluated the predictive accuracy of linear and DL techniques in two important small fruits or berries: strawberry and blueberry. The two datasets contained a total of 1,358 allopolyploid strawberry (2n=8x=112) and 1,802 autopolyploid blueberry (2n=4x=48) individuals, genotyped for 9,908 and 73,045 SNP markers, respectively, and phenotyped for five agronomic traits each. DL depends on numerous parameters that influence performance and optimizing hyperparameter values can be a critical step. Here we show that interactions between hyperparameter combinations should be expected and that the number of convolutional filters and regularization in the first layers can have an important effect on model performance. In terms of genomic prediction, we did not find an advantage of DL over linear model methods, except when the epistasis component was important. Linear Bayesian models were better than Convolutional Neural Networks for the full additive architecture, whereas the opposite was observed under strong epistasis. However, by using a parameterization capable of taking into account these non-linear effects, Bayesian linear models can match or exceed the predictive accuracy of DL. A semiautomatic implementation of the DL pipeline is available at https://github.com/lauzingaretti/deepGP/.

4.1 Introduction

Deep Learning (DL) techniques comprise a heterogeneous collection of Machine Learning algorithms which have excelled at many prediction tasks, and this is a very active area of research [1–3]. All DL algorithms employ multiple neuron layers and numerous architectures have been proposed: Multiple Layer Perceptrons (MLPs), Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) [4] and others. DL is relatively straightforward to implement (https://keras.io/why-use-keras/) but optimum performance depends on an adequate hyperparameter choice, which is not trivial and requires considerable computational resources [5,6]. Although previous, limited evidence does not show a consistent advantage of DL over penalized linear methods for Genomic Prediction (GP) purposes [7–12], more efforts are needed to fully understand the behavior and potential constraints and capabilities of DL in GP scenarios.

Genomic Selection (GS) is the breeding strategy consisting in predicting complex traits using genomic-wide genetic markers. The idea was developed to overcome the limitations of Marker-Assisted Selection (MAS) and was formalized by Meuwissen *et al.* [13]. While MAS establishes a model with only the markers with significant associations, genomic selection includes all, or most available markers, for genomic prediction (GP), irrespective of their effect and its significance. Due to the decrease in genotyping costs, genomic selection is becoming the standard tool in many

plant and animal breeding programs [14–18]. There are an increasing number of successful applications of genomic selection in diploid and polyploid organisms where its use has generated important genetic gains by improving the accuracy of breeding value prediction and dramatically reducing generation intervals [19–23].

In any scenario, GP poses statistical challenges since the number of markers is usually much larger than the number of individuals, i.e., the so-called large *p* (number of features) small *n* (sample size) paradigm [24,25]. In this context, statistical methods require either shrinkage, variable selection or a combination of both [26]. Most GP methods are based on linear models, such as Genomic Best Linear Unbiased Prediction (GBLUP) [27], the Bayesian GP family [13,25] or LASSO [26]. In GBLUP, all marker effects are assumed to be normally distributed with equal variance and a homogeneous shrinkage is induced, whereas Bayesian models are more flexible and differential shrinkages and/or variable selection can be applied to distinct marker subsets. Note that these methods are linear and, in contrast to DL, have not been designed to model non-additive genetic effects (such as dominance or epistasis); however, these effects can be incorporated in the model with appropriate parameterizations.

One potential advantage of DL for GP over standard methods is that the whole genetic merit, including all non-additive effects, can potentially be predicted without the need to partition all effects. This is an interesting property for clonally propagated outcrossing species, because genomes can be asexually reproduced from single plants once the desirable individual is found. It should also be a promising strategy in polyploids, although their complex genetic structure has delayed the availability of whole genome markers and of specific analytic tools for, e.g. SNP calling [28–30]. A few studies have demonstrated the potential advantages of GS in allo and autopolyploids [29,31–35], although its implementation is still in its infancy.

When non-additive effects are investigated, there are two important points that need to be considered for higher ploidy levels: i) there is a portion of the intra-locus allele interaction (i.e., dominance) that is passed to the progeny (particularly full-sibs), and ii) the definition of non-additive effects is more complex than in diploids as higher order interaction exist [36]. Thus, methodologies that could model the whole genetic merit without restrictive assumptions could facilitate and improve the prediction for polyploid species, making DL an attractive choice for genomic prediction. In practice, DL aims at predicting the whole genetic merit, including interactions irrespective of their origin.

Among the polyploid species, strawberries (*Fragaria x ananassa*) and blueberries (*Vaccinium corymbosum*) are considered two of the most important soft fruit commodities. Considered a rich source of vitamins and minerals, fruit markets for both species have experienced a global increase in production and consumption over the past decade (https://www.nass.usda.gov/Publications/Todays_Reports/reports/ncit0619.pdf). To ensure that production and fruit quality meet the global demand, genetic improvement, and particularly GP, has a role to play in maximizing the utility, diversity and yield of resources. In this sense, previous experimental assessments performed in blueberry [32,35] and strawberry [29] have proven the feasibility of incorporating genomic selection to either accelerate the pace or improve the efficiency of breeding programs. From a genetic standpoint, one important difference between

both species is its inheritance pattern. Cultivated strawberry (*Fragaria x ananassa)* is an allo-octoploid hybrid plant originated by cross between two wild octoploid species *F. chiloensis* and *F.virginiana* [37] both descendants of *Fragaria* diploid species; referred as allopolyploids, meiosis is mainly dictated by preferential pairing, exhibiting a diploid-like (or disomic) segregation. In contrast, blueberry is a tetraploid organism originated from genome duplication within the same species. In autopolyploids, the meiotic pairing is mainly described by forming either random bivalents or multivalent during the division. Since the molecular mechanisms in auto and allopolyploids are quite complex, comparing new algorithms is a relevant issue to the prospect of GP in these and other polyploid species.

In this study, we evaluated the performance of deep learning for genomic prediction in two important horticultural species: allo-octoploid strawberry and auto-tetraploid blueberry. We complement the empirical study with simulations to understand better the impact of genetic architecture on DL performance. Given the complexity of implementing DL, we also provide a guideline on best practices for hyperparameter tuning and evaluate its importance in terms of predictive ability. To facilitate reproducibility of these methods, a python-based package for semiautomatic DL implementation, including auto and allopolyploid organisms have been made available at https://github.com/lauzingaretti/deepGS/.

4.2 Materials and methods

4.2.1 Plant Material and Genotypes

Predictive performances were compared in two polyploid species (blueberry and strawberry), for a series of traits with presumably contrasting genetic architecture. A summary of both experimental data sets is presented in the Table 4.1.

Table 4.1: Summary of blueberry and strawberry experimental data sets used in this paper.

	Strawberry (allopolyploid)	**Blueberry (autopolyploid)**
Ploidy	2n = 8x = 112	2n = 4x = 48
No. observations	1,358 (1,233 unique genotypes)	1,802
No. SNPs	9,908	73,045
Traits analyzed	• Soluble solid content (brix) • Average fruit weight (AveWtT) • Total marketable weight (MktWtT) • Early marketable yield (MktWtE) • Percentage of culled fruit (CullsTPer).	• Firmness • Fruit Size • Weight • Yield • Scar
Main reference	Gezan et al. (2017)	Amadeu et al. (2019) and Oliveira et al. (2019)

Regarding strawberry, we used 1,233 unique genotypes which correspond to five advanced selection trials (T2, T4, T6, T8 and T10) from the strawberry breeding program at the University of Florida, Institute of Food and Agricultural Sciences (USA). These advanced trials were planted in five consecutive seasons and were given an even code starting with season 2013-2014 as T2 and ending with season 2018-2019 as T10. The number of lines in each trial was 217, 240, 236, 272 and 393 for T2, T4, T6, T8 and T10, respectively. Some of the genotypes in the last trial T10 were

already tested in earlier trials, making the total number of observations sum up to 1358 (instead of 1,233). Plants were genotyped with the Axiom IStraw90 SNP array [38]. After quality control, in which those markers with minor allele frequencies (MAF) < 5% and with missing marker data > 5% were eliminated, 9,908 polymorphic SNP markers were available. A total of five yield and fruit quality traits were evaluated in each trial: soluble solid content (brix), average fruit weight (AveWtT), total marketable yield (MktWtT), early marketable yield (MktWtE) and percentage of culled (unmarketable) fruit (CullsTPer). Additional details for T2 and T4 can be found in Gezan *et al.* [29].

The blueberry population used in this study encompasses one cycle of the University of Florida blueberry breeding program's recurrent selection and comprised 1,802 lines from 117 full-sib families. The population was originated from 146 parents that presented superior phenotypic performance (cultivars and advanced stage of breeding). Individuals were evaluated for five yield and fruit quality-related traits: Firmness, Fruit Size, Weight, Yield, and Picking Scar, which were collected during two production seasons. Phenotypes were pre-corrected for fixed year effects, as detailed in [32,35]. A total of 73,045 SNPs was obtained using sequence capture by Rapid Genomics (Gainesville, FL), after aligning the reads against the high-quality "Draper" genome assembly [39] as described in Benevenuto *et al.* [40]. Marker filtering followed these criteria: biallelic, mean coverage > 40, minimum allele frequency > 0.01; maximum missing data = 0.5%; minimum quality = 20. Also, individuals with more than 50% missing data were removed, missing genotypes were simply imputed with the mean. Tetraploid genotypes were called and the allele dosages were inferred with the updog R package [41]. Standard genotype calling with updog allows inferring genotypes according to the number of allele copies, and genotypes can be coded say 0,1,2,3,4. In addition, as in [32], here we considered a set of 'diploidized' genotypes that were obtained pooling all heterozygous genotypes in a single class, i.e., genotypes above 0,1,2,3,4 can be recoded as 0,1,1,1,2. The rationale is that there can be incertitude on the exact number of allele copies in heterozygous genotypes.

The GP methods evaluated in this study were assessed by true validation, which was obtained by splitting data into a training and a validation dataset. In the strawberry dataset, we considered that predicting performance of the last stage lines (T10) is the most interest application for the industry and therefore the population was divided between training (T2, T4, T6 and T8 trials) and validation (T10) subsets with 965 and 393 lines, respectively. In the case of blueberry data, all samples were equally important and 30% of randomly sampled genotypes were assigned to the validation set. Predictive ability (PA) was defined as the correlation between observed and predicted phenotypes in the validation set; prediction was computed from parameters estimated in the training dataset only.

4.2.2 Genetic structure and heritability inference

Potential genetic structure was assessed by Principal Component Analysis (PCA) using all genotypes. Since genetic architecture may have an impact on GP performance and on the optimum GP model [42], additive and non-additive genetic features were assessed by computing variance components from the model:

$$\boldsymbol{y} = \boldsymbol{\mu 1} + \boldsymbol{a} + \boldsymbol{d} + \boldsymbol{e} + \boldsymbol{\varepsilon}, \qquad \text{[Eq. 4.1]}$$

where the vector $\boldsymbol{y}$ represents the adjusted phenotype, $\boldsymbol{\mu 1}$ is the intercept, $\boldsymbol{a} \sim N(\mathbf{0}, \boldsymbol{A}\sigma_a^2)$, $\boldsymbol{d} \sim N(\mathbf{0}, \mathbf{D}\boldsymbol{\sigma}_d^2)$, and $\boldsymbol{e} \sim N(\mathbf{0}, \mathbf{E}\sigma_e^2)$ are the additive, dominant and epistatic effects, respectively, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$ is the residual component. Matrices **A** and **D** were obtained using AGHmatrix package [43] for both strawberry (as diploid) and blueberry (autotetraploid) species. For diploids, **A** and **D** were computed using [27] and [44] methods, respectively. In fact, $\mathbf{A} = \frac{\mathbf{ZZ'}}{2\sum_j p_j(1-p_j)}$, where **Z** is the matrix that contains the centered individual genotype values and $\mathbf{D} = \frac{\mathbf{MM'}}{4\sum_j[p_j(1-p_j)]^2}$ is the dominance matrix, where the **M** elements are $-2p_j^2, 2p_j(1-p_j), -2(1-p_j)^2$ for genotypes 0, 1 and 2, respectively. In the case of ploidy = 4, **D** was obtained as in [28]. The additive x additive epistatic matrix (**E**) considered is the Hadamard product of additive effects, i.e. $\mathbf{A} \odot \mathbf{A}$ [45]. Posterior distributions of genetic parameters were obtained using Reproducing Kernel Hilbert Spaces (RKHS) regression with BGLR package [25]. The additive, dominance and epistatic ratios were estimated as: $\hat{h}_a^2 = s_a^2/(s_a^2 + s_d^2 + s_e^2 + s_\varepsilon^2)$, $\hat{h}_d^2 = s_d^2/(s_a^2 + s_d^2 + s_e^2 + s_\varepsilon^2)$ and $\hat{h}_e^2 = s_e^2/(s_a^2 + s_d^2 + s_e^2 + s_\varepsilon^2)$; where s_i^2 the i^{th} mean posterior estimates of σ^2 as in Eq. 4.1. We used both training and validation datasets combined in this stage, since this is purely a descriptive analysis and the values obtained are not employed in the later prediction stages.

4.2.3 Penalized Linear Methods

We compared the prediction performance of DL models with two well-established linear methods: Bayesian Lasso (BL, [13] and Bayesian Ridge Regression (BRR, [46]. In these models, the trait can be expressed as:

$$\boldsymbol{y} = \boldsymbol{\mu 1} + \boldsymbol{g} + \boldsymbol{\varepsilon}, \qquad \text{[Eq. 4.2]}$$

where $\boldsymbol{\mu 1}$ is the overall mean, $\boldsymbol{g} = \boldsymbol{X\beta}$, $\boldsymbol{X}$ is the genotypes' matrix and $\boldsymbol{\beta}$ is a vector of marker effects. In BRR, prior distributions of marker effects $\boldsymbol{\beta}$ are $N(\mathbf{0}, \boldsymbol{I}\sigma_\beta^2)$, whereas the prior distributions for $\boldsymbol{\beta}$ in BL have a Laplace distribution, i.e., $p(\boldsymbol{\beta}|\, \lambda, \sigma_\varepsilon^2) = \frac{\lambda}{2\sigma_\varepsilon^2}\exp\left(-|\boldsymbol{\beta}|\frac{\lambda}{\sigma_\varepsilon^2}\right)$. Note that the Laplace distribution does not remove markers so, contrary to its frequentist counterpart, BL is not a variable selection approach. Each model was fitted by using only phenotypes from the training subset. The models were run using the BGLR package [25] with a Gibbs sampler algorithm for a total of 6,000 cycles, discarding the first 1,000 samples for burn-in.

The above parameterization assumes additivity of effects, although linear models can address non-linear relationships if properly parameterized. Non-linear interactions can be modeled by expressing $\boldsymbol{g}$ (Eq. 4.2) in a general way, i.e., $\boldsymbol{g} = \boldsymbol{\Omega\omega}$ where $\boldsymbol{\Omega}$ (centered and scaled) is a matrix of dummy variables that indicates the number of copies of the reference allele ranging from 0 to the ploidy level [28,31]. This model is, in principle, a good parameterization to account for non-linear

interactions and we will refer to it as BRR general model (BRR-GM), since Bayes ridge regression was used. For more details, see [31,47].

Non-linearity can also be managed by means of RKHS regression [48] as an alternative to a linear regression for capturing complex interactions. This model considers $\boldsymbol{g}$ in Eq. 4.2 as $N(\mathbf{0}, \boldsymbol{K}\sigma_g^2)$ with $K(x_i, x_{i'}) = \exp\left(-\frac{h\|x_i - x_{i'}\|^2}{p}\right)$, a kernel function where h is de bandwidth parameter controlling how fast the covariance function drops with the distance between pairs of markers and $\|x_i - x_{i'}\|^2$ is the Euclidean distance between any two pairs of genotypes. This parameterization induces a general matrix of genetic covariance between markers. The key point here is that the kernel can model non-linear relationships because it is a non-linear transformation of the distances between the input variables. Empirical evidence confirms that it is an accurate approach to predict phenotypes of complex traits [49–51]. BRR-GM and RKHS were only implemented for strawberry and simulated scenarios, since it was in strawberry where we found the trait with the largest epistasis component, as described below.

4.2.4 Deep Learning (Convolutional Neural Networks)

Deep Learning (DL) has been described as a universal learning approach able to solve supervised, semi-supervised and unsupervised problems. Several DL architectures have been proposed, such as Multiple Layer Perceptrons (MLPs), Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs) and Reinforcement Learning (RL). The Figure 4.1 shows a generic pipeline to evaluate DL in a GP context.

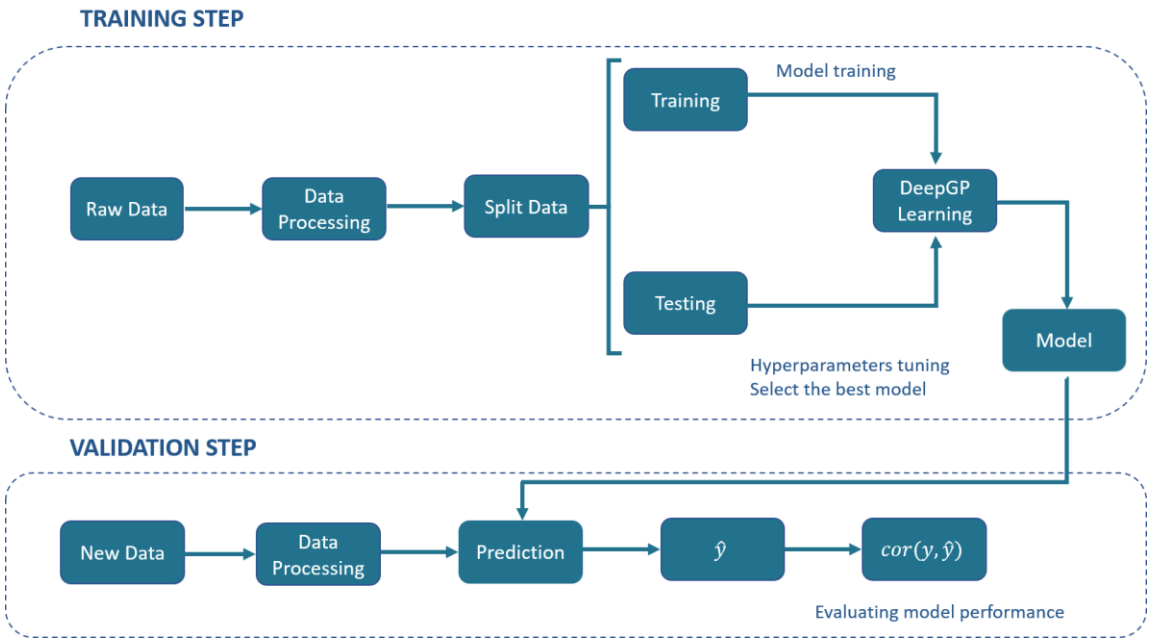


Figure 4.1. A generic Deep Learning (DL) pipeline for Genomic Prediction (GP) purposes. The general process includes the training and validation steps. In the training step, data are split into training and testing, DL hyperparameters are optimized by internal cross-validation with the test set and the model with the best Predictive Ability (PA) is chosen. In the validation step, the model PA is evaluated using a new set of data.

In our previous experiment ([52]), CNNs were the best performing methods and therefore are the only ones discussed here. The advantage of CNNs in a GP context is that they can model the correlation between adjacent input variables, that is, linkage disequilibrium between nearby SNPs. This is done via a mathematical operation called convolution [53]. A typical CNN is made up of 'convolutional layers', 'pooling', 'flatten' and 'dense' fully connected layers (Figure 4.2). In the 'convolutional layer', an operation called convolution is performed along the input of predefined width and strides, which are known as 'kernel' and 'filter' in the DL jargon, respectively. From a mathematical view, a convolution *s(t)* is a function that can be defined as an 'integral transform' [54]:

$$s(t) = (f * k)(t) = \int k(t-x)f(x)dx \qquad \text{[Eq. 4.3]}$$

where one of the functions (k or f) in (Eq. 4.3) must be a kernel. Assuming that the kernel is represented by k, the convolution is the transformation of f (input data in the DL context) into $s(t)$. The operation is just the weighted sum of an infinite number of copies f shifting over the kernel. The discrete version of Eq. 4.3 follows naturally as:

$$s(t) = (f * k)(t) = \sum_x k(t-x)f(x) \qquad \text{[Eq. 4.4]}$$

One of the main advantages of convolution networks is their capability to reduce the number of operations, i.e., the hyperparameters to be estimated. As usual, an activation function (generally non-linear) is applied after each convolution to produce the output layer. Finally, 'pooling' layers reduces dimension and achieves a smoother representation, summarizing adjacent neurons by computing their maximum or mean.

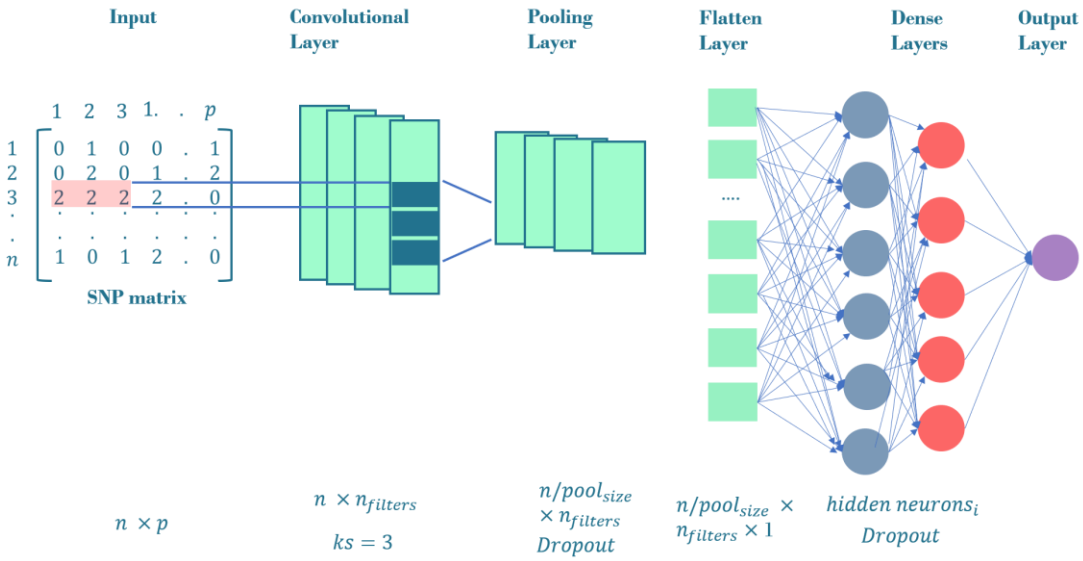


Figure 4.2 General CNN architecture employed in our workflow. The input layer is a SNP matrix of size n x p, where n is the number of training set and p, the number of SNPs. The convolutional layer consists on a $n_{filters}$ convolution followed by a max-pooling layer with $pool_{size}$ = 3 and an optional dropout. The outputs of max-pooling layer are joined together into one vector by flattening. All the neurons in the flatten layer are fully connected to the first dense layer. We tune the network using i dense layers with a variable number of hidden neurons in the respective hidden layers. The output of these dense layers is the prediction layer that uses linear function as activation. The neurons in convolutional and dense layers use relu, tanh or linear function as activations.

4.2.5 Hyperparameter optimization

Since DL depends on numerous parameters that influence performance, optimizing hyperparameter can be a critical unresolved step, which relies heavily on heuristics. Hence, it is surprising that many DL applications in GP have not paid enough attention to this problem [9][11,55,56]. Several approaches have been proposed for hyperparameter tuning [7,57–60]. Here, DL architectures were optimized using Talos (Autonomia Talos, 2019), which works combining all parameters in a grid. Talos can choose the best model either maximizing the predictive accuracy or minimizing the error; the former criterion was employed here. Since the approach can be expensive as the number of hyperparameters increases, a random search is the best strategy in practice[53][53][53][53][53][53][53][53][53]. This rule evolves a list of CNN models for each phenotypic trait. We optimized the following hyperparameters (values considered within parentheses): activation function (relu, tanh, linear), number of filters (16, 32, 64, 128), regularization (i.e., weight decay in DL terminology, 0, 0.1, 0.01, 0.001), learning rate (0.1, 0.01, 0.001, 0.0025), number of neurons in fully connected layer (4, 8, 12, 16), number of hidden layers (1,5,10), and dropout (0, 0.01, 0.1, 0.2).

Talos output is the accuracy for each hyperparameter combination; we then used hyperparameter values as independent variables and accuracy as target variable to run a random forest algorithm, which allowed us to compute the hyperparameter value importance, measured as the decrease in Gini's coefficient when adding the given hyperparameter. This hyperparameter importance can be then used as guide to improve interpretability. The R package randomForest [61] was employed for this analysis.

The DL algorithms used in this study were implemented in Keras [62] and Tensorflow [63] and run on a GPU equipped Linux workstation. A generic script is publicly available at https://github.com/lauzingaretti/deepGS/.

4.2.6 Simulation

We studied the impact of genetic architecture on prediction performance by simulation using the actual observed strawberry genotypes, assessing predictive performance with the same T10 strawberry genotypes (and genotypic data) as in the real experiment, except that phenotypic responses were simulated. Three contrasting genetic architectures were considered:

1- Additive: 200 randomly chosen SNPs were considered as causal loci. No dominance was simulated. Total individual genetic value was the sum of effects across loci.
2- Epistatic: 100 epistatic pairs of SNPs were randomly sampled. Epistasis was multiplicative by pairs, i.e., the genotype was the product of individual genotypes in each pair. Total genetic value was the sum of effects across pairs of loci.

3- Mixed: 80 individual additive SNPs and 60 epistatic SNP pairs were randomly chosen. Total genetic value was the sum of effects across pairs of loci and individual additive loci.

Allele substitution effects were sampled from a gamma distribution $\Gamma(\alpha = 1, \beta = 0.2)$ The trait was obtained adding the genetic value to an environmental normal residual. Environmental variance was chosen such that broad-sense heritability was set to 0.50. For each genetic architecture, five replicates were run. We compared BRR, BRR-GM, RKHS and DL. DL architectures were specifically optimized to each phenotypic trait, since no universal architecture is able to make accurate predictions for all cases.

4.3 Results

4.3.1 Population structure and genetic parameters

No clear population structure was observed, neither in the strawberry nor in the blueberry datasets (Figure S1). Note that genetic relationships between trials in strawberry data are rather uniform, irrespective of whether they are successive seasons or not. This, together with the fact that little genotype by environment (or year) interaction was observed [29], suggests a favorable scenario for GP.

Heritability estimates in strawberry are slightly different from those obtained in the same material by [29] since here we used additional data and we removed genotypes tested since here we used additional material and we removed genotypes tested more than once on different seasons. Nevertheless, in agreement with previous results [29,32] narrow-sense heritabilities were moderate, ranging from 0.25 to 0.35 for most strawberry (Figure 4.3) and blueberry (Figure 4.4) traits, except for strawberry average fruit weight ($h_a^2 = 0.58$). The degree of dominance found was quite low in general, especially in strawberry. An exception was blueberry yield, where dominant and epistatic variances were similar to the additive variance (Figure 4.4e). A remarkable case is percentage of culled fruit (CulsTPer) in strawberry, where the epistatic ratio (18%) was only slightly smaller than the additive one (25%, Figure 4.3e).

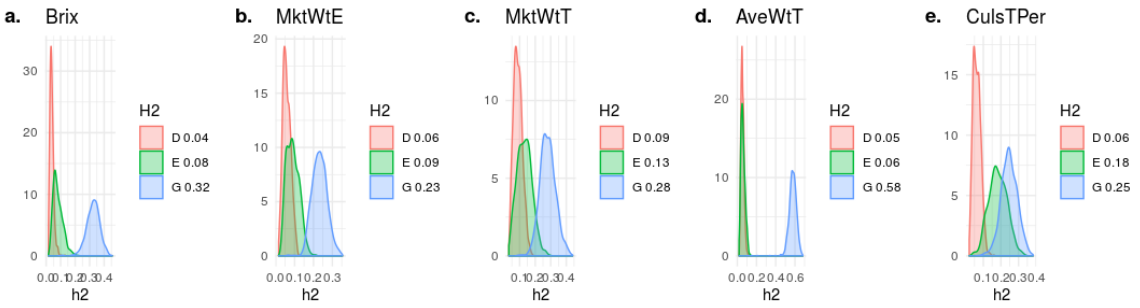


Figure 4.3 Posterior distributions of additive (blue), dominant (red), epistasis (green) fractions of variance in octoploid strawberry: (a) soluble solid content (brix); (b) early marketable yield (MktWtE); (c) total marketable yield (MktWtT); (d) average fruit weight (AveWtT); and (e) percentage of culled fruit (CullsTPer). Note the scale may vary along traits.

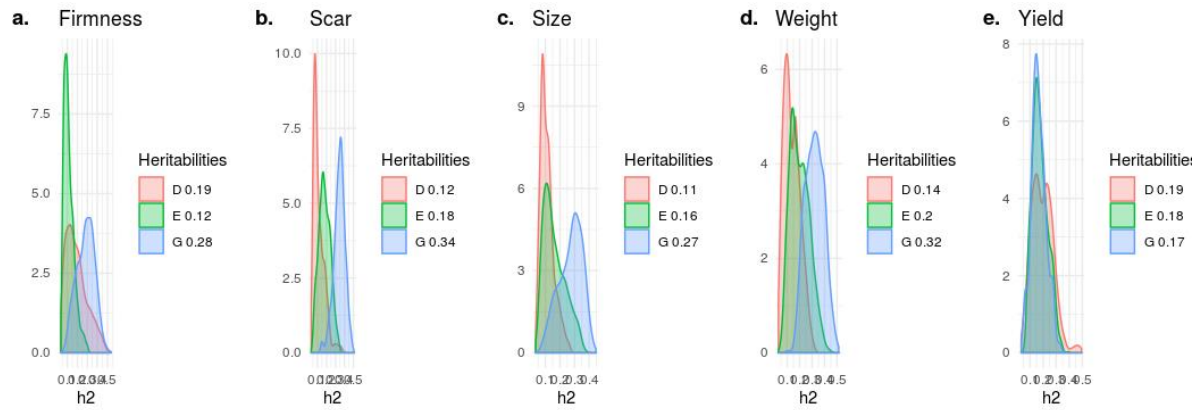


Figure 4.4 Posterior distributions of additive (blue), dominant (red), epistasis (green) fractions of variance in blueberry obtained with the tetraploid genotypes: (a) Firmness; (b) Scar; (c) Size; (d) Fruit Weight and (e) Yield. Note the scale may vary along traits.

4.3.2 Hyperparameter Importance

CNN hyperparameters were optimized for each strawberry trait separately. Figure 4.5a shows the importance of each hyperparameter obtained from random forest by regressing the model predictive accuracies (obtained by an inner cross-validation) on all hyperparameter values combinations. Interestingly, the number of filters was overall the most relevant factor, whereas other factors such as learning rate, whose importance has been claimed in the literature as critically important [64–66], played only a minor role. We also observed that the effect of each hyperparameter depends on the layer, e.g., regularization or dropout were more important in first than in deep layers.

In Figure 4.5a, the 'trait' effect was excluded since it cannot be controlled by the experimenter, although it was by far the most influential variable. This is illustrated in Figure 4.5b, which shows the distribution of accuracies for each trait studied. Not only maximum accuracies varied across trait, the profiles were also extremely different, usually multimodal. This suggests interactions between hyperparameter combinations, and it also indicates that trait–specific optimization should be performed whenever possible.

Figure 4.5 illustrates the kind of complex interactions that we observed in hyperparameter optimization. For instance, Figures 4.5c,e show the distinct influence of activation functions in percentage of culled fruit (Figure 4.5c) and brix (Figure 4.5e). Although 'relu' activation function has been suggested as the activation of choice in recent DL literature [65,67], here we observed that linear or even sigmoid-like hyperbolic tangent (tanh) seemed to be a safer choice overall. It is relevant to note that interactions were clearly observed for some hyperparameters, such as the number of filters. For CulsTPer, either 16 or 128 filters resulted in optimum accuracies, although they were also associated with the worst hyperparameter combinations. In contrast, either 32 or 64 filters are to be preferred for average weight in strawberry (Figure 4.5f).

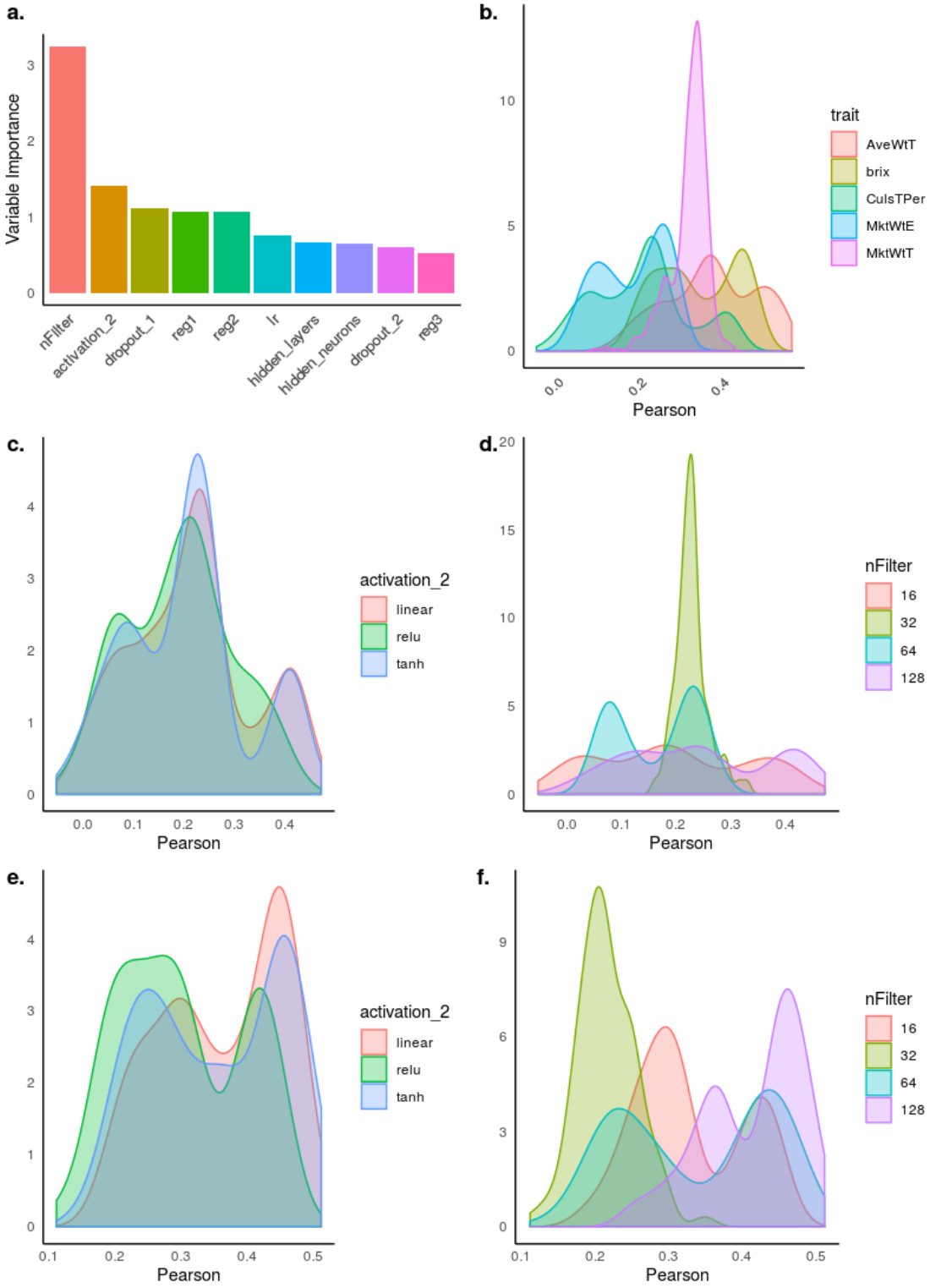


Figure 4.5 Hyperparameter influence on predictive accuracy in strawberry. Accuracy is defined as correlation between observed and predicted phenotypes by internal cross-validation. a) Hyperparameter importance obtained from a random forest algorithm. nFilter: number of filters in the convolutional layer, activation_2, activation function in layer 2; reg_i, regularization in i-th layer; dropout_i, dropout rate in i-th layer; lr, learning rate; hidden_neurons, number of neurons in inter-mediate layers; hidden_layers, number of intermediate layers. b) Distribution of accuracies along hyperparameter combinations for each phenotype. c) Accuracies as a function of activation function for percentage of culls. d) Accuracies as a function of number of filters for percentage of culls. e) Distribution of accuracies as a function of activation for brix. f) Distribution of accuracies as a function of number of filters for average fruit weight (AveWtT).

The final sets of hyperparameters for strawberry and blueberry phenotypes are indicated in Tables S1 and S2, respectively. Overall, our study shows that shallow architectures are more competitive than deep architectures in terms of PA, since the majority of models only included one CNN layer. The number of filters -in combination with dropout- has a large effect in the PA but is highly dependent of the trait. For instance, all optimal architectures for strawberry contain 128 convolutions but this is much more variable in the case of blueberry, with a range between 16 and 128 convolutional operations. As for the fully connected layers, the situation is less clear, and no obvious pattern is observed. We can highlight some characteristics though, for example, the number of hidden fully connected layers is quite variable but only a few neurons (4, 8, 12) are preferable in most of the architectures. As also reported in [68], combining weight decay and dropout regularization is an efficient option to increase PA. Finally, the best overlapping stride was 1 and optimum window size was 3 in the convolutional layer, confirming Bellot *et al.* [7] results.

4.3.3 Comparing Deep Learning with Bayesian Penalized Linear Models

Figure 4.6a shows observed predictive abilities for each of the five GP methods compared: BL, BRR, BRR-GM, RKHS, and CNNs in strawberry. When averaged over traits in the strawberry species, PAs were 0.43, 0.43, 0.44, 0.44 and 0.44 for each of the five methods, respectively. By trait, the BRR-GM was best in AveWtT prediction, BL, BRR and RKHS for MktWtE, RKHS and BRR-GM for MkWtT, whereas CNN performed best in brix and percentage of culled fruit. In all, nevertheless, there were no important differences between methods except in percentage of culled fruit. For this trait, CNN was ~20% better than any linear model method. Interestingly, this trait was also the one with the largest epistatic component and exhibited a modest additive component (Figure 4.3e).

As for the blueberry phenotypic traits, we found no differences between GP methods BL and BRR (average PA = 0.42), whereas CNNs were somewhat underperforming (average PA = 0.40). The most remarkable result in blueberry is that CNN performance was barely affected by the ploidy level employed to build the genetic relationship matrix. In fact, the 'diploid' option seemed more robust than the tetraploid one, except in fruit yield, the only trait that was measured using a rating scale.

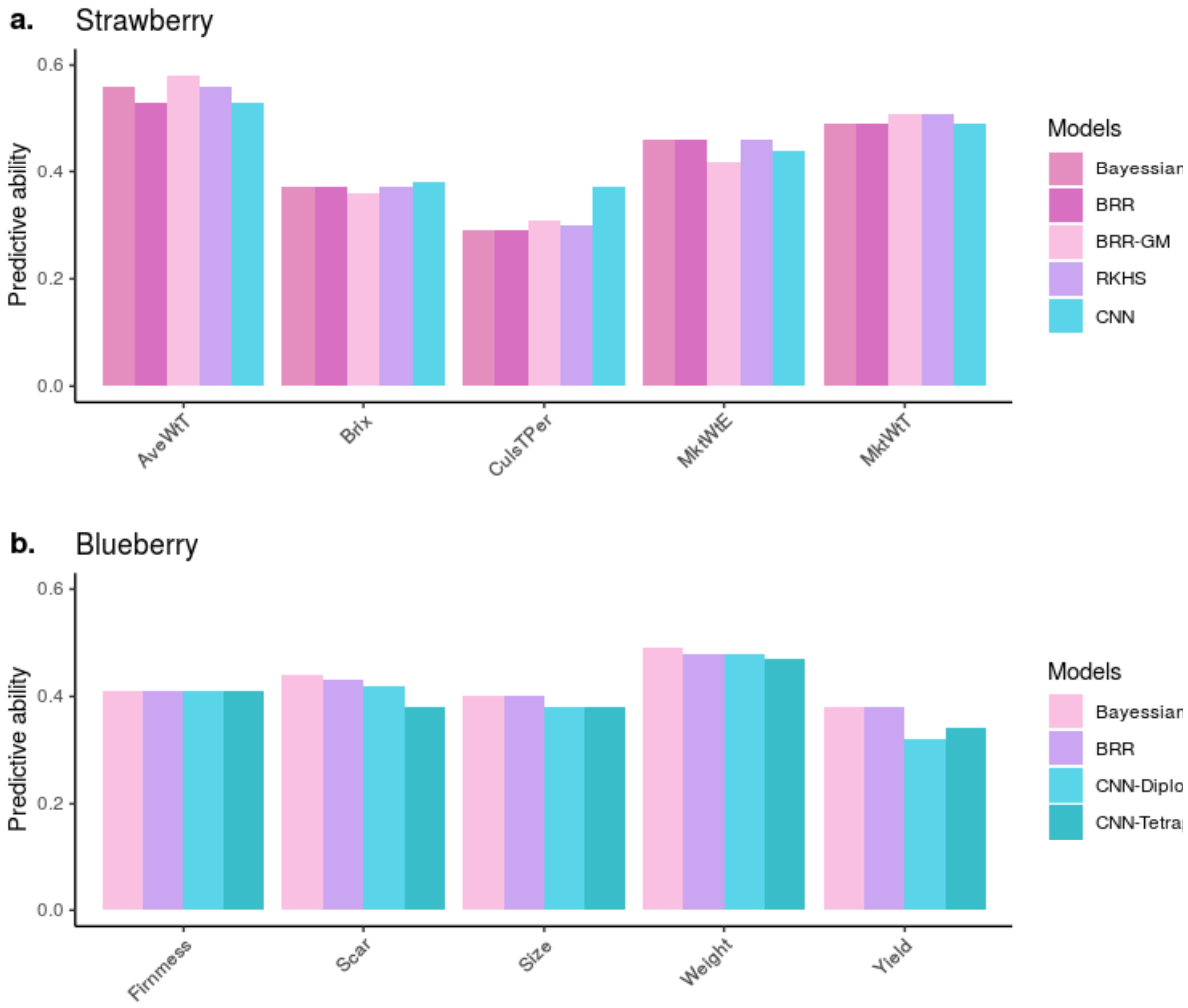


Figure 4.6 Predictive ability (PA) measured as correlation between observed and predicted phenotypes in the validation dataset in strawberry (a) and blueberry (b). Bayesian linear models (lasso and BRR) PAs in blueberry were computed with tetraploid genotypes, but were almost identical to those obtained with the diploidized ones.

4.3.4 Simulation study

Table 4.2 presents the main simulation results and Table S3, the CNN architectures used for computing the PA in each replicate. Some interesting remarks can be made from these simulations. First, although biased, the variance component estimates do detect whether epistasis is important: h_e^2 estimates are larger than the narrow-sense heritability in the presence of complete epistasis. Results are far less clear when only a fraction of loci show epistasis. But the most relevant result is that, as we hypothesized, predictive accuracies of CNN and additive penalized methods were affected by genetic architecture. BRR and RKHS were better than CNNs for the pure additive architecture, whereas the opposite was observed with pure epistasis. However, BRR-GM, which accounts for non-linear relationships, was better than either CNNs or pure additive linear models in most of the studied cases.

Table 4.2: Posterior distribution means of variance component estimates ($\hat{h}_2$) and predictive ability (in simulated data using Bayes Ridge Regression (BRR), general model BRR (BRR-GM), Reproducing Kernel Hilbert Space regression (RKHS) and Convolutional Neural Networks (CNN).

		Genetic parameter estimates			**Predictive ability (PA)**			
Replicate	**Architecture**	$\hat{h}_a^2$	$\hat{h}_d^2$	$\hat{h}_e^2$	**BRR**	**BRR-GM**	**RKHS**	**CNN**
1	Additive	0.29	0.16	0.06	0.57	0.60	0.57	0.59
2	Additive	0.16	0.21	0.06	0.35	0.43	0.35	0.32
3	Additive	0.26	0.25	0.05	0.52	0.58	0.51	0.51
4	Additive	0.24	0.23	0.06	0.42	0.52	0.43	0.40
5	Additive	0.35	0.11	0.05	0.42	0.47	0.43	0.38
6	Mixed	0.17	0.19	0.06	0.33	0.44	0.33	0.30
7	Mixed	0.10	0.11	0.08	0.24	0.26	0.22	0.24
8	Mixed	0.16	0.10	0.08	0.29	0.33	0.31	0.28
9	Mixed	0.13	0.16	0.06	0.26	0.30	0.26	0.25
10	Mixed	0.22	0.07	0.07	0.40	0.42	0.40	0.43
11	Epistatic	0.11	0.11	0.21	0.23	0.29	0.24	0.25
12	Epistatic	0.11	0.11	0.33	0.31	0.37	0.32	0.34
13	Epistatic	0.12	0.09	0.23	0.34	0.38	0.35	0.32
14	Epistatic	0.05	0.13	0.21	0.21	0.34	0.23	0.28
15	Epistatic	0.10	0.11	0.15	0.21	0.31	0.23	0.21

4.4 Discussion

Supervised DL methods are examples of predictive modelling, consisting of approximating a mapping function (*f*) from input (***X***) to output (***y***) variables [69]. These problems include classification or regression tasks, to use the Machine Learning jargon. Numerous successful applications of DL in classification contexts have been published, e.g. pattern recognition [70–73] and Natural Language Processing (NLP) [74]. The DL implementation in regression tasks is less abundant and the benefit of using these methods remains uncertain [7,12,75]. Most Genomic Prediction (GP) problems fall into the regression task due to the complex nature of quantitative traits [76]. So far, GP problems have been mainly addressed using penalized linear models [20,77]. More recently, the prediction of complex traits from genetic data is receiving attention from DL users [9,12,55]. The present work aim was to study the strengths and weaknesses of applying Convolutional Neural Networks (CNN) to GP problems in polyploid species. CNN networks are attractive for addressing these problems, as they can accommodate situations where input variables are distributed along a space pattern, as with the case of SNPs.

Implementing GS in polyploids is challenging. In allopolyploids, genetic analyses have been traditionally implemented assuming diploidy, taking advantage of the fact that systems present disomic inheritance. High predictive performances have been observed in a variety of

allopolyploid species (e.g. cotton, strawberry, wheat) and traits ([22,29,78]. Recently, the importance of accounting for the contribution of subgenomes – potentially expressing epistatic effects – was considered in wheat, which shed light on the importance of accounting for this source of variation within the GP models [79]. However, the scenario is even more complex in autopolyploid species. Even with the recent advances in genotyping and sequencing technologies, the amount of genomic information, and understanding, in most autopolyploid species is still limited when compared to allopolyploid crops. One of the challenges is resolving the allelic dosage of individual locus [80,81]. From a quantitative genetics standpoint, we emphasize that polyploid species might present higher degrees of complete and partial intra-locus interactions than diploids [82,83]. Here, the interest of investigating DL methods in polyploids is to take advantage of its non-linearity and less restrictive assumptions for GP in comparison to the traditional linear model-based methods.

Previous studies [7,9,12,55] have not shown clear advantages of DL over linear model GP, as conventional models were competitive in terms of prediction accuracy (PA), with their added benefit of being faster and with more biological interpretability. However, DL could be better suited to explore non-linear components than linear models, especially when genotypes can be transmitted integrally, as occurs with asexual propagation. Certainly, the weak performance of classical additive models in the presence of non-additive variance (e.g. Figure 4.6 for percentage of culled fruit) confirms the relevance of developing methodologies that can incorporate non-linearity [29,84]. This purpose can be attained by different approaches. The simplest one is to incorporate a general matrix into the linear models made up of dummy variables. This model contains as many degrees of freedom as ploidy level per locus and allowing for any interaction structure between alleles [31,35]. RKHS models [48,50,51] are also able to capture complex interaction patterns in a relatively straightforward manner. Alternatively, a CNN can be implemented using simply the raw data. Our analyses suggest that DL can perform better than additive and RKHS models for traits where the epistatic component is important and where narrow-sense heritability is low (e.g. percentage of culled fruit, Figure 4.6). The simulation study performed in this work (Table 2) suggested that BRR including additivity, dominance and the general dummy matrix described above can improve upon CNNs when the non-additive component is important, although CNNs were better than strict additive linear models. Additional analyses with a wider range of phenotypic traits, genetic structures and in larger datasets are needed to validate our results.

An underlying goal of our work was to investigate the effect of accounting for allele dosage in a GP context. Owing to the complex nature of polyploids, genotype calling can be a challenge and 'diploidization', i.e., considering a polyploid genome as diploid is usual [30]. Some studies have recently investigated the effect of accounting the ploidy level in prediction accuracy in polyploids [33–35,85,86]. As in these previous results [32,34], here we found that 'diploidization', in which all heterozygous genotypes are pooled, is as efficient and accurate as polyploid genotyping for prediction purposes, albeit it is trait dependent. Therefore, we conjecture that genomic selection, particularly for low levels of ploidy, can pay off in polyploids even with simplified genotyping strategies. We need to be careful though as this approach may not be equally appropriate for all levels of ploidy and heterozygosity. For instance, this might be an issue with sugarcane (with ploidy starting from 2n=20) as most individuals will be heterozygous.

It is traditionally thought that DL requires extremely large datasets to be trained effectively [71,87,88]. However, this and related works [7,9,12,55] have shown that DL performance in GP is comparable to those of linear methods. Furthermore, the largest dataset analyzed so far with DL for prediction (~100k individuals) did not show a consistent advantage of DL [7]. Therefore, it seems that is the trait what really influences the success of DL and it appears not so critical the size of the dataset. This does not preclude, of course, that a large N is needed to advance in our knowledge on best GP strategies. In fact, a larger N can be especially recommended in clonally propagated species. It is well known that an efficient breeding program tests a low number of crosses with a high number of genotypes in each of them. A cross would need to be tested if not much information is available though. Numerous clonally propagated species of agricultural interest are polyploids, leading to high heterozygosity, non-linear interactions and scarce prior knowledge about the crosses. In this scenario, as many cross-combinations as feasible should be produced to ensure the discovery and evaluation of the best genotypes [89]. The actual balance will depend on the level of epistasis and dominance. If dominance is large, then the best clone would be within families with good performance; if dominance is low, this is not necessarily so.

A drawback of DL models is that they lack biological or process interpretability and neither feature selection nor feature importance are obvious. In our opinion, GP algorithms are not too useful for providing biological insight into the genetic basis of phenotypes; genome wide association studies should be more appropriate. In all, our results suggest that DL performance improve as non-additive variance increases, a situation is usually encountered in fitness related traits.

DL hyperparameter tuning is critical and difficult, especially in terms of computational resources. Our analysis allows us to provide some generic recommendations though. First, we and others [7,12,55] concluded that the predictive accuracy is mainly dependent of the trait, i.e., the architecture needs to be tuned for each trait individually. Second, here we show that the popular relu activation function is not necessarily a universally valid activation function, that interactions between hyperparameter combinations should be expected and that the number of convolutional filters and regularization in the first layers can have an important effect into the model performance (Figure 4.5). In general, we and other authors [52,90] have reported that a shallow network is the best scenario in most cases. Nevertheless, DL can still be attractive because it does not require feature engineering, a critical step in most Machine Learning methods. A further strength of DL is its flexibility, e.g., it allows to define latent variables by using autoencoder or embedding as a generative latent variable model. In addition, networks, even if shallow, can model complex relationships employing any non-linear activation function.

Overall, there is no evidence that applying DL in GP applications necessarily improves the prediction accuracy upon that of classical linear model methods. PA depends on the trait and is affected by many factors; no one algorithm is uniformly better for all species and traits [91,92]. PA usually decays if heritability is low or in the presence of high epistatic effects. Even under these conditions though, Bayesian models were better than CNNs in almost all cases (Table S1, S2, Table 2). Even if performance of DL for GP is not outstanding, we cannot ignore that plant breeding is based on both genotyping and phenotyping, and that high throughput phenotyping is critical for genomic dissection of complex traits [93]. Imaging and computer vision can be employed to

measure the physiological, growth, development, and other phenotypic properties of plants with the advantage of being fast, non-invasive and a low-cost strategy [94], hyperspectral imaging is useful to measure plant traits under say disease progression [95], infrared thermography is able to scan temperature and transpiration; NMR (nuclear magnetic resonance spectroscopy) and mass spectrometry (MS) are applied in plants metabolite evaluation [96]. These examples should be an ideal scenario to Neural Networks as they involve imaging at high scale, complex and heterogeneous datasets with multiple variables and outcome. In summary, we believe that the enormous amount of data that can be automatically recorded revolutionizing plant breeding and the flexible nature of Neural Networks makes them promising for meeting this future challenge.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

MPE conceived and supervised research. VMW, PRM, LFO, SAG and LFF contributed experimental data. LMZ developed software and performed research. LMZ and MPE wrote the initial manuscript draft. All authors contributed to discussion and to writing the final draft.

Funding

LMZ was supported by a PhD grant from the Ministry of Economy and Science (MINECO, Spain), by the MINECO grant AGL2016-78709-R to MPE and from the EU through the BFU2016-77236-P (MINECO/AEI/FEDER, EU) and the "Centro de Excelencia Severo Ochoa 2016-2019" award SEV-2015-0533. VMW and LFO were supported by the US Department of Agriculture/National Institute of Food and Agriculture Specialty Crop Research Initiative (SCRI) project 'RosBREED: Combining disease resistance with horticultural quality in new rosaceous cultivars' under Award Number 2014-51181-22378.

Acknowledgments

The VMW lab acknowledges Dr. Sujeet Verma for curation of strawberry SNP data.

4.5 Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00025/full#supplementary-material

Data Availability Statement

The strawberry dataset analyzed for this study can be obtained from V.M. Whitaker (vwhitaker@ufl.edu); the blueberry dataset can be obtained from P.R. Muñoz (p.munoz@ufl.edu) or Dryad Digital Repository (accession number doi: 10.5061/dryad.kd4jq6h) and

https://gsajournals.figshare.com/articles/Supplemental_Material_for_de_Bem_Oliveira_et_al_2019/7728365.

4.6 References

[1] S. Min, B. Lee, S. Yoon, Deep learning in bioinformatics, Brief. Bioinform. 18 (2017) 851–869. https://doi.org/10.1093/bib/bbw068.

[2] S.T. Namin, M. Esmaeilzadeh, M. Najafi, T.B. Brown, J.O. Borevitz, Deep phenotyping: deep learning for temporal phenotype/genotype classification, Plant Methods. 14 (2018) 66. https://doi.org/10.1186/s13007-018-0333-4.

[3] S. Pattanayak, Unsupervised Learning with Restricted Boltzmann Machines and Auto-encoders, in: Pro Deep Learn. with TensorFlow, Apress, Berkeley, CA, 2017: pp. 279–343. https://doi.org/10.1007/978-1-4842-3096-1_5.

[4] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444. https://doi.org/10.1038/nature14539.

[5] S.R. Young, D.C. Rose, T.P. Karnowski, S.-H. Lim, R.M. Patton, Optimizing deep learning hyper-parameters through an evolutionary algorithm, in: Proc. Work. Mach. Learn. High-Performance Comput. Environ. - MLHPC '15, ACM Press, New York, New York, USA, 2015: pp. 1–5. https://doi.org/10.1145/2834892.2834896.

[6] M. Chan, D. Scarafoni, R. Duarte, J. Thornton, L. Skelly, Learning network architectures of deep CNNs under resource constraints, in: IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., 2018: pp. 1784–1791. https://doi.org/10.1109/CVPRW.2018.00222.

[7] P. Bellot, G. de Los Campos, M. Pérez-Enciso, Can Deep Learning Improve Genomic Prediction of Complex Human Traits?, Genetics. 210 (2018) 809–819. https://doi.org/10.1534/genetics.118.301298.

[8] O. González-Recio, G.J.M. Rosa, D. Gianola, Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits, Livest. Sci. 166 (2014) 217–231. https://doi.org/10.1016/j.livsci.2014.05.036.

[9] W. Ma, Z. Qiu, J. Song, Q. Cheng, C. Ma, DeepGS: Predicting phenotypes from genotypes using Deep Learning, BioRxiv. (2017) 241414. https://doi.org/10.1101/241414.

[10] O.A. Montesinos-López, J. Martín-Vallejo, J. Crossa, D. Gianola, C.M. Hernández-Suárez, A. Montesinos-López, P. Juliana, R. Singh, A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding, G3 Genes, Genomes, Genet. 9 (2019) 601–618. https://doi.org/10.1534/g3.118.200998.

[11] O.A. Montesinos-López, A. Montesinos-López, J. Crossa, D. Gianola, C.M. Hernández-Suárez, J. Martín-Vallejo, Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits, G3 Genes, Genomes, Genet. 8 (2018) 3829–3840. https://doi.org/10.1534/g3.118.200728.

[12] A. Montesinos-López, O.A. Montesinos-López, C.M. Hernández-Suárez, D. Gianola, J. Crossa, Multi-environment Genomic Prediction of Plant Traits Using Deep Learners with Dense Architecture, G3 Genes, Genomes, Genet. (2018) g3.200740.2018. https://doi.org/10.1534/g3.118.200740.

[13] T.H.E. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, Genetics. 157 (2001) 1819–1829. https://doi.org/11290733.

[14] R. Bernardo, Molecular markers and selection for complex traits in plants: Learning from the last 20 years, Crop Sci. 48 (2008) 1649–1664. https://doi.org/10.2135/cropsci2008.03.0131.

[15] J. Crossa, P. Pérez, J. Hickey, J. Burgueño, L. Ornella, J. Ceró N-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, D. Bonnett, K. Mathews, Genomic prediction in CIMMYT maize and wheat breeding programs, Heredity (Edinb). 112 (2013) 48–60. https://doi.org/10.1038/hdy.2013.16.

[16] J.M. González-Camacho, G. de los Campos, P. Pérez, D. Gianola, J.E. Cairns, G. Mahuku, R. Babu, J. Crossa, Genome-enabled prediction of genetic values using radial basis function neural networks, Theor. Appl. Genet. 125 (2012) 759–771. https://doi.org/10.1007/s00122-012-1868-9.

[17] T. Meuwissen, B. Hayes, M. Gddard, Accelerating Improvement of Livestock with Genomic Selection, Annu. Rev. Anim. Biosci. 1 (2013) 221–237. https://doi.org/10.1146/annurev-animal-031412-103705.

[18] G.R. Wiggans, J.B. Cole, S.M. Hubbard, T.S. Sonstegard, Genomic Selection in Dairy Cattle: The USDA Experience, Annu. Rev. Anim. Biosci. 5 (2017) 309–327. https://doi.org/10.1146/annurev-animal-021815-111422.

[19] H. Castillo-Juárez, G.R. Campos-Montes, A. Caballero-Zamora, H.H. Montaldo, Genetic improvement of Pacific white shrimp [Penaeus (Litopenaeus) vannamei]: perspectives for genomic selection, Front. Genet. 06 (2015) 93. https://doi.org/10.3389/fgene.2015.00093.

[20] J.J. Crossa, Y. Beyene, S. Kassa, P. Pérez, J.M. Hickey, C. Chen, G. de los Campos, J. Burgueño, V.S. Windhausen, E. Buckler, J.-L. Jannink, M.A. Lopez Cruz, R. Babu, Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing., G3 (Bethesda). 3 (2013) 1903–26. https://doi.org/10.1534/g3.113.008227.

[21] J. Duangjit, M. Causse, C. Sauvage, Efficiency of genomic selection for tomato fruit quality, Mol. Breed. 36 (2016). https://doi.org/10.1007/s11032-016-0453-3.

[22] P. Juliana, J. Poland, J. Huerta-Espino, S. Shrestha, J. Crossa, L. Crespo-Herrera, F.H. Toledo, V. Govindan, S. Mondal, U. Kumar, S. Bhavani, P.K. Singh, M.S. Randhawa, X. He, C. Guzman, S. Dreisigacker, M.N. Rouse, Y. Jin, P. Pérez-Rodríguez, O.A. Montesinos-López, D. Singh, M. Mokhlesur Rahman, F. Marza, R.P. Singh, Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics, Nat. Genet. 51 (2019) 1530–1539. https://doi.org/10.1038/s41588-019-0496-6.

[23] I. de B. Oliveira, M.F.R. Resende, L.F.V. Ferrão, R.R. Amadeu, J.B. Endelman, M. Kirst, P.R. Munoz, Genomic prediction of autotetraploids, G3 Genes, Genomes, Genet. 9 (2019) 1189–98.

[24] G. de los Campos, J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, M.P.L. Calus, Whole-genome regression and prediction methods applied to plant and animal breeding, Genetics. 193 (2013) 327–345. https://doi.org/10.1534/genetics.112.143313.

[25] P. Pérez, G. De Los Campos, Genome-wide regression and prediction with the BGLR statistical package, Genetics. 198 (2014) 483–495. https://doi.org/10.1534/genetics.114.164442.

[26] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, J. R. Stat. Soc. Ser. B. 58

(1996) 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[27] P.M. VanRaden, Efficient Methods to Compute Genomic Predictions, J. Dairy Sci. 91 (2008) 4414–4423. https://doi.org/10.3168/jds.2007-0980.

[28] A.T. Slater, N.O.I. Cogan, J.W. Forster, B.J. Hayes, H.D. Daetwyler, Improving genetic gain with genomic selection in autotetraploid potato, Plant Genome. 9 (2016).

[29] S.A. Gezan, L.F. Osorio, S. Verma, V.M. Whitaker, An experimental validation of genomic selection in octoploid strawberry, Hortic. Res. 4 (2017) 16070. https://doi.org/10.1038/hortres.2016.70.

[30] P.M. Bourke, R.E. Voorrips, R.G.F. Visser, C. Maliepaard, Tools for Genetic Studies in Experimental Populations of Polyploids, Front. Plant Sci. 9 (2018) 513. https://doi.org/10.3389/fpls.2018.00513.

[31] F. Enciso-Rodriguez, D. Douches, M. Lopez-Cruz, J. Coombs, G. de los Campos, Genomic selection for late blight and common scab resistance in tetraploid potato (Solanum tuberosum), G3 Genes, Genomes, Genet. 8 (2018) 2471–2481. https://doi.org/10.1534/g3.118.200273.

[32] I. de Bem Oliveira, M.F.R. Resende, L.F. V. Ferrão, R.R. Amadeu, J.B. Endelman, M. Kirst, A.S.G. Coelho, P.R. Muñoz, Genomic Prediction of Autotetraploids; Influence of Relationship Matrices, Allele Dosage, and Continuous Genotyping Calls in Phenotype Prediction, G3 Genes, Genomes, Genet. 9 (2019) g3.400059.2019. https://doi.org/10.1534/g3.119.400059.

[33] M. Nyine, B. Uwimana, N. Blavet, E. Hřibová, H. Vanrespaille, M. Batte, V. Akech, A. Brown, J. Lorenzen, R. Swennen, J. Doležel, Genomic prediction in a multiploid crop: Genotype by environment interaction and allele dosage effects on predictive ability in Banana, Plant Genome. 11 (2018). https://doi.org/10.3835/plantgenome2017.10.0090.

[34] M.L. Zingaretti, A. Monfort, M. Pérez-Enciso, pSBVB: A Versatile Simulation Tool To Evaluate Genomic Selection in Polyploid Species., G3 (Bethesda). 9 (2019) 327–334. https://doi.org/10.1534/g3.118.200942.

[35] R.R. Amadeu, L.F. V. Ferrão, I. de B. Oliveira, J. Benevenuto, J.B. Endelman, P.R. Munoz, Impact of Dominance Effects on Autotetraploid Genomic Prediction, Crop Sci. 0 (2019) 0. https://doi.org/10.2135/cropsci2019.02.0138.

[36] T.C. Osborn, J. Chris Pires, J.A. Birchler, D.L. Auger, Z.J. Chen, H.S. Lee, L. Comai, A. Madlung, R.W. Doerge, V. Colot, R.A. Martienssen, Understanding mechanisms of novel gene expression in polyploids, Trends Genet. 19 (2003) 141–147. https://doi.org/10.1016/S0168-9525(03)00015-5.

[37] G.A. Hancock, J. F., Sjulin, T. M., and Lobos, Strawberries, in: Temp. Fruit Crop Breed., Springer, Dordrecht., Wallingford, UK, 2008: pp. 393–437.

[38] N. V Bassil, T.M. Davis, H. Zhang, S. Ficklin, M. Mittmann, T. Webster, L. Mahoney, D. Wood, E.S. Alperin, U.R. Rosyara, H. Koehorst-vanc Putten, A. Monfort, D.J. Sargent, I. Amaya, B. Denoyes, L. Bianco, T. van Dijk, A. Pirani, A. Iezzoni, D. Main, C. Peace, Y. Yang, V. Whitaker, S. Verma, L. Bellon, F. Brew, R. Herrera, E. van de Weg, Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry Fragaria × ananassa, BMC Genomics. 16 (2015) 155. https://doi.org/10.1186/s12864-015-1310-1.

[39] M. Colle, C.P. Leisner, C.M. Wai, S. Ou, K.A. Bird, J. Wang, J.H. Wisecaver, A.E. Yocca, E.I. Alger, H. Tang, Z. Xiong, P. Callow, G. Ben-Zvi, A. Brodt, K. Baruch, T. Swale, L.

Shiue, G.Q. Song, K.L. Childs, A. Schilmiller, N. Vorsa, C. Robin Buell, R. Vanburen, N. Jiang, P.P. Edger, Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry, Gigascience. 8 (2019). https://doi.org/10.1093/gigascience/giz012.

[40] J. Benevenuto, L.F. V. Ferrão, R.R. Amadeu, P. Munoz, How can a high-quality genome assembly help plant breeders?, Gigascience. 8 (2019) 1–4. https://doi.org/10.1093/gigascience/giz068.

[41] D. Gerard, L.F.V. Ferrão, A.A.F. Garcia, M. Stephens, Genotyping polyploids from messy sequencing data, Genetics. 210 (2018) 789–807.

[42] H.D. Daetwyler, M.P.L. Calus, R. Pong-Wong, G. de los Campos, J.M. Hickey, Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking, Genetics. 193 (2013) 347–365. https://doi.org/10.1534/genetics.112.147983.

[43] R.R. Amadeu, C. Cellon, J.W. Olmstead, A.A.F. Garcia, M.F.R. Resende, P.R. Muñoz, AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: A blueberry example, Plant Genome. 9 (2016). https://doi.org/10.3835/plantgenome2016.01.0009.

[44] Z.G. Vitezica, L. Varona, A. Legarra, On the additive and dominant variance and covariance of individuals within the genomic selection scope, Genetics. 195 (2013) 1223–1230.

[45] C.R. Henderson, Best linear unbiased prediction of performance and breeding value, Biometrics. (1984) 172–192. papers3://publication/uuid/627506AA-ACB7-491A-B468-9A3B5C2A52EC.

[46] D. Gianola, Priors in whole-genome regression: The Bayesian alphabet returns, Genetics. 194 (2013) 573–596. https://doi.org/10.1534/genetics.113.151753.

[47] R.R. Amadeu, L.F. V. Ferrão, I. de B. Oliveira, J. Benevenuto, J.B. Endelman, P.R. Munoz, Impact of Dominance Effects on Autotetraploid Genomic Prediction, Crop Sci. 0 (2019) 0. https://doi.org/10.2135/cropsci2019.02.0138.

[48] D. Gianola, R.L. Fernando, A. Stella, Genomic-assisted prediction of genetic value with semiparametric procedures., Genetics. 173 (2006) 1761–76. https://doi.org/10.1534/genetics.105.049510.

[49] G. de Los Campos, D. Gianola, G.J. Rosa, K.A. Weigel, J. Crossa, Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods, Genet. Res. (Camb). 92 (2010) 295–308. https://doi.org/10.1017/S0016672310000285.

[50] G. de los Campos, D. Gianola, G.J.M. Rosa, Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation1, J. Anim. Sci. 87 (2009) 1883–1887. https://doi.org/10.2527/jas.2008-1259.

[51] D. Gianola, J.B.C.H.M. van Kaam, M.A. Toro, Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits., Genetics. 178 (2008) 2289–303. https://doi.org/10.1534/genetics.107.084285.

[52] P. Bellot, G. de los Campos, M. Pérez-Enciso, Can deep learning improve genomic prediction of complex human traits?, Genetics. 210 (2018) 809–819. https://doi.org/10.1534/genetics.118.301298.

[53] A. Goodfellow, I., Bengio, Y., Courville, Deep Learning, MIT Press Cambridge, 2016.

[54] D. V Widder, The Convolution Transform, Bull. Am. Math. Soc. 60 (1954) 444–456. https://doi.org/10.1090/S0002-9904-1954-09828-2.

[55] O.A. Montesinos-López, J. Martín-Vallejo, J. Crossa, D. Gianola, C.M. Hernández-Suárez, A. Montesinos-López, P. Juliana, R. Singh, New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes, G3 Genes, Genomes, Genet. 9 (2019) 1545–1556. https://doi.org/10.1534/g3.119.300585.

[56] W. Ma, Z. Qiu, J. Song, Q. Cheng, C. Ma, DeepGS: Predicting phenotypes from genotypes using Deep Learning, BioRxiv. (2017) 241414. https://doi.org/10.1101/241414.

[57] M. Cho, C. Hegde, Reducing the Search Space for Hyperparameter Optimization Using Group Sparsity, in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., Institute of Electrical and Electronics Engineers Inc., 2019: pp. 3627–3631. https://doi.org/10.1109/ICASSP.2019.8682434.

[58] S. Rajaraman, S. Jaeger, S.K. Antani, Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images, PeerJ. 7 (2019) e6977. https://doi.org/10.7717/peerj.6977.

[59] N.Q.K. Le, T.T. Huynh, E.K.Y. Yapp, H.Y. Yeh, Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles, Comput. Methods Programs Biomed. 177 (2019) 81–88. https://doi.org/10.1016/j.cmpb.2019.05.016.

[60] Y.J. Yoo, Hyperparameter optimization of deep neural network using univariate dynamic encoding algorithm for searches, Knowledge-Based Syst. 178 (2019) 74–83. https://doi.org/10.1016/j.knosys.2019.04.019.

[61] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News. 2 (2002) 18–22.

[62] F. Chollet, Keras: Deep Learning library for Theano and TensorFlow, Https://Keras.Io. 7 (2015) T1. url: https://keras. io/k.

[63] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, S. ... & Ghemawat, TensorFlow: Large-scale machine learning on heterogeneous systems, Softw. Available from Tensorflow. Org. 1 (2015). https://doi.org/10.1016/0076-6879(83)01039-3.

[64] V.S. Bawa, V. Kumar, Linearized sigmoidal activation: A novel activation function with tractable non-linear characteristics to boost representation capability, Expert Syst. Appl. 120 (2019) 346–356. https://doi.org/10.1016/J.ESWA.2018.11.042.

[65] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: ICML Work. Deep Learn. Audio, Speech Lang. Process., 2013. https://www.semanticscholar.org/paper/Rectifier-Nonlinearities-Improve-Neural-Network-Maas/367f2c63a6f6a10b3b64b8729d601e69337ee3cc (accessed October 11, 2019).

[66] J. Feng, S. Lu, Performance Analysis of Various Activation Functions in Artificial Neural Networks, in: J. Phys. Conf. Ser., 2019. https://doi.org/10.1088/1742-6596/1237/2/022030.

[67] F. Pouladi, H. Salehinejad, A.M. Gilani, Deep Recurrent Neural Networks for Sequential Phenotype Prediction in Genomics, ArXiv Prepr. ArXiv1511.02554. (2016). https://arxiv.org/pdf/1511.02554.pdf (accessed March 4, 2019).

[68] P. Waldmann, Approximate Bayesian neural networks in genomic prediction, Genet. Sel. Evol. 50 (2018) 70. https://doi.org/10.1186/s12711-018-0439-1.

[69] K. Hornik, M. Stinchcombe, H. White, Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, Neural Networks. 3 (1990) 551–

560. https://doi.org/10.1016/0893-6080(90)90005-6.

[70] B. Drayer, T. Brox, Training deformable object models for human detection based on alignment and clustering, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2014: pp. 406–420. https://doi.org/10.1007/978-3-319-10602-1_27.

[71] M. Liang, X. Hu, Recurrent convolutional neural network for object recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2015: pp. 3367–3375. https://doi.org/10.1109/CVPR.2015.7298958.

[72] A. Işin, C. Direkoğlu, M. Şah, Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods, in: Procedia Comput. Sci., Elsevier, 2016: pp. 317–324. https://doi.org/10.1016/j.procs.2016.09.407.

[73] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

[74] L. Deng, Y. Liu, Deep Learning in Natural Language, 2018. https://doi.org/10.1007/978-981-10-5209-5_11.

[75] C.B. Azodi, A. McCarren, M. Roantree, G. de los Campos, S.-H. Shiu, Benchmarking algorithms for genomic prediction of complex traits, BioRxiv. (2019) 614479. https://doi.org/10.1101/614479.

[76] T.F. MacKay, Q & A: Genetic analysis of quantitative traits, J. Biol. 8 (2009) 23. https://doi.org/10.1186/jbiol133.

[77] G. De Los Campos, H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, J.M. Cotes, Predicting quantitative traits with regression models for dense molecular markers and pedigree, Genetics. 182 (2009) 375–385. https://doi.org/10.1534/genetics.109.101501.

[78] W. Gapare, S. Liu, W. Conaty, Q.H. Zhu, V. Gillespie, D. Llewellyn, W. Stiller, I. Wilson, Historical datasets support genomic selection models for the prediction of cotton fiber quality phenotypes across multiple environments, G3 Genes, Genomes, Genet. 8 (2018) 1721–1732. https://doi.org/10.1534/g3.118.200140.

[79] N. Santantonio, J.L. Jannink, M. Sorrells, A low resolution epistasis mapping approach to identify chromosome arm interactions in allohexaploid wheat, G3 Genes, Genomes, Genet. 9 (2019) 675–684. https://doi.org/10.1534/g3.118.200646.

[80] P.M. Bourke, R.E. Voorrips, R.G.F. Visser, C. Maliepaard, Tools for Genetic Studies in Experimental Populations of Polyploids, Front. Plant Sci. 9 (2018) 513. https://doi.org/10.3389/fpls.2018.00513.

[81] D. Gerard, L.F.V. Ferrão, A.A.F. Garcia, M. Stephens, Genotyping Polyploids from Messy Sequencing Data, Genetics. 210 (2018) 789–807. https://doi.org/10.1534/genetics.118.301468.

[82] L.F. V. Ferrão, J. Benevenuto, I. de B. Oliveira, C. Cellon, J. Olmstead, M. Kirst, M.F.R. Resende, P. Munoz, Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a GWAS Context, Front. Ecol. Evol. 6 (2018). https://doi.org/10.3389/fevo.2018.00107.

[83] A. Gallais, Quantitative genetics and breeding methods in autopolyploids plants, 2003.

[84] U. Ober, W. Huang, M. Magwire, M. Schlather, H. Simianer, T.F.C. Mackay, Accounting for Genetic Architecture Improves Sequence Based Genomic Prediction for a Drosophila

Fitness Trait, PLoS One. 10 (2015) e0126880. https://doi.org/10.1371/journal.pone.0126880.

[85] J.B. Endelman, C.A.S. Carley, P.C. Bethke, J.J. Coombs, M.E. Clough, W.L. da Silva, W.S. De Jong, D.S. Douches, C.M. Frederick, K.G. Haynes, D.G. Holm, J.C. Miller, P.R. Muñoz, F.M. Navarro, R.G. Novy, J.P. Palta, G.A. Porter, K.T. Rak, V.R. Sathuvalli, A.L. Thompson, G.C. Yencho, Genetic variance partitioning and Genome-Wide prediction with allele dosage information in autotetraploid Potato, Genetics. 209 (2018) 77–87. https://doi.org/10.1534/genetics.118.300685.

[86] L.A.D.C. Lara, M.F. Santos, L. Jank, L. Chiari, M.M. De Vilela, R.R. Amadeu, J.P.R. Dos Santos, G.S. Da Pereira, Z.B. Zeng, A.A.F. Garcia, Genomic selection with allele dosage in Panicum maximum Jacq, G3 Genes, Genomes, Genet. 9 (2019) 2463–2475. https://doi.org/10.1534/g3.118.200986.

[87] B. Alipanahi, A. Delong, M.T. Weirauch, B.J. Frey, Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning, Nat. Biotechnol. 33 (2015) 831–838. https://doi.org/10.1038/nbt.3300.

[88] H.Y. Xiong, B. Alipanahi, L.J. Lee, H. Bretschneider, D. Merico, R.K.C. Yuen, Y. Hua, S. Gueroussov, H.S. Najafabadi, T.R. Hughes, Q. Morris, Y. Barash, A.R. Krainer, N. Jojic, S.W. Scherer, B.J. Blencowe, B.J. Frey, RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease., Science. 347 (2015) 1254806. https://doi.org/10.1126/science.1254806.

[89] W. Grüneberg, R. Mwanga, M. Andrade, J. Espinoza, Selection methods. Part 5: Breeding clonally propagated crops., Plant Breed. Farmer Particip. (2009) 275–322. http://www.cabdirect.org/abstracts/20103075062.html.

[90] P. Waldmann, Approximate Bayesian neural networks in genomic prediction, Genet. Sel. Evol. 2018 501. 50 (2018) 1–9. https://doi.org/10.1186/s12711-018-0439-1.

[91] Q. Hu, C.S. Greene, Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics, BioRxiv. (2018) 385534. https://doi.org/10.1101/385534.

[92] Pérez-Enciso, Zingaretti, A Guide on Deep Learning for Complex Trait Genomic Prediction, Genes (Basel). 10 (2019) 553. https://doi.org/10.3390/genes10070553.

[93] J.N. Cobb, G. DeClerck, A. Greenberg, R. Clark, S. McCouch, Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement, Theor. Appl. Genet. 126 (2013) 867–887. https://doi.org/10.1007/s00122-013-2066-0.

[94] N. Fahlgren, M.A. Gehan, I. Baxter, Lights, camera, action: high-throughput plant phenotyping is ready for a close-up, Curr. Opin. Plant Biol. 24 (2015) 93–99. https://doi.org/10.1016/J.PBI.2015.02.006.

[95] S. Bergsträsser, D. Fanourakis, S. Schmittgen, M. Cendrero-Mateo, M. Jansen, H. Scharr, U. Rascher, HyperART: non-invasive quantification of leaf traits using hyperspectral absorption-reflectance-transmittance imaging, Plant Methods. 11 (2015) 1. https://doi.org/10.1186/s13007-015-0043-0.

[96] J. Hong, L. Yang, D. Zhang, J. Shi, Plant metabolomics: An indispensable system biology tool for plant science, Int. J. Mol. Sci. 17 (2016). https://doi.org/10.3390/ijms17060767.

Chapter 5

Automatic fruit morphology phenome and genetic analysis: An application in the octoploid strawberry

Laura M. Zingaretti[1], Amparo Monfort[1,2], Miguel Pérez-Enciso[1,3]

Affiliations

1. Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Barcelona, Spain.
2. IRTA (Institut de Recerca i Tecnologia Agroalimentàries), Barcelona, Spain.
3. ICREA, Passeig de Lluís Companys 23, 08010 Barcelona, Spain.

E-mail addresses:

m.lau.zingaretti@gmail.com (LMZ)
amparo.monfort@irta.cat (AM)
miguel.perez@uab.es (MPE)

Correspondence:

Laura M. Zingaretti (m.lau.zingaretti@gmail.com)

Keywords: Fruit Geometric-Morphometrics, Heritability, Multivariate shape descriptors, Variational Autoencoders.

Submitted to Plant Phenomics (December 2020) with positive review

Abstract

Automatizing phenotype measurement will decisively contribute to increase plant breeding efficiency. Among phenotypes, morphological traits are relevant in many fruit breeding programs, as appearance influences consumer preference. Often, these traits are manually or semi-automatically obtained. Yet, fruit morphology evaluation can be enhanced using fully automatized procedures and digital images provide a cost- effective opportunity for this purpose. Here, we present an automatized pipeline for comprehensive phenomic and genetic analysis of morphology traits extracted from internal and external strawberry (*Fragaria x ananassa*) images. The pipeline segments, classifies and labels the images, extracts conformation features, including linear (area, perimeter, height, width, circularity, shape descriptor, ratio between height and width) and multivariate (Fourier Elliptical components and Generalized Procrustes) statistics. Internal color patterns are obtained using an autoencoder to smooth out the image. In addition, we develop a variational autoencoder to automatically detect the most likely number of underlying shapes. Bayesian modeling is employed to estimate both additive and dominance effects for all traits. As expected, conformational traits are clearly heritable. Interestingly, dominance variance is higher than the additive component for most of the traits. Overall, we show that fruit shape and color can be quickly and automatically evaluated and are moderately heritable. Although we study strawberry images, the algorithm can be applied to other fruits, as shown in the GitHub repository https://github.com/lauzingaretti/DeepAFS.

5.1 Introduction

Demographic pressure and climate change are two of the major challenges of the 21st century. The worldwide population continues growing exponentially and it is expected to reach ~9.8 x 10^9 in 2050 [1]. Climate change generated by greenhouse gas emissions is possibly the greatest threat, as it is leading to extreme weather conditions, increasing areas of drought, and species extinction, among others [2–4]. In this adverse context, food production needs to be increased significantly. Increasing food production is not enough though. Breeding programs should also consider food safety and environmental care among their objectives [5,6].

Artificial breeding is a main responsible for the dramatic rise in food production we have witnessed for over a century. The main goal of plant and animal breeding is to utilize genetic variability of complex traits to increase performance and optimize use of resources. A current bottleneck in plant breeding programs is the evaluation of hundreds of lines under different environmental conditions [7,8]. Plant breeding involves both genomics and phenomics, i.e., the expression of a genome in given environments. While available technologies can routinely and inexpensively scan the genome, high-throughput phenotypic characterization remains a difficult task [9,10]. Automatizing phenotype measurement is then needed to increase the pace of artificial selection and is, unsurprisingly, one of the main targets of 'Precision Agriculture' [11,12].

The term 'phenomics' or 'phenometrics' was coined by Schork [13] as an attempt to understand events happening in between full genome and clinical endpoints phenotypes in complex human diseases. The expression quickly spread to animal and plant breeding research as a concept that bridges the gap between genotypes and the 'end-phenotypes'. Although the term phenomics was devised in line with 'genomics', that is, to describe the whole phenome of any organism, note the phenome varies over time and between cells or tissues, and can never be fully portrayed [14].

Although electronics applied to agriculture has a long history, a window of opportunities has emerged in the phenomics field with recent improvements in robotics, electronics, and computer science. The subjective, time- consuming and often destructive human data collection is being replaced by miniaturized, cheap sensors, digital cameras, cell phones, unmanned aerial vehicles, mass spectrometry, among others, that allow collecting hundreds of phenotype data objectively and inexpensively [9,15–17]. The challenge now is to develop new and improved analytical tools, capable of transforming this wealth of data into valuable knowledge [15]. This is a rapidly evolving field, and numerous software and pipelines to automatize phenotype collection are already available [18–22]. Many of these tools focus on the analysis of root images and, as far as we know, require more user intervention than we propose, making it impractical to analyze hundreds of images.

Digital images are among the cheapest and most widely available type of data. Imaging allows assessing morphological traits, which are highly relevant in numerous plant breeding schemes, since they can critically affect consumer acceptance especially in fruits [23–25]. Nevertheless, consumer preferences on appearance traits differ around the world and between communities. Like most traits, fruit shape is determined by genetic and environmental factors such as flower morphology or insect- mediated pollination [26,27]. In all, morphological traits are among those with the highest heritability, which has allowed breeders to rapidly modify shape, size, and color patterns of agricultural products [20,28–30].

Although numerous works have been developed in the area of fruit morphology, most of them have focused in the inheritance of linear measures, e.g., diameter, perimeter, circularity [20,31–33]. By definition, however, morphological traits are highly dimensional. Computing only linear, univariate phenotype leads to a loss of information by extremely simplifying the features of a shape [28,34]. The use of geometric-morphometric approaches for shape analysis is warranted [35]. Further, fruit shape has been traditionally evaluated subjectively [36], but can be enhanced by resorting to automatized procedures. For instance, hundreds of fruit pictures can be routinely and inexpensively collected, even in the field, with a cell phone camera. Automatized image processing and analysis can then dramatically change the way shape and color traits are collected and characterized.

Here, we present a comprehensive phenomic and genetic analysis pipeline for fruit morphology automatic analysis. Two main issues are addressed: 1) converting the raw data (fruit images) into a processed curated database, and 2) designing an efficient analysis workflow to analyze the fruit shape and color phenome. Finally, genetic parameters are automatically inferred from pedigree information. We apply the pipeline to images of cultivated strawberry (*Fragaria x ananassa*) fruits.

In addition to previous similar works in strawberry, e.g., Feldman et al. [18], we provide a wholly automatized pipeline and new tools to analyze shape and color patterns.

5.2 Materials and Methods

5.2.1 Plant material and imaging acquisition

Lines employed are part of the strawberry breeding program of PLANASA company (https://planasa.com/en/) and are routinely used to develop new elite genotypes. The experiment consisted of 24 crosses between 30 parental lines of *F. x ananassa*. We evaluated 20 randomly chosen lines per cross for all but 2 crosses, for which we chose 19 lines at random. A total of 478 seedlings and 30 parentals genotypes were evaluated (Supp. Table 1). Shape varied between the cultivars studied, e.g., circular, ellipsoid or rhomboid and color ranged from white to dark red.

Strawberries were grown in plastic semi-tunnel using standard cultivation practices in South West Spain (Huelva, 37° 16' 59" N, 7° 9' 18" W). Fruits were collected from two individual plants of each line at the end of April 2018 in only one harvest event. Fruits from both plants were pooled in the photographs. We took images of 1 to 7 sliced fruits per genotype using a Nikon D80 digital camera. Samples were laid on a black surface, with the camera positioned at 35 cm height. The focal length was 18 mm, the manual aperture was f/8, and the exposure time was 1/8 second. Illumination consisted of two white light sources at both sides of the camera. In total, we took 508 images of 3872×2592 pixels that contained all external and internal side of fruits and the label for each genotype in the same image.

5.2.2 Preprocessing and segmentation

The first step in the pipeline is to segment and recognize the objects, since each raw image contains internal and external fruits, a rule, a coin and a printed genotype (the strawberry line) label. Image segmentation is needed for obtaining meaningful morphometric and color information. However, most of available technologies to determine the boundaries of the objects at the pixel level are usually semi-automatic and time-consuming [37–40]. Our fully automatic python-based pipeline takes the images of each strawberry line and outputs a curated database of square images (1000 px) and reads the genotype label (Fig. 5.1). The https://github.com/lauzingaretti/DeepAFS/blob/main/main.ipynb explains how to apply the most expensive part of this workflow to alternative experiments. Note that after creating a curated database, standard multivariate analysis can be easily run using R/Python tools to shapes evaluation.

For segmentation, the three-channel digital signals (RGB/BGR) are converted into grayscale and blurred using Gaussian filtering of size 5, to remove undesirable noise. The histogram information is used for image binarization, i.e., splitting the background and foreground. Here, we binarized the image using simply the mean value of the pixel as a threshold. The pipeline also allows Otsu thresholding [41], which is designed to automatically define the threshold by minimizing the "overlap" between two classes. After binarization, we performed erosion and dilation, the former

shrinks the edges and the latter makes the image region grow. Finally, the algorithm extracts the Regions of Interests (ROI) and determines whether it is a strawberry or an image label. The color pattern analysis allows us to classify the internal or external part if a fruit image. We here apply a k-means clustering based on the information about the color mean, color standard deviation and the ration between them for all the fruits, i.e. we compute these 3 features for all the fruits and then we classify these observations into 2 clusters to split the internal and external part of the fruits. For the labels, the Optical Character Recognition (OCR) algorithm from PyTesseract library (https://pypi.org/ project/pytesseract/) is used to read the genotype name and automatically label the image into the database. As a result, the algorithm delivers a curated database of 508 folders labeled with the name of each genotype, and subfolders containing either the internal or external strawberry pictures (Fig. 5.1, Algorithm 1 in Suppl. Info). All fruits are stored in square images (1000 px size or user-defined), with the fruits placed in the center and filled with black pixels.

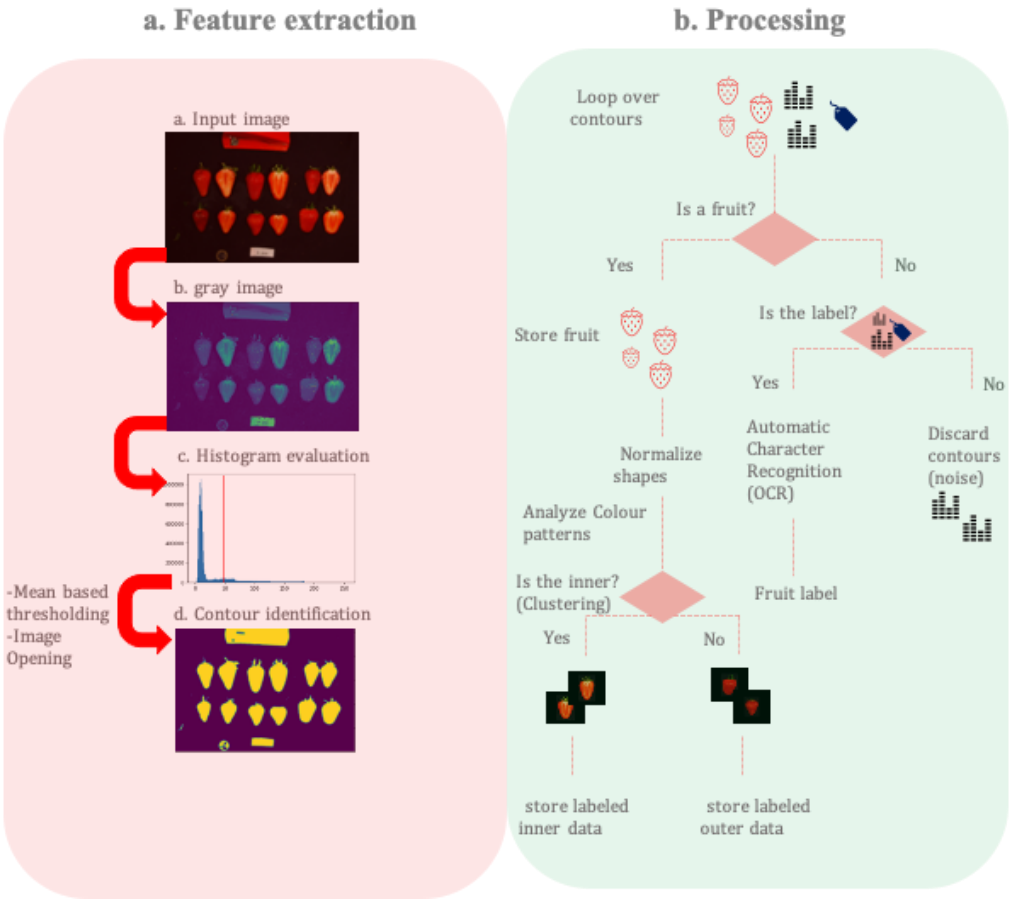


Figure 5.1. Workflow for automatic segmentation and label recognition from strawberry images. (a) Feature extraction. (b) Feature preprocessing and database generation.

5.2.3 Automatic fruit phenotyping

Once masks for either internal or external fruit images are obtained, an automatic phenotyping procedure is run for inside or outside parts separately (Fig. 5.2). Classical linear descriptors, multivariate and deep learning techniques are combined from a novel perspective to dissect a variety shape and color patterns. If pedigree or marker information is available, a genetic analysis can be employed to estimate variance components for each of the fruit phenotypes. In the following, we describe the main methods implemented in the pipeline of Fig. 5.2.

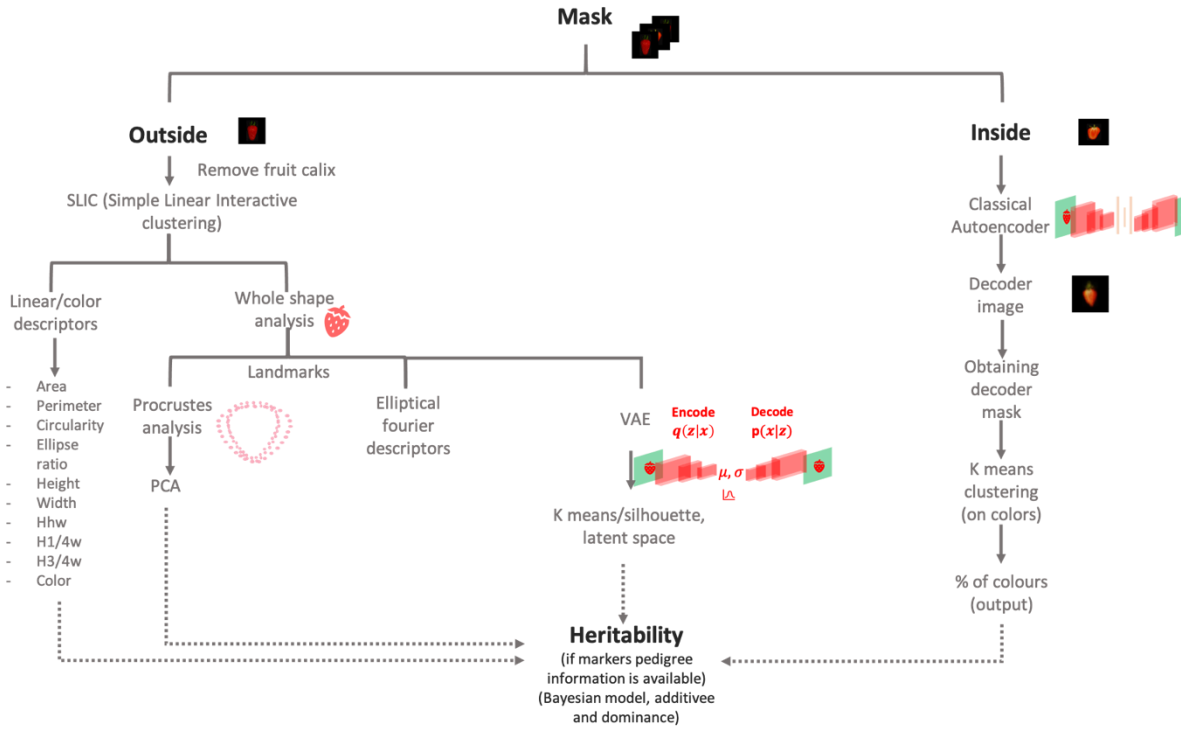


Figure 5.2. Data analysis workflow. The input is the segmented internal and external fruit images from workflow in Fig. 1. External images are used for linear and multi-dimensional shape analysis through different standard and machine learning approaches, including deep learning. Internal images are used to estimate the color pattern of the internal fruit section. Additive and dominance genetic components of each of the extracted morphometric and color phenotypes are estimated using Bayesian Linear Modeling using either pedigree or DNA marker information. Code available at https://github.com/lauzingaretti/DeepAFS

5.2.4 Autoencoder and k-means to infer internal color patterns

We used an 'autoencoder' (AE) network to perform an unsupervised clustering of the internal images. An autoencoder (Fig. 5.3a) is an unsupervised machine learning technique that applies backpropagation to train a neural network where the outputs are the same values as the inputs [42]. The AE gives new insight into image analysis by learning the structure about the data, i.e., it is not designed to copy an exact replicate of the input but instead to learn the repeatable and most useful properties.

We used a convolutional AE, as convolutional operations are especially suited for image analysis [42,43]. These layers create a feature map from the input image, preserving the relationships between pixels in the original space (Fig. 5.3a). Each convolution outputs a scored- filtered image, where a high score means a perfect match between the original and filtered image. The output layer is obtained by applying the Rectified Linear Unit (ReLU) activation function. Finally, as usual in any convolutional architecture, a max- pooling layer shrinks the output size and achieves a smoother representation, summarizing adjacent neuron outputs by computing their maximum (see accompanying GitHub).

The decoded images from an AE architecture are less noisy than the original ones, making it easier to detect repeatable/consistent color patterns. Our approach consists in taking five colors as reference: a class for the background (black) and four classes for the internal fruit color patterns,

including calyx. The four 'reference classes' were: "orange-like" (198, 99, 35, in RGB coordinates), "quasi-red" (184, 46, 8), "pale" (194, 144, 78), and "green" (76, 75, 20) for sepals. We then perform a k-means clustering with k=4 after removing the background and we assigned each cluster to the nearest reference color using the Euclidean distance between the average color of each cluster in the sample and the reference coordinates. As a result of this step, the surface of each of 1900 strawberry images is split into four categories of colors.

5.2.5 Superpixel algorithm to remove the calyx

Some of the fruit pictures contains sepals that interfere with fruit shape quantification and need to be removed prior to estimating shape parameters. For that purpose, we applied the Simple Linear Interactive Clustering (SLIC) algorithm [44] from the Python skimage library. SLIC is based on the 'superpixel' concept. Basically, a superpixel is a group of pixels sharing perceptual and semantic information, e.g., the pixels in a superpixel are grouped together because of their color or texture features. The iterative algorithm starts with regularly spaced K-centers at a given distance, user defined as S, which are then relocated in the direction of the lowest gradient in a 3x3 neighborhood window to avoid being at the edges of the image. Further, a pixel is assigned to a given cluster if its distance to the cluster's center is smaller than the distance to the other centers in the search area, as determined by S. Finally, the centers are recalculated by averaging all the pixels belonging to the superpixel. The iterative process ends when the residual error (distance between previous centers and recomputed ones) does not exceed a fixed threshold. SLIC outputs a set of meaningful clusters, splitting the background, the calyx and the fruit. Knowing that all our fruits are centered in the image (the segmentation procedure outlined in Fig. 5.1 ensured that every image was centered), the superpixel containing the central pixel matches with the fruit.

5.2.6 Univariate phenotypes: linear descriptors

Numerous object shape descriptors exist in the literature. Particularly for fruits, a controlled vocabulary was established in Brewer et al. [20]. Here, we implement a custom script to compute some standard linear measures: circularity, solidity, shape aspect [32], ellipse ratio [20], fruit perimeter and area, fruit width at 25% height, fruit width at 75% height, fruit width at 50% of height, total height and maximum width. Circularity is a measure of the degree of roundness of a given object, defined as the ratio between the area of a given object and the area of a circle with the same convex perimeter, i.e., a value near one means a "globe" o "circular" shape. Solidity is the ratio between the area of the object and the area of the convex hull of a given shape. Most of the linear descriptors used here are standard in fruit shape analyses [18,20,32,39,45].

The external fruits color was measured using the CIELAB space, where L indicates the luminosity, and a and b are the chromatic coordinates. The variation on the index a indicates the transition between green to red, where a higher value means a redder object. Variations in b reflect the change between yellow and blue color, i.e., a higher b value refers to a 'bluer' object

5.2.7 Generalized Procrustes Analysis (GPA)

Shape is usually defined as all the geometric information that remains unchanged after filtering out the location, scale and rotation effects of a given object [28]. The above shape linear descriptors are standard in the literature but do not provide a whole shape portrayal. Alternatively to linear descriptors, shape variations can be described using 'pseudo-landmarks' [35], which identify points around the outline of the object. Here 50 pseudo-landmarks were defined as the intersection between 50 equally spaced conceptual lines starting from the centroid and the fruit contour (Fig. 5.4a). Next we performed a Procrustes analysis [46]. The Procrustes analysis aims at finding the transformation T such that given two matrices X_1, X_2, the product X_2T best matches X_1. The Generalized Procrustes Analysis (GPA) is an extension of the method devised to align many matrices simultaneously [46]. In morphometric analysis, this is done by averaging the distance between all the landmarks on a target shape and the corresponding points on a reference. The pseudo-landmarks of the samples can then be analyzed as a multivariate object using, for instance, a principal component analysis (PCA). In addition, the pseudo-landmark variability gives insight on the most important regions that determine the differences between shapes. We used the Momocs [47] and geomorph [48] R packages to run these analyses.

5.2.8 Elliptical Fourier Descriptors

An alternative approach to morphometric analysis is Elliptical Fourier transformation [49]. This method describes a closed curve as a sum of sine and cosine functions of growing frequencies. As its name suggests, Fourier harmonics are ellipses, and a larger number of harmonic means that more ellipses are fitted to a given contour. The second-order harmonic is simply one ellipse with the values of sine and cosine components for the x and y-axis, respectively. As the strawberry fruit is a relatively simple shape, four harmonics were enough to describe all the shapes in the database, giving a total of 16 coefficients. A PCA of the Fourier components can also be employed to quantify morphometric variability, as in procrustes analysis. Geomorph [48] R package was employed for this purpose.

5.2.9 Conditional Variational Autoencoders (VAE) to cluster shapes

Fruit shape can also be addressed from a completely different angle, such as obtaining clusters of shapes to objectively classify fruits in groups of similar morphology [18]. A standard approach consists of flattening the image and grouping the raw data, treating each pixel as a feature. Unfortunately, clustering algorithms are not exempt from the "curse of the dimensionality" problem [50] and they perform poorly as the number of analyzed dimensions increases, especially if noise is high.

A natural way to solve the aforementioned issue is to apply a dimension reduction technique before clustering. Although the classical autoencoders seem to be a good option, as shown above, AEs were conceived to perform a non-linear and not isometric dimensionality reduction, and thus they do not preserve the geometrical properties of the original space [51]. Unlike traditional autoencoders, variational autoencoders (VAEs) [52–54] preserve distances and, importantly, are generative models (Fig. 5.3b). The main difference between AE and VAE is that the latter encodes

the input as a distribution over a latent space. Basically, given an input x, VAE creates a latent distribution $p(z|x)$ and the input reconstruction $d(z)$ is obtained after sampling z from the latent representation $z \sim p(z|x)$. The VAE does not only force the latent space to be continuous, it can also generate meaningful information, even with images that it has never been seen before.

The key aspect in VAE training lies in the loss function, which includes a "reconstruction" and a "regularization" term. The former is the usual loss or the joint log-likelihood between the true and the VAE output, whereas the second is the entropy corresponding to the Kulback-Leibler divergence [42] between the latent distribution $N(\mu_x, \sigma_x)$ and the standard normal distribution $N(0,1)$. Without incorporating a regularization, the VAE behaves as AE, where the latent space is neither complete nor continuous. Regularization forces the latent distribution to be close to the normal standard, generating a continuous space of low variance centered in the origin, which is suitable for data clustering and generation [42].

Here, we run standard k-means clustering of the latent space, with k varying between 2 and 9 groups. We chose a maximum k=9 given that up to nine strawberry shapes have been proposed in the literature, in particular in the Japanese market [55]. We assessed the cluster robustness using the silhouette index [56]. This index determines how well each object fits into its cluster, taking into account intra and between classes variations. The index ranges between -1 and 1, and a value close 1 means that the cluster is compact and homogeneous. Importantly, the combination of VAE and clustering also allows us to use conditional VAE to generate the expected fruit pertaining to a specific group.

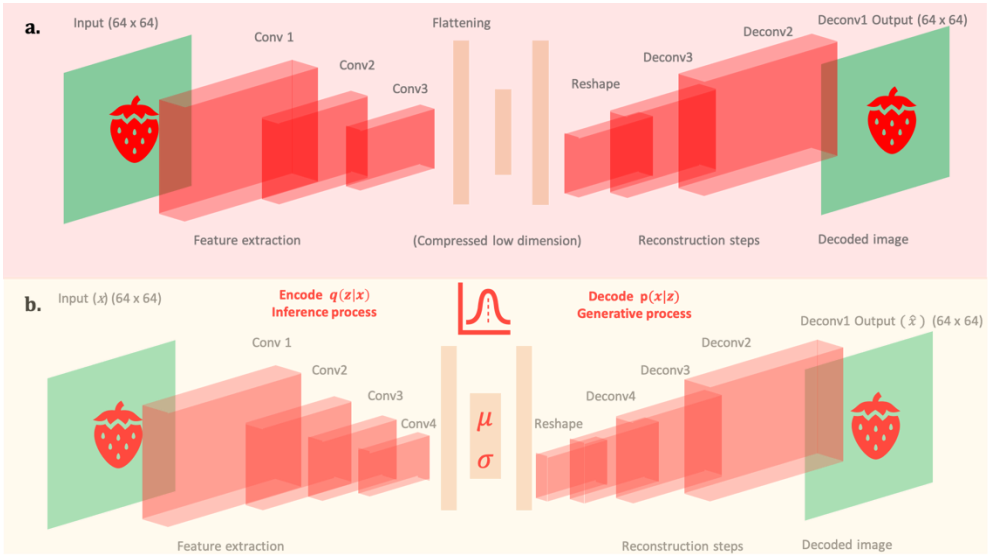


Figure 5.3. Autoencoder architectures. (a.) Architecture of convolutional autoencoder applied to the internal fruit images. (b) Architecture of convolutional variational autoencoder applied to external fruit. Unlike classical autoencoders, variational autoencoders are generative process as they learn the parameters of a distribution, instead of the feature representation. The last network was trained using image of size 64x64, the encoder step consisted on 4 convolutional layers with kernel size equal to 3 and the linear rectified 'relu' as activation function to perform feature extraction, see details in github account. Finally, the convolution output is flattened, and the mean and sigma parameters are extracted from a dense layer. In the last network, the decoder step starts with a vector sampled from the latent distribution as input and reconstructs the input by performing deconvolution operations. The last deconvolution uses the sigmoid as activation function. The loss function is the Kulback-Leibler (KL) divergence, which consists of both, a 'reconstruction' and a 'regularization' term. The first network is a classical autoencoder, which uses the classical Mean Squared Error (MSE) as loss function.

5.2.10 Genetic parameter inference

Genetic parameters determine how successful will be artificial selection and are therefore a critical parameter of any plant breeding scheme. Heritability (h^2) is the proportion of phenotypic variance explained by the genetic variation [57]. To estimate h^2, the degree of resemblance between relatives using the pedigree was used (see Supp. Table 2). Take linear model

$$\boldsymbol{y} = \mu + \boldsymbol{a} + \boldsymbol{d} + \boldsymbol{\varepsilon}, \quad \text{[Eq. 5.1]}$$

where $\boldsymbol{y}$ represents the phenotypes vector, averaged for each genotype, μ is the intercept, $\boldsymbol{a} \sim N(\mathbf{0}, \mathbf{A}\boldsymbol{\sigma}_a^2)$, $\boldsymbol{d} \sim N(\mathbf{0}, \mathbf{D}\boldsymbol{\sigma}_d^2)$, and $\boldsymbol{e} \sim N(\mathbf{0}, \mathbf{E}\boldsymbol{\sigma}_e^2)$ are the additive, dominance effects, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$ is the residual component, respectively; $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{D} = \{d_{ij}\}$ are the additive and dominance covariance matrices, respectively. Both **A** and **D** can be computed recursively from the pedigree [58]. In the presence of marker information, **A** and **D** can be computed as specified in [59,60], and implemented in [61] but statistical inference is otherwise identical. Posterior distributions of the genetic parameters were obtained using Reproducing Kernel Hilbert Spaces (RKHS) regression with the BGLR package [62]. The additive and dominance variance fraction were estimated as $\hat{h}_a^2 = s_a^2/(s_a^2 + s_d^2 + s_\varepsilon^2)$ and $\hat{h}_d^2 = s_d^2/(s_a^2 + s_d^2 + s_\varepsilon^2)$, where s_i^2 is the mean posterior estimate of σ_i^2.

5.3 Results

5.3.1 Shape descriptors

Shape linear descriptors, color in CIELAB scale, pseudo-landmarks and Elliptical Fourier transforms for fruit shape were computed for the 1920 external images output from pipeline in Fig. 5.1 and Algorithm S1. Fig. 5.4d shows the minimum and maximum consensus for shape superimposition, suggesting that shapes vary between a "globose-like" to an "elongated-like" form in these samples. The standard deviation of the first PCA from GPA coordinates (Supp. Fig. 1) of tip, neck and both sides around the neck are above the mean (Fig. 5.4c). This suggests that these regions are responsible for the main shape variations in strawberry, in agreement with Feldmann et al. [18]. Supp. Fig. 1 shows the fruit shape variations from the Procrustes Principal Component Analysis (Proc-PCA). The first principal component describes the variations between 'elongated' to 'globose' like. Observations with a negative score on that component correspond to elongated fruits, while those who have positive scores are 'globose'-like fruits. A permutation-based Procrustes analysis of variance to assess the effect of the crosses on the fruit shape. The p-value obtained after 100 permutations shows a significant effect of the lines, i.e., genotypes, in the fruit shape ($p<0.01$), suggesting that the shape is heritable (Supp. Table 3).

We also set a fourth order elliptical Fourier to describe the main strawberry shape variations (see Supp. Figs. 2 and 3). As in the Procrustes Analysis, variations in the first principal component of the elliptical analysis show that the strawberry shapes vary between "globose-like" to "elongated-like" (See a few examples in Supp. Fig. 7). Similarly, the first component from Elliptical -PCA analysis can also be used as a "morphological" descriptor. A k- means clustering using the two first

PCA components of Fourier transform similarly detects the two previously defined groups of shapes when setting k=2 (Supp. Fig. 5).

Alternatively, one can directly identify the number of different shapes from a collection of images. We used a VAE (Fig. 5.3b) to automatically discover the optimal number of shapes in our database, which again was k=2 (Supp. Fig. 5,6). About 35% of the strawberries belong to the 'globose-like' shape, whereas the remaining fruits were classified as 'elongated-like' (Fig. 5.5a,b).

Fig. 5.4d shows a PCA on the linear descriptors, where the color of each sample is proportional to the predicted cluster probability. A dark color corresponds to a fully elongated shape and a light blue, to a fully round fruit. Note that shape gradient is mainly observed along the second principal component. Interestingly, the most influential variables in this component are the fruit ratio between main and minor ellipse axis, the circularity and solidity coefficients (Fig. 5.4f). All of these are shape related variables. It is not surprising that solidity and circularity are highly correlated, since the convex hull area increases when a shape digresses from a circle (circularity), and solidity approaches zero. The area, perimeter and height are quasi-independent of the aforementioned descriptors and are not related with the shape clusters.

5.3.2 Color descriptors

For the external side color in our dataset, *L* channel ranged between 7.01 and 118.30, mean of 75.54, *b* channel ranged between 127.9 and 184.8, mean value of 167.1, and *a* channel had a mean of 175.4, ranging between 128.8 and 192.6.

Estimating the color of the internal fruit is more challenging than that of external parts, as it fluctuates in a wider range of patterns. Fig. 5.6 shows the estimated percentages of each reference color for four chosen strawberries. Note the percentage of "quasi-red" is zero and most of the fruit is computed as "pale" (~95%) for the first two, whitish fruits. Two colors, "quasi-red" and "orange-like", predominate in the third fruit. Finally, the last fruit is almost red, as can be verified from the estimated quasi-red value (99%).

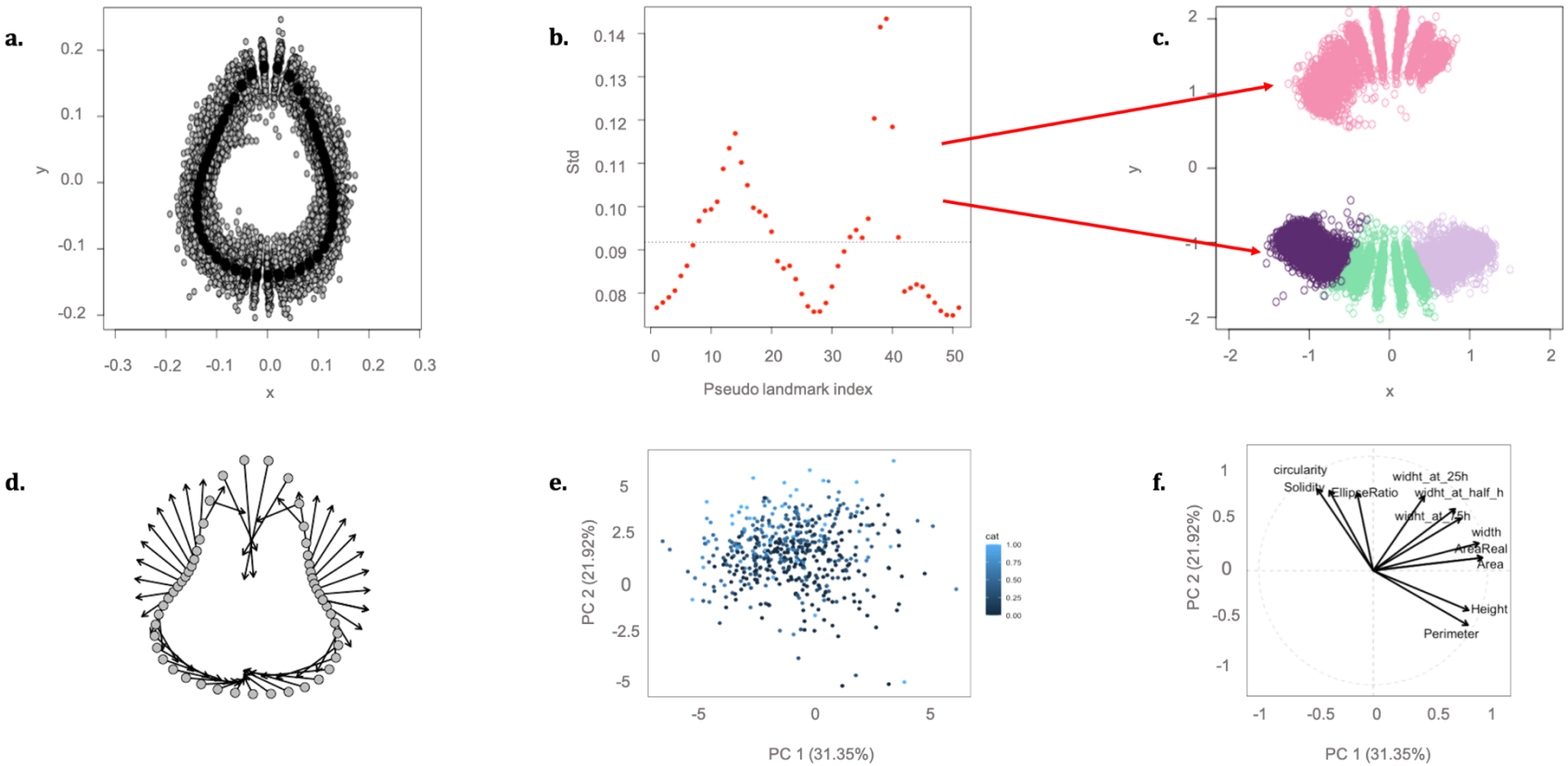


Figure 5.4. Summary of main analyses performed. (a) Generalized Procrustes Analysis output: landmarking superimposition for all external fruit shapes. (b) Standard deviation of each of the 50 landmarks, the dotted line parallel to x-axis corresponds to the average standard regression coefficient. Landmarks with a coefficient above the average are the most variable regions. (c) The most variable regions, which determine fruit shape, are the tip, the neck and both sides around the neck. (d) Two extreme Procrustes analysis plots: minimum and maximum consensus for the fruit shape. (e) PCA of all linear shape descriptors, each dot represents a different sample and the color is proportional to the predicted proportion of fruit to each category from the clusters obtained by Variational Autoencoder. (f) Relationship between the linear shape variables from the PCA analysis.

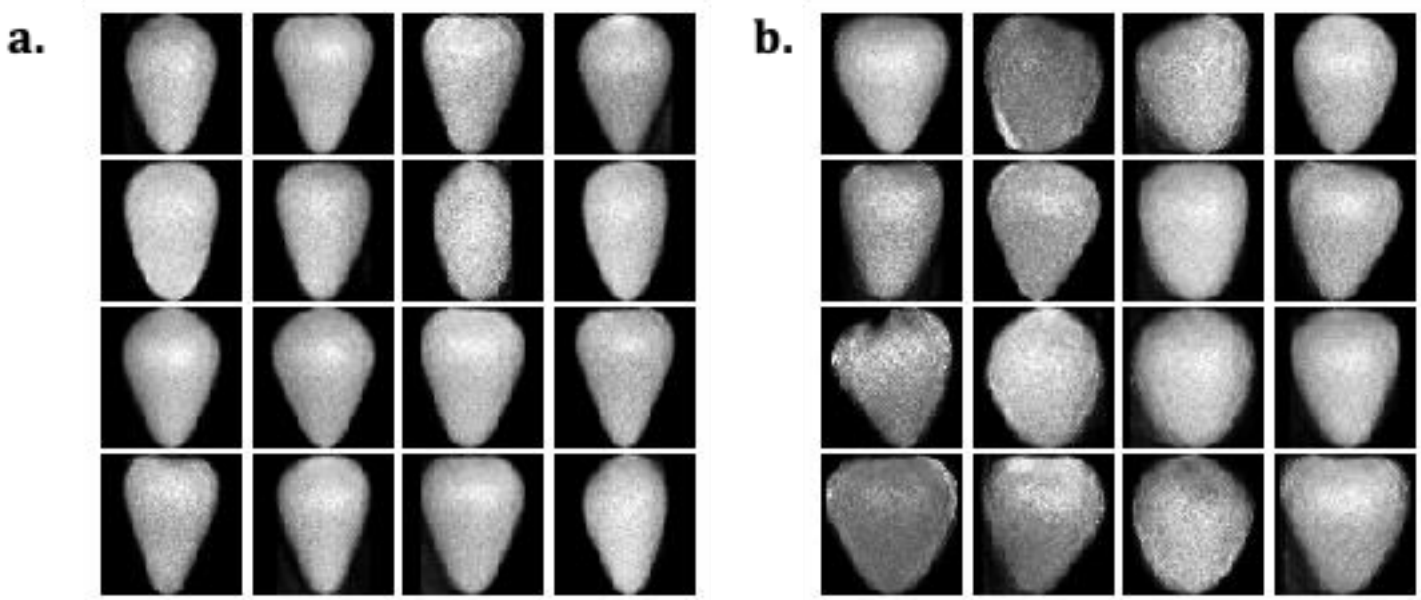


Figure 5.5. Images generated using the variational autoencoder combined with k-means in the latent space with k=2. (a) Images from the "elongated-like" cluster (b) Images from "globose-like" cluster.

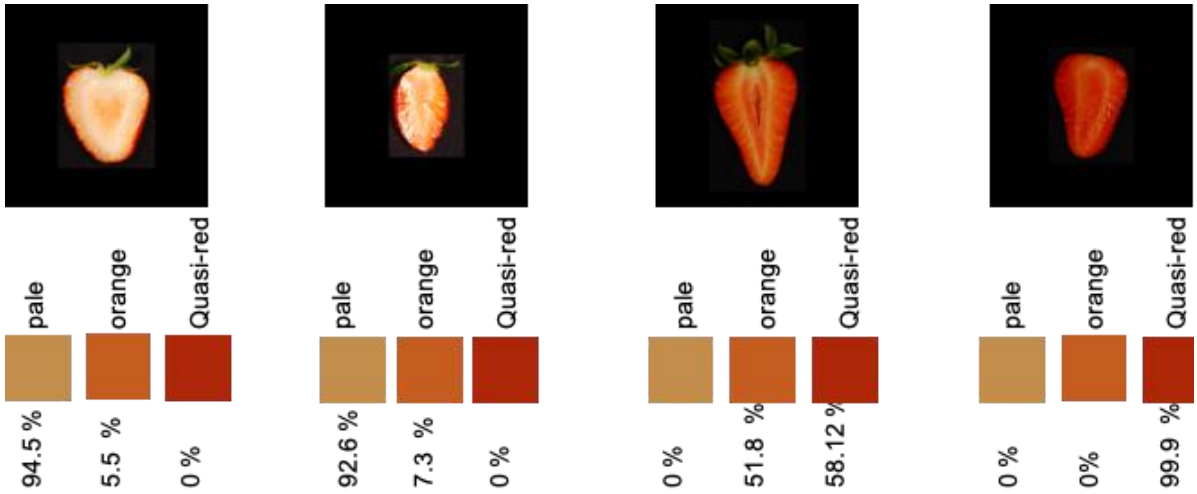


Figure 5.6. Estimated percentage of each of the three reference colors in four picked strawberries.

5.3.3 Genetic parameter estimation

Figure 5.7 shows the Bayesian estimates of heritability for all automatically extracted traits. We used the pedigree information to compute the additive and dominance relationship matrices, since we did not have genotypes. Like many polyploid species, strawberry is clonally propagated [63]. Inferring the dominance component in these cases is critical, as clonal propagation allows a straightforward utilization of gene interaction [64]. Interestingly, we found that dominance variance was higher than the additive component for most of the traits. The sum of both components $\hat{h}_a^2 + \hat{h}_d^2$ ranged between 0.4 to 0.6, indicating than the traits are clearly heritable. The ellipse ratio, and the ratio between height and width were the most heritable characters, exhibiting an important additive component. Elliptical Fourier components, as well the percentage of fruits of each of both categories obtained from VAE also have a high heritability, for both additive and dominance components. Regarding the internal color, we find that the pale color has an important dominance component.

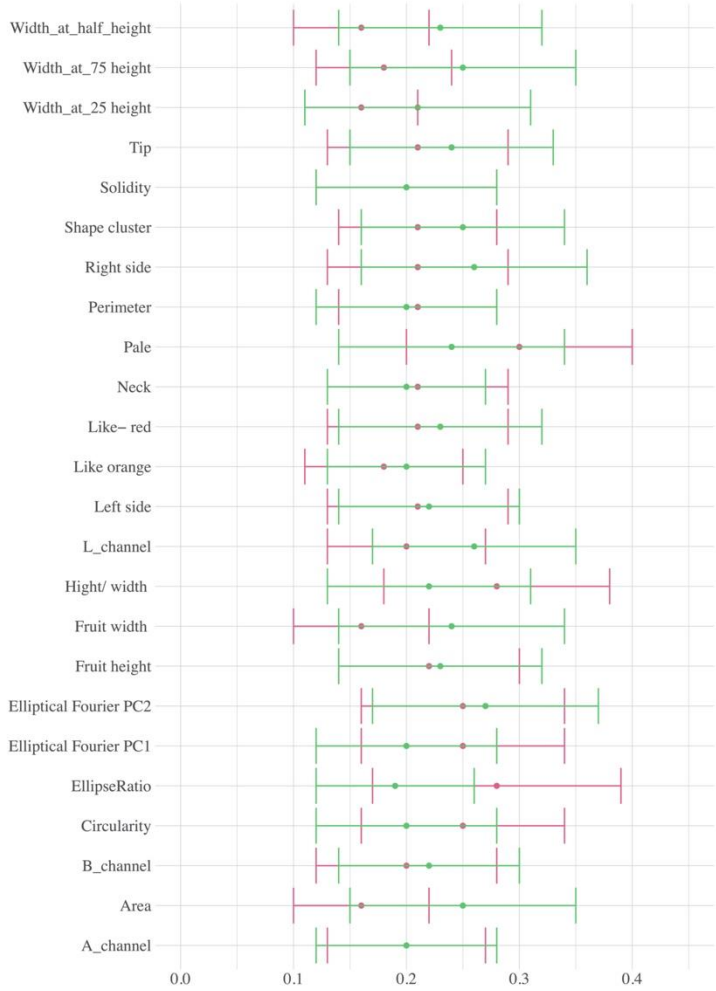


Figure 5.7. Estimation of additive (h2a) and dominance (h2d) variance fractions for measured traits.

5.4 Discussion

Over the last decades, plant and animal breeding programs have benefited from the development and cost reduction on genomic technologies [65,66]. Breeding nevertheless depends of both genotype and phenotype, and our ability of characterizing the latter is much more limited compared to the former [10,67]. In fact, one of the biggest challenges of 'Precision Agriculture' is to transform large-scale datasets collected with sensors into phenotypic measurements that can be used for genetic improvement.

Consumer attitudes are increasingly shaping agricultural practices. In the case of fruits, consumer preferences are based primarily on fruit appearance. However, measuring this trait is not straightforward, as it is a complex mixture of shape and color patterns. A crucial aspect for improving appearance is then to characterize the color and shape of the fruits in an inexpensive and fast way. In this paper, we deliver a fully automatized pipeline that analyzes fruit appearance as complex multivariate data. While this is not the first study characterizing fruit shape variations, our procedure is quite more automatized than their predecessors as requires minimal human intervention [18,20,40]. It also incorporates new features such as the use of variational

autoencoders (VAE) to automatically detect the most likely number of underlying shapes or to clustering the internal color.

The pipeline presented here or previous efforts to automatize fruit morphology measurement by Feldman et al. [18] are important steps to increase agriculture efficiency. They are by no means sufficient, and additional developments are warranted. A first limitation is that algorithms need to be trained in the specific dataset that will be used in production and can sometimes be difficult to generalize to different scenarios. A second limitation concerns the phenotypes measured. For instance, uniformity of shape and lack of blemishes (like depressions or creases) significantly impact the value of the product but were not studied here. Uniformity of fruits can be easily quantified, e.g., measuring the dispersion along landmarks (Fig. 5.4a) whereas irregularities in color that may mean fruit damage can be more challenging. In the lab, as done here, perhaps a suitable color clustering to associate color patterns with fruit damages could be envisaged. To be really useful, however, fruit damages should be evaluated once the product has been packaged, prior or after distribution, which would need distinct code from that employed here. The number of seeds is also important economically, but we found very high resolution is needed to quantify them. Finally, 3D approaches have also been evaluated in fruits, including strawberry [68,69]. 3D imaging is far more demanding in terms of sample collection and computationally than 2D[70,71]. This hamper using 3D technologies as massively as 2D, although 3D has a number of advantages, mainly a far more realistic and comprehensive fruit representation. For instance, Li et al. [72] utilize 3D imaging to assess fruit uniformity and show it can be characterized by combining up to six linear parameters.

Our algorithm requires images being taken on a homogeneous black or white surface and field images are not allowed. To compare the shapes and colors, all shots must be taken in the same conditions, using the same digital camera placed at the same height and setting the same parameters: focal length, manual aperture, exposure time and lighting. Scanned images are also allowed but the same scanning conditions must be followed in all images.

Although two-D digital images are among the easiest phenotypes to collect, analyzing them can be challenging, partly because objects boundaries must be determined, a process known as feature extraction. Numerous classical [41,44,73] and deep learning approaches [74,75] have been developed in computer vision and image processing to meet this objective. Here, we combined some of these methods to automatically segment fruit snapshots and read the fruit label. The main approach we used is not new, as it is based in an algorithm developed in the late seventies [41]. However, we resort to novel techniques in order to remove undesirable image noise [76], and we characterize color pattern or classify fruits through a variational autoencoder (Fig. 5.3) [42].

In this work, we characterize shape and color variations using several complementary methods, from naïve linear descriptors to multivariate and deep learning techniques. It is important to point out that results from all approaches are consistent, and suggest that the fruits in our database can be classified in two groups, "globose-like" and "elongated-like" (Supp. Figs. 5,6). We determine that the most variable regions are the neck, neck-sides and the tip of the fruit (Fig. 5.4b,c). The "shape" linear descriptor, i.e., the ratio between fruit shape and height, is a good morphological descriptor (Fig. 5.4e,f) and is as discriminative as more complex multivariate characterizations. An ANOVA on the Procrustes coordinates shows that genotype is significant (p-value<0.01, Supp Tab. 2), another indirect indication that shape is heritable.

Shapes can be classified using standard clustering techniques with the number of clusters k previously specified, as shown by Feldmann et al. in strawberry [18]. Our results are in agreement with these authors' in that we also find that shape is heritable and that a few components may be

needed to classify shapes (Fig. 5.4f). In addition to that approach, here we propose a completely unsupervised manner based on variational autoencoders (Fig. 5.3). The advantage of this analysis is that shape discovery can be automatized, but also that is capable of generating shapes not seen before. Predicting shapes and appearance of new genotypes can be a powerful tool to design new crosses, as the breeder can evaluate the average shape but also their variability in morphology. To our knowledge, VAEs have not been utilized for these purposes yet.

Describing internal color patterns is challenging, mainly because color is a quantitative multi-channel character. We addressed this problem by defining three reference colors named as "quasi-red", "orange-like" and "pale". We then automatically determined the percentage of color corresponding to each of these reference colors for each fruit using an autoencoder for fruit denoising and a k-means for segmentation. The algorithm calculates the Euclidean distance between the three RGB coordinates obtained by means of clustering to the target color coordinates and classifies the cluster as belonging to one of the three targets whose distance is minimal. The colors patterns are satisfactorily dissected, as can be seen in some picked images from the database (Fig. 5.6).

The phenotype results from a complex interaction between the genotype and environmental factors. Portraying the phenotypes would not be worthwhile for breeding if the desirable characters could not be transmitted to the progeny. Thus, quantifying the heritability of all of these traits is crucial. Typically, genetic variance is decomposed in additive and non- additive effects [77]. Clonally propagated species like strawberry allows direct utilization of dominance and epistatic interaction. We used Bayesian modeling to estimate both additive and dominance effects. As can be observed in Fig. 5.7 and Supp. Table 4, most traits are moderately heritable, and a high degree of variance is explained by the dominance component. In this scenario, prediction accuracy in genomic selection could possibly increase by including dominance in the model [63].

Nevertheless, data are from a single sampling season, making it not possible to estimate the variance caused by genotype x environment (GxE). Therefore, heritabilities reported are likely overestimates. Further, the pedigree utilized considered only parents and offspring, while parents themselves are related, which was ignored except in a subset of parents. The effect in this case should be smaller than that of GxE and should affect the variance of the estimates rather than bias, since relationships decrease quadratically with generation, and most information is contained in closest relatives [78].

We estimated heritabilities using pedigree information, but a similar study could be carried out if genetic markers were available. This would have the extra benefit of allowing to perform Genome Wide Association studies (GWAS) and to implement genomic selection [63,77]. It is straightforward to implement these features in our pipeline. Association studies in humans, apple or tomato have revealed genes or markers associated with human craniofacial shape [34,79,80], leaf variation [81] and tomato morphology [29]. To the best of our knowledge, there is not a similar study in strawberry and there is still a long way to go to fully unravel the genetic basis of strawberry shape [82].

5.5 Conclusion

There is a need to develop analysis pipelines for plant high-throughput phenotyping, suitable to automate processes that are often subjective and time consuming. Our workflow establishes a proof of concept in strawberry morphometrics, which can be transferred to other visual phenotypes and fruits with relatively minor modifications. We developed a python-based pipeline

(https://github.com/lauzingaretti/DeepAFS/blob/main/main.ipynb) that shows how to apply our methodology to other fruits like apple, tomatoes, citrus and prunus. This code is able to automatically read the fruit image, to segment it and to compute some linear and color descriptors. This code also allows to save the segmented images into a pre-defined folder, as well as the fruit outline reference points used for posterior multivariate comparison. Overall, our results show that, although fruit shape is made up of a complex set of traits, it can be quickly and automatically evaluated and is moderately heritable (Figs. 5.1,5.2,5.7). Future improvements are still needed as, e.g., image segmentation is not always simple in field conditions and many additional phenotypes are of commercial interest (e.g., uniformity, blemishes, among others). Future improvements should also address additional technological developments such as spectral and MIR images [17] and 3D imaging [72]. Finally, a word of caution: the user should be aware that artificial intelligence tools need through training in the specific conditions on which they are going to be employed and that optimizing algorithms may not be that simple.

Acknowledgments

The authors would like to thank Planasa for providing the strawberry fruits under the Planasa-IRTA collaboration contract, headed by AM. We thank the reviewers for their constructive, pertinent and complementary comments, and the editor for choosing them. LMZ was supported by a PhD grant from the Ministry of Economy and Science (MINECO, Spain). Work funded by the MINECO grants AGL2016-78709-R and PID2019-108829RB-I00 to MPE, and by the CERCA Programme / Generalitat de Catalunya.

We acknowledge financial support from the Spanish Ministry of Science and Innovation-State Research Agency (AEI), through the "Severo Ochoa Programme for Centres of Excellence in R&D" SEV-2015-0533 and CEX2019-000902-S.

Author Contributions

LMZ, AM and MPE conceived research. AM provided data. LMZ developed methods and code. LMZ and MPE wrote the manuscript with help from AM.

Conflict of Interest

The author declares no conflict of Interest.

Data Availability

Code is available at https://github.com/lauzingaretti/DeepAFS.

5.6 Bibliography

1. Hunter MC, Smith RG, Schipanski ME, Atwood LW, Mortensen DA. Agriculture in 2050: Recalibrating targets for sustainable intensification [Internet]. Bioscience. Oxford University Press; 2017 [cited 2020 Sep 22]. p. 386–91. Available from: https://pennstate.pure.elsevier.com/en/publications/agriculture-in-2050-recalibrating-targets-for-sustainable-intensi
2. Carnicer J, Coll M, Ninyerola M, Pons X, Sánchez G, Peñuelas J. Widespread crown condition decline, food web disruption, and amplified tree mortality with increased climate change-type

drought. Proc Natl Acad Sci U S A. 2011;108:1474–8.
3. Wernberg T, Smale DA, Tuya F, Thomsen MS, Langlois TJ, De Bettignies T, et al. An extreme climatic event alters marine ecosystem structure in a global biodiversity hotspot. Nat Clim Chang. Nature Publishing Group; 2013;3:78–82.
4. Hegerl GC, Hanlon H, Beierkuhnlein C. Climate science: Elusive extremes. Nat Geosci. Nature Publishing Group; 2011;4:142–3.
5. Porfirio LL, Newth D, Finnigan JJ, Cai Y. Economic shifts in agricultural production and trade due to climate change. Palgrave Commun [Internet]. Palgrave Macmillan Ltd.; 2018 [cited 2020 Sep 22];4:1–9. Available from: https://www.nature.com/articles/s41599-018-0164-y
6. Lobos GA, Camargo A V., del Pozo A, Araus JL, Ortiz R, Doonan JH. Editorial: Plant Phenotyping and Phenomics for Plant Breeding. Front Plant Sci [Internet]. Frontiers Media S.A.; 2017 [cited 2020 Sep 22];8:2181. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2017.02181/full
7. Pieruschka R, Schurr U. Plant Phenotyping: Past, Present, and Future. Plant Phenomics [Internet]. 2019 [cited 2020 Sep 18];2019:1–6. Available from: https://spj.sciencemag.org/plantphenomics/2019/7507131/
8. Fasoula DA, Fasoula VA. Gene Action and Plant Breeding. Plant Breed Rev. John Wiley & Sons, Inc.; 2010. p. 315–74.
9. Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M. Plant Phenomics, From Sensors to Knowledge. Curr. Biol. Cell Press; 2017. p. R770–83.
10. Awada L, Phillips PWB, Smyth SJ. The adoption of automated phenotyping by plant breeders. Euphytica. Springer Netherlands; 2018;214.
11. Mahlein A-K. Present and Future Trends in Plant Disease Detection. Plant Dis. 2016;100:1–11.
12. Stafford J V. Implementing precision agriculture in the 21st century. J Agric Eng Res. 2000;76:267–75.
13. Schork NJ. Genetics of Complex Disease Approaches, Problems, and Solutions. Am J Respir Crit Care Med. 1997.
14. Großkinsky DK, Svensgaard J, Christensen S, Roitsch T. Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap [Internet]. J. Exp. Bot. Oxford University Press; 2015 [cited 2020 Sep 18]. p. 5429–40. Available from: https://academic.oup.com/jxb/article/66/18/5429/482901
15. Zhao C, Zhang Y, Du J, Guo X, Wen W, Gu S, et al. Crop phenomics: Current status and perspectives. Front. Plant Sci. Frontiers Media S.A.; 2019.
16. Gracia-Romero A, Kefauver SC, Fernandez-Gallego JA, Vergara-Díaz O, Nieto-Taladriz MT, Araus JL. UAV and Ground Image-Based Phenotyping: A Proof of Concept with Durum Wheat. Remote Sens [Internet]. MDPI AG; 2019 [cited 2020 Sep 22];11:1244. Available from: https://www.mdpi.com/2072-4292/11/10/1244
17. Ruckelshausen A, Busemeyer L. Toward digital and image-based phenotyping. Phenomics Crop Plants Trends, Options Limitations. 2015.
18. Feldmann MJ, Hardigan MA, Famula RA, López CM, Tabb A, Cole GS, et al. Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry. Gigascience. 2020;9:1–17.
19. Turner SD, Ellison SL, Senalik DA, Simon PW, Spalding EP, Miller ND. An automated image analysis pipeline enables genetic studies of shoot and root morphology in carrot (daucus carota l.). Front Plant Sci. 2018;871:1–17.
20. Brewer MT, Lang L, Fujimura K, Dujmovic N, Gray S, Van Der Knaap E. Development of a controlled vocabulary and software application to analyze fruit shape variation in tomato and other plant species. Plant Physiol [Internet]. American Society of Plant Biologists; 2006 [cited 2020 Oct 5];141:15–25. Available from: www.plantphysiol.org/cgi/doi/10.1104/pp.106.077867.
21. Das A, Schneider H, Burridge J, Ascanio AKM, Wojciechowski T, Topp CN, et al. Digital

imaging of root traits (DIRT): A high-throughput computing and collaboration platform for field-based root phenomics. Plant Methods. BioMed Central; 2015;11:1–12.
22. Seethepalli A, Guo H, Liu X, Griffiths M, Almtarfi H, Li Z, et al. RhizoVision Crown: An Integrated Hardware and Software Platform for Root Crown Phenotyping. Plant Phenomics. 2020;2020:1–15.
23. Jaeger SR, Machín L, Aschemann-Witzel J, Antúnez L, Harker FR, Ares G. Buy, eat or discard? A case study with apples to explore fruit quality perception and food waste. Food Qual Prefer. Elsevier Ltd; 2018;69:10–20.
24. Gilbert JL, Olmstead JW, Colquhoun TA, Levin LA, Clark DG, Moskowitz HR. Consumer-assisted selection of blueberry fruit quality traits. HortScience [Internet]. American Society for Horticultural Science; 2014 [cited 2020 Sep 23];49:864–73. Available from: http://www.
25. Lewers KS, Newell MJ, Park E, Luo Y. Consumer preference and physiochemical analyses of fresh strawberries from ten cultivars. Int J Fruit Sci [Internet]. Taylor and Francis Inc.; 2020 [cited 2020 Sep 23];1–24. Available from: https://www.tandfonline.com/doi/full/10.1080/15538362.2020.1768617
26. González M, Baeza E, Lao JL, Cuevas J. Pollen load affects fruit set, size, and shape in cherimoya. Sci Hortic (Amsterdam). Elsevier; 2006;110:51–6.
27. Klatt BK, Holzschuh A, Westphal C, Clough Y, Smit I, Pawelzik E, et al. Bee pollination improves crop quality, shelf life and commercial value. Proc R Soc B Biol Sci [Internet]. Royal Society; 2014 [cited 2020 Sep 23];281:20132440. Available from: https://royalsocietypublishing.org/doi/10.1098/rspb.2013.2440
28. Peter Klingenberg C. Evolution and development of shape: integrating quantitative approaches. Nat Publ Gr [Internet]. 2010 [cited 2020 Jun 22]; Available from: www.nature.com/reviews/genetics
29. Rodríguez GR, Muños S, Anderson C, Sim SC, Michel A, Causse M, et al. Distribution of SUN, OVATE, LC, and FAS in the tomato germplasm and the relationship to fruit shape diversity. Plant Physiol. 2011;156:275–85.
30. Monforte AJ, Diaz A, Caño-Delgado A, Van Der Knaap E. The genetic basis of fruit morphology in horticultural crops: Lessons from tomato and melon. J Exp Bot. 2014;65:4625–37.
31. Brewer MT, Moyseenko JB, Monforte AJ, Van Der Knaap E. Morphological variation in tomato: A comprehensive study of quantitative trait loci controlling fruit shape and development. J Exp Bot. 2007;58:1339–49.
32. Rashidi M, Keshavarzpour F. Classification of Apple Size and Shape Based on Mass and Outer Dimensions. J Agric Environ Sci [Internet]. 2010 [cited 2020 Oct 5];9:618–21. Available from: http://faostat.fao.org.
33. Mezghani N, Zaouali I, Amri WB, Rouz S, Simon PW, Hannachi C, et al. Fruit morphological descriptors as a tool for discrimination of Daucus L. germplasm. Genet Resour Crop Evol. 2014;61:499–510.
34. Claes P, Liberton DK, Daniels K, Rosana KM, Quillen EE, Pearson LN, et al. Modeling 3D Facial Shape from DNA. PLoS Genet. 2014;10.
35. Dryden, Ian L., Mardia K V. Statistical shape analysis. Wiley series in probability and statistics; 1998.
36. UPOV. Strawberry: Guidelines for the conduct of tests for distinctness, uniformity and stability. Upov. 2012;1–26.
37. Schindelin J, Rueden CT, Hiner MC, Eliceiri KW. The ImageJ ecosystem: An open platform for biomedical image analysis. Mol Reprod Dev [Internet]. John Wiley & Sons, Ltd; 2015 [cited 2020 Oct 1];82:518–29. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/mrd.22489
38. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: An open-source platform for biological-image analysis. Nat. Methods. 2012. p. 676–82.
39. Darrigues A, Hall J, Van Der Knaap E, Francis DM, Dujmovic N, Gray S. Tomato analyzer-

color test: A new tool for efficient digital phenotyping. J Am Soc Hortic Sci. 2008;133:579–86.
40. Gehan MA, Fahlgren N, Abbasi A, Berry JC, Callen ST, Chavez L, et al. PlantCV v2: Image analysis software for high-throughput plant phenotyping. PeerJ [Internet]. PeerJ Inc.; 2017 [cited 2020 Oct 1];2017:e4088. Available from: http://plantcv.danforthcenter.org/pages/data.html
41. Otsu N. THRESHOLD SELECTION METHOD FROM GRAY-LEVEL HISTOGRAMS. IEEE Trans Syst Man Cybern. 1979;SMC-9:62–6.
42. Goodfellow, I., Bengio, Y., Courville A. Deep Learning. MIT Press. MIT Press Cambridge; 2016.
43. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
44. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC Superpixels.
45. Diaz-Garcia L, Covarrubias-Pazaran G, Schlautman B, Grygleski E, Zalapa J. Image-based phenotyping for identification of QTL determining fruit shape and size in American cranberry (Vaccinium macrocarpon L.). PeerJ. 2018;2018:1–19.
46. Gower JC. Generalized procrustes analysis. Psychometrika. 1975;40:33–51.
47. Bonhomme V, Picq S, Gaucherel C, Claude J. Momocs: Outline analysis using R. J Stat Softw. 2014;56:1–24.
48. Adams DC, Otárola-Castillo E. Geomorph: An r package for the collection and analysis of geometric morphometric shape data. Methods Ecol Evol [Internet]. John Wiley & Sons, Ltd; 2013 [cited 2020 Oct 15];4:393–9. Available from: https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12035
49. Kuhl FP, Giardina CR. Elliptic fourier features of a closed contour. Comput Graph Image Process. 1982;18:236–58.
50. Bellman R. Dynamic Programming Princeton University Press [Internet]. Princeton, NJ. 1957 [cited 2020 Oct 13]. Available from: https://press.princeton.edu/books/paperback/9780691146683/dynamic-programming
51. Gropp A, Atzmon M, Lipman Y. ISOMETRIC AUTOENCODERS. arXiv Prepr arXiv200609289. 2020;
52. Kingma DP, Welling M. Auto-encoding variational bayes. 2nd Int Conf Learn Represent ICLR 2014 - Conf Track Proc. 2014.
53. Kingma DP. Fast Gradient-Based Inference with Continuous Latent Variable Models in Auxiliary Form. 2013 [cited 2020 Oct 13]; Available from: http://arxiv.org/abs/1306.0733
54. Rezende DJ, Mohamed S, Wierstra D. Stochastic Back-propagation and Variational Inference in Deep Latent Gaussian Models. Proc 31st … [Internet]. 2014 [cited 2020 Oct 13];32:1278–86. Available from: http://jmlr.org/proceedings/papers/v32/rezende14.html
55. Ishikawa T, Hayashi A, Nagamatsu S, Kyutoku Y, Dan I, Wada T, et al. Classification of strawberry fruit shape by machine learning. Int Arch Photogramm Remote Sens Spat Inf Sci - ISPRS Arch [Internet]. 2018 [cited 2020 Oct 13]. p. 463–70. Available from: https://doi.org/10.5194/isprs-archives-XLII-2-463-2018
56. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
57. Falconer, Douglas S and Mackay, Trudy FC and Frankham R. Introduction to quantitative genetics (4th edn). Trends Genet. 1996;12:280.
58. Nazarian A, Gezan SA. GenoMatrix: A Software Package for Pedigree-Based and Genomic Prediction Analyses on Complex Traits. J Hered [Internet]. Oxford University Press; 2016 [cited 2020 Oct 13];107:372–9. Available from: /pmc/articles/PMC4888442/?report=abstract
59. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. Elsevier Inc.; 2008;91:4414–23.
60. Vitezica ZG, Varona L, Legarra A. On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. Genetics [Internet]. 2013 [cited 2019 Oct 11];195:1223–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24121775
61. Amadeu RR, Cellon C, Olmstead JW, Garcia AAF, Resende MFR, Muñoz PR, et al.

AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. 2016; Available from: https://github.com/prmunoz/AGHmatrix/blob/master/
62. Pérez P, De Los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics [Internet]. 2014;198:483–95. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25009151
63. Gezan SA, Osorio LF, Verma S, Whitaker VM. An experimental validation of genomic selection in octoploid strawberry. Hortic Res [Internet]. 2017;4:16070. Available from: http://www.nature.com/articles/hortres201670
64. Grüneberg W, Mwanga R, Andrade M, Espinoza J. Selection methods. Part 5: Breeding clonally propagated crops. Plant Breed Farmer Particip [Internet]. 2009;275–322. Available from: http://www.cabdirect.org/abstracts/20103075062.html
65. Robertsen CD, Hjortshøj RL, Janss LL. Genomic selection in cereal breeding. Agronomy. MDPI AG; 2019.
66. J. Crossa, P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, Campos G de los. Genomic selection in plant breeding: Methods, models, and perspectives. Trends Plant Sci. 2017;22:961–75.
67. Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M. Plant Phenomics, From Sensors to Knowledge. Curr. Biol. Cell Press; 2017. p. R770–83.
68. He JQ, Harrison RJ, Li B. A novel 3D imaging system for strawberry phenotyping. Plant Methods. 2017;13:1–8.
69. Li B, Cockerton HM, Johnson AW, Karlström A, Stavridou E, Deakin G, et al. Defining strawberry shape uniformity using 3D imaging and genetic mapping. Hortic Res. 2020;7.
70. Wahabzada M, Paulus S, Kersting K, Mahlein A-K. Automated interpretation of 3D laserscanned point clouds for plant organ segmentation. BMC Bioinformatics [Internet]. BioMed Central Ltd.; 2015 [cited 2020 Oct 8];16:248. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0665-2
71. Paulus S. Measuring crops in 3D: Using geometry for plant phenotyping. Plant Methods [Internet]. BioMed Central; 2019;15:1–13. Available from: https://doi.org/10.1186/s13007-019-0490-0
72. Li B, Cockerton HM, Johnson AW, Karlström A. Defining Strawberry Uniformity using 3D Imaging. 2020;44.
73. Unit MI. A review on image segmentation techniques. 1993;26.
74. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN.
75. Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans Pattern Anal Mach Intell [Internet]. 2017 [cited 2019 Oct 11];39:2481–95. Available from: https://ieeexplore.ieee.org/document/7803544/
76. Mao X-J, Shen C, Yang Y-B. Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections. 2016;1–17. Available from: http://arxiv.org/abs/1606.08921
77. Zingaretti LM, Gezan SA, Ferrão LF V., Osorio LF, Monfort A, Muñoz PR, et al. Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species. Front Plant Sci [Internet]. 2020;11:1–14. Available from: https://www.frontiersin.org/article/10.3389/fpls.2020.00025/full
78. Lourenco DAL, Misztal I, Tsuruta S, Aguilar I, Lawlor TJ, Forni S, et al. Are evaluations on young genotyped animals benefiting from the past generations? J Dairy Sci [Internet]. Elsevier; 2014;97:3930–42. Available from: http://dx.doi.org/10.3168/jds.2013-7769
79. Claes P, Roosenboom J, White JD, Swigut T, Sero D, Li J, et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. Nat Genet [Internet]. Springer US; 2018;50:414–23. Available from: http://dx.doi.org/10.1038/s41588-018-0057-4
80. Galvánek M, Furmanová K, Chalás I, Sochor J. Automated facial landmark detection, comparison and visualization. Proc - SCCG 2015 31st Spring Conf Comput Graph. 2015;7–14.

81. Migicovsky Z, Li M, Chitwood DH, Myles S. Morphometrics reveals complex and heritable apple leaf shapes. Front Plant Sci. 2018;8:1–14.
82. Gaston A, Osorio S, Denoyes B, Rothan C. Applying the Solanaceae Strategies to Strawberry Crop Improvement. Trends Plant Sci. 2020;25:130–40.

Chapter 6

Estimating conformational traits in dairy cattle with DeepAPS: a two-step Deep learning Automated Phenotyping and Segmentation approach

Jessica Nye[1,*], **Laura M. Zingaretti**[1,*], and Miguel Pérez-Enciso[1,2]

1. Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Barcelona, Spain.
2. ICREA, Passeig de Lluís Companys 23, 08010 Barcelona, Spain.

*These authors contributed equally to the work

Estimating Conformational Traits in Dairy Cattle With DeepAPS: A Two-Step Deep Learning Automated Phenotyping and Segmentation Approach

Correspondence:

Jessica Nye:
Dr.Jessica.Nye@gmail.com
Laura Zingaretti:
m.lau.zingaretti@gmail.com
Miguel Pérez-Enciso
miguel.perez@uab.es

Keywords: image analysis, morphology, phenomics, image mask, deep learning, dairy cattle.

Frontiers in Genetics

Front Genet. 2020 May 21;11:513. doi: 10.3389/fgene.2020.00513. PMID: 32508888; PMCID: PMC7253626.
2020, May 21

Abstract

Assessing conformation features in an accurate and rapid manner remains a challenge in the dairy industry. While recent developments in computer vision have greatly improved automated background removal, these methods have not been fully translated to biological studies. Here, we present a composite method (DeepAPS) that combines two readily available algorithms in order to create a precise mask for an animal image. This method performs accurately when compared with manual classification of proportion of coat color with an adjusted $R^2 = 0.926$. Using the output mask, we are able to automatically extract useful phenotypic information for fourteen additional morphological features. Using pedigree and image information from a web catalog (www.semex.com), we estimated high heritabilities (ranging from $h^2 = 0.18 – 0.82$), indicating that meaningful biological information has been extracted automatically from imaging data. This method can be applied to other datasets and requires only a minimal number of image annotations (~50) to train this partially supervised machine-learning approach. DeepAPS allows for the rapid and accurate quantification of multiple phenotypic measurements while minimizing study cost. The pipeline is available at https://github.com/lauzingaretti/deepaps.

6.1 Introduction

Breeding programs depend on large-scale, accurate phenotyping, which is also critical for genomic dissection of complex traits. While the genome of an organism can be characterized, e.g., with high density genotyping arrays, the 'phenome' is much more complex and can never be fully described, as it varies over time and changes with the environment [1]. The cost of genotyping continues to drop, but there is still a need for improvements in obtaining high-performance phenotypes at a lower cost [2]. In cattle, the number of phenotypes recorded in traditional breeding schemes is relatively small, because its recording is expensive. For instance, yearly milk yield is usually inferred by extrapolation using a few lactation measurements, whereas actual milk production can now be measured individually and daily using automated milking robots.

In addition to milk yield, dairy cattle breeders are interested in conformational traits. These metrics are not only relevant aesthetically but can also have an important influence on an animal's breeding value. Body conformation is associated with dairy performance [3] and longevity, which strongly contributes to lifetime milk production [4]. Milk production is positively correlated with udder size [5]. The highest negative economic impact for dairy farmers is caused by lameness either due to leg malformations or injury [6,7]. Extracting the detailed conformational phenotypes which may impact progeny success are likewise time consuming and costly to collect, and in the absence of quantitative tools, farmers often evaluate morphometric measurements qualitatively.

The emergence of modern sensor technologies, such as Unmanned Aerial Vehicles (UAV) combined with simple digital cameras [8], mass spectroscopy, robotics, and hyper-spectral images [9], among others, have revolutionized breeding programs, mainly in plants, allowing for non-invasive evaluation of multiple complex traits. Although in animal breeding their application is

more scarce, modern livestock farming is beginning to benefit from access to these inexpensive sensor tools. Now, it is possible to remotely monitor behavior [10–12] and animal welfare [13], assess movement [14], measure body confirmation [15,16], quantify individual food intake [12,13,17], maintain an optimum environment [18], or decrease instances of stillbirths [19,20]. These automated measurements rely on temperature [18–20], pressure [13,17], movement [14], and visual sensors [10–12,15,16].

As several remote monitoring schemes are based on digital images or video, automated image analysis techniques are urgently needed to quantify traits of interest [21]. Applying image analysis to breeding programs is not new, however many of these methods largely depend on time consuming image-by-image processing facilitated by the researcher (as in [22–24]). The few automated resources currently implemented for cattle analyses require complicated set-ups and costly equipment [14,16]. This is not surprising as accurately quantifying phenotypic information is one of the most challenging aspects in biology [25–27].

The availability of new algorithms based on machine learning has revolutionized computer vision, impacting a wide range of fields that rely on computers to analyze images, with the potential to optimize herd care and improve animal and plant breeding program outcomes [11,12,16]. These recent advances have led to precise object detection and semantic segmentation in complex images [28–30].

Here we show how automatically parsed web-based catalog datasets can be converted into useful information by automatically inferring genetic parameters of several morphological measurements in dairy cattle. We combined web scraping, deep learning, and statistical techniques in order to achieve our objective. The proposed methodology is a mixture between a supervised deep learning approach, Mask R-CNN [31] and an unsupervised algorithm [32] which can achieve highly precise automatic image segmentation. After removing the background, phenotypic information, including coat color and body conformational traits can be easily quantified. Lastly, we demonstrate the potential applications of this method in other datasets. We assert that our work could constitute a good proxy for using inexpensive and non-invasive computer vision techniques into the dairy cattle breeding programs.

6.2 Materials and Methods

6.2.1 Image Collection

Images of bulls were collected through web-scraping using the python library Beautiful Soup [33].Images from sire catalogs of six Artificial Insemination companies were collected. We additionally automatically collected bull images from one semen provider (www.semex.com) and those of identified familial relationships (daughters, dams, granddams, and great granddams) where possible. We downloaded a total of 1,819 images. These images ranged in size between 339 to 879 pixels and 257 to 672 pixels for width and height, respectively. The animals are Holstein with patched black and white bodies, but some images are red Holstein. Individuals ranged in color from all white, all black, all brown, to a mixture of the colors. The images were flipped so that all

animals faced the right side of the image using ImageMagick version 7.0.9-0 convert -flop function. The animals are standing in front of dynamic backgrounds including forest, field, snow, water, and straw. All images contained only one animal, and sometimes contained a person or an arm.

6.2.2 Automated Segmentation

One of the most challenging tasks in computer vision is instance segmentation, i.e., the identification of boundaries of objects at the pixel level [32], whereas object classification, i.e., to determine if an object belongs to certain class is relatively simpler. R-CNN [28], a deep learning approach, as well as Fast R-CNN [34], Faster R-CNN [35] or Mask R-CNN [36] are widely used to solve this task. Although these methods are efficient, they are not accurate enough for some purposes since the obtained segmentation often removes parts of the object of interest or contains parts of the background.

We applied a two-step procedure to automatically segment the animal's profile as accurately as possible. The composite method begins by using Mask R-CNN [36], which has three outputs for each candidate object in an input image (Figure 6.1A): a class label (say 'cow'), bounding box offset or region of interest (RoI), and the object mask consisting of an approximate layout of a spatial object. As in the original Mask R-CNN, we used the annotated image database common objects in context (COCO, http://cocodataset.org; [37]) to train the algorithm, and select the class codes for cow. In short, Mask R-CNN is a deep learning algorithm that consists of two steps: first, it proposes regions within the image that may contain objects of interest and, second, generates a mask for every detected object. The latter step consists of a binary classification of pixels, either a pixel belongs to the object or to the background. For more details about this method readers can consult, e.g., https://towardsdatascience.com/computer-vision-instance-segmentation-with-mask-r-cnn-7983502fcad1, https://engineering.matterport.com/splash-of-color-instance-segmentation-with-mask-r-cnn-and-tensorflow-7c761e238b46 or should refer to [36]. Figure 6.1B shows the applied mask predicted by Mask R-CNN, this mask removes the majority of the background, but also removes parts of the cow's body making it necessary for the development of our two-step composite method. We used the implementation of Mask R-CNN in https://github.com/matterport/Mask_RCNN.
After the RoI and class labels are extracted, we select only the RoI for our desired object (i.e., the bull or cow). This allows us to remove some of the background and obtain a smaller, less noisy image. As explained above, the Mask R-CNN segmentation was not accurate enough for our purposes (Figure 6.1B). Therefore, we passed the RoI and predicted mask to a modified version of the unsupervised image segmentation algorithm from [32]. We used the code available at https://github.com/kanezaki/pytorch-unsupervised-segmentation. The original algorithm relies on separating pixels from each other and grouping them into distinct clusters based on color and texture. The underlying assumptions of this model are that: 1) pixels of similar features should be clustered together, 2) spatially continuous pixels should be clustered together, and 3) the number of clusters should be large. This is achieved by applying a linear classifier which groups pixels into different clusters based on their features. The difference between the original algorithm and ours is we do not try to maximize the total number of clusters, but instead we merely improve upon the mask generated by Mask R-CNN based on pixel identity. This makes more effective the algorithm to run, since the algorithm applied to the whole original image was not completely

satisfactory. This proceeds by self-training the network through back propagation, by alternating between two stages: 1) forward super pixel refinement, and 2) backward gradient descent. Much like any supervised approach this is achieved by calculating the cross-entropy loss between network and cluster labels, then back propagating the error rates used to update the convolutional filter parameters. Backpropagation is a popular and clever method used in deep learning. It allows computing the gradient of the loss function very efficiently by using the chain rule for derivatives, which greatly simplifies optimization in complex models.

After refinement through the unsupervised algorithm, we obtained a relatively precise mask for our input image (Figure 6.1C). However, the unsupervised clustering still can confound the foreground and the background. We then apply an additional filter to the mask, median blur function from OpenCV [38], removing small islands that have been mislabeled during the clustering step (Figure 6.1D). We lastly apply the mask by coloring all pixels predicted to be in the background by a solid color (Figure 6.1E).

To extract the proportion of and average color(s) from each cluster, we apply k-means using the scikit-learn library [39].To measure anatomical features, we extract only the outline of the desired object from the mask (Figure 6.1F) using the edge detection algorithm developed by Canny [40] and implemented in OpenCV [38]. After extracting the edge, we apply one more filter to remove any islands that may remain using the remove_small_objects function from the morphology package available from scikit-learn [39]. Now that the input image has been reduced down to just the object outline, we can take advantage of common conformational features of the underlying data, and extract pixel coordinates. For example, we extracted the coordinate of the pixel closest to the bottom left corner which corresponds to the back foot of the cow. We proceeded in this way to extract 13 coordinates from each animal (Figure 6.1G). We then calculate the distance in pixels between various points, effectively extracting body confirmations automatically. The

fourteen conformational traits are described in Supplementary Figure 1. Code for the whole pipeline is available at https://github.com/lauzingaretti/deepaps.

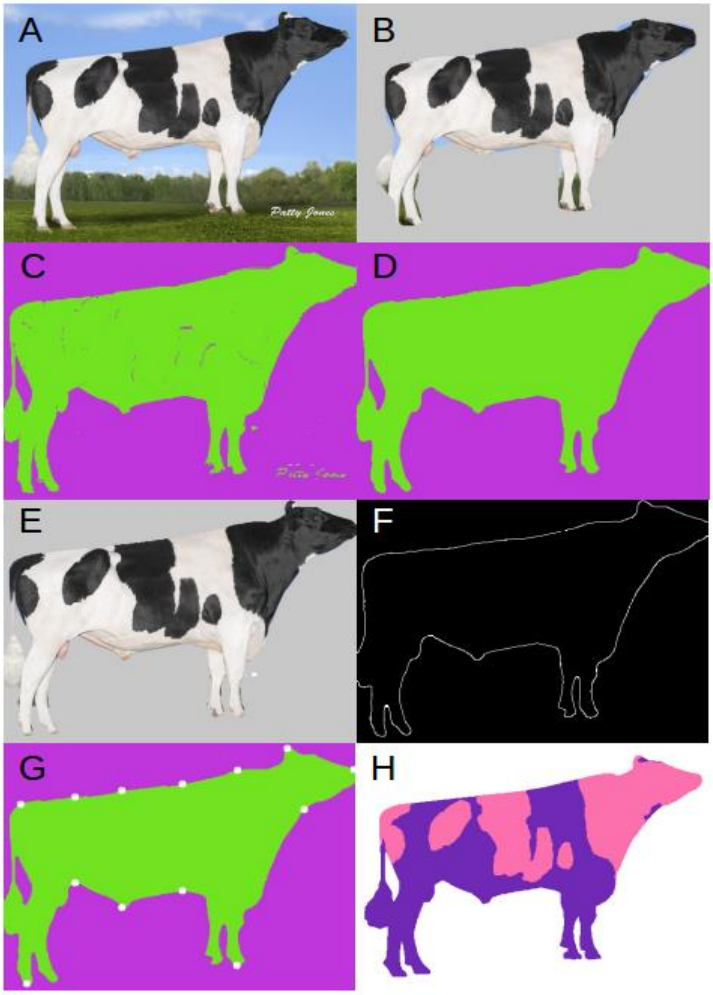


Figure 6.1. Example input and outputs. A: Original input image. B: Mask R-CNN applied mask. C: DeepAPS raw output. D: Final output of DeepAPS after all applied filters. E: Final DeepAPS mask applied to input image. F: Outline extraction of original input image. G: Extracted landmark coordinates. H: Manual color segmentation. Image from Semex.

6.2.3 Manual Segmentation

To check how accurate the automated segmentation was, we manually segmented N = 481 images that were not of Semex origin. We used Kanezaki's demo.py program [32] in python3.6 [41] using default parameters. The output images were opened in the image processing software GIMP (https://www.gimp.org/), and the background was manually changed from the colored cluster to white (Figure 6.1H). To extract the color clusters, we calculated the proportion of color clusters

in each image by using k-means as above, and manually matched each color cluster to the original picture and removed the proportion of background.

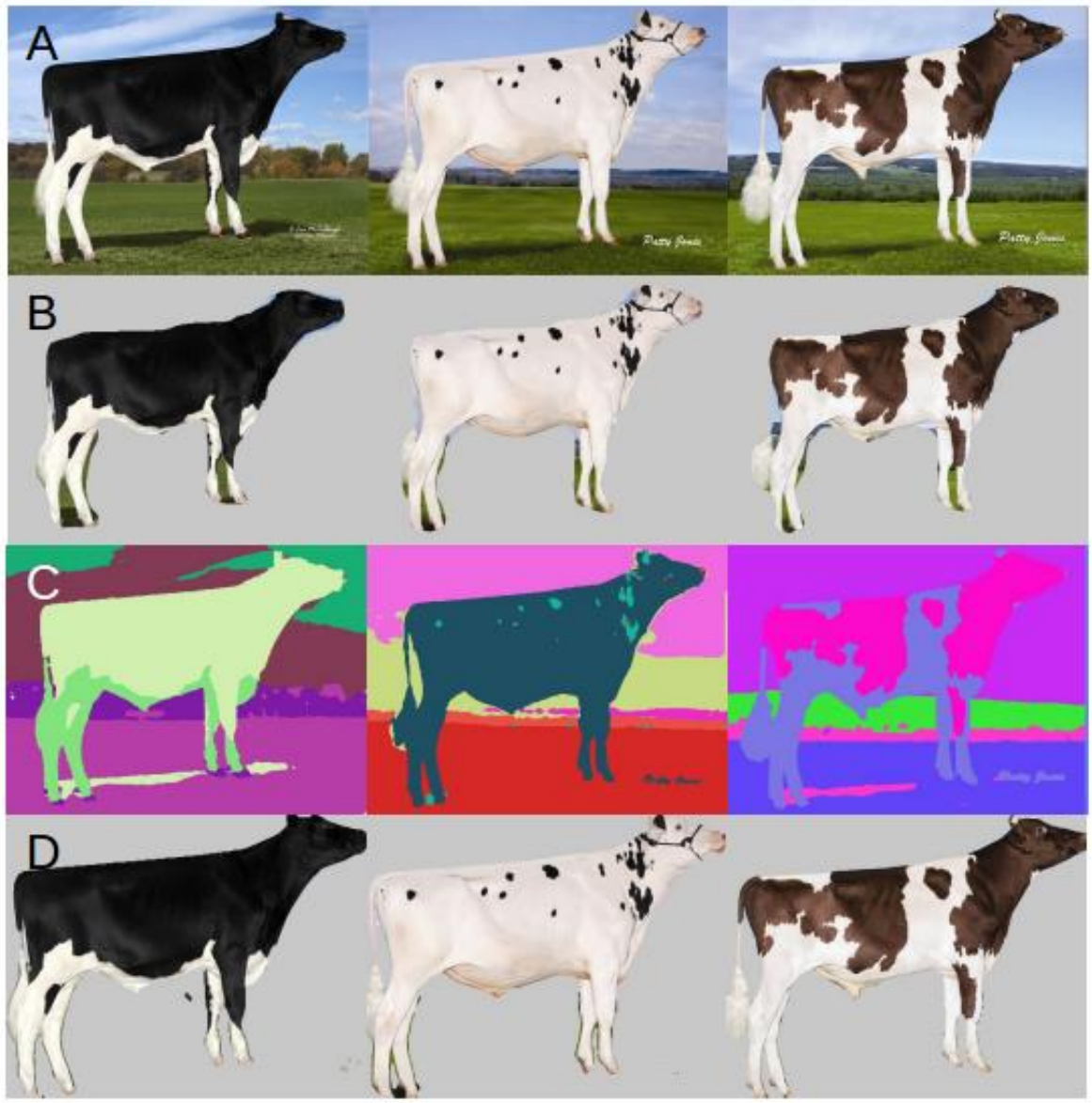


Figure 6.2. Example input and outputs. A: Original input image. B: Mask R-CNN applied mask. C: DeepAPS raw output. D: Final output of DeepAPS after all applied filters. E: Final DeepAPS mask applied to input image. F: Outline extraction of original input image. G: Extracted landmark coordinates. H: Manual color segmentation. Image from Semex.

6.2.4 Genetic parameters

To calculate heritabilities for the measured phenotypes, we extracted pedigree information and constructed a relationship matrix for each bull whenever possible. This was done by automatic web scraping in the sire catalog website, where we identified bull id, any relative type (i.e., daughter, dam, granddam, and great granddam), and their images. From the list of bull and relatives' ids, we computed the standard numerator relationship matrix, which contains the genetic relationships

assuming an infinitesimal model. Bayesian estimates of heritability were calculated with the R 3.5.2. [42] package BGLR [43] using default priors. One thousand Gibbs iterations were performed. Our sample sizes were N = 1,338 for proportion of white and N = 1,062 for morphological characteristics. The difference in sample size is due to removing any image with a missing coordinate.

6.2.5 Application to other datasets

To assess the applicability to other datasets, we chose two other objects that had been annotated in the COCO database [37], horse and giraffe, as well as two objects that had not been annotated, butterfly and duck. We downloaded 50 images from the internet that had the license set to 'labeled for noncommercial reuse' for horse and giraffe and 100 images for butterfly and duck. For the unannotated objects we annotated 50 of the images using VGG Image Annotator (VIA; [44]). These annotations were used to train a model in Mask R-CNN using the starting weights of the COCO database [37].The model was trained for 20 epochs and default parameters. Using either the COCO or custom model, DeepAPS was applied, and the composite mask was visually assessed for accuracy.

6.3 Results

We first visually compared the masks generated by the three methods that were applied to our entire dataset of 1,819 images (Figure 6.2A). When we used the supervised algorithm Mask R-CNN and applied the mask to the input images (Figure 6.2B), we observed in all cases parts of the cow body were removed along with the background (i.e., tail, nose, ear, and hoof). These masks are not satisfactorily precise to extract morphological measurements. The unsupervised segmentation by back propagation (Figure 6.2C) often separates the precise border between cow and background, but that this method on its own is not automated. Each output image would still need to be processed separately in order to match which body parts were grouped into each color cluster. DeepAPS (Figure 6.2D) across our input dataset produces a more accurate mask than Mask R-CNN and a fully automated mask, which the unsupervised approach fails to do.

In order to assess how accurately we were able to extract the true coat color percentage from each image, we compared manual and automated color segmentation. Our test set consists of 481 manually annotated images. After removing the background, we clustered each bull into one- or two-color components and extracted the percentage of dark and light colors in the coats. The automated method reports a highly accurate color segmentation with an adjusted $R^2 = 0.926$ (Figure 6.3A) when compared to manual segmentation (Figure 6.3B-D). The images that fall out as outliers belong to one of two groups, the majority of the outliers have small image sizes (less than 400 x 400 pixels), and therefore the quality was not sufficient to accurately separate the body into two color classes, the second group were bulls with a two-toned body color, in which the legs

were of a different color than the body. In these cases, the algorithm has difficulty in separating the dark-colored legs from the dark background.

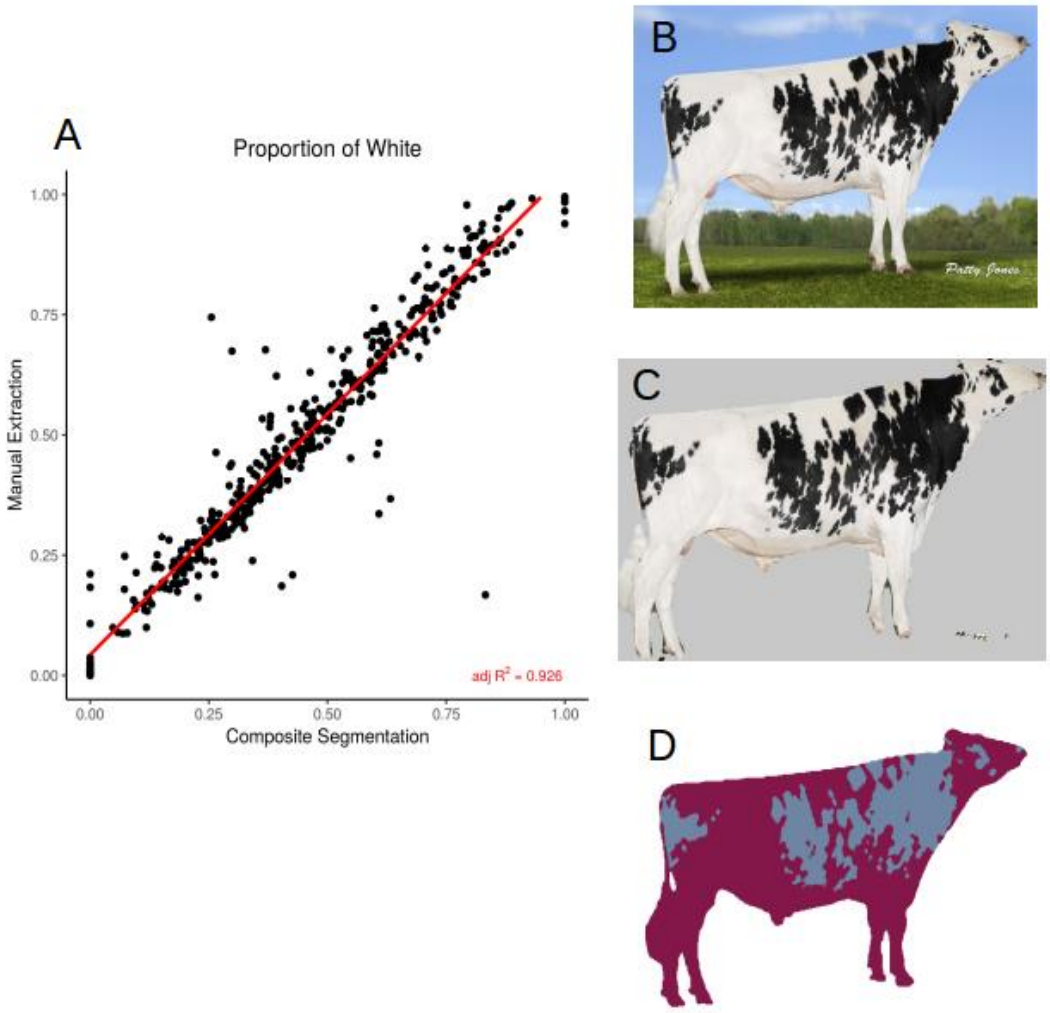


Figure 6.3. A: Correlation (adjusted $R^2 = 0.926$) between manual and automated color segmentation of 481 images. B: Example input image. C: Applied DeepAPS output mask. D: Manual color segmentation. Image from Semex.

Because the mask recovered after using this composite method is so precise, we could extract coordinates of 13 points located around the outline of the cow body (Figure 6.1G and Supplementary Figure 1) which allowed for measurements of 14 body conformation distances (see Supplementary Figure 2 for phenotypic distributions). Next, we estimated heritability using 1,338 images of related animals, in which we had partial information about great granddam, granddam, dam, bull, and daughter relationships. Our relationship matrix consists of 689 families, with an average of 2.6 individuals per family. Figure 6.4 shows the 15 posterior distributions of the heritability calculations and lists average values. Coat color proportion has the highest calculated heritability $h^2 = 0.82$, followed by body area (triangle) $h^2 = 0.43$, body area (polygon) $h^2 = 0.38$, and cow body length $h^2 = 0.34$. These values are similar to previously published results [22,45].

These high heritability measurements indicate foremost that the meaningful genetic information can be quickly and easily extracted from imaging and pedigree data available online.

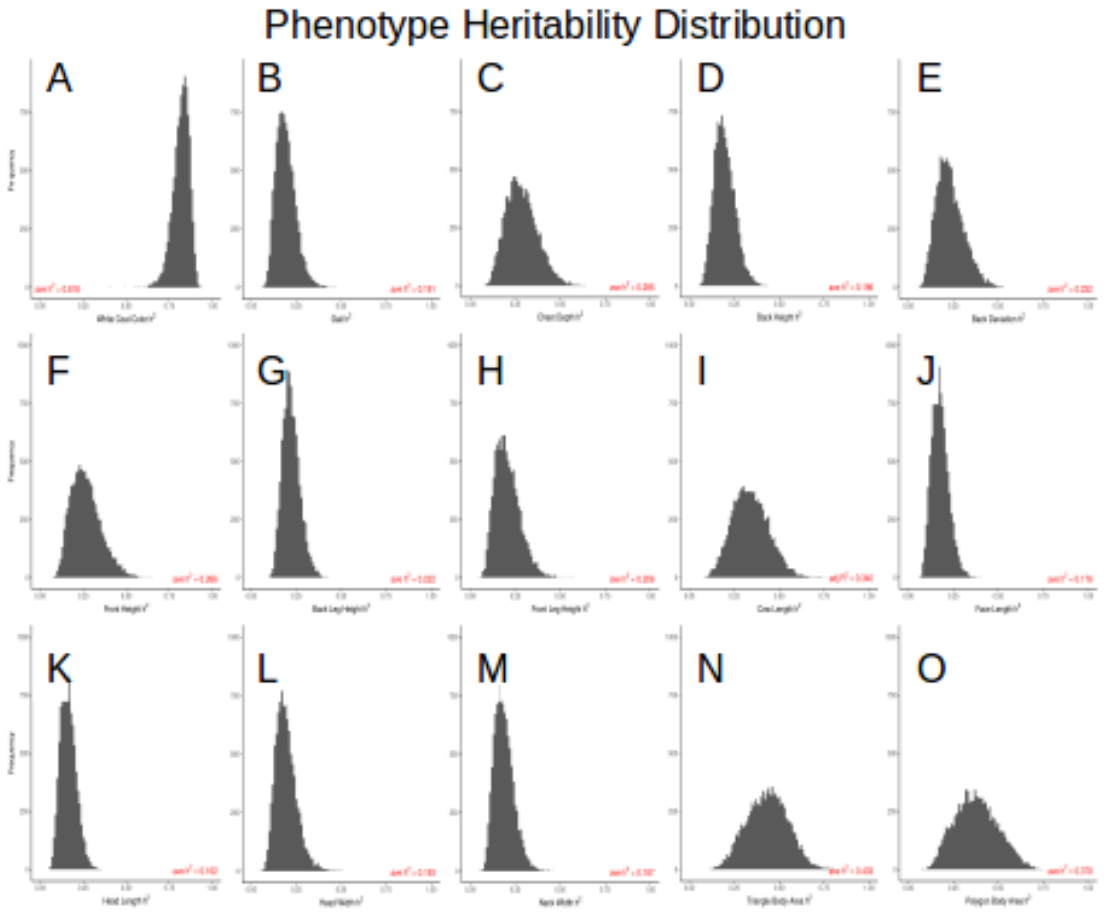


Figure 6.4. Posterior distributions of calculated heritabilities for all 15 measured phenotypes (N = 1,338 for proportion of white and N = 1,062 for morphological features). See supplementary Figure 1 for morphological measures used.

To assess whether this method is robust to the type and quality of the underlying data, we downloaded images from the internet of horse, giraffe, butterfly, and duck. These images were randomly collected, and we had no control over quality, size, lighting, or background. We also wanted to test how many input annotations are required to produce a robust mask using DeepAPS. Because the two-step method uses back propagation in order to refine the predicted mask generated from the machine learning algorithm, we hypothesized that fewer annotations would be needed. Therefore, we annotated 50 images for the butterfly and duck datasets, as they were not pre-annotated in the COCO database. We found that overall, our composite method preforms accurately (Figure 6.5). The masks generated from the thousands of annotations from the COCO dataset were precise (Figure 6.5A and 6.5B), while those based on only fifty annotations were still far more accurate than using any currently available method (Figure 6.5C and 6.5D). These results

together indicate this method is robust to input data and can still preform reliably despite being trained by few instances, making it a promising tool for automatic morphological analyses.

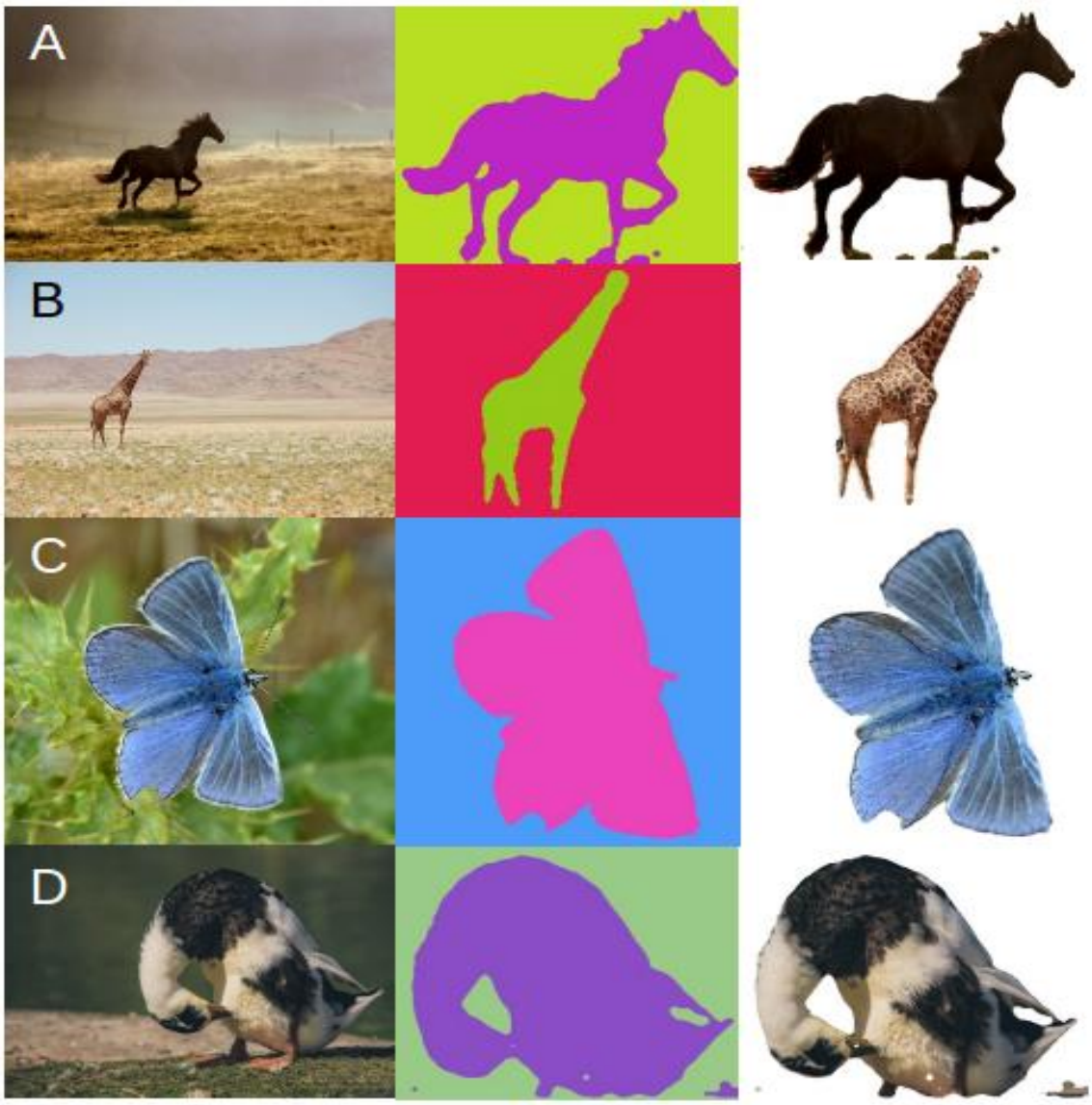


Figure 6.5. Application of DeepAPS method to four additional datasets. A: Horses and B: Giraffes trained using the COCO database. C: Butterflies and D: Ducks trained using 50 custom annotations.

6.4 Discussion

In recent decades, there have been vast improvements in molecular and statistical methods applied to animal and plant breeding. While modern livestock studies typically involve the analysis of entire genomes and/or vast number of polymorphic sites [46], high throughput phenotyping is lagging, especially in animal breeding. Often, phenotypic variation is explored today in the same manner as it was done decades ago, using simple quantifications such as length, number, categorical classifications, etc. [1,25,47]. Phenomics is extremely important in breeding programs in particular, as the desired outcome is a change in a phenotype. As phenotypes are formed by a complex process involving multiple genes, is dependent on the environment, and dynamic overtime, collecting

multiple descriptive statistics can make relating genotype to phenotype more feasible and, importantly, more meaningful.

Images are among the easiest to collect data and are underutilized. Here we combine two of the state-of-the-art image analysis tools, the supervised Mask R-CNN [36] and unsupervised segmentation [32] in order to automatically extract phenotypic measurements accurately. Not only can we create a precision mask but can cluster and segment the underlying colors and automatically measure body confirmation. Accurate image segmentation remains the most challenging part of computer vision. The ability of DeepAPS to separate the animal from multiple background types at the pixel level out preforms, for our purposes, the available algorithms currently published [32,36].

The validity and speed of this method allows for multiple quantitative morphological traits to be implemented in breeding programs. Despite the success of ongoing dairy breeding programs [48], including more and accurately quantified measurements has the potential to result in further improvements [49,50]. Furthermore, this method uses standard side-view stud images which are inexpensive to generate and store. Our presented method eliminates the high cost of phenotype collection while maintaining quality and can contribute to lowering the cost of conformational measurement collections.

Our analyses were performed on images scrubbed from the internet. As such, we had no control over backgrounds, lighting, image size, or quality. Despite the dynamic input data on which we tested DeepAPS, we were able to produce high quality masks and phenotypic measurements in most cases (Figure 6.4). Furthermore, the heritability rates we calculated from over 1,000 images of related individuals broadly agree with published results, indicating that our method accurately captures underlying information. Hayes *et al.* [22] estimated heritability of coat color percentage by manual quantification and reported a heritability of $h^2 = 0.74$ in N=327 bulls; remarkably, we found similar estimates ($h^2 = 0.81$), even if our pedigree information was quite incomplete. The reported heritability of back leg height is nearly identical to previous reports ($h^2 = 0.22$ vs. 0.21; [45]). Nevertheless, estimates of two other reported conformational heritabilities were somewhat lower: chest depth $h^2 = 0.28$ vs. 0.37 and height $h^2 = 0.27$ vs 0.42 [45]; perhaps because actual metrics analyzed here are not exactly those used in previous studies and because we cannot obtain absolute values (e.g., height in meters), since there is not a common scale across images. In all, this proof of concept shows how genetic parameters could be estimated using solely data that are already available on the web. For practical applications, more accurate estimates suitable for breeding programs could be obtained, e.g., combining SNP genotyping data with automatic image analyses from larger datasets.

While imaging data is fast and simple to collect as well as inexpensive to store, the most burdensome stage of image analysis is the generation of image annotations. We found that this method is able to leverage the publicly available COCO database and apply it to new and different

problem sets. Allowing for the creation of an accurate object mask based only on a training set of 50 instances (Figure 6.5), which is remarkably low for any machine learning approach.

This method has the potential to allow for imaging data to be easily and quickly applied to high-throughput studies, which can be highly useful and improve extant breeding programs. We provide a combined deep learning algorithm that results in highly accurate segmentation of animal profiles, which is necessary for further processing in applications related to conformational measurements. Nevertheless, we are well aware that much work remains to be done in the area. For instance, software to accurately quantify a number of additional conformational features, such as udder metrics or movement, using different angle pictures or videos should be developed. Software should also be optimized for speed and be able to analyze high-resolution pictures.

Acknowledgments

We thank comments and suggestions from José Antonio Jiménez (CONAFE, Valdemoro, Spain) and Michael Louis from Semex. LMZ was supported by a PhD grant from the Ministry of Economy and Science (MINECO, Spain), by the MINECO grant AGL2016-78709-R to MPE and from the EU through the BFU2016-77236-P (MINECO/AEI/FEDER, EU) and the "Centro de Excelencia Severo Ochoa 2016-2019" award SEV-2015-0533.

6.5 Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00513/full#supplementary-material

6.6 References

[1] D. Houle, D.R. Govindaraju, S. Omholt, Phenomics: the next challenge, (2010). https://doi.org/10.1038/nrg2897.

[2] F. Tardieu, L. Cabrera-Bosquet, T. Pridmore, M. Bennett, Plant Phenomics, From Sensors to Knowledge, Curr. Biol. 27 (2017) R770–R783. https://doi.org/10.1016/j.cub.2017.05.055.

[3] P. Guliński, K. Młynek, Z. Litwińczuk, E. Dobrogowska, Heritabilities of and genetic and phenotypic correlations between condition score and production and conformation traits in Black-and-White cows., Undefined. (2005).

[4] A. Sawa, M. Bogucki, S. Krężel-Czopek, W. Neja, Relationship between Conformation Traits and Lifetime Production Efficiency of Cows, ISRN Vet. Sci. 2013 (2013) 1–4. https://doi.org/10.1155/2013/124690.

[5] K. Jean-Pierre Mingoas, J. Awah-Ndukum, H. Dakyang, P. André Zoli, Effects of body conformation and udder morphology on milk yield of zebu cows in North region of Cameroon, (2017). https://doi.org/10.14202/vetworld.2017.901-905.

[6] L.E. Green, J. Borkert, G. Monti, N. Tadich, Associations between lesion-specific lameness and the milk yield of 1,635 dairy cows from seven herds in the Xth region of Chile and implications for management of lame dairy cows worldwide, Anim. Welf. (2010) 419–427.

[7] Å.M. Sogstad, O. Østerås, T. Fjeldaas, Bovine claw and limb disorders related to reproductive performance and production diseases, J. Dairy Sci. 89 (2006) 2519–2528.

https://doi.org/10.3168/jds.S0022-0302(06)72327-X.

[8] S.C. Kefauver, R. Vicente, O. Vergara-Díaz, J.A. Fernandez-Gallego, S. Kerfal, A. Lopez, J.P.E. Melichar, M.D. Serret Molins, J.L. Araus, Comparative UAV and Field Phenotyping to Assess Yield and Nitrogen Use Efficiency in Hybrid and Conventional Barley, Front. Plant Sci. 8 (2017). https://doi.org/10.3389/fpls.2017.01733.

[9] N. Fahlgren, M.A. Gehan, I. Baxter, Lights, camera, action: high-throughput plant phenotyping is ready for a close-up, Curr. Opin. Plant Biol. 24 (2015) 93–99. https://doi.org/10.1016/J.PBI.2015.02.006.

[10] O. Guzhva, H. Ardö, A. Herlin, M. Nilsson, K. Åström, C. Bergsten, Feasibility study for the implementation of an automatic system for the detection of social interactions in the waiting area of automatic milking stations by using a video surveillance system, Comput. Electron. Agric. 127 (2016) 506–509. https://doi.org/10.1016/j.compag.2016.07.010.

[11] N. Zehner, J.J. Niederhauser, M. Schick, C. Umstatter, Development and validation of a predictive model for calving time based on sensor measurements of ingestive behavior in dairy cows, Comput. Electron. Agric. 161 (2019) 62–71. https://doi.org/10.1016/j.compag.2018.08.037.

[12] B. Foris, A.J. Thompson, M.A.G. von Keyserlingk, N. Melzer, D.M. Weary, Automatic detection of feeding- and drinking-related agonistic behavior and dominance in dairy cows, J. Dairy Sci. 102 (2019) 9176–9186. https://doi.org/10.3168/jds.2019-16697.

[13] G. Beer, M. Alsaaod, A. Starke, G. Schuepbach-Regula, H. Müller, P. Kohler, A. Steiner, Use of Extended Characteristics of Locomotion and Feeding Behavior for Automated Identification of Lame Dairy Cows, PLoS One. 11 (2016) e0155796. https://doi.org/10.1371/journal.pone.0155796.

[14] N. Chapinal, A.M. de Passillé, M. Pastell, L. Hänninen, L. Munksgaard, J. Rushen, Measurement of acceleration while walking as an automated method for gait assessment in dairy cattle, J. Dairy Sci. 94 (2011) 2895–2901. https://doi.org/10.3168/jds.2010-3882.

[15] T. Van Hertem, V. Alchanatis, A. Antler, E. Maltz, I. Halachmi, A. Schlageter-Tello, C. Lokhorst, S. Viazzi, C.E.B. Romanini, A. Pluk, C. Bahr, D. Berckmans, Comparison of segmentation algorithms for cow contour extraction from natural barn background in side view images, Comput. Electron. Agric. 91 (2013) 65–74. https://doi.org/10.1016/j.compag.2012.12.003.

[16] X. Song, E.A.M. Bokkers, P.P.J. van der Tol, P.W.G. Groot Koerkamp, S. van Mourik, Automated body weight prediction of dairy cows using 3-dimensional vision, J. Dairy Sci. 101 (2018) 4448–4459. https://doi.org/10.3168/jds.2017-13094.

[17] U. Braun, T. Tschoner, M. Hässig, Evaluation of eating and rumination behaviour using a noseband pressure sensor in cows during the peripartum period, BMC Vet. Res. 10 (2014) 195. https://doi.org/10.1186/s12917-014-0195-6.

[18] C.-S. Chen, W.-C. Chen, Research and Development of Automatic Monitoring System for Livestock Farms, Appl. Sci. 9 (2019) 1132. https://doi.org/10.3390/app9061132.

[19] C. Palombi, M. Paolucci, G. Stradaioli, M. Corubolo, P.B. Pascolo, M. Monaci, Evaluation of remote monitoring of parturition in dairy cattle as a new tool for calving management, BMC Vet. Res. 9 (2013). https://doi.org/10.1186/1746-6148-9-191.

[20] V. Ouellet, E. Vasseur, W. Heuwieser, O. Burfeind, X. Maldague, Charbonneau, Evaluation of calving indicators measured by automated monitoring devices to predict the onset of calving in Holstein dairy cows, J. Dairy Sci. 99 (2016) 1539–1548. https://doi.org/10.3168/jds.2015-10057.

[21] A.L.N. Zhang, B.P. Wu, C.X.H. Jiang, D.C.Z. Xuan, E.Y.H. Ma, F.Y.A. Zhang, Development and validation of a visual image analysis for monitoring the body size of sheep, J. Appl. Anim. Res. 46 (2018) 1004–1015. https://doi.org/10.1080/09712119.2018.1450257.

[22] B.J. Hayes, J. Pryce, A.J. Chamberlain, P.J. Bowman, M.E. Goddard, Genetic Architecture

of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits, PLoS Genet. 6 (2010) e1001139. https://doi.org/10.1371/journal.pgen.1001139.

[23] D.F.M. Cortes, R.S. Catarina, G.B. de A. Barros, F.A.S. Arêdes, S.F. da Silveira, G.A. Ferreguetti, H.C.C. Ramos, A.P. Viana, M.G. Pereira, Model-assisted phenotyping by digital images in papaya breeding program, Sci. Agric. 74 (2017) 294–302. https://doi.org/10.1590/1678-992X-2016-0134.

[24] A. Rosero, L. Granda, J.L. Pérez, D. Rosero, W. Burgos-Paz, R. Martínez, J. Morelo, I. Pastrana, E. Burbano, A. Morales, Morphometric and colourimetric tools to dissect morphological diversity: an application in sweet potato [Ipomoea batatas (L.) Lam.], Genet. Resour. Crop Evol. 66 (2019) 1257–1278. https://doi.org/10.1007/s10722-019-00781-x.

[25] D. Houle, C. Pélabon, G.P. Wagner, T.F. Hansen, Measurement and meaning in biology, Q. Rev. Biol. 86 (2011) 3–34. https://doi.org/10.1086/658408.

[26] M. V. Boggess, J.D. Lippolis, W.J. Hurkman, C.K. Fagerquist, S.P. Briggs, A. V. Gomes, P.G. Righetti, K. Bala, The need for agriculture phenotyping: "Moving from genotype to phenotype," J. Proteomics. 93 (2013) 20–239. https://doi.org/10.1016/j.jprot.2013.03.021.

[27] M.M. Rahaman, D. Chen, Z. Gillani, C. Klukas, M. Chen, Advanced phenotyping and phenotype data analysis for the study of plant growth and development, Front. Plant Sci. 6 (2015) 619. https://doi.org/10.3389/fpls.2015.00619.

[28] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., IEEE Computer Society, 2014: pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.

[29] J. Han, D. Zhang, G. Cheng, N. Liu, D. Xu, Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey, IEEE Signal Process. Mag. 35 (2018) 84–100. https://doi.org/10.1109/MSP.2017.2749125.

[30] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377. https://doi.org/10.1016/j.patcog.2017.10.013.

[31] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, n.d.

[32] A. Kanezaki, Unsupervised image segmentation by backpropagation, in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., 2018: pp. 1543–1547. https://doi.org/10.1109/ICASSP.2018.8462533.

[33] L. Richardson, Beautiful soup documentation, April. (2007).

[34] R. Girshick, Fast R-CNN, 2015. https://github.com/rbgirshick/ (accessed March 18, 2021).

[35] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, 2015. https://github.com/ (accessed March 17, 2021).

[36] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 386–397. https://doi.org/10.1109/TPAMI.2018.2844175.

[37] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2014: pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

[38] G. Bradski, A. Kaehler, Learning OpenCV: Computer Vision in C++ with the OpenCV Library, 2nd ed., O'Reilly Media, Inc., 2013.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[40] J. Canny, A Computational Approach to Edge Detection, IEEE Trans. Pattern Anal. Mach.

Intell. PAMI-8 (1986) 679–698. https://doi.org/10.1109/TPAMI.1986.4767851.
[41] G. Van Rossum, F.L. Drake, The Python Language Reference Manual, Network Theory Ltd., 2011.
[42] R Core Team, R: A Language and Environment for Statistical Computing, (2017). https://www.r-project.org/.
[43] P. Pérez, G. De Los Campos, Genome-wide regression and prediction with the BGLR statistical package, Genetics. 198 (2014) 483–495. https://doi.org/10.1534/genetics.114.164442.
[44] A. Dutta, A. Zisserman, The VIA Annotation Software for Images, Audio and Video, MM 2019 - Proc. 27th ACM Int. Conf. Multimed. (2019) 2276–2279. https://doi.org/10.1145/3343031.3350535.
[45] T. Pritchard, M. Coffey, R. Mrode, E. Wall, Genetic parameters for production, health, fertility and longevity traits in dairy cows, Animal. 7 (2013) 34–46. https://doi.org/10.1017/S1751731112001401.
[46] G.R. Wiggans, J.B. Cole, S.M. Hubbard, T.S. Sonstegard, Genomic Selection in Dairy Cattle: The USDA Experience, Annu. Rev. Anim. Biosci. 5 (2017) 309–327. https://doi.org/10.1146/annurev-animal-021815-111422.
[47] J.B. Cole, G.R. Wiggans, L. Ma, T.S. Sonstegard, T.J. Lawlor, B.A. Crooker, C.P. Van Tassell, J. Yang, S. Wang, L.K. Matukumalli, Y. Da, Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows, BMC Genomics. 12 (2011) 408. https://doi.org/10.1186/1471-2164-12-408.
[48] G.R. Wiggans, J.B. Cole, S.M. Hubbard, T.S. Sonstegard, Genomic Selection in Dairy Cattle: The USDA Experience, Annu. Rev. Anim. Biosci. 5 (2017) 309–327. https://doi.org/10.1146/annurev-animal-021815-111422.
[49] M. Goddard, Genomic selection: Prediction of accuracy and maximisation of long term response, Genetica. 136 (2009) 245–257. https://doi.org/10.1007/s10709-008-9308-0.
[50] O. Gonzalez-Recio, M.P. Coffey, J.E. Pryce, On the value of the phenotypes in the genomic era, J. Dairy Sci. 97 (2014) 7905–7915. https://doi.org/10.3168/jds.2014-8125.

Chapter 7

Link-HD: a versatile framework to explore and integrate heterogeneous microbial communities

Laura M Zingaretti[1,2*], Gilles Renand[3], Diego P. Morgavi[4], Yuliaxis Ramayo-Caldas[3,5]

1. Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Barcelona, Spain.
2. UNVM, Villa María, Córdoba, Argentina.
3. GABI, INRA, AgroParisTech, Universi-té Paris-Saclay, 78352 Jouy-en-Josas, France.
4. INRA, Herbivore Research Unit, Saint Genès-Champanelle, France 5Animal Breeding and Genetics Program, IRTA, 08140 Caldes de Montbui, Spain.

Bioinformatics, 36(7), 2020, 2298–2299
doi: 10.1093/bioinformatics/btz862
Advance Access Publication Date: 18 November 2019
Applications Note

Systems biology

Link-HD: a versatile framework to explore and integrate heterogeneous microbial communities

Correspondence:

Laura M. Zingaretti, e-mail: m.lau.zingaretti@gmail.com

Bioinformatics

36(7), 2020, 2298–2299doi: 10.1093/bioinformatics/btz862
Advance Access Publication Date: 2019, Nov 18

Abstract

Motivation: We present Link-HD, an approach to integrate multiple datasets. Link-HD is a generalization of STATIS-ACT ('Structuration des Tableaux A Trois Indices de la Statistique – Analyse Conjointe de Tableaux'), a family of methods designed to integrate information from heterogeneous data. Here, we extend the classical approach to deal with broader datasets (e.g., compositional data), methods for variable selection, and taxon-set enrichment analysis.

Results: The methodology is demonstrated by integrating rumen microbial communities from cows for which methane yield (CH_4y) was individually measured. Our approach reproduces the significant link between rumen microbiota structure and CH_4 emission. When analyzing the TARA's ocean data, Link-HD replicates published results, highlighting the relevance of temperature with members of phyla Proteobacteria on the structure and functionality of this ecosystem.

Availability: The source code, examples, and a complete manual are freely available in GitHub https://github.com/lauzingaretti/LinkHD.

7.1 Introduction

The reduction of 'omics' technology cost now enables collection of data from multiple sources. This allows researchers to simultaneously study several datasets and investigate their relationship with complex traits. The integration of these heterogeneous datasets is not trivial and several statistical methods have been developed to address this challenge **[1–4]**. In particular, the amalgamation of multiple microbial ecosystems poses unique challenges as these are compositional and sparse data. MixKernel **[5]** is a well-known tool designed to integrate heterogeneous datasets including microbial communities, but no method to perform a taxonomic enrichment analysis is available. Another popular integrative approach is MOFA **[1],** however, it is unable to deal with compositional data.

Here, we present Link-HD, a tool to integrate and explore multiple microbial communities based on STATIS **[6]**, a family of multivariate methods to integrate multiple datasets. Link-HD generalizes STATIS with Regression Biplot **[7]**, clustering, differential abundance, enrichment taxonomic analysis, and visualization tools. Link-HD analyzes distance tables computed from numerical, categorical, or compositional data as a generalization of multidimensional scaling **[8]**. Furthermore, Link-HD performs variable selection and can link the obtained common sub-space with phenotype information.

7.2 Methods

Like STATIS, Link-HD aims to compare and analyze the relationships between datasets with a shared set of observations or variables. However, our package was specifically designed to integrate microbial communities and incorporate distances and transformations to deal with compositional data [9]. The method is implemented in three main phases (Figure 7.1).

1. **Inter-structure step:** The algorithm first assesses the similarity between transformed distance tables using the vector correlation coefficient (Rv) [10], which can be interpreted as a general "vector covariance" between matrices, i.e., this step evaluates similarity between the disparate dataset**s.**

2. **Compromise step:** Next, the 'compromise' matrix is calculated, which is a weighted sum of each cross-product matrix. This step involves an optimization problem since the weights are chosen to maximize the correlation between the compromise matrix and each individual component.

3. **Intra-structure step:** Finally, the compromise matrix is evaluated through a Principal Component Analysis. The coordinates of the common elements are projected into a low rank space, where the relationships between them can be easily interpreted.

Variable selection is tackled by two alternative approaches: 1) by projecting all the input variables into the compromise through a general Biplot formulation [7]; and 2) by computing the differential abundance of features between clusters of samples. A novelty of Link-HD is its ability to aggregate the selected variables at several taxonomic levels and to establish whether that level is enriched using a cumulative hypergeometric distribution. This function also allows users to add a custom OTUs list. Finally, the SPIEC-EASI [11] tool can be used to visualize variable interactions.

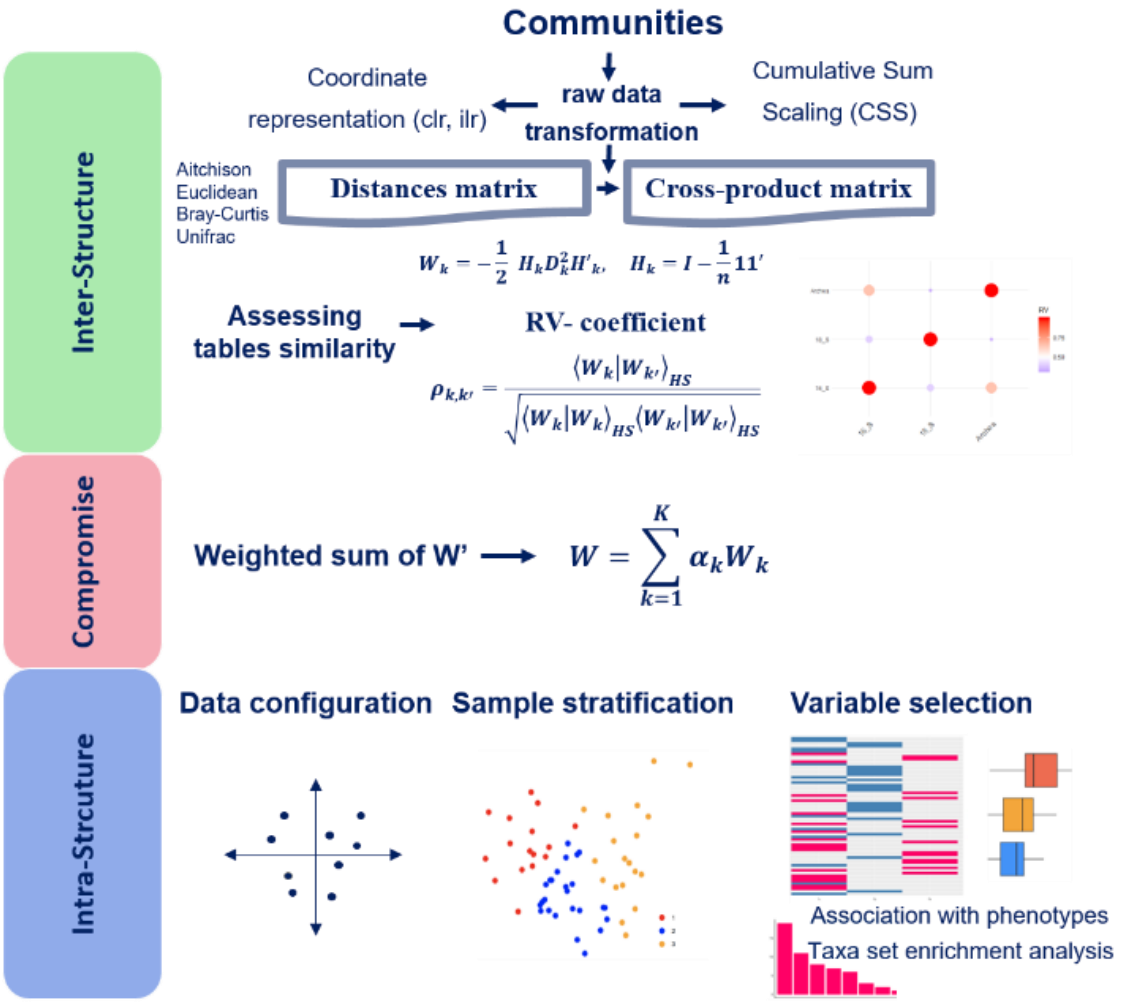


Figure 7.1. **Link-HD Workflow.** In the Inter-structure step, raw data are transformed using cumulative sum scaling (CSS) or centered log ratio (CLR), and the correlation coefficient (Rv) is computed. The second step is the compromise (W) and, finally, the intra-structure step involves the eigen-decomposition of W. Observations can be clustered and methods for selecting variables and association with phenotypes are available.

7.3 Case studies

We illustrate our approach with rumen microbial [12], TARA's Ocean expedition [13], and transcriptome NCI-60 cell line datasets [14].

In the rumen study, we integrated Bacteria, Archaea, and Protozoa from 65 Holstein cows. Link-HD was able to reproduce previous results [12,15,16], showing a link between the structure of the rumen microbiota and CH_4 emission. We also identify microbial markers associated to CH_4. In the TARA's example, Link-HD replicates the relevant role of temperature and Proteobacteria phyla on the structure of this ecosystem, as described in [5]. Finally, we show the potential of

Link-HD to integrate other omics layers by using transcriptome NCI-60 cell lines. Link-HD recapitulates the reported data structure [17] and ontology analysis reveals several cancer-related pathways.

In all, our results demonstrate that Link-HD is robust in combining several heterogeneous data types. A detailed description of these case studies and the theory behind Link-HD is available at https://lauzingaretti.github.io/LinkHD/. Of note, Link-HD is ~4 times more computationally efficient than mixKernel in a 3.3 GHz Intel Core i7 CPU with 8 GB of RAM.

7.4 Conclusions

We have developed an R package to integrate multiple microbial communities and other 'omics' layers combining a plethora of statistical methods in a fast, simple, and flexible way.

Funding

LMZ is recipient of a PhD grant from Ministry of Economy and Science, Spain associated with 'Centro de Excelencia Severo Ochoa 2016-2019' award SEV-2015-0533 to CRAG. YRC was funded by Marie Skłodowska-Curie grant (P-Sphere) agreement No 6655919 (EU).
Conflict of Interest: none declared.

7.5 References

[1] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J.C. Marioni, F. Buettner, W. Huber, O. Stegle, Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets, Mol. Syst. Biol. 14 (2018) e8124.

[2] C. Meng, B. Kuster, A.C. Culhane, A. Moghaddas Gholami, A multivariate approach to the integration of multi-omics datasets., BMC Bioinformatics. 15 (2014) 162. https://doi.org/10.1186/1471-2105-15-162.

[3] J. Mariette, N. Villa-Vialaneix, Unsupervised multiple kernel learning for heterogeneous data integration, Bioinformatics. 34 (2017) 1009–1015.

[4] F. Rohart, A. Eslami, N. Matigian, S. Bougeard, K.-A. Lê Cao, MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms, BMC Bioinformatics. 18 (2017) 128. https://doi.org/10.1186/s12859-017-1553-8.

[5] J. Mariette, N. Villa-Vialaneix, Unsupervised multiple kernel learning for heterogeneous data integration, Bioinformatics. 34 (2018) 1009–1015. https://doi.org/10.1093/bioinformatics/btx682.

[6] H.L. des Plantes, Structuration des tableaux à trois indices de la statistique: théorie et application d'une méthode d'analyse conjointe, Université des sciences et techniques du Languedoc, 1976.

[7] C.J.F. Ter Braak, JC Gower and DJ Hand, Biplots. Monographs on Statistics and Applied Probability, Psychometrika. 62 (1997) 457–460.

[8] H. Abdi, D. Valentin, S. Chollet, C. Chrea, Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications, Food Qual. Prefer. 18 (2007) 627–640. https://doi.org/10.1016/j.foodqual.2006.09.003.

[9] J. Aitchison, The Statistical Analysis of Compositional Data, J. R. Stat. Soc. Ser. B. 44 (1982) 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

[10] Y. Escoufier, Le traitement des variables vectorielles, Biometrics. (1973) 751–760.

[11] Z.D. Kurtz, C.L. Müller, E.R. Miraldi, D.R. Littman, M.J. Blaser, R.A. Bonneau, Sparse and Compositionally Robust Inference of Microbial Ecological Networks, PLOS Comput. Biol. 11 (2015) e1004226. https://doi.org/10.1371/journal.pcbi.1004226.

[12] Y. Ramayo-Caldas, L. Zingaretti, M. Popova, J. Estellé, A. Bernard, N. Pons, P. Bellot, N. Mach, A. Rau, H. Roume, M. Perez-Enciso, P. Faverdin, N. Edouard, D. Ehrlich, D.P. Morgavi, G. Renand, Identification of rumen microbial biomarkers linked to methane emission in Holstein dairy cows, J. Anim. Breed. Genet. (2019) jbg.12427. https://doi.org/10.1111/jbg.12427.

[13] S. Sunagawa, L.P. Coelho, S. Chaffron, J.R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D.R. Mende, A. Alberti, F.M. Cornejo-Castillo, P.I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J.M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B.T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T.O. Tara Oceans coordinators, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M.B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S.G. Acinas, P. Bork, Ocean plankton. Structure and function of the global ocean microbiome., Science. 348 (2015) 1261359. https://doi.org/10.1126/science.1261359.

[14] W.C. Reinhold, M. Sunshine, H. Liu, S. Varma, K.W. Kohn, J. Morris, J. Doroshow, Y. Pommier, CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set., Cancer Res. 72 (2012) 3499–511. https://doi.org/10.1158/0008-5472.CAN-12-1370.

[15] S. Kittelmann, C.S. Pinares-Patiño, H. Seedorf, M.R. Kirk, S. Ganesh, J.C. McEwan, P.H. Janssen, Two Different Bacterial Community Types Are Linked with the Low-Methane Emission Trait in Sheep, PLoS One. 9 (2014) e103171. https://doi.org/10.1371/journal.pone.0103171.

[16] R. Danielsson, J. Dicksved, L. Sun, H. Gonda, B. Müller, A. Schnürer, J. Bertilsson, Methane Production in Dairy Cows Correlates with Rumen Methanogenic and Bacterial Community Structure, Front. Microbiol. 8 (2017) 226. https://doi.org/10.3389/fmicb.2017.00226.

[17] C. Meng, A.M. Gholami, Multiple Co-inertia Analysis of Multiple OMICS Data using omicade4, (2014) 1–6.

Chapter 8

General Discussion

The main motivation behind this research was to show the rewards of using Machine Learning (ML) techniques for the new breeding challenges in the plant and animal industry. We have followed an extensive roadmap, covering many topics and going through an important collection of agronomic problems. The application of ML to questions in breeding is not only promising but has already enhanced a broad range of problems[1–4]. Despite all the progress made, there are still gaps to be filled and issues that deserve more attention. For example, the application of Genomic Selection (GS) (a method using prediction machines) in plant breeding and, in particular, in polyploid species, i.e., in numerous plant breeding programs, is still immature. High-throughput phenotyping is another factor compromising genetic progress in breeding programs. It requires the development of specific software and data management tools, which of course rely on ML technologies. It has a more or less extensive track record in plants [5] but its application in animal breeding has been scarce.

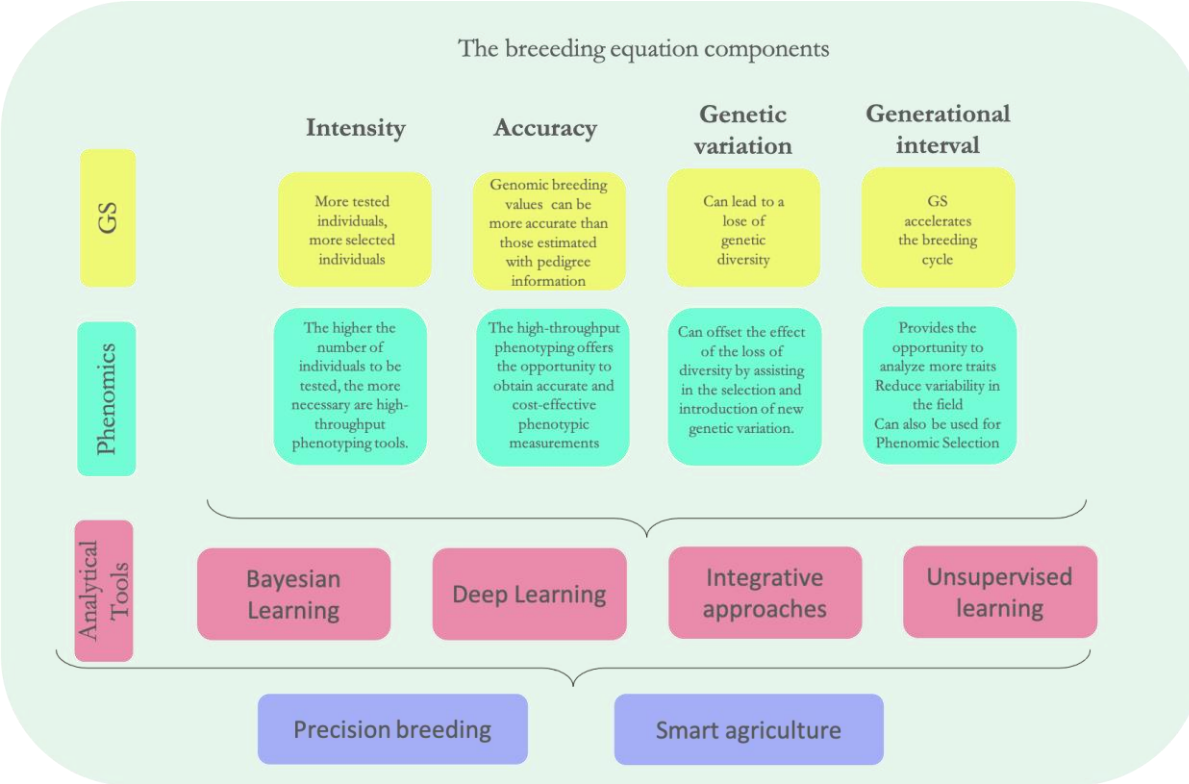


Fig. 8.1. Overview of the contribution of genomic selection and phenomics to the components of the breeding equation. Machine learning-based analysis techniques are central to analyze the huge amount of heterogeneous data being generated. Technological advances are enabling the massive collection of genomics and phenomics data at low cost, so the development of analysis methods and publicly accessible resources are key to the next generation of agriculture.

The breeding equation (i.e., how to enhance the genetic gain) is at the heart of any breeding program [6]. In this equation, the change in trait mean per year (R) is expressed as $R = \frac{i \times d_g \times r}{L}$,

where d_g is the amount of genetic variation, i is the intensity of selection, r is the accuracy of selection and L is the generation interval. It provides a quantitative framework for the identification of bottlenecks in breeding programs, as well as for the development of strategies to cope with these shortcomings, and its optimization requires the use of both, genetic and phenomics tools (Fig. 8.1).

Enhancing breeding involves enhancing the breeding equation. The contribution of GS programs to all of the components is fairly clear but massive phenotyping also plays a strategic role. (see Fig. 8.1). Breeders had been making selections based on phenotypes for a long time when molecular markers were discovered. The genetic gain is now dependent on the rapid evolution of phenomics, which particularly implies technological and techniques advances. GS and high-throughput phenotyping share their strong dependence on the innovations in data processing and robotics, making fundamental the contribution of disciplines such as computer science, machine learning, and statistics [7–10]. Let us explain the impact of GS and phenomics on each component of the breeding equation in further detail below.

Selection intensity (i) is determined by selection rate, which is the proportion of the population to be selected, so increasing the population size, automatically increases this value. Since genotyping selection is faster and cheaper than phenotyping selection, GS strategies can be used to increase i [11]. There is one important detail to keep in mind: increasing population size automatically requires selecting large populations efficiently, so high-throughput phenotyping becomes crucial [7].

It is well known that genetic variation (d_g) may be affected by GS since increasing selection pressure leads to a loss in variability [11]. To cope with this issue, breeders usually introduce external germplasm, especially in plants, but high-throughput phenotyping is also decisive for the efficient selection of new genetic variation that can be incorporated into breeding programs for the long-term sustained genetic benefit [7,8]. The continuous inflow of new variation created by mutation cannot be neglected either, as shown by long-term selection experiments[12,13].

The accuracy in the estimation of the breeding values is another factor that can be enhanced by both GS and automatic phenotyping. The Genomic Estimated Breeding Values (GEBV) can be more accurate than those estimated with pedigree information [11], but there is always room for improvement as most of the routinely collected traits depend on manual measurements, which are subjective, prone to human error, and lack repeatability. The high-throughput phenotyping offers the opportunity to obtain accurate and cost-effective measurements, which combined with massive genotyping can maximize the genetic signal, improving the effectivity of GS [7].

GS accelerates the breeding cycle and is perhaps the factor that contributes most to genetic gain because it allows the generation interval (L) to be greatly shortened. Phenomics may not provide a direct contribution in this respect but it is clear that, indirectly, it may benefit from having more (secondary) traits to analyze and from the reduction of variability in the field.

Overall, breeders are facing new challenges and they are moving from field phenotyping to phenomics, multiple- omics, and automatic, non-destructive screening of agronomic traits. Our work has addressed some bottlenecks in new breeding targets by leveraging ML and programming. The rationale behind all our papers is that ML (comprising Bayesian Learning, Deep Learning, Integrative approaches, computer vision techniques) can empower breeding. Not only phenomics but any omics (like metabolomics, epigenomics, transcriptomics, and microbiomes) have become standard [14]. However, the development of methods to integrate all these heterogeneous data sources has lagged behind. Therefore, in this thesis, we have proposed a method for the integration of multiple microbiome data and their association with a target trait. Although our example only addresses the integration of the microbiome, this approach could be used with alternative data, including massive field phenotypes.

In addition to addressing the problem of data integration, we focused on the application of GS in polyploid species, which has a much less extensive tradition than in diploid organisms or animals, and in the development of tools for automatic recording and evaluation of shape and color traits in fruits and livestock. We have covered both plant and animal breeding, as we strongly believe that, despite the differences, breeding has the same ultimate goal in common (i.e., to increase the genetic gain) and ML techniques offer the opportunity to treat them from a unified framework.

8.1 Machine Learning in plant and animal breeding

The enormous amount of data available today in any area of knowledge is making ML the hot topic behind countless scientific projects. The success of ML techniques is grounded on two main pillars: 1) ML comprises not just one, but a wide variety of data analysis methods, ranging from unsupervised (pattern recognition) to supervised (feature predictions) tasks, and 2) ML can analyze heterogeneous datasets, which can be very valuable in combining multiple data sources (SNP, gene annotation, metabolic pathways, expression data, microbiome data, etc.) [15,16].

The power of ML lies in the fact that be used to obtain the best predictive performance but also for unsupervised learning, which can deliver a comprehensive insight into the data structure. ML are the natural tools to confront the 'large p small n' problem caused by the high-throughput data collection era. As stated above, in this research work the classical methods used in Genomic Selection in animal breeding (the Bayes- alphabet like methods, [17]) are considered under the ML umbrella as they are predictive machines, while their inferential outcomes may be misleading due to the curse of the dimensionality [17].

The new explosion of data is forcing us to rethink data analysis strategies in breeding, as well as breeding objectives. MLs can drive breeding in many ways, as they are very flexible; but building an ML system is challenging, involving several steps and requiring specific expertise. Sometimes, its success critically depends on the availability of enough data, model training can be difficult and computationally expensive, while optimization is not trivial [18]. One of the main disadvantages of ML relies on the lack of interpretability, i.e., while ML are powerful predictors, their biological interpretation is quite difficult as they do not naturally provide a way to perform inference [16,19]. Another challenge facing ML is the development of tools capable of modeling genotype-

environment interactions and providing alternative ways that allow complex traits to be measured secondarily, for example inferring crop yield using images [16].

MLs open a window of opportunity for "data-driven decision agriculture": the data from electronic devices and the internet revolution in agriculture are enabling the implementation of precision farming or precision agriculture. Precision livestock farming can help in livestock management, health surveillance, production, welfare, and environmental footprint reduction. In plant breeding, meanwhile, precision farming is making it possible to detect pests, diseases, and weeds before outbreaks, is enabling effective site management of plant diseases, and customization of fertilizer, herbicide, and irrigation application in response to special pattern variability [20]. Although we have tried here to address some of the issues where ML can improve new breeding objectives, there is still much work to be done and more research in this area is urgently needed.

8.2 Handling the breeding equation in polyploids

GS assumes that a dense enough marker map can be used to select the favorable allele at each QTL without actually identifying them [21]. GS is a mature technology and is standard in many breeding programs [22–27]. Despite the outstanding success, there is still a gap in the literature regarding modeling strategies for polyploid organisms[28], including accounting for allelic dosage, for non-additive effects (i.e., dominance or epistasis), and the variance partition[29].

Polyploid organisms, i.e., those with a high level of gene redundancy, are usually classified into two categories: allo-polyploids and auto-polyploids. While the genomes of the former are well-differentiated, the latter have identical or closely related genomes [30]. This difference has practical consequences since allopolyploids can be treated as diploids, i.e., the different pairs of genomes have a strong pairing barrier, deriving bivalent formation [31]. In autotetraploid, in turn, the chromosomes are all homologous, making it possible a multivalent formation (i.e., the pairing of more than two chromosomes), which could lead to a phenomenon known as double reduction (i.e., sister chromatids segregating in the same gamete)[30,32].

GS has been applied to allopolyploids, taking advantage of their disomic inheritance, which allows them to be treated as diploids [33,34]. However, the differentiation between allo and auto-polyploids is sometimes blurred and mixed behavior between the two phenomena remains possible (i.e., 'mixosomic' inheritance, [35]) [36]. Therefore, the development of specific methodologies and analysis tools for polyploids is not only interesting from a theoretical point of view but could also bring a genetic gain in breeding. Here, we have contributed to this topic by addressing these three interrelated issues: is genomic prediction affected by the underlying genetic architecture? Is it possible to account for the effect of allelic dose on genomic prediction? how could non-linear interactions induce by polyploidy be modeled?

The first and second questions were approached harnessing the benefits of simulations to tackle complex genomic problems. We have evaluated different underlying genetic architectures for the `sugar content` trait in allopolyploid strawberry as follows: by taking 100 QTNs randomly chosen from any of the sugar-pathways SNPs (i.e., segregating in more than one homeologous group) or

by considering 100 QTNs only in those that behave as 'diploid' SNPs (i.e., we simulated the trait considering that it has a disomic inheritance). We also simulated a trait with QTNs randomly chosen throughout the genome. We envisaged two extreme gene- actions (fully additive and complete dominance) for each trait, while the genomic relationships matrices were computed using the three options available in pSBVB software (i.e., considering the full allele dosage varying between 0 and 8, using only the "diploids-like" SNPs and through the Haploid-like option, where only one full homozygous can be distinguished from the remaining genotypes). We also compared the predictive ability of genomic information with that based only on pedigree. Our simulation indicated that the predictive accuracy of GS in polyploids may critically depend on the genetic architecture, but not so much on how we compute the genomic relationship matrices.

We are not the only ones who have addressed GS-related issues in polyploids. For instance, the calculation of an additive genomic relationship matrix accounting for the allele dosage has been advanced in several studies [37–39]. In AGHmatrix [39], the authors provide a way for estimating the autotetraploid' double reduction coefficient when the relationship matrix is computed from the pedigree information following the theory developed by Kerr et al. [40]. Although these approaches are indeed interesting, they focus only on autotetraploid. Our software is more comprehensive, as the genomic relationship matrices can be computed using three different methods and take into account any ploidy level. It is a simulation tool, not just a relationship matrix generator, that can simulate preferential or not preferential pairing, making it suitable for both, auto and allopolyploids. Yet, our tool has some limitations: it can neither simulate epistasis nor compute the genomic relationship matrix to account for epistatic interactions, as provided in other studies [37,38].

The discussion of whether it is worth paying more for sequencing to obtain true allelic dosage rather than "diploidization" deserves special attention. It has been shown that polyploidy can create phenotypic variation through the allele dosage [41], so it is not surprising that taking it into account may influence predictive accuracy in GS. In our study, we found no difference in predictive ability when taking into account allelic dose could be influenced by the technology used as we will explain below. We used data from Genotyping-by-Sequencing (GBS), a method that pools many samples into a single library for sequencing: it has the advantage of being cheaper than other tools (e.g., SNP arrays), but its accuracy may be compromised if the depth of reads is poor. This issue has been addressed in the literature [38,42,43] concluding that there is an interaction between the allele dosage and the trait, i.e. observing an advantage -in terms of predictive accuracy- for certain traits and a disadvantage for others.

The third question we sought to discuss is how to model the complex relationships induced by polyploids and whether they can have an effect on prediction accuracy. We decided to apply deep learning machines since DL and, in particular convolutional neural networks (CNNs), are promising strategies to account for nonlinearity. We approach the problem by comparing DL with Bayesian machine learning (BL). The potential advantage of nonlinear BL over DL lies in the fact that it can shed light on the genetic architecture of the trait, allowing the estimation of their variance components, albeit they are not orthogonal and may be inflated. Our findings show that BL accounting for nonlinearity performs almost as complex DL techniques.

In line with previous studies [42,44,45], we found that models accounting for nonadditive effects do not show a marked improvement –in terms of predictive accuracy- over the simpler ones (Bayesian Lasso, Bayesian Ridge Regression), which only take into account additive effects. These findings provide a valuable practical conclusion from the point of view of the breeder since it would be enough to use simple frameworks, which not only are easier to understand but are less computationally demanding.

GS programs in polyploids could be implemented by assembling a training population of potential parents, as the use of family information has been shown to improve predictive accuracy [46]. Simulation tools such as pSBVB are crucial in this regard, as can be used for in silico evaluation of the next breeding population. Although further research is still needed, the findings during the last years show that GS is promising for polyploids and may be implemented similarly to diploids.

8.3 Towards digital image-based phenotyping

As discussed, GS represents only one side of the coin in breeding. We cannot forget that breeding is based on genomics and phenomics and that is precisely the reason why we have also studied the application of automatic image analysis, which allows high throughput phenotyping at a low cost.

Many devices can be used to image plants and animals for different purposes. For instance, thermal infrared cameras (TIR) quantify the leaf temperature using a technology that detects long-wave infrared radiation emitted, they can monitor plant disease [47] and, are also used in animal breeding for reproduction, thermoregulation, animal welfare, or milking process [48]. NIR (near-infrared wavelengths) cameras can measure leaf thickness or water content and can predict developmental, tolerance, and productivity traits [49]. Hyperspectral cameras can measure hundreds of spectral bands between 350 and 2500 nm and can measure drought tolerance and yield [50–52], animal color vision [53], among others. Fluorescence cameras measure chlorophyll contents, making them suitable for massive phenotyping of dynamic photosynthesis and photoprotection processes in leaves [54] or early detection of infections [55]. Finally, visible-spectrum (RGB) cameras (i.e., the conventional camera that everyone has on their cell phones) are the most basic tool for high-throughput phenotyping systems that use images of plants or animals to capture information because they are inexpensive and easily accessible. These devices are used to measure color, morphology, and geometrical aspects [5] in fruit, roots, leaves, conformational traits in animals, etc.

It is easy to see that the assortment of tools available can lead to the creation of diverse high-throughput phenotyping platforms. Each type of sensor has its own characteristics, making it difficult to construct a common analysis process. In this research work, we have only focused on the analysis of images obtained by visible spectrum cameras, suitable for the fast and accurate acquisition of morphological analysis. In livestock, appearance, and shape-related phenotypes are particularly important in dairy cattle breeding [56]. It is also a global highly interconnected activity and one of the largest livestock industries. On the other hand, shape and color are important traits in plant breeding, especially those that produce fruits, as they are closely related to quality and product value [57]. The linear descriptors (e.g., fruit height, width) may be deficient to measure a

complex and multidimensional trait such as fruit anatomy, where the underlying issue is to elucidate all the genetic and non-genetic aspects responsible for its variations. However, in many programs, shape evaluation is still a manual task, which is not only time-consuming but also inaccurate.

Analyzing shape variations has many corners and we have attempted to address of them. Images are cheap and easy to obtain but analyzing them may be a challenge since the first essential and important step of low-level vision is segmentation (i.e., the process of partitioning the image into some non-intersecting regions). Segmentation transforms the image into meaningful information (i.e., containing the boundaries of the objects) but is a difficult task, especially when the image background is non-homogeneous. There are many successful tools based on Deep Learning analysis that can segment automatically image with any degree of complexity (i.e., a complex background) [58–61], but most of them are supervised algorithms requiring a lot of annotated images. The problem is also not so simple if the image were taken under controlled conditions (i.e., a homogeneous background), because while segmentation is much simpler, it is also not obvious how to automate the analysis of hundreds of images simultaneously. In this regard, there are many examples in the literature to analyze fruit and leaves shape or plant roots [57,62–68], however, they are at best based on specialized software macros, which are only useful for segmenting the database images under consideration or can only analyze one image at a time, which is impractical if the task comprises hundreds of snapshots.

A secondary aspect we considered is how to quantify shape variations. Many morphological traits can be quantified by single measurements (length of the fruit, number of petals in a flower, etc.), but more complex ones require more complex measurements, such as proportions and relative positions of parts. On the other hand, geometric morphometrics [69,70] is a discipline that analyzes shape variation and its covariation, which can detect and visualize shape differences more clearly than classical (linear) approaches by taking advantage of multivariate statistical methods for superimposing landmark configurations of all samples in a common coordinate system.

We approached these questions in two different ways. The first one consisted of the analysis of fruit shape and color variations, which was exemplified using images of strawberries but can be used for other fruits and vegetables. In this research work, we developed an analysis workflow and a software pipeline to automatically segment and create a curated fruit database and we also take advantage of multivariate analysis and deep learning techniques for a full analysis of morphologic variations. The second work analyzed cow images where the bottleneck lies in the segmentation of complex images with non-homogeneous backgrounds. Most of the literature addressed this problem with relatively high success using Deep Supervised Learning, an expensive process requiring a large number of (annotated) training samples. Our novel purpose solves this problem using a two-step methodology: 1) We remove most of the background, obtaining the Region of Interest (ROI) of the cows' images through Mask R-CNN [61], a well- tested classification tool trained with the COCO database[71], which contains 123,287 training and validation images with 886,284 instances from 91 different categories, including cows and 2) We segmented the ROI using an unsupervised learning algorithm.

These studies have both strengths and weaknesses. Among the former, we can highlight the fact that the segmentation algorithms are based on unsupervised or semi-supervised learning, making these tools more accessible to other researchers and breeders, both methods are relatively fast and easy to use and can be applied to hundreds of images simultaneously, providing several morphological measurements. We are aware that our works are incomplete. There are many more morphological traits (e.g., leg and foot angles in cows, more relative measurements in fruits, specific measurement of internal fruit anatomy) that should be considered further. The genetical analysis to estimate the heritability of the shape variations could be more precise using molecular marker information rather than the pedigree of one or two generations. The problems of fruit shape classification and color characterization remain open. The work analyzing fruit shape is not suitable for images of fruits taken in the field and it is an issue that disserves more attention, as it could be important for automated fruit picking or fruit discarding, as well as for early disease detection. The algorithms are far from being perfect and there is significant room for further improvement of the techniques.

Most of the current research focused on morphological and structural phenotypes. However, the phenome is dynamic, varying throughout the time-life of the organisms, thus timing determination (germination, the emergence of leaves, flowers, and fruit, growth traits in animals) can provide crucial information and is an open issue. Phenomics prediction is another open issue, at least in plant breeding, where spectral images can be used to predict yield, or for predicting the biomass of the fully grown plant from early developmental stage plants [51,72]. It could complement or be an alternative to GS. It becomes a substitute in the absence of genotypic information that is more expensive than the obtention of hyperspectral images and complements GS through cost-effective screening (and filtering) of clones, allowing the application of GS only on a limited number of candidates. [49].

8.4 The future of data integration in plant and animal breeding

The advent of Next Generation Sequence (NGS) has changed the way scientific studies are carried out, enabling the collection of an enormous volume of molecular data cheaply. These techniques allow breeders to study various sources of biological information from a given organism, including the collection of genome sequencing, RNA-seq (i.e., transcriptomic), metabolite, protein, and microbiome information, among others [73,74]. NGS replaced 'old' microarrays for 'modern' RNA-seq in transcriptomic analysis, while microbiome studies shifted away from the detection of species through microarrays to the sequencing of small-subunit (16S) ribosomal RNA gene-sequence-based surveys of bacterial communities that reside in the animals' stomach, gut, etc., in the rhizosphere, endosphere, phyllosphere for plants or to the direct sequencing of microbiome metagenome DNA [75].

We have referred to phenomics throughout the manuscript, arguing that the breeding equation and the new agricultural challenges do not lie only on genotypes but also in the 'phenome', which is nothing but the sum of all organism phenotypes [76], i.e. a combination of multiple layers of information. Although the "phenome" cannot be reduced only to the 'omics' obtained by NGS technologies, as it also comprises morphology and behavioral traits, all the 'omics' are a part of the whole phenome.

The ultimate goal of data integration is to combine multiple sources of information to gain insight into the biological system. For example, ML-based data integration tools have been successfully used to predict gene function in animal breeding and to forecast 2D and 3D structure in proteome annotation and protein-protein interaction in plants and animals [18]. It is well known that the variation of complex traits may be subject to the interplay between many 'omics' layers. That is, 'omics' are promising for next-generation GS, as they may be better at predicting phenotype than SNPs alone [14].

The omics-data integration methods comprise a wide variety of tools that usually focus on two main inter-related issues. The first one is the classification of samples (plants or animals) in different subtypes based on the multi-omics profile (note that it is only a reduced view as it could also include behavioral or morphological information), while the second issue is the prediction of biomarkers related with the underlying discovered structure. Most of the tools available to integrate multiple omics data [77–79] were designed to deal with continuous microarrays chips; however, the next-generation sequence technologies pose analytical and computational challenges. Albeit the differences between the platforms and source of information, all the data generated by NGS rely on the same principle consisting in the estimation of the abundances of unique sequences, which are constrained by the library size. That is, the NGS data are compositional and the library size limiting abundances are uninformative because they contain no population information [80–82]. Unlike microarrays, NGS data are discrete and the number of zero counts can be large, especially in the case of the microbiome.

All microorganisms (bacterial, archaea, fungi, etc.) living in a given host ecosystem constitute their microbiota, while the genes they encode are the associated microbiome, which is currently one of the most popular NGS layers being studied [83,84]. Several studies have shown that the rhizosphere microbiome composition affects plant health [85,86] or that the microbiota has a key role in health status and growth characters in pigs [87,88]. The study of the microbiota in ruminants is of special interest, as shown by the innumerable works on the subject, which highlight the relationship between methane production and the host-microbiota [89–93]. Elucidating biomarkers associated with methane production may contribute to the development of new 'micro-driven' breeding programs capable of providing a sustainable solution to increase efficiency and reduce emissions from ruminant livestock.

Here we have developed a multi-omics data integration tool suitable for compositional and sparse data. Our research work was intended solely for the analysis of multiple-microbiome data and therefore incorporates ideas well suited for this context, such as differential abundances or hypergeometric tests to elucidate enriched family (genus) among the selected variables. However, the generic nature of the tool makes it easily extensible to combine not only the microbiome but also other 'omics' layers, as it can integrate multiple dissimilarity (distance) matrices into a common subspace. Our method could even be used for forecasting, i.e., the matrix generated by combining the various 'omics' layers could predict a trait of interest.

The success of any integrative approach depends heavily on good data processing, including the normalization of data filtering and elimination of batch effect, among others. We have made an

important contribution in this regard because we have fine-tuned the methods for NGS. The variable selection provided by our tool also has a major impact since it helps to refine biological hypotheses (e.g., the presence of genera/families associated with higher methane emission levels). Our proposal is not exempt from the limitations provided for their linear nature (i.e., is just a linear combination of several matrices) and the fact that can only integrate structured data (i.e., matrices generated from the same individuals). We believe that new developments enabling the combination of structured and unstructured data could contribute to gain biological insights. Finally, deep learning-based methods deserve more attention in the general problem of integrating various subspaces, as they are suitable for exploring their underlying nonlinear relationships.

The main message derived from this and related research is that new technologies are shifting the course of breeding strategies. The 'big data' paradigm is held by two major pillars: the huge amount of data being generated today and their heterogeneous nature [94]. Breeders need to adapt to this digital revolution and harness technological advances. Additional efforts are needed in the development of data analysis tools but also in the collection and management of data. Hence, agriculture is and will be sustained by ML workflows capable of performing data-driven decision approaches, i.e., methods that allow breeders efficient utilization of highly heterogeneous and complex data. We should not lose sight of the fact that the knowledge generated only has value for breeding if it can be translated into practice, i.e., transformed into genetic gain under the paradigm of sustainable Agriculture [7,14].

8.5 Bibliography

[1] R. McDowell, Genomic selection with deep neural networks, 2016. https://doi.org/10.31274/etd-180810-5600.

[2] O. González-Recio, G.J.M. Rosa, D. Gianola, Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits, Livest. Sci. 166 (2014) 217–231. https://doi.org/10.1016/j.livsci.2014.05.036.

[3] O.A. Montesinos-López, J. Martín-Vallejo, J. Crossa, D. Gianola, C.M. Hernández-Suárez, A. Montesinos-López, P. Juliana, R. Singh, A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding, G3 Genes, Genomes, Genet. 9 (2019) 601–618. https://doi.org/10.1534/g3.118.200998.

[4] J. Crossa, G. de Los Campos, P. Pérez, D. Gianola, M. Banziger, H.J. Braun, Predicting quantitative traits with regression models for dense molecular markers, Genetics. 182 (2010) 1–25. https://doi.org/10.1534/genetics.109.101501.

[5] N. Fahlgren, M.A. Gehan, I. Baxter, Lights, camera, action: high-throughput plant phenotyping is ready for a close-up, Curr. Opin. Plant Biol. 24 (2015) 93–99. https://doi.org/10.1016/J.PBI.2015.02.006.

[6] J.L. Lush, Animal bredding plans - - Google Libros, n.d. https://books.google.es/books/about/Animal_bredding_plans.html?id=4xttAAAAMAAJ&redir_esc=y (accessed March 10, 2021).

[7] J.L. Araus, S.C. Kefauver, M. Zaman-Allah, M.S. Olsen, J.E. Cairns, Translating High-Throughput Phenotyping into Genetic Gain, Trends Plant Sci. (2018). https://doi.org/10.1016/j.tplants.2018.02.001.

[8] L. Cabrera-Bosquet, J. Crossa, J. von Zitzewitz, M.D. Serret, J. Luis Araus, High-throughput Phenotyping and Genomic Selection: The Frontiers of Crop Breeding

Converge, J. Integr. Plant Biol. 54 (2012) 312–320. https://doi.org/10.1111/j.1744-7909.2012.01116.x.

[9] L. Araus, J.E. Cairns, Field high-throughput phenotyping : the new crop breeding frontier, 19 (2014). https://doi.org/10.1016/j.tplants.2013.09.008.

[10] D. Boichard, M. Brochard, New phenotypes for new breeding goals in dairy cattle, Animal. 6 (2012). https://doi.org/10.1017/S1751731112000018.

[11] J.M. Hickey, T. Chiurugwi, I. Mackay, W. Powell, Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery, Nat. Genet. 49 (2017) 1297–1303. https://doi.org/10.1038/ng.3920.

[12] R.T. Odell, The Morrow plots: a century of learning, [Urbana, Ill.]: Agricultural Experiment Station, College of Agriculture~…, 1982.

[13] W.G. Hill, Understanding and using quantitative genetic variation, (n.d.). https://doi.org/10.1098/rstb.2009.0203.

[14] A.L. Harfouche, D.A. Jacobson, D. Kainer, J.C. Romero, A.H. Harfouche, G. Scarascia Mugnozza, M. Moshelion, G.A. Tuskan, J.J.B. Keurentjes, A. Altman, Accelerating Climate Resilient Plant Breeding by Applying Next-Generation Artificial Intelligence, Trends Biotechnol. (2019). https://doi.org/10.1016/j.tibtech.2019.05.007.

[15] M. Pérez-Enciso, Animal Breeding learning from machine learning, J. Anim. Breed. Genet. 134 (2017) 85–86. https://doi.org/10.1111/jbg.12263.

[16] A.D.J. van Dijk, G. Kootstra, W. Kruijer, D. de Ridder, Machine learning in plant science and plant breeding, IScience. 24 (2021) 101890. https://doi.org/10.1016/j.isci.2020.101890.

[17] D. Gianola, Priors in whole-genome regression: The Bayesian alphabet returns, Genetics. 194 (2013) 573–596. https://doi.org/10.1534/genetics.113.151753.

[18] C. Ma, H.H. Zhang, X. Wang, Machine learning for Big Data analytics in plants, Trends Plant Sci. 19 (2014) 798–808. https://doi.org/10.1016/j.tplants.2014.08.004.

[19] M.P. Enciso, J.P. Steibel, Phenomes : the current frontier in animal breeding, Genet. Sel. Evol. (2021) 1–10. https://doi.org/10.1186/s12711-021-00618-1.

[20] R. Sharma, S.S. Kamble, A. Gunasekaran, V. Kumar, A. Kumar, A systematic literature review on machine learning applications for sustainable agriculture supply chain performance, Comput. Oper. Res. 119 (2020). https://doi.org/10.1016/j.cor.2020.104926.

[21] M. Goddard, Genomic selection: Prediction of accuracy and maximisation of long term response, Genetica. 136 (2009) 245–257. https://doi.org/10.1007/s10709-008-9308-0.

[22] G. de los Campos, J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, M.P.L. Calus, Whole-genome regression and prediction methods applied to plant and animal breeding, Genetics. 193 (2013) 327–345. https://doi.org/10.1534/genetics.112.143313.

[23] M. Shamshad, A. Sharma, The Usage of Genomic Selection Strategy in Plant Breeding, Next Gener. Plant Breed. (2018). https://doi.org/10.5772/intechopen.76247.

[24] D. Barabaschi, A. Tondelli, F. Desiderio, A. Volante, P. Vaccino, G. Valè, L. Cattivelli, Next generation breeding, Plant Sci. 242 (2015) 3–13. https://doi.org/10.1016/j.plantsci.2015.07.010.

[25] Y. Zhao, M. Gowda, W. Liu, T. Würschum, H.P. Maurer, F.H. Longin, N. Ranc, J.C. Reif, Accuracy of genomic selection in European maize elite breeding populations, Theor. Appl. Genet. 124 (2012) 769–776. https://doi.org/10.1007/s00122-011-1745-y.

[26] G. Charmet, E. Storlie, F.X. Oury, V. Laurent, D. Beghin, L. Chevarin, A. Lapierre, M.R. Perretant, B. Rolland, E. Heumez, L. Duchalais, E. Goudemand, J. Bordes, O. Robert, Genome-wide prediction of three important traits in bread wheat, Mol. Breed. 34 (2014) 1843–1852. https://doi.org/10.1007/s11032-014-0143-y.

[27] A.J. Lorenz, S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, J.L. Jannink, Genomic Selection in Plant Breeding. Knowledge and Prospects., 2011. https://doi.org/10.1016/B978-0-12-385531-2.00002-5.

[28] E. Sverrisdóttir, S. Byrne, E.H.R. Sundmark, H.Ø. Johnsen, H.G. Kirk, T. Asp, L. Janss, K.L. Nielsen, Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing, Theor. Appl. Genet. 130 (2017) 2091–2108. https://doi.org/10.1007/s00122-017-2944-y.

[29] L. Varona, A. Legarra, M.A. Toro, Z.G. Vitezica, Non-additive effects in genomic selection, Front. Genet. 9 (2018) 78-.

[30] G.L. Stebbins, Types of Polyploids: Their Classification and Significance, Adv. Genet. 1 (1947) 403–429. https://doi.org/10.1016/S0065-2660(08)60490-3.

[31] X. Yang, Y. Lv, X. Pang, C. Tong, Z. Wang, X. Li, S. Feng, C.M. Tobias, R. Wu, A unifying framework for bivalent multilocus linkage analysis of allotetraploids, Brief. Bioinform. 14 (2013) 96–108. https://doi.org/10.1093/bib/bbs011.

[32] B.A. Ronald Fisher, B. Oliver, B. Ltd, Theory of Inbreeding PRINTED AND PUBLISHED IN GREAT BRITAIN, n.d.

[33] S.A. Gezan, L.F. Osorio, S. Verma, V.M. Whitaker, An experimental validation of genomic selection in octoploid strawberry, Hortic. Res. 4 (2017) 16070. https://doi.org/10.1038/hortres.2016.70.

[34] P. Juliana, J. Poland, J. Huerta-Espino, S. Shrestha, J. Crossa, L. Crespo-Herrera, F.H. Toledo, V. Govindan, S. Mondal, U. Kumar, S. Bhavani, P.K. Singh, M.S. Randhawa, X. He, C. Guzman, S. Dreisigacker, M.N. Rouse, Y. Jin, P. Pérez-Rodríguez, O.A. Montesinos-López, D. Singh, M. Mokhlesur Rahman, F. Marza, R.P. Singh, Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics, Nat. Genet. 51 (2019) 1530–1539. https://doi.org/10.1038/s41588-019-0496-6.

[35] D.E. Soltis, C.J. Visger, D. Blaine Marchant, P.S. Soltis, Polyploidy: Pitfalls and paths to a paradigm, Am. J. Bot. 103 (2016) 1146–1166. https://doi.org/10.3732/ajb.1500501.

[36] P.M. Bourke, R.E. Voorrips, R.G.F. Visser, C. Maliepaard, Tools for genetic studies in experimental populations of polyploids, Front. Plant Sci. 9 (2018). https://doi.org/10.3389/fpls.2018.00513.

[37] A.T. Slater, N.O.I. Cogan, J.W. Forster, B.J. Hayes, H.D. Daetwyler, Improving genetic gain with genomic selection in autotetraploid potato, Plant Genome. 9 (2016). https://doi.org/10.3835/plantgenome2016.02.0021.

[38] J.B. Endelman, C.A.S. Carley, P.C. Bethke, J.J. Coombs, M.E. Clough, W.L. da Silva, Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato, Genetics. 209 (2018) 77–87.

[39] R.R. Amadeu, C. Cellon, J.W. Olmstead, A.A.F. Garcia, M.F.R. Resende, P.R. Muñoz, AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: A blueberry example, Plant Genome. 9 (2016). https://doi.org/10.3835/plantgenome2016.01.0009.

[40] R.J. Kerr, L. Li, B. Tier, G.W. Dutkowski, T.A. McRae, Use of the numerator relationship matrix in genetic analysis of autopolyploid species, Theor. Appl. Genet. 124 (2012) 1271–1282. https://doi.org/10.1007/s00122-012-1785-y.

[41] S. Yadav, P. Jackson, X. Wei, E.M. Ross, K. Aitken, E. Deomano, F. Atkin, B.J. Hayes, K.P. Voss-Fels, Accelerating Genetic Gain in Sugarcane Breeding Using Genomic Selection, Agronomy. 10 (2020) 585. https://doi.org/10.3390/agronomy10040585.

[42] I. de Bem Oliveira, M.F.R. Resende, L.F. V. Ferrão, R.R. Amadeu, J.B. Endelman, M. Kirst, A.S.G. Coelho, P.R. Munoz, Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction, G3 Genes, Genomes, Genet. 9 (2019) 1189–1198. https://doi.org/10.1534/g3.119.400059.

[43] M. Nyine, B. Uwimana, N. Blavet, E. Hřibová, H. Vanrespaille, M. Batte, V. Akech, A. Brown, J. Lorenzen, R. Swennen, J. Doležel, Genomic Prediction in a Multiploid Crop: Genotype by Environment Interaction and Allele Dosage Effects on Predictive Ability in

Banana, Plant Genome. 0 (2018) 0. https://doi.org/10.3835/plantgenome2017.10.0090.
[44] R.R. Amadeu, L.F. V. Ferrão, I. de B. Oliveira, J. Benevenuto, J.B. Endelman, P.R. Munoz, Impact of Dominance Effects on Autotetraploid Genomic Prediction, Crop Sci. 0 (2019) 0. https://doi.org/10.2135/cropsci2019.02.0138.
[45] F. Enciso-Rodriguez, D. Douches, M. Lopez-Cruz, J. Coombs, G. de los Campos, Genomic selection for late blight and common scab resistance in tetraploid potato (Solanum tuberosum), G3 Genes, Genomes, Genet. 8 (2018) 2471–2481. https://doi.org/10.1534/g3.118.200273.
[46] I. de Bem Oliveira, R.R. Amadeu, L.F.V. Ferrão, P.R. Muñoz, Optimizing whole-genomic prediction for autotetraploid blueberry breeding, Heredity (Edinb). 125 (2020) 437–448. https://doi.org/10.1038/s41437-020-00357-x.
[47] C.M. Ortiz-Bustos, M.L. Pérez-Bueno, M. Barón, L. Molinero-Ruiz, Use of Blue-Green Fluorescence and Thermal Imaging in the Early Detection of Sunflower Infection by the Root Parasitic Weed Orobanche cumana Wallr., Front. Plant Sci. 8 (2017) 833. https://doi.org/10.3389/fpls.2017.00833.
[48] I. KNÍŽKOVÁ Petr KUNC, G. Alp Kağan GÜRDİL Yunus PINAR Kemal Çağatay SELVİ OMU, APPLICATIONS OF INFRARED THERMOGRAPHY IN ANIMAL PRODUCTION, Anadolu Tarım Bilim. Derg. 22 (2007) 329–336. https://doi.org/10.7161/anajas.2007.22.3.329-336.
[49] R. Rincent, J.P. Charpentier, P. Faivre-Rampant, E. Paux, J. Le Gouis, C. Bastien, V. Segura, Phenomic selection is a low-cost and high-throughput method based on indirect predictions: Proof of concept on wheat and poplar, G3 Genes, Genomes, Genet. 8 (2018) 3961–3972. https://doi.org/10.1534/g3.118.200760.
[50] F.M. Aguate, S. Trachsel, L.G. Pérez, J. Burgueño, J. Crossa, M. Balzarini, D. Gouache, M. Bogard, G. de los Campos, Use of Hyperspectral Image Data Outperforms Vegetation Indices in Prediction of Maize Yield, Crop Sci. 57 (2017) 2517–2524. https://doi.org/10.2135/cropsci2017.01.0007.
[51] M. Lopez-Cruz, E. Olson, G. Rovere, J. Crossa, S. Dreisigacker, S. Mondal, R. Singh, G. de los Campos, Regularized selection indices for breeding value prediction using hyper-spectral image data, Sci. Rep. 10 (2020) 1–12. https://doi.org/10.1038/s41598-020-65011-2.
[52] S. Trachsel, T. Dhliwayo, L.G. Perez, J.A.M. Lugo, M. Trachsel, Estimation of physiological genomic estimated breeding values (PGEBV) combining full hyperspectral and marker data across environments for grain yield under combined heat and drought stress in tropical maize (Zea mays L.), PLoS One. 14 (2019). https://doi.org/10.1371/journal.pone.0212200.
[53] N.E. Nevala, T. Baden, A low-cost hyperspectral scanner for natural imaging and the study of animal colour vision above and under water, Sci. Rep. 9 (2019) 1–14. https://doi.org/10.1038/s41598-019-47220-6.
[54] L. McAusland, J.A. Atkinson, T. Lawson, E.H. Murchie, High throughput procedure utilising chlorophyll fluorescence imaging to phenotype dynamic photosynthesis and photoprotection in leaves under controlled gaseous conditions, Plant Methods. 15 (2019) 109. https://doi.org/10.1186/s13007-019-0485-x.
[55] E. Bauriegel, H. Brabandt, U. Gärber, W.B. Herppich, Chlorophyll fluorescence imaging to facilitate breeding of Bremia lactucae-resistant lettuce cultivars, Comput. Electron. Agric. 105 (2014) 74–82. https://doi.org/10.1016/j.compag.2014.04.010.
[56] T. Van Hertem, V. Alchanatis, A. Antler, E. Maltz, I. Halachmi, A. Schlageter-Tello, C. Lokhorst, S. Viazzi, C.E.B. Romanini, A. Pluk, C. Bahr, D. Berckmans, Comparison of segmentation algorithms for cow contour extraction from natural barn background in side view images, Comput. Electron. Agric. 91 (2013) 65–74. https://doi.org/10.1016/j.compag.2012.12.003.

[57] M.J. Feldmann, M.A. Hardigan, R.A. Famula, C.M. López, A. Tabb, G.S. Cole, S.J. Knapp, Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry, Gigascience. 9 (2020) 1–17. https://doi.org/10.1093/gigascience/giaa030.

[58] A.F.A. Fernandes, E.M. Turra, É.R. de Alvarenga, T.L. Passafaro, F.B. Lopes, G.F.O. Alves, V. Singh, G.J.M. Rosa, Deep Learning image segmentation for extraction of fish body measurements and prediction of body weight and carcass traits in Nile tilapia, Comput. Electron. Agric. 170 (2020) 105274. https://doi.org/10.1016/j.compag.2020.105274.

[59] S. Mao, Y. Li, Y. Ma, B. Zhang, J. Zhou, Kai Wang, Automatic cucumber recognition algorithm for harvesting robots in the natural environment using deep learning and multi-feature fusion, Comput. Electron. Agric. 170 (2020). https://doi.org/10.1016/j.compag.2020.105254.

[60] K. Lin, L. Gong, Y. Huang, C. Liu, J. Pan, Deep Learning-Based Segmentation and Quantification of Cucumber Powdery Mildew Using Convolutional Neural Network, Front. Plant Sci. 10 (2019) 155. https://doi.org/10.3389/fpls.2019.00155.

[61] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, n.d.

[62] M.T. Brewer, J.B. Moyseenko, A.J. Monforte, E. Van Der Knaap, Morphological variation in tomato: A comprehensive study of quantitative trait loci controlling fruit shape and development, J. Exp. Bot. 58 (2007) 1339–1349. https://doi.org/10.1093/jxb/erl301.

[63] M.T. Brewer, L. Lang, K. Fujimura, N. Dujmovic, S. Gray, E. Van Der Knaap, Development of a controlled vocabulary and software application to analyze fruit shape variation in tomato and other plant species, Plant Physiol. 141 (2006) 15–25. https://doi.org/10.1104/pp.106.077867.

[64] Z. Migicovsky, M. Li, D.H. Chitwood, S. Myles, Morphometrics reveals complex and heritable apple leaf shapes, Front. Plant Sci. 8 (2018) 1–14. https://doi.org/10.3389/fpls.2017.02185.

[65] M.A. Gehan, N. Fahlgren, A. Abbasi, J.C. Berry, S.T. Callen, L. Chavez, A.N. Doust, M.J. Feldman, K.B. Gilbert, J.G. Hodge, J.S. Hoyer, A. Lin, S. Liu, C. Lizárraga, A. Lorence, M. Miller, E. Platon, M. Tessman, T. Sax, PlantCV v2: Image analysis software for high-throughput plant phenotyping, PeerJ. 2017 (2017) e4088. https://doi.org/10.7717/peerj.4088.

[66] J. Schindelin, C.T. Rueden, M.C. Hiner, K.W. Eliceiri, The ImageJ ecosystem: An open platform for biomedical image analysis, Mol. Reprod. Dev. 82 (2015) 518–529. https://doi.org/10.1002/MRD.22489.

[67] A. Das, H. Schneider, J. Burridge, A.K.M. Ascanio, T. Wojciechowski, C.N. Topp, J.P. Lynch, J.S. Weitz, A. Bucksch, Digital imaging of root traits (DIRT): A high-throughput computing and collaboration platform for field-based root phenomics, Plant Methods. 11 (2015) 1–12. https://doi.org/10.1186/s13007-015-0093-3.

[68] J.A. Atkinson, M.P. Pound, M.J. Bennett, D.M. Wells, Uncovering the hidden half of plants using new advances in root phenotyping, Curr. Opin. Biotechnol. 55 (2019) 1–8. https://doi.org/10.1016/j.copbio.2018.06.002.

[69] D.C. Adams, E. Otárola-Castillo, Geomorph: An r package for the collection and analysis of geometric morphometric shape data, Methods Ecol. Evol. 4 (2013) 393–399. https://doi.org/10.1111/2041-210X.12035.

[70] D.E. Slice, Geometric morphometrics, Annu. Rev. Anthropol. 36 (2007) 261–281. https://doi.org/10.1146/annurev.anthro.34.081804.120613.

[71] T. Lin, C.L. Zitnick, P. Doll, Microsoft COCO : Common Objects in Context, (n.d.) 1–15.

[72] C. Edlich-Muth, M.M. Muraya, T. Altmann, J. Selbig, Phenomic prediction of maize hybrids, BioSystems. 146 (2016) 102–109. https://doi.org/10.1016/j.biosystems.2016.05.008.

[73] S.G. Crandall, K.M. Gold, M. del M. Jiménez-Gasco, C.C. Filgueiras, D.S. Willett, A multi-

omics approach to solving problems in plant disease ecology, PLoS One. 15 (2020) e0237975. https://doi.org/10.1371/journal.pone.0237975.

[74] Y. lan YANG, R. ZHOU, K. LI, Future livestock breeding: Precision breeding based on multi-omics information and population personalization, J. Integr. Agric. 16 (2017) 2784–2791. https://doi.org/10.1016/S2095-3119(17)61780-5.

[75] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D.R. Mende, G.R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H.B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E.G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W.M. de Vos, S. Brunak, J. Doré, M. Antolín, F. Artiguenave, H.M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K.U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Mérieux, R. Melo Minardi, C. M'rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S.D. Ehrlich, P. Bork, Enterotypes of the human gut microbiome, Nature. 473 (2011) 174–180. https://doi.org/10.1038/nature09944.

[76] D. Houle, D.R. Govindaraju, S. Omholt, Phenomics: the next challenge, (2010). https://doi.org/10.1038/nrg2897.

[77] F. Rohart, B. Gautier, A. Singh, K.-A. Lê Cao, mixOmics: An R package for 'omics feature selection and multiple data integration, PLOS Comput. Biol. 13 (2017) e1005752. https://doi.org/10.1371/journal.pcbi.1005752.

[78] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J.C. Marioni, F. Buettner, W. Huber, O. Stegle, Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets, Mol. Syst. Biol. 14 (2018) e8124.

[79] C. Meng, B. Kuster, A. Culhane, A.M. Gholami, A multivariate approach to the integration of multi-omics datasets., BMC Bioinformatics. (2013).

[80] G.B. Gloor, J.M. Macklaim, V. Pawlowsky-Glahn, J.J. Egozcue, Microbiome datasets are compositional: And this is not optional, Front. Microbiol. 8 (2017). https://doi.org/10.3389/fmicb.2017.02224.

[81] T.P. Quinn, I. Erb, M.F. Richardson, T.M. Crowley, Understanding sequencing data as compositions: an outlook and review, (n.d.). https://doi.org/10.1093/bioinformatics/bty175.

[82] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M.W. Szcześniak, D.J. Gaffney, L.L. Elo, X. Zhang, A. Mortazavi, A survey of best practices for RNA-seq data analysis, Genome Biol. 17 (2016) 13. https://doi.org/10.1186/s13059-016-0881-8.

[83] M. Gopal, A. Gupta, Microbiome selection could spur next-generation plant breeding strategies, Front. Microbiol. 7 (2016) 1971. https://doi.org/10.3389/fmicb.2016.01971.

[84] S.A. Huws, C.J. Creevey, L.B. Oyama, I. Mizrahi, S.E. Denman, M. Popova, R. Muñoz-Tamayo, E. Forano, S.M. Waters, M. Hess, I. Tapio, H. Smidt, S.J. Krizsan, D.R. Yáñez-Ruiz, A. Belanche, L. Guan, R.J. Gruninger, T.A. McAllister, C.J. Newbold, R. Roehe, R.J. Dewhurst, T.J. Snelling, M. Watson, G. Suen, E.H. Hart, A.H. Kingston-Smith, N.D. Scollan, R.M. do Prado, E.J. Pilau, H.C. Mantovani, G.T. Attwood, J.E. Edwards, N.R. McEwan, S. Morrisson, O.L. Mayorga, C. Elliott, D.P. Morgavi, Addressing Global Ruminant Agricultural Challenges Through Understanding the Rumen Microbiome: Past, Present, and Future., Front. Microbiol. 9 (2018) 2161. https://doi.org/10.3389/fmicb.2018.02161.

[85] G. Berg, The plant microbiome and its importance for plant and human health, Front.

Microbiol. 5 (2014) 1. https://doi.org/10.3389/fmicb.2014.00491.

[86] J.M. Chaparro, A.M. Sheflin, D.K. Manter, J.M. Vivanco, Manipulating the soil microbiome to increase soil health and plant fertility, Biol. Fertil. Soils. 48 (2012) 489–499. https://doi.org/10.1007/s00374-012-0691-4.

[87] Y. Ramayo-Caldas, N. Mach, P. Lepage, F. Levenez, C. Denis, G. Lemonnier, J.-J. Leplat, Y. Billon, M. Berri, J. Doré, C. Rogel-Gaillard, J. Estellé, Phylogenetic network analysis applied to pig gut microbiota identifies an ecosystem structure linked with growth traits, ISME J. 10 (2016) 2973–2977. https://doi.org/10.1038/ismej.2016.77.

[88] J.M. Fouhse, R.T. Zijlstra, B.P. Willing, The role of gut microbiota in the health and disease of pigs, Anim. Front. 6 (2016) 30–36. https://doi.org/10.2527/af.2016-0031.

[89] G.F. Difford, D.R. Plichta, P. Løvendahl, J. Lassen, S.J. Noel, O. Højberg, A.-D.G. Wright, Z. Zhu, L. Kristensen, H.B. Nielsen, B. Guldbrandtsen, G. Sahana, Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows, PLOS Genet. 14 (2018) e1007580. https://doi.org/10.1371/journal.pgen.1007580.

[90] N. Cancino, J. Romero, S. Huws, E. Vargas, Effect of dietary olive oil and palm oil on rumen bacterial composition in dairy cows, Access Microbiol. 1 (2019). https://doi.org/10.1099/acmi.ac2019.po0576.

[91] S.A. Huws, C.J. Creevey, L.B. Oyama, I. Mizrahi, S.E. Denman, M. Popova, R. Muñoz-Tamayo, E. Forano, S.M. Waters, M. Hess, I. Tapio, H. Smidt, S.J. Krizsan, D.R. Yáñez-Ruiz, A. Belanche, L. Guan, R.J. Gruninger, T.A. McAllister, C.J. Newbold, R. Roehe, R.J. Dewhurst, T.J. Snelling, M. Watson, G. Suen, E.H. Hart, A.H. Kingston-Smith, N.D. Scollan, R.M. do Prado, E.J. Pilau, H.C. Mantovani, G.T. Attwood, J.E. Edwards, N.R. McEwan, S. Morrisson, O.L. Mayorga, C. Elliott, D.P. Morgavi, Addressing Global Ruminant Agricultural Challenges Through Understanding the Rumen Microbiome: Past, Present, and Future, Front. Microbiol. 9 (2018) 2161. https://doi.org/10.3389/fmicb.2018.02161.

[92] Y. Ramayo-Caldas, L. Zingaretti, M. Popova, J. Estellé, A. Bernard, N. Pons, P. Bellot, N. Mach, A. Rau, H. Roume, M. Perez-Enciso, P. Faverdin, N. Edouard, D. Ehrlich, D.P. Morgavi, G. Renand, Identification of rumen microbial biomarkers linked to methane emission in Holstein dairy cows, J. Anim. Breed. Genet. (2019) jbg.12427. https://doi.org/10.1111/jbg.12427.

[93] R. John Wallace, G. Sasson, P.C. Garnsworthy, I. Tapio, E. Gregson, P. Bani, P. Huhtanen, A.R. Bayat, F. Strozzi, F. Biscarini, T.J. Snelling, N. Saunders, S.L. Potterton, J. Craigon, A. Minuti, E. Trevisi, M.L. Callegari, F.P. Cappelli, E.H. Cabezas-Garcia, J. Vilkki, C. Pinares-Patino, K.O. Fliegerová, J. Mrázek, H. Sechovcová, J. Kopečný, A. Bonin, F. Boyer, P. Taberlet, F. Kokou, E. Halperin, J.L. Williams, K.J. Shingfield, I. Mizrahi, A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions, Sci. Adv. 5 (2019) 8391–8394. https://doi.org/10.1126/sciadv.aav8391.

[94] J. Fan, F. Han, H. Liu, Challenges of Big Data analysis, Natl. Sci. Rev. 1 (2014) 293–314. https://doi.org/10.1093/nsr/nwt032.

Chapter 9

General Conclusions

- Performance of Genomic Selection (GS) in polyploids critically depends on the underlying genetic architecture of the trait. In turn, how the genomic relationship matrix is computed (i.e., accounting for the full level of ploidy or not) does not affect the accuracy of Genomic Prediction (GP). 'Diploidization', where all heterozygous genotypes are pooled, is as efficient and accurate as polyploid genotyping for prediction purposes.

- A Bayesian model capable of accounting for nonlinearity behaves in much the same way as complex deep learning techniques for predicting complex traits in polyploids with the advantage of providing an estimate (although not orthogonal) of the different sources of variation.

- The performance of deep learning modeling is highly dependent on the hyperparameter tuning and must be adjusted for each trait individually. In most cases, shallow networks are the best architecture.

- Shape-related phenotypes in fruits and animals can be automatically retrieved through automatic analysis of digital images. A major advantage is that the provided methods only used unsupervised or semi-supervised segmentation for feature extraction. We show shape-related phenotypes are moderately heritable.

- Deep generative networks are promising tools to generate synthetic fruit images, which could have important implications for breeding, as it is an easy way to synthetically produce the appearance of a given organism (e.g., fruit, animal) conditioned to a genotype.

- We provide a tool to integrate the microbiome and other omics layers, taking into account the sparse and non-continuous nature of sequence data, which was able to identify microbial signatures related to methane emission in cows. We propose a test to carry out 'taxa set enrichment analysis', which facilitates the biological interpretation of the "microbial signatures".

Annexes

Supplementary Material Chapter 5

Algorithm Sup 1:

This is the pseudo-code used for automatic morphology analysis for strawberry images. The code can analyze any fruit shape in similar conditions (homogeneous background) and it can be easily extended for custom purposes.

Algorithm 1: Create a segmented fruit database from raw data

n: number of images to process.
for i=1 to n **do**:

1. Read image
2. Convert RGB/BGR image into a grayscale image.
3. Smooth image using Gaussian filtering
4. Binarize image using mean based adaptative thresholding methodology
5. Apply erosion + dilation (Opening operation)
6. To obtain image contours

sh=[] (empty list)
for c in contours **do**:

1. Obtain *h, w* (contour height and width)
2. Obtain x, y (contour position in the main image)
3. **If** 1.1<(h/w)<3 (What is the expected aspect for fruits?) :
   sh+=c
   Analyze sh color pattern (to determine whether it is the inside or outside of the fruits) (Skip this step if your fruits are all inner (outer))
4. Get the ROI (Region of interest) from the RGB/BGR image.
5. Create an equal size image for each contour

7. Use OCR (Optical character recognition) to read the image label
8. Output a folder named by (7) containing the sample images (it splits inner/outer if necessary)

Algorithm available at https://github.com/lauzingaretti/DeepAFS

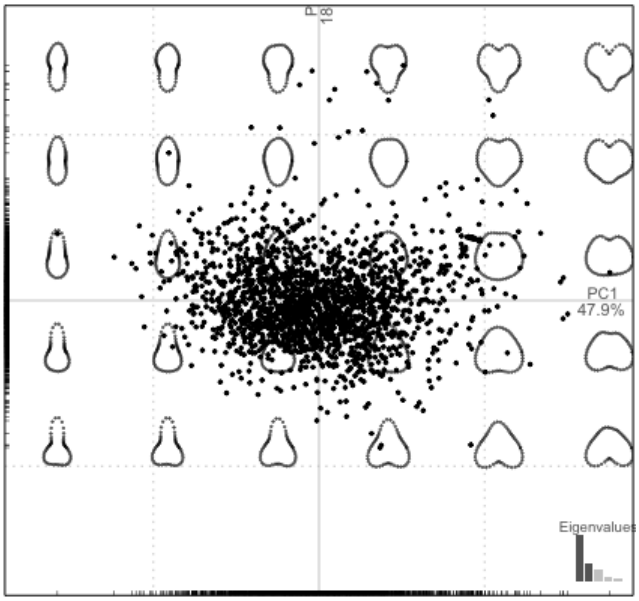
Cross number	Male	Female	Number of seedlings
1	P-01	P-06	20
2	P-27	P-06	20
3	P-16	P-06	20
4	P-15	P-06	20
5	P-13	P-11	20
6	P-25	P-11	20
7	P-31	P-11	20
8	P-11	P-19	20
9	P-22	P-16	20
10	P-17	P-16	20
11	P-28	P-23	20
12	P-16	P-23	20
13	P-03	P-07	20
14	P-24	P-07	20
15	P-08	P-07	20
16	P-12	P-02	20
17	P-04	P-09	20
18	P-23	P-09	20
19	P-20	P-10	20
20	P-30	P-10	20
21	P-18	P-10	19
22	P-21	P-10	19
23	P-14	P-05	20
24	P-18	P-29	20
Total number of seedlings			478
Total number of parentals			30
Total number of individuals			508

Supp Table 1. Scheme of the crosses used in the experiment. This consisted of 24 crosses among 30 parentals, with 20 individuals for all the crosses, except for 2 that only contained 19 individuals.

Male	Female	Seedling
P-23	P-24	P-01
P-08	P-03	P-02
P-32	P-39	P-03
unknown	unknown	P-04
unknown	unknown	P-05
P-33	P-40	P-06
P-34	P-43	P-07
P-35	P-44	P-08
unknown	unknown	P-09
P-21	P-41	P-10
P-45	P-42	P-11
P-36	P-41	P-12
unknown	unknown	P-13
P-37	P-12	P-14
unknown	unknown	P-15
P-24	P-48	P-16
P-38	P-08	P-17
P-08	P-46	P-18
P-03	P-12	P-19
unknown	unknown	P-20
P-33	P-42	P-21
unknown	unknown	P-22
P-33	P-32	P-23
unknown	unknown	P-24
P-24	P-42	P-25
unknown	unknown	P-27
P-03	P-46	P-28
P-36	P-47	P-29
unknown	unknown	P-30
unknown	unknown	P-31

Supp Table 2. Scheme of the pedigree used in the analysis.

Procrustes PCA: Procrustes Principal Component analysis (Proc-PCA) on fruit shape. We evaluated the effect of the crosses on the fruit shape through a Procrustes analysis of variance using residual randomization permutation procedure with 101 permutations.



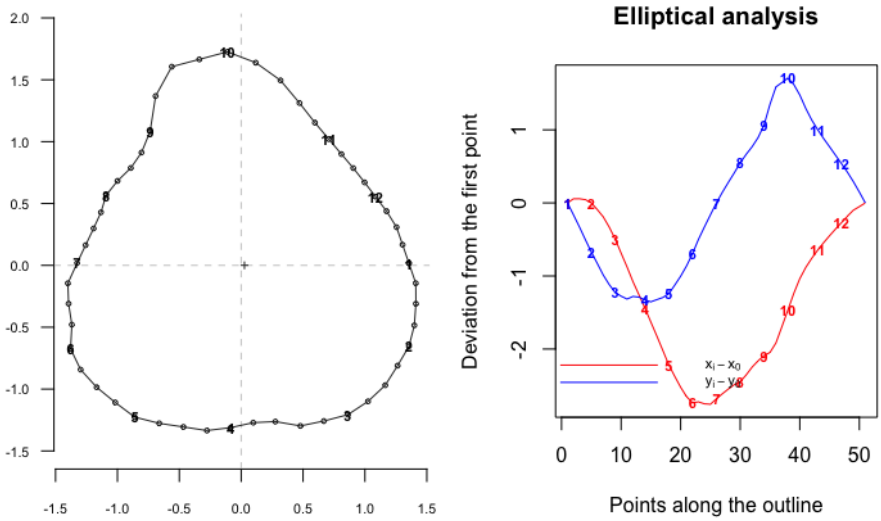
Supp Fig. 1 Output plot from Procrustes Principal Component Analysis (Proc-PCA). The analysis shows a variation between 'elongated' and 'globose'-like shape.

	Df	SS	MS	Rsq	F	Z.	Pr(>F)
crosses	24	0.5388	0.0224483	0.05525	4.5615	7.4238	0.009901 **
Residuals	1872	9.2126	0.0049213	0.94475			
Total	1896	9.7514					

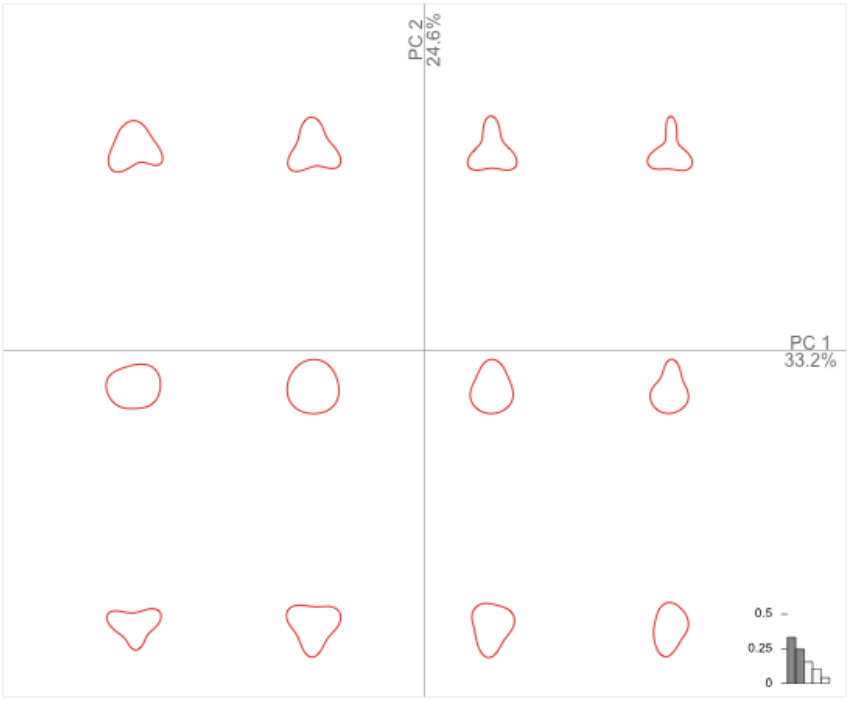
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Supp Table 3. Output from Procrustes ANOVA, which evaluates the effect of the crosses on fruit shape.

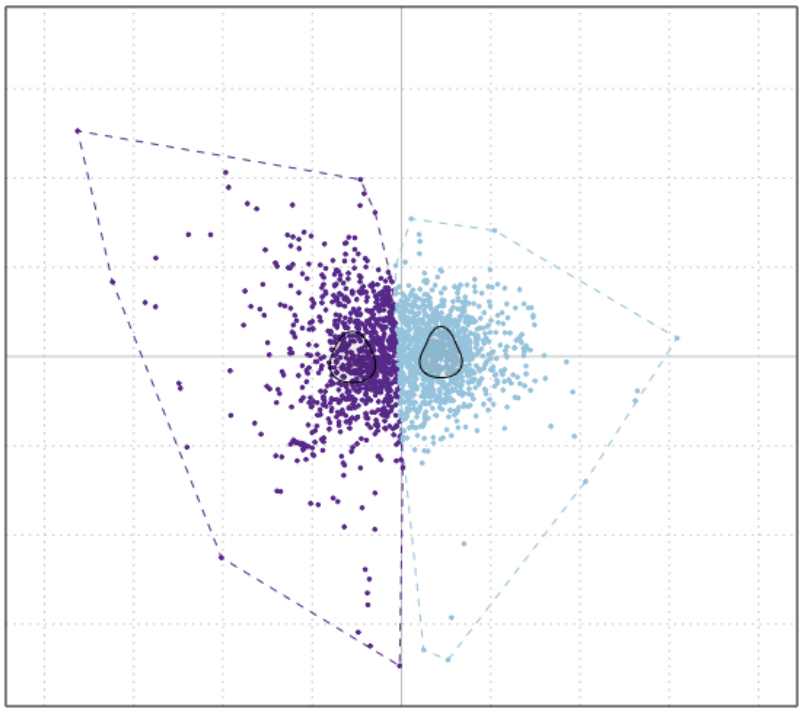
Fourier Analysis



Supp Fig. 2. Output from elliptical Fourier analysis.

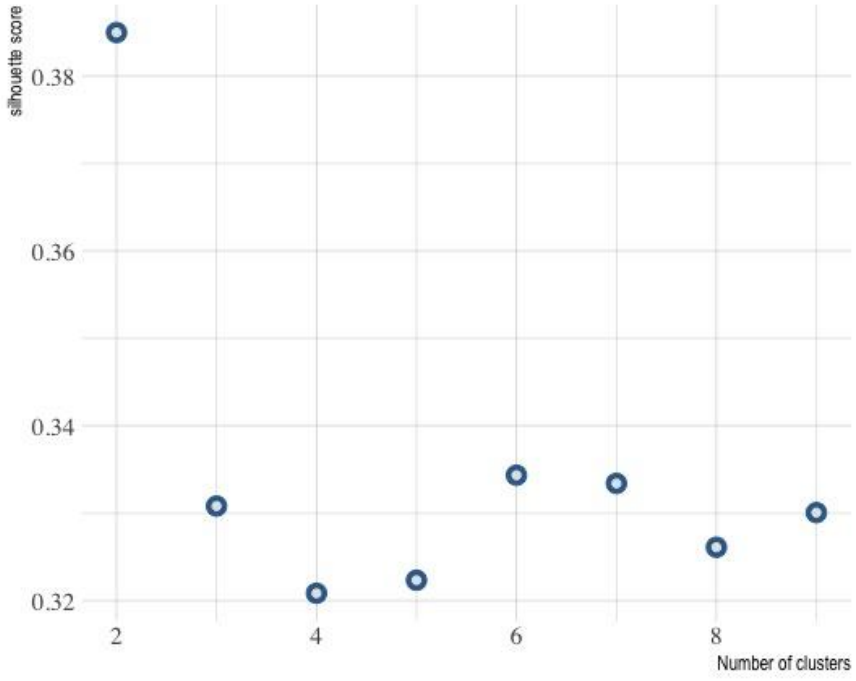


Supp Fig. 3. Shape variation derived from PCA on Elliptical Fourier Analysis.

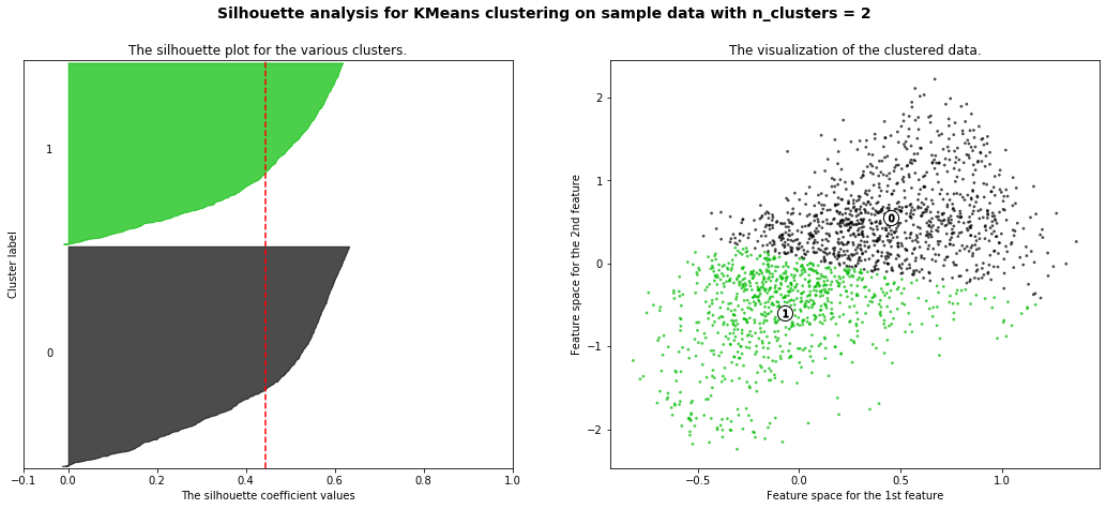


Supp Fig. 4. Clustering on the Elliptical Fourier components characterizing fruit shapes. When cluster number is set to two, the two characteristics shape are 'elongated' and 'globose' like, as shown by the black contours.

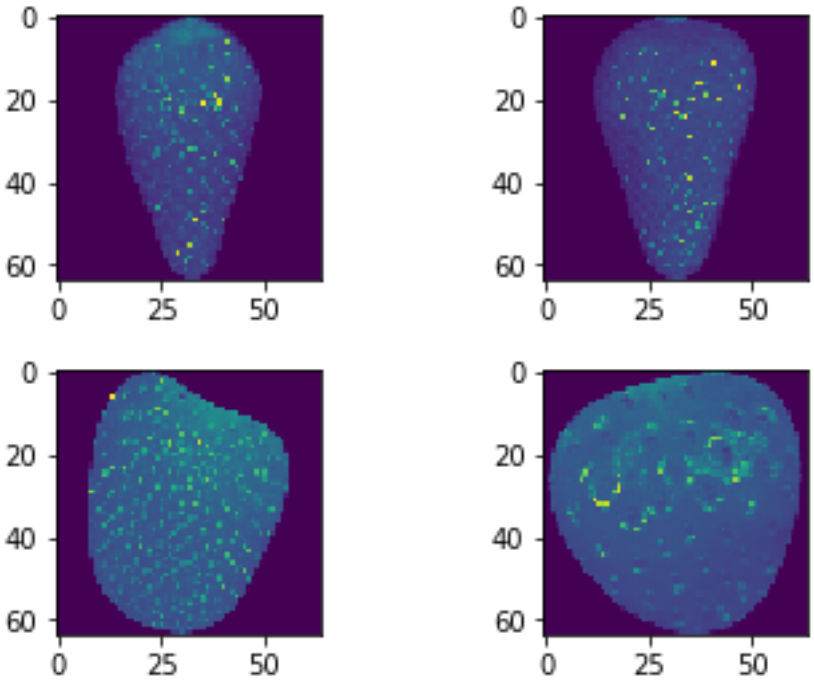
Variational Autoencoder to describe shape categories: We discovered shape categories from the latent space generated by a variational autoencoder deep neural network. The details are in the main manuscript. Here, we present the silhouette score of the clusters on the latent space varying between 2 to 9, the k-means clustering and two representative fruits of each cluster.



Supp Fig. 5 Silhouette analysis for number of clusters on shape latent space from Variational autoencoders output. The index is the mean silhouette score through all the clusters. The optimal number of clusters is 2, i.e. there are two main shapes for strawberries in this database.



Supp Fig. 6 Left: silhouette plot for each cluster (0 and 1). Right: visualization of the clustered data in the latent space.



Supp Fig. 7 The upper panel shows two examples of fruit belonging to the cluster 0, i.e. these are 'elongated' fruits. The lower panel show two examples of fruit on the cluster 1, these fruits have a 'globose' appearance.

Heritability evaluation

Trait	h2a	h2a_sd	h2d	h2d_sd
Like- red	0.21	0.08	0.23	0.09
Pale	0.30	0.10	0.24	0.10
Like orange	0.18	0.07	0.20	0.07
Shape cluster	0.21	0.07	0.25	0.09
A_channel	0.20	0.07	0.20	0.08
B_channel	0.20	0.08	0.22	0.08
L_channel	0.20	0.07	0.26	0.09
Fruit height	0.22	0.08	0.23	0.09
Fruit width	0.16	0.06	0.24	0.10
Widht_at_75 height	0.18	0.06	0.25	0.10
Widht_at_25 height	0.16	0.05	0.21	0.10
widht_at_half_height	0.16	0.06	0.23	0.09
Area	0.16	0.06	0.25	0.10
Perimeter	0.21	0.07	0.20	0.08
Solidity	0.20	0.08	0.20	0.08
circularity	0.25	0.09	0.20	0.08
EllipseRatio	0.28	0.11	0.19	0.07
Hight/ width	0.28	0.10	0.22	0.09
Tip	0.21	0.08	0.24	0.09
Neck	0.21	0.08	0.20	0.07
Left side	0.21	0.08	0.22	0.08
Right side	0.21	0.08	0.26	0.10
Elliptical Fourier PC1	0.25	0.09	0.20	0.08
Elliptical Fourier PC2	0.25	0.09	0.27	0.10

Suppl. Table 4 Heritability values for all the shape and color related traits, the additive and dominant components were evaluated.

RELATED PUBLICATIONS BY THE AUTHOR

1. Perez-Enciso, M., **Zingaretti, L. M.**, Ramayo-Caldas, Y., & de los Campos, G. (2020). Opportunities and limits of combining microbiome and genome data for complex trait prediction. *bioRxiv*. **doi:** https://doi.org/10.1101/2020.10.05.325977. Submitted to Genetics Selection Evolution.

2. Godia, M. Ramayo-Caldas Y., **Zingaretti, L.M**., López, S., Rodríguez-Gil J.E., Yeste, M., Sánchez, A., Clop A.. (2020). A pilot RNA-seq study in 40 pietrain ejaculates to characterize the porcine sperm microbiome. *Theriogenology*, *157*, 525-533 https://doi.org/10.1016/j.theriogenology.2020.08.001

3. Ramayo-Caldas, Y., Prenafeta-Boldú, F., **Zingaretti, L. M**., Gonzalez-Rodriguez, O., Dalmau, A., Quintanilla, R., & Ballester, M. (2020). Gut eukaryotic communities in pigs: diversity, composition and host genetics contribution. Animal Microbiome, 2, 1-12. https://doi.org/10.1186/s42523-020-00038-4

4. Pérez-Enciso, M., Ramírez-Ayala, L. C., and **Zingaretti, L. M.** (2020). SeqBreed: a python tool to evaluate genomic prediction in complex scenarios. Genetics Selection Evolution, 52, 1-9. https://gsejournal.biomedcentral.com/articles/10.1186/s12711-020-0530-2.

5. Ramayo-Caldas, Y., **Zingaretti, L.**, Popova, M., Estellé, J., Bernard, A., Pons, N., Bellot, P., Mach, N., Rau, A., Roume, H., Perez-Enciso, M., Faverdin, P., Edouard, N., Ehrlich, D., Morgavi, D. P., & Renand, G. (2020). Identification of rumen microbial biomarkers linked to methane emission in Holstein dairy cows. *Journal of Animal Breeding and Genetics*, *137*(1), 49-59. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6972549/pdf/JBG-137-49.pdf.

6. Pérez-Enciso, M., & **Zingaretti, L. M.** (2019). A Guide on Deep Learning for Complex Trait Genomic Prediction. Genes, 10(7), 553. https://doi.org/10.3390/genes10070553

7. Felipe V, Morgante CA, Somale PS, Varroni F, **Zingaretti ML**, Bachetti RA, Correa SG, Porporatto C. (2017). Evaluation of the biofilm forming ability and its associated genes in Staphylococcus species isolates from bovine mastitis in Argentinean dairy farms. Microbial pathogenesis, 104, 278-286. https://www.sciencedirect.com/science/article/pii/S0882401016301401?via%3Dihub.

Acknowledgments

Este trabajo no hubiese sido posible sin el inestimable apoyo de muchísimas personas con las que tuve la suerte de cruzarme a lo largo de toda mi vida. Quiero agradecer especialmente a quienes confiaron en mí para realizar este proyecto, Miguel y Amparo. No sólo por la confianza inicial, pero también por su tiempo, paciencia y consejos, que me han ayudado a crecer tanto personal como profesionalmente. Las palabras son escasas para expresar la gratitud que siento por la oportunidad brindada. Ha sido un camino extraordinario y, aunque el aprendizaje es continuo e infinito, me voy muy feliz con el crecimiento experimentado en estos cuatro años, ¡muchísimas gracias!

Yuliaxis merece un párrafo especial, no sólo por haberme adoptado en casa al poco de llegar aquí, sino también por su confianza para trabajar en equipo casi desde el primer momento en que nos conocimos, por haber sido un ejemplo como profesional y, especialmente, por su amistad incondicional que es impagable, ¡aún cocinando todos los risottos del mundo!

A Gustavo por recibirme en la estancia en MSU, porque en poco tiempo me ha enseñado muchísimo y porque es un ejemplo de profesional para mi: ¡Muchas Gracias! Mil gracias a todos los chicxs del grupo de Quantgen lab en MSU -y algunas extensiones- (Fer, Pía, Gabriel, Ana, Agus, Bea, Marquitos, Paulino y flia, Filipe, Alexa, Ben) porque hicieron de mi estancia en EE. UU. una experiencia inolvidable.

A los "viejos" compañerxs del grupo de genética animal (Lino, Elies, Emilio, Dani, MG, Marta, Manu, Jordi, Tai, Alice, Jessica, Pau, Lourdes, Ioanna, Dailu, Jesús, Yron) porque ha sido un placer compartir este tiempo con ustedes, me llevo aprendizajes y anécdotas de cada uno. Estoy segura de que este camino no hubiese sido el mismo sin su presencia. A los nuevos compañerxs (Magi, Lien, Cristina, Giacomo), con quiénes he tenido menos oportunidades de compartir debido a la situación actual, quiero desearles que disfruten al máximo la aventura de la tesis. A veces parece un camino arduo, pero visto en retrospectiva, nos deja un aprendizaje muy fructuoso y nos hace crecer en múltiples sentidos. Por último, decirles a todos que siempre pueden contar conmigo.

A Ray, Anto, Yuli, Lino, Moni, Vane, Raquel, Elenetta y Katy quiero agradecerles porque han sido los amigos que encontré de este lado del "charco". Aunque seguramente no se los digo a menudo, qué sepan que los quiero mucho y que no tengo palabras para darles las gracias por todos los momentos compartidos y el apoyo brindado. A Elli y Hans por las cervezas de domingo en Sant Andreu, un placer hablar de viajes y aventuras ciclistas con ustedes.

Gracias a mi familia y amigxs en Argentina porque, aunque estén muy lejos físicamente, sé que estamos muy cerca, y me lo hacen sentir en cada momento. No saben cuánto los echo de menos (o sí, porque se los digo cada vez que se presenta la ocasión), espero poder retribuirles un poquito de todo el cariño que recibo permanentemente y ¡sepan que tengo muchas ganas de cocinarles más asaditos!

Extiendo el agradecimiento a MINECO que financió este trabajo, a través de la acreditación de excelencia Severo Ochoa del CRAG. ¡Por más y mejor inversión en ciencia y educación! No puedo olvidarme de todo el personal CRAG e IRTA (pre-docs, IPs, postdocs, personal de apoyo) porque uno no trabaja solo y la ayuda del grupo es inestimable. Agradezco también a Planasa por los datos brindados para algunos de nuestros trabajos; a Vance, Patricio, Salvador, Luis, Luis Felipe por la colaboración en el segundo trabajo de esta tesis y el aporte de material; así como a Diego y Gilles del INRA por la colaboración que hemos establecido para realizar el último trabajo presentado aquí.

Por último -y no menos importante- quiero agradecer a la educación pública de Argentina en la que me he formado desde el jardín de infantes hasta la universidad y el posgrado y en la que tuve el enorme privilegio de trabajar durante casi 8 años. Tengo una gratitud enorme. La educación pública y gratuita es una de las banderas y pilares de nuestro país. Tenemos que defenderla con ahínco porque es una de las pocas herramientas igualadoras que permanece y ha resistido a tantos procesos de saqueos y pérdidas de derechos sociales que hemos experimentado a lo largo de tantas décadas.