Botanical Sciences 100 (Special Issue): S83-S109. 2022
DOI: 10.17129/botsci.3112

Received: February 8, 2022, Accepted: June 29, 2022
On line first: August 25, 2022

Genetics / Genética

Perspectives in plant evolutionary genetics: A field guide in 15 "easy steps" to modern tools in evolutionary genetics and genomics

Perspectivas en genética evolutiva de plantas: Una guía de campo en 15 "sencillos pasos" a las herramientas modernas de la genética evolutiva y la genómica

Luis E. Eguiarte[1]*, Erika Aguirre-Planter[1], Gabriela Castellanos-Morales[2], Valeria Souza[1,3]

[1] Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico
[2] Departamento de Conservación de la Biodiversidad, El Colegio de la Frontera Sur, Unidad Villahermosa, Villahermosa, Tabasco, Mexico
[3] Centro de Estudios del Cuaternario de Fuego-Patagonia y Antártica (CEQUA), Punta Arenas, Chile
* Author for correspondence: fruns@unam.mx

Abstract

Plant genomes contain huge troves of information, and nowadays molecular tools to analyze genomes are less expensive and keep improving. In this review, we aimed to produce a "roadmap" to take advantage of this explosion of molecular methods and opportunities. We explain how to decide which strategies are adequate for a given evolutionary or taxonomic problem by describing 15 possible (and in some cases nonconsecutive) steps to take advantage of all the genomic resources drawing from the ever-increasing studies. We describe how to obtain an adequate genome sequence given our study species and objectives and discuss if we need to also obtain a transcriptome and additional "omic" data (*i.e.*, proteome, metabolome, epigenome, microbiome). We analyze what is needed to conduct population genomics studies in terms of genomic methods and sampling strategies and discuss the pangenome concept. In addition, we present some considerations about how to estimate population genetics parameters and how to analyze geographic differentiation, inbreeding and gene flow. We examine ideas and methods on how to estimate natural selection and local adaptation, how to detect candidate genes, how coalescent analyses can help in these studies, the importance of genomic information for conservation studies and to understand adaptability to climate change. We assess the use of these methods in domestication studies and in understanding how form and function can be inferred from genes; likewise, how to use the genomic information for improvement of cultivated plants. We also review how can we use these methods in phylogenomic studies.

Keywords: Adaptation, Coalescence, Conservation genomics, Massive parallel sequencing, Pangenome, Population genomics.

Resumen

Los genomas vegetales contienen gran cantidad de información y las herramientas moleculares para analizarlos son cada vez más económicas y continúan mejorando. Presentamos un "mapa de ruta" para aprovechar esta explosión de métodos moleculares. Explicamos cómo decidir cuál es la estrategia adecuada para abordar un problema evolutivo o taxonómico a partir de 15 posibles (y en ocasiones no consecutivos) pasos para aprovechar los recursos genómicos disponibles. Describimos cómo obtener una secuencia genómica adecuada considerando la especie y los objetivos del estudio, y discutimos si se requiere obtener datos de transcriptoma u "ómicos" adicionales (p. ej., proteoma, metaboloma, epigenoma, microbioma). Analizamos qué se necesita para hacer estudios de genómica poblacional en términos de los posibles métodos genómicos y las estrategias de muestreo y discutimos el concepto del pangenoma. Además, presentamos algunas consideraciones sobre cómo estimar parámetros genéticos poblacionales y cómo analizar la diferenciación genética, la endogamia y el flujo génico. Examinamos ideas y métodos para estimar la selección natural y la adaptación local, para detectar genes candidatos, y reflexionamos sobre cómo los análisis coalescentes pueden ayudar en estos estudios, así como sobre la importancia de la información genómica en estudios de conservación y para comprender la adaptabilidad ante el cambio climático. Evaluamos cómo la forma y la función pueden inferirse a partir de los genes. Adicionalmente, discutimos como el uso de la información genómica puede servir para el mejoramiento de plantas cultivadas y en estudios de domesticación. También revisamos el uso de estos métodos en estudios filogenómicos.

Palabras clave: Adaptación, Coalescencia, Genómica de poblaciones, Genómica para la conservación, Pangenoma, Secuenciación paralela masiva.

The genomes of organisms contain huge troves of information. It is well known that genomes contain the information which allows organisms to function, grow, survive, and reproduce. But it is not so obvious for everyone that they also contain the best available information about the organisms' evolutionary histories and adaptations. For these reasons, the study of genomes is critical for any modern biological project.

In particular, for botanists, the exploration of genomes is a bottomless chest of important information about their phylogeny, on how and when the species and populations evolved, which populations belong or not to a given species or lineage, about the genes relevant for their local adaptation and how they have changed, and of the evolutionary processes involved in both their ancient and recent evolution. Genomes also store information that allow us to know how and when populations expanded or contracted, how they were affected by climatic changes during glaciations and other past global climatic processes, and how they may change in the future. And very importantly, the study of genomes can help us design better strategies to preserve the extant genetic variation found in populations and species and in this way contribute to the long-term conservation of biodiversity.

However, this almost endless treasure of genomic information is not just standing there to be directly used by us, the common botanists. As it is not easy to understand this information, we need to knead and disentangle it to reveal its secrets: the genomes are not "written" in a clear language or even in a standard alphabet or typography. Like the legendary libraries from ancient times, the information is there, but it is dispersed in vast archives without any catalogue, and often this information is partial, sometimes it is written or coded in strange and poorly known alphabets or in forgotten handwriting styles that are still very difficult to decipher by us; in many cases the "manuscripts" are fragmented, or they have even been scratched and rewritten on top. But now we can decipher most of these genomic documents and even read the apparently removed or scratched out information. Indeed, the reading and understanding of the complex libraries that represent the genomes of microbes, fungi, animals and plants has been advancing at accelerated rates in recent years, by taking advantage of a large set of molecular and informatic advances and tools that were reviewed recently in a very succinct and clear way (Nature Milestones 2021).

Nowadays, both the molecular and informatic tools to analyze genomes are less expensive and keep improving constantly. New methods allow to sequence complete genomes at lower costs in short time and with better coverage and reliability, and recently developed informatic tools allow now to efficiently assemble, analyze, and compare them with less pain and effort. It has been a long road since Mendel's modest crosses of peas between 1856 and 1863 that started the study of genetics, and when Charles Darwin set up our fundamental ideas about evolution and natural selection based on natural history observations and pigeon breeding in 1859 (Darwin 1859). It is daunting to consider that these central contributions to science were made just a little bit over 150 years ago.

In this review we aimed to produce a "roadmap" or "field guide" to the perplexed botanist, intrigued but at the same time scared of all this explosion of molecular and bioinformatic methods, ideas, and research strategies and opportunities that they should not miss! These opportunities are particularly important for the botanists like us, living in regions of the world that are rich in local flora, represented by thousands of interesting, but poorly known species —many of them of economic or ecological value— but in many cases these researchers have limited economic resources for conducting their studies. In consequence, many of us, botanists in very diverse countries, not only have limited funding but also, up to this moment, little or no experience in using these new genetic tools.

Here, we will briefly explain and help decide which strategies are adequate for understanding a given evolutionary or taxonomic problem in the clearest possible way, and guide the lost botanists through the labyrinth of the current methods and research paradigms, by describing and commenting the basic possible and in some cases nonconsecutive steps or research questions that any botanist could follow to take advantage of all the new genomic and future resources that are and will be available. Also, we will try to explain how to use this genomic information and its potential, but also its limitations in a realistic perspective, drawing from the ever-increasing studies in these fields, and in particular of the information that we know better, *i.e.*, the studies conducted in our labs, mainly in Mexico with Mexican plants, as a modest effort to celebrate the 100th volume of our beloved journal, *Botanical Science*s, formerly known as *Boletín de la Sociedad Botánica de México*.

Step 1. (Optional) Obtain a "good" genome sequence

The best case scenario is to have a reference genome for any genomic study, so that we can interpret all the information in terms of both functional (for instance, a given gene, which environmental or ecological problem may solve or be involved with) or genetic-evolutionary (if the gene is neutral and tells us about the effective population size or migration, or if it is adaptive and tells us about natural selection and adaptation) perspectives. This first step may sound a very big and daunting step for most readers! And indeed, this was the case 30 years ago, considering all the time, money and effort that represented the most famous and well-publicized genome project —the first sequence of the human genome— that took from the initial launch, in 1991, 11 years to produce a couple of independent "first drafts" (Venter *et al.* 2001, International Human Genome Sequencing Consortium 2001), but… do not despair fellow botanists!

The original human genome projects were indeed complicated and very, very expensive, but they also opened many research avenues that prompted the development of many methods and strategies. Genomics has since developed standard techniques that are easier to follow, and more importantly, genomic studies are now orders of magnitude cheaper and faster. For many plant model organisms, genomes have been accumulating since the first plant genome was published in 2000 (*Arabidopsis thaliana*; The Arabidopsis Genome Initiative 2000). Plant genomes are very variable in size, due to the tendency of plants to duplicate their genomes, via polyploidization events. For instance, *Arabidopsis thaliana* was selected because of its small genome size, 0.115- 0.211 Giga pairs of bases (Gbp = a billion base pairs), but now it has been possible to sequence the genome of the saguaro, *Carnegiea gigantea*, which is ca. 1.4 Gpb (Copetti *et al.* 2017), maize, *Zea mays* subspecies *mays*, which is variable but ca. 2.4 Gpb (Díez *et al.* 2013), or even far larger genomes as in the case of pines*,* reaching for instance 22 Gpb in *Pinus taeda* (Zimin *et al.* 2014).

Hence, the smaller the genome, the easier it will be to sequence it and later to analyze it. A good example is one of the first plant genomes sequenced in Mexico, the one of the carnivorous plant *Utricularia gibba*. With a genome size of only ca. 0.082 Gbp, this genome is a beautiful example of compact architecture and evolution of a minute plant genome (Ibarra-Laclette *et al*. 2013).

If you are working with a model organism for which a genome is available, or if you are working with a close relative of a sequenced organism, then you can use the available genome directly as a reference genome. Or you can sequence your organism at a low coverage, and then use the reference genome to help you assemble your genome with less effort and greater confidence.

If you do not have the genome of a close relative —as happens in most studies of our plants— a preliminary step is to estimate the size of the genome, as this information is critical both to evaluate how much you need to sequence to obtain an adequate coverage (a good coverage is usually considered to be of more than 30×, but it also depends on the size and complexity of the genome: the more complex and larger, the more coverage is needed).

Then, you need to evaluate the quality of your final assemblage, for instance if your assembled genome is larger or shorter than the estimated genome size, or if there is something wrong with the sequencing, in your analysis, or with the initial estimate of the genome size. Also, it is better if your assembled genome is in contiguous and large fragments; at best, each fragment should correspond to each one of the chromosomes of your studied species.

An additional and relevant complexity that we need to consider is the fact that in many plant species there can be a high variability in the size of the genomes of different populations; for instance, in maize and teosinte genomes can vary in ca. 30 % (Díez *et al.* 2013).

There are different possible ways to find out the genome size of your organism. One is from the databases including genome size information for many species, as the Kew genome size database (cvalues.science.kew.org). Otherwise, the most common methods used to estimate genome size are related to flow cytometry (see Díez *et al.* 2013, Bourge *et al.* 2018). A very large genome size has been an important handicap in the study of some plant groups, in particular conifers, which have genomes of ca. 22 Gbp but that can be as large as 34 Gbp in *Pinus ayacahuite* (Grotkopp *et al.* 2004); consequently, there are few reliable assembled genomes for this important genus.

For most plants, even a modest laboratory can now afford to obtain a competent first draft of a genome, usually by using hybrid methods. For instance, you can employ a joint sequencing using on one hand the Illumina platform —which produces many millions of short sequences, each one between 150-300 bp (www.illumina.com) at a low price, giving a good coverage— and on the other hand, PacBio or Oxford Nanopore technologies, which allow sequencing in a continuous way to obtain very long DNA fragments up to 25,000 bp and sometimes even larger (www.pacb.com; nanoporetech.com). These latter platforms sometimes generate high errors in sequencing, but these errors can later be corrected with the Illumina sequences, and these technologies keep improving.

There are other data and methods that can help in the assembly of genomes. If older and reliable data of chromosome counts or karyotypes are available, they can be useful to check the assembled genomes (as we mentioned above, if the assembled genome is indeed very good, the largest sections should correspond to each chromosome). It is even better if older genome mapping data (*i.e.*, QTLs and related analyses) are available —obtained with microsatellites, RFLPs or even isoenzymes— using controlled crosses; these data can be very useful in guiding the bioinformatic effort. Also, genomes of related plants that maintain their order in the genes, *i.e.*, synteny, can be extremely useful to help guide in a new (called *de novo*) genome assembly (see for instance, Barrera-Redondo *et al.* 2021).

Currently, recent modern methods like optical mapping, Hi-C and other strategies can help assemble the genomes at the chromosome level (Cosgrove 2021, LaFlamme 2021), and even obtain the sequences in the centromeres and telomeres, which has been very difficult in the past given the high number of repeated sequences in these sections of the genomes (Wrighton 2021). Once you have the genome sequences, it is also useful to have the transcriptome, as we will see below.

Step 2. (Optional) Do we need to obtain a transcriptome?

While having the information of the complete genome is interesting and relevant as we explained above, sometimes we are not very concerned about the large regions of the genome that are less informative. These less informative regions can be stretches of the genome without genes, or that sometimes comprise many repetitions of ribosomal genes (to which belong the famous ITS sequences used in phylogenetic and population studies in plants), or copies of transposable elements, pseudogenes and other sequences, sometimes called "junk-DNA". In this case, a useful strategy is to also sequence the transcriptome, which includes all the expressed genes. Or in other cases, we may prefer to only have a transcriptome instead of the complete genome, as to obtain a single transcriptome is usually cheaper to sequence and requires less bioinformatic efforts to assemble than the complete genome (since it is smaller than a genome); and in theory, from the transcriptome we can find the relevant genes involved in adaptation (although in some cases the expression of a given gene may be given by other DNA regions that are not expressed).

The transcriptome consists of the sequences of the expressed (transcribed) messenger RNA (mRNA) that can be extracted from the different plant tissues. In each tissue, different transcripts (mRNAs), can be found, as in each tissue different genes are expressed (that is why tissues are different). The main problem is that it is very easy for the mRNA to degrade, and it is also easy to contaminate the samples with mRNA from bacteria, other plants, or even from the researchers.

On the other hand, it is cumbersome to conduct a detailed transcriptome analysis of the total expression profile of a plant, since you need not only to consider most or many different tissues of a plant but also to analyze different biological replicates (*i.e.*, different plants) and at different developmental stages (*i.e.*, seeds, seedlings, and adult plants), depending on the objective of your study. You will also need to evaluate the expression under different environmental conditions —preferably contrasting— as a given genotype could express different genes (and thus produce different phenotypes) in different environments (see for example Figueroa-Corona *et al.* 2021). Additionally, technical replicates and controls are needed, as different RNA extraction experiments and methods could yield different qualities and types of mRNA, thus biasing the results. Then the RNA sequences have to be translated back into DNA with the reverse transcriptase enzyme, and then the DNA is sequenced in massive sequencing platforms, usually Illumina, and different runs could yield different sequences and coverages.

But do not panic, because as a tool for guiding our genomic studies, we usually only need the transcriptome from a few tissues that would give us the majority of the most common expressed genes. In this way we can know the minimum and important genes that should be detected in our complete genomes. In addition, the transcriptome information can be used to help annotate (identify) the main protein genes in our genome of interest (see for instance our study in pumpkins, Barrera-Redondo *et al.* 2019).

Step 3 (Optional) Evaluate if you need additional "omic" studies: proteome, metabolome, epigenome, and microbiome supporting studies

Most evolutionary, ecological, and evolutionary genomic studies will only need a reference genome and maybe an auxiliary transcriptome to improve the analyses, but if for example we want to get a deeper understanding about the relationship between the genotype and phenotype, the role of phenotypic plasticity in adaptation or the response of organisms to stress or to environmental change, to mention a few, other "omic" studies can be useful or relevant.

For instance, the *proteome* allows to know the proteins that are actually produced, not only the mRNA. Proteome methodologies are based on sophisticated chemical techniques, including separation by HPLC and later identification of the compounds using methods such as mass spectrometry and reference libraries. And now even more "inclusive" molecular methods are available that allow analyzing the lipids, sugars, other carbohydrates and other metabolites. These are analyzed using *metabolomics* strategies, using again sophisticated separation methods, detection by mass spectrometry or other methods and reference libraries.

Another set of relevant tools can be the *epigenetic studies*. The idea is that, while the DNA is very stable, some changes associated to the DNA, including methylation and histone modifications, can affect the expression of the genes, and these changes can sometimes be inherited. For instance, in methylation, adenine and cytosine are modified including a methyl group, and if this happens in a regulatory region, the change can suppress the expression of the involved gene (see review in Barrera-Redondo *et al.* 2020).

Currently, it is possible to sequence the total genome and perform an analysis that recognizes the methylated genome. This method, called "Whole Genome Bisulfite Sequencing", has been used in some population studies using a reduced representation strategy (Liu *et al.* 2012, van Gurp *et al.* 2016, Paun *et al.* 2019).

In addition, we can conduct *microbiome* studies in our plants. While the field is still expanding, it is clear that in many cases —as it happens with animals— their associated microbes (bacteria, archaea, fungi, protists, and viruses) may be important to plant functioning. The associated bacteria represent part of the "extended phenotype" of a plant, and together with the plant, they form what is sometimes called the "plant holobiont", a single functional and ecological unit, which is part of the eco-evolutionary feedback and niche construction of the plants (Vandenkoornhuyse *et al.* 2015, Borges 2017, Compant *et al.* 2019).

A well-known example of associated bacteria relevant to plant fitness are the nitrogen fixing bacteria associated with many Fabaceae species —in particular *Rhizobium* spp.— or Actinobacteria (in particular of the genus *Frankia*) associated with other plant lineages, and even cyanobacteria, associated with cycads and with the angiosperm genus *Gunnera* (De Bruijn 2015). Other relevant and well known (but still poorly understood) associations are the mycorrhiza and the rest of the complex root microbiome of land plants (van der Heijden *et al*. 2015).

The modern study of the microbiome and of the associated microbes is a direct descendent of the classic molecular ecology and metagenomics methods, involving the extraction of total DNA of a sample, and either amplifying a marker gene, usually 16S for bacteria and archaea, and ITS for fungi and protists (Eguiarte *et al.* 2007). For recent reviews of this subject see Compant *et al.* (2019), Fitzpatrick *et al.* (2020) and Trivedi *et al.* (2020).

An interesting recent example analyzing the complexities of the soil bacterial communities associated with plants and seasonal changes is the study of Rebollar *et al.* (2017) on the milpa, the traditional maize management plots of Mexico and Mesoamerica involving not only maize but also other plants, in particular beans (*Phaseolus* spp.) and squashes (*Cucurbita* spp.). Another relevant paper on particular bacterial groups in the milpa is Aguirre-von-Wobeser *et al.* (2018).

For studies on the role of microbes in floral evolution and pollination, we direct the reader to the recent studies of Rebolleda-Gómez & Ashman (2019) and Rebolleda-Gómez *et al.* (2019); and regarding the role of the microbiome in adaptation to drought in squash roots, to Hernández-Álvarez *et al.* (2022).

Step 4: What do you need for conducting population genomics studies? I. Genetic aspects: different strategies to acquire the genomic data (and also, as an optional goal, to obtain the pangenome of a species or a group of species)

What do you need to conduct a population genomics study? A few years ago, the short answer to this question was "A lot of money!", but now costs have become much lower, to the point that many labs can now carry out a genomic study.

This is a very important step in any evolutionary genomics study, as it represents the acquisition of the basic information for conducting most of the following possible steps. For this step, you have to decide the sampling of your populations along two main axes: the molecular axis, involving basic questions such as which sequencing method and platform to use (based on costs and availability), the desired coverage (how many times on average a single site in the genome is sequenced), and other methodological questions; and the ecological axis: which and how many populations to analyze, and how many and which individuals. For both axes, the "right" answers depend on the details of your research question: what do you want to ask your experimental system? *i.e.*, the evolutionary (for instance, if we want to evaluate the effective population sizes, or the species limits), functional (as a possible example, the physiological adaptations used to survive in an extreme environment), ecological (for instance, to find signals of selection, or how past climate has affected populations), or a taxonomic question (to better understand if a set of populations conform a single or several species, and if different, when did they evolve, for example, or in comparative phylogenomic studies).

Let us first consider the possible sequencing strategies we can use. Nowadays, different versions of what we call limited or reduced representation sequencing of the genome have been implemented. The idea is that we do not need to sequence the complete genome of each individual, but for most questions it is sufficient to have parts of their genomes sequenced, so the costs, sequencing time and efforts to analyze data can be drastically reduced. With these strategies, we can obtain thousands or even millions of genetic markers, usually called SNPs, single nucleotide polymorphisms. A SNP is a change in a site of the DNA sequence of one base to another, for instance C to G, etc. SNPs are found along the entire genome, including coding and non-coding regions.

There are different strategies to obtain these reduced representation SNPs data (*i.e.*, see Davey *et al.* 2011, Andrews *et al.* 2016). Many strategies are based in first cutting the DNA with restriction enzymes and different combinations of enzymes to obtain fragments of given sizes (*i.e.*, 300–500 bp). Then these fragments are sequenced in platforms like Illumina, the sequences starting from the cut-end of the used restriction enzyme. If the restriction enzyme recognizes a short sequence, then it will cut the genome many times, resulting in many fragments (and each fragment will have a lower coverage); if the enzyme recognizes longer sequences, then it makes less cuts and less fragments are produced (and the coverage of each fragment will be higher). The genome coverage obtained will be a function of the total size of the genome, but the desired coverage partly depends on the question of the study. In some cases —for instance for knowing the general genetic structure, effective population, and migration patterns— you only need some neutral SNPs, as most of them will show a common history. But if you want to find SNPs under selection and particular adaptations, you will need to have as many SNPs as possible. Different restriction enzymes and combinations of them can be used (Davey *et al.* 2011), as some combinations will give you more SNPs, but also may increase the cost of the study.

Indeed, there are many variants of these reduced representation methods, the most common being GBS, RADseq, ddRADseq, DArTseq, nextRAD (Andrews *et al.* 2016, Guerra-García *et al.* 2017, Aguirre-Liguori *et al.* 2019a, 2020, Arteaga *et al.* 2020, Barrera-Redondo *et al.* 2020, 2021). In some variants, like in nextRAD, each fragment obtained with the restriction enzyme is first amplified by PCR, which has the advantage that the amount and quality of DNA per individual can be low, but they can incorporate PCR artifacts.

One limitation of these reduced representation methods is that even if you can find many polymorphic sites (*i.e.*, SNPs), it is possible that there are some biases, in particular because these methods cannot detect a SNP if a mutation happens in the restriction site. Another possible problem is that the number of high quality and the reliability of the SNPs that can be obtained depend on the particular genomic platform and strategy used (*i.e.*, GBS, RADtags, RADseq, etc.), and also on the genome size (as mentioned above, the larger the genome, the more expensive and complicated are the analyses) and the used coverage (*i.e.*, the amount of sequence you managed to obtain, given the size of the genomes and of your research budget) (Andrews & Luikart 2014, Andrews *et al.* 2016). For example, for larger genomes a ddRADseq —which uses a double digestion with restriction enzymes (a common cut site and a less common cut site)— has been used to obtain higher coverage, but with a lower representation of genomic regions (Peterson *et al*. 2012).

There is a methodological variation that may solve these problems and reduce costs; instead of analyzing (sequencing) every single individual by itself, it involves pooling the DNA of all (or of several) sampled individuals of a given population, as we will explain and illustrate later. This has the advantage that economic resources can be used to obtain a better coverage of all the populations. The main disadvantages are, on one hand, that you need to be careful that the amount and quality of each DNA's individual is the same or similar, to avoid biasing the estimates. The other disadvantage is that the analyses of the levels of genetic variation in the populations using standard measures (as the expected heterozygosity) and the inbreeding levels (*i.e.*, F_{IS} and equivalents) cannot be performed (Ferretti *et al.* 2013, Schlötterer *et al.* 2014, Fustier *et al*. 2017).

If you do not have a reference genome, you can still advance in the population genomic analyses using the SNPs as a set of anonymous genetic markers, and later conducting linkage disequilibrium analyses to avoid oversampling a given region of the genome, and other strategies to estimate genotyping error (Mastretta-Yanes *et al.* 2014). These anonymous analyses can be adequate for many studies but limit the information that you can obtain later. Nevertheless, sometimes you can annotate interesting SNPs by using other not so closely related genomes, and genomic databases (many associated to the NCBI GeneBank), especially for the SNPs considered as candidates after conducting selection tests (see below), or by using a transcriptome sequence of your particular species (which, as we explained above, is easier and less expensive than obtaining the complete genome), that should allow you to analyze if an expressed gene is codified by the genomic region where a SNP was detected.

There are other possibilities to explore the genomes without completely sequencing them. One strategy is called the "exome capture" method. In this strategy, first a genome or a transcriptome of the target species is needed, and from it you design PCR primers for a subset of proteins (*i.e.*, the exome) (Heyduk *et al.* 2016a, b). This set of primers is then used to amplify and then sequence the products by Sanger or in a parallel massive platform like Illumina. This strategy has the advantage of allowing you to identify the particular genes involved. Similar methods can be used for different sets of genomes or genomic regions, as for instance for analyzing the so called "ultra-conserved genes" and similar regions that are especially useful in animals to compare groups of very divergent organisms, *i.e.*, to conduct phylogenomic analyses that we will discuss later (Faircloth *et al.* 2012, Hime *et al.* 2021).

For some model organisms you can use (or even develop) commercially available "chips" that detect thousands or even millions of SNPs, which in some cases can be used in non-model related wild species. They are very commonly used now in human and animal studies. For instance, the Mexican human populations were analyzed by Moreno-Estrada *et al.* (2014) using Affymetrix 6.0 and Affymetrix 500K arrays, to obtain 909,622 SNPs. In plants, we used the Illumina MaizeSNP50 Genotyping BeadChip chip, that we will call hereafter the "50 K chip", developed in maize, to study the wild populations of the ancestor of maize, the teosintes, both for the *Zea mays mexicana* and the *Z. m. parviglumis* subspecies, to successfully analyze more than 33 thousand SNPs (Aguirre-Liguori *et al.* 2017, 2019a, b, 2020, Moreno-Letelier *et al.* 2020).

Another possible strategy to study population genomics is to actually sequence the complete genomes, but to a lower coverage, either of each single individual or pooling the DNA of individuals and sequencing for the complete populations (as mentioned above). This can provide thousands of shared SNPs to analyze, with good coverage and reliability, as was done in the teosintes by Fustier *et al*. (2017).

Alternatively, instead of using the complete genomes you can analyze the transcriptomes of different individuals (Xu *et al*. 2016, Zaidem *et al.* 2019). The advantage of this procedure is that the total number of proteins is relatively small (compared with the extent of the total genomes), ca. 25-35 thousand, so the sequencing effort drops considerably. The main problem of conducting population transcriptomics studies is that RNA extraction can be complicated and expensive, in part because it is both very easy for the RNA to become contaminated during the extraction procedure, and also because of the "fragility" of RNA, which degrades fast, since ribonucleases are everywhere in the environment. Again, the advantage is that all the studied genes can be annotated and interpreted in a genetic and, in many cases, in a physiological context. The disadvantage of this approach is that you do not know anything about the non-coding, truly neutral parts of the genome, which can be important for the expression of the genes or in molecular evolution studies.

An optional analysis related to this step is the "pangenome". This concept was initially proposed by Tettelin *et al*. (2005) for bacteria. The idea is that each time you sequence a new genome in a given species, you find more genes until you reach a plateau. If this is the case, you have what we call a "closed pangenome". But, apparently, in some bacterial lineages, such as *Escherichia coli*, the number of genes keep increasing, given the extremely large population size of this cosmopolitan bacteria, and also due to the input of new genes to the gene pool from other bacterial lineages, a process called "horizontal gene transfer"; in these cases, we have an "open pangenome". The study of the pangenomes, though difficult as it involves obtaining many genomes, is interesting from many evolutionary perspectives, as it allows to evaluate all the different functional strategies (in terms of different genes, solving different environmental and other ecological problems) used in a given group of organisms (species, genus, families) and to analyze the selective patterns and ecological correlations in different environments, for instance.

Advances in the study of the pangenome at the species and genus levels in plants have been reported, in particular in plant species of commercial interest and model systems, such as tomato and sunflower, as reviewed recently by Barrera-Redondo *et al*. (2020). For instance, we are working at this moment in a long-term project attempting to assemble the pangenome of the *Cucurbita* genus.

Given the sequencing and informatic advances, it seems that we will soon be able to conduct population genomics using complete genomes (see for example Fustier *et al.* 2017; Cornejo *et al.* 2018). This will allow us not only to identify changes of some SNPs in the genome, but to know exactly what is happening in the genome at population levels: how many duplications have occurred, and which ones are found in different populations, the missing regions in some of the genomes, and rearrangements like inversions and translocations in different chromosomes that inhibit recombination, as well as changes not only in coding but also in regulatory regions. This will allow for detailed population and evolutionary analysis, like gene flow estimation, effective population size analyses, and in particular accurate studies on how and when natural selection has acted.

Although the idea of eventually using complete genomes for population studies is very attractive, at this moment it is not yet practical, not only because costs are still very high, but because the needed informatic analyses are daunting. In addition, if you follow this completist path, you will eventually want to have not only the complete genome of many individuals, but also their transcriptomes and the epigenetic analysis of the methylated genome, also you may want to conduct a phenotypic description of the individuals, if possible in different environments… so it can become a never ending study, and even in the best case scenario, the resources —economic and time— are going to be limited.

Step 5. What do you need for conducting population genomics? II. Ecological and sampling aspects: how many individuals and populations?

The short and good answer is that simulation and empirical studies show that in many cases you need sample sizes that are smaller than those used when analyzing a lower number of genetic markers, as when we analyze populations using allozymes/isozymes, microsatellites or dominant markers as RAPDs, ISSRs or AFLPs, as reviewed and analyzed in Aguirre-Liguori *et al*. (2020). This is in part because having thousands of markers (SNPs) along the genomes (over-) compensates the lower number of individuals to give an accurate estimate of the diversity levels in the genomes. For instance, for isozymes and microsatellite studies our "golden standard" was to reach at least 30 individuals per population, a criterion based on the minimum number of samples required to estimate the frequency of the relatively uncommon alleles, and in part derived from population ecology and basic statistical methods that equate "more than 30" to "almost infinite", based on old statistical tables for tests, like the *t*-test and χ^2-test. But the minimum number of individuals needed per population depends on the levels of genetic variation, and the details of the distribution of genetic variation within and among populations.

An important study on this sampling issue is reported in Aguirre-Liguori *et al.* (2020) and other comparative and simulations studies cited therein. Aguirre-Liguori *et al.* (2020) used simulations to compare the effects of different sampling schemes in teosinte (wild maize) using three different data sets: one of 33,454 SNPs from the Illumina 50 K chip, another of 9,735 SNPs derived from a pooled-sample populations data set obtained from a version of GBS, and 22 nuclear microsatellite loci. In general, and depending on which analysis you wanted to perform, either genetic differentiation (measured as D_{ST}), diversity levels, or levels of inbreeding, the optimal or recommended numbers of individuals and of populations varied, but usually genomic methods will need less individuals than microsatellite-based studies to yield consistent and reliable estimates (Aguirre-Liguori *et al.* 2020). Another issue to consider is the number and distribution of sampled populations or sampling strategy, which will largely depend on the type of question we wish to answer; yet it is recommended to sample as many populations as possible and to cover environmental heterogeneity (De Mita *et al.* 2013, Lotterhos & Whitlock 2015).

Indeed, other studies have shown that by having complete genomes, few samples (individuals) even as low as a single genome, can be used to perform some assessment of the changes in the effective population size of a species, using programs such as PSMC (Li & Durbin 2011); for an example see Liu *et al.* (2021).

Step 6. Population genomic analyses: estimating the relevant parameters and geographic differentiation

Once you have your genomic data set obtained from any of the strategies described above, it is important to first calculate the standard estimates of genetic variation, for instance the expected heterozygosities, and π, also known as the nucleotide diversity (the average number of nucleotide differences per site between two DNA sequences, see Hedrick 2011, page 106), and/or θ (a measure of genetic diversity based on the number of segregating -variable- sites in a DNA alignment, see Hedrick 2011, page 303) at both population and species levels. Genomic data can also be used to obtain neutrality estimates, in particular the Tajima's D test (Tajima 1989) and the related family of tests, like Fu & Li (1993) and Fu (1997), etc., all of which indicate if the genetic variation seems to adjust to a neutral process, or if a given gene has signals of purifying or directional selection (if the variants -*i.e.*, alleles- are less common than they should be according to the neutral theory), or balancing selection (if some variants are more common than expected). These tests can also give a general idea of the demographic behavior of the species in the past, for instance if it has been expanding (if the variants are less common) or contracting (if some variants are more common). These results represent hypothesis that can later be explored in detail by the powerful coalescence analyses, as we will see in Step 10.

It is also usually very important to estimate if there are differences in the frequencies of the alleles in each locus (in this case, SNPs) among populations. Two different set of strategies are possible. One is to use *a priori* the original data from our sampling (*i.e.*, as we defined the populations when sampling), and obtaining different estimates usually related to F_{ST} (the among population differentiation, Hedrick 2011). You could also directly compare the allelic frequencies by using explicit population genetic tests —for instance, the Workman & Niswander (1970) test— or use standard statistical tests, like ANOVA.

Basically, the F_{ST} and related tests give a value of 0 —or a value not significantly different from 0— if all the populations have the same (or very similar) allelic frequencies; higher values indicate higher genetic differentiation, until reaching 1, which usually means that the different populations do not share any allele at all. Obviously, the pattern of sharing or not alleles can be different among the thousands or even millions of SNPs that are obtained using genomic data. This is in part because there is random variation among loci, and in part because some of the loci/SNPs can be under different selection regimes (or the genes closely linked to them), and these differences are the basic foundations for some of the natural selection tests that will be reviewed later, in Step 9.

Instead of using *a priori* defined populations, we can analyze the data in an agnostic way, and let the data speak for themselves. Again, two different and complementary ways are available to analyze the SNP data. One way is to use multivariate statistics and visualize the data using standard multivariate methods. For instance, you can conduct a principal component analysis (PCA) or a similar statistical tool (like factor or discriminant analyses, etc.), and use the first two, three, or more principal components to visualize which individuals cluster together, and from this clustering define *a posteriori* the populations. Usually, the data will separate in different clusters, and each cluster would group together similar samples, and if the patterns make geographic and/or biological sense each cluster can be considered in further analyses as a single genetic unit. This procedure can be complemented with other analyses, for instance the DAPC, the Discriminant Analysis of the Principal Components (Jombart *et al.* 2010), which selects the sets of variables with the higher discriminant power, and in some cases can give better resolution than the basic PCA. Another possible strategy is to cluster the data using the Euclidean distance of the SNPs among all the possible pairs of analyzed individuals, and then use classic clustering methods, like the UPGMA (Unweighted Pair Group Method with Arithmetic Means, see Hedrick 2011, pages 343-347) or NJ (Neighbor-Joining, see Hedrick 2011, pages 343-347), to choose a differentiation level to define the populations; but remember that the differentiation levels can be highly hierarchical, and sometimes the decided level can be very arbitrary.

A group of strategies to define the limits of populations relies on the use of Bayesian algorithms to separate and define groups of organisms, like the one implemented in the software Structure (Pritchard *et al*. 2000). Similar programs and related strategies in some cases are more efficient in terms of computing time when having thousands or even millions of SNPs, thousands of individuals and many populations, such as the program Admixture (Alexander *et al.* 2009).

Nevertheless, a problem with all these strategies is the decision of how many groups ("populations") should the data be divided into, given the hierarchical structure of the geographic differentiation we commented above. A popular test used along the program Structure is the test of Evanno *et al.* (2005) and its on-line implementation by Earl & vonHoldt (2012), or for the Admixture program, a cross validation test (Alexander *et al.* 2009). Most researchers now recommend analyzing different numbers of "K" (groups) that may suggest different numbers of hierarchical partitions (the first partitions separate the main genetic groups, then other partitions divide these large groups into subsequent smaller categories) or "populations", depending on the total spatial distributions of the samples, the suspected number of populations, etc. and to explore and even present in the paper the results of different partitions (Janes *et al.* 2017).

Step 7. Estimate the levels of inbreeding, as it is usually very informative for the ecology and evolution of the species

Now that you know the levels of variation and the levels of structure/differentiation, what can you do with your data? One set of further possible analyses are related to the F_{IS} estimate and are used to evaluate the levels of inbreeding within each population. For instance, in plants, inbreeding can be caused either by self-pollination and by crossing among relatives, and the level of inbreeding is one of the most important determinants of other evolutionary parameters of the organisms, including their genetic structure, effective population size and the relative role of natural selection (Hamrick & Godt 1990).

There are several powerful methods to estimate the different components of inbreeding using genomic data (see for instance David *et al*. 2007). In some cases, given the large set of genetic information we can obtain now, it is possible to distinguish self-pollination from other forms of inbreeding, and even start analyzing the relatedness of the individuals in the samples using different programs, as the ones mentioned above (for an example using microsatellites in teosinte, see Gasca-Pineda *et al.* 2020).

Nevertheless, a note of caution is relevant here: in some cases, depending on the sequencing platforms, and the employed reduced representation techniques, the number of heterozygotes can be mismeasured (usually they are underestimated). In many cases, at a given genomic site, due to low sequencing coverage, low quality of the sequences, or even low quality or quantity of the DNA, it is not possible to evaluate in a reliable way if an individual is heterozygous, and thus in the analyses it will appear to have only one type of base (*i.e.*, to be homozygous). Therefore, the estimates of genetic variation and all other analyses, in particular those related to F_{IS} and other inbreeding statistics may be biased, usually suggesting an excess of homozygotes and thus yielding high but spurious values of inbreeding.

If it is important for your project to have accurate estimates of the inbreeding levels of the populations, it may be useful to have additional data sets to verify the estimates, for instance, by obtaining parallel microsatellite data for the same individuals. Microsatellites are genetic markers that have many possible allelic forms as they have high mutation levels, so they are very good at detecting heterozygotes, and thus can help distinguish self-pollination from other forms of inbreeding (see for instance Gasca-Pineda *et al.* 2020). An alternative path is to remove from your database all the loci/SNPs with lower coverage and only use for the analysis the loci that are more reliable, as you do not need to have thousands of sites for estimating this population parameter.

Step 8. Gene flow estimates

Once you have defined the populations, estimated the levels of genetic variation and differentiation among populations, as well as the inbreeding levels within each population, usually you would like to make inferences about gene flow, both recent and historical. Some programs and analyses are useful for a first estimate, for instance Bayesass (Wilson & Rannala 2003), Migrate (Beerli *et al.* 2019), or NewHybrids (Anderson & Thompson 2002). For recent examples of the use of some of these programs, see Gasca-Pineda *et al.* (2020) and Martínez-González *et al.* (2021).

A very useful approach to understand and visualize the role of gene flow in the evolution of a species is by using Treemix (Pickrell & Pritchard 2012). This program involves a two-step analysis: first you estimate a general genealogy of the population that would result from random changes in the allelic frequencies, *i.e.*, evolution produced only by genetic drift. Then, the program estimates the direction, relative time, and magnitude of possible events of gene flow based on your data that are not explained by the genetic drift only scenario. We have used this approach to disentangle the role of gene flow in the evolution of teosinte (Aguirre-Liguori *et al.* 2019a) and in our analyses of the origin of cultivated maize (Moreno-Letelier *et al.* 2020).

Step 9. Analyze data for signals of natural selection, candidate genes, local adaptation

To many of us, this is perhaps the most exciting of all steps. The idea is that, by comparing the distribution of the genetic variation within and among populations, we can infer the genetic targets of natural selection. In other words, we can, in principle, find the genes involved in the process of natural selection and adaptation, *i.e.*, the sites in the genome (or closely linked sites) that will allow us to finally understand the genetic basis and the fine details of the process of adaptation. This detail of understanding not even Darwin allowed himself to fantasize (well, Darwin had a very embryonic idea of the bases of heredity, but nevertheless, he would have loved to get a grasp on this level of understanding).

There are many possible strategies in which we can use our genomic data to infer selection and/or adaptation. Usually, it is important before conducting the selection analyses, to first analyze the data for genetic structure. If there is strong genetic structure, you have to analyze each genetic group of populations separately, because if you fail to do this, the supposed detected signals of selection may be just the result of the general differentiation (in theory, mostly due to genetic drift). For instance, in the study of teosinte of Aguirre-Liguori *et al.* (2017, 2019a, b), it was critical to first separate the data of the two different subspecies of teosinte and after this, we could search for signals of the selective patterns and loci.

Thereafter, one possible strategy is to analyze, locus by locus, the differences in allelic frequencies, as it is done by the F_{ST} and related analyses. The idea stems from the classic work of Lewontin & Krakauer (1973) and it is pretty straightforward: most genes (SNPs in genomic studies) are "neutral" and should display similar F_{ST} (differentiation) values among them —but not exactly the same value, as each one is diverging by random genetic drift processes— but for some genomic sites, the SNPs variants (the alleles) will be very different among populations, because the selection process is different and has changed their allelic frequencies accordingly. The SNPs with contrasting differentiation in their allelic frequencies from the rest of the SNPs are the possible (*i.e.*, candidate) genes involved in local adaptation, or there are genes under selection in their genomic neighborhood (linked loci). In addition, it should be possible to also find other SNPs where the allelic frequencies are very similar among populations, suggesting genes under strong purifying or even balancing selection.

There are many strategies to conduct these "outlier" loci analyses (Hoban *et al.* 2016, Ahrens *et al.* 2018). A popular and useful tool has been the Bayescan program (Foll & Gaggiotti 2008). There are other related programs, including Bayescenv (de Villemereuil & Gaggiotti 2015) and Bayenv 2.0 (Coop *et al*. 2010) that can help associate the outlier loci with environmental variables. More recently, other similar tools, such as pcadapt (Luu *et al.* 2017) and LFMM 2 (Caye *et al.* 2019) are useful and complementary, as they use different strategies and algorithms (Ahrens *et al.* 2018). Nowadays, it is a standard procedure to use different programs and consider as the most probable SNPs under selection the sites detected by more than one of these algorithms (see Barrera-Redondo *et al.* 2021 for a recent example).

In a few words, in these selection analyses, from the list of outlier genes, you try to find out to which environmental and ecological conditions the different alleles correlate, for instance, soil, pH, temperature, precipitation, soil nitrogen and phosphorous, etc., as conducted in Aguirre-Liguori *et al.* (2017, 2019a), or with changes related to the domestication (selection) process, as the recent published study on the pumpkins by Barrera-Redondo *et al.* (2021), to cite some studies conducted in Mexico that we know very well and can serve as a model of possible studies, among a growing set of papers.

These outlier-based selection analyses can actually be performed even if you lack a reference genome, but it is better to have a genome to correctly annotate the SNPs, *i.e.*, to know the genes where they belong, or close linked genes (that may be the true targets of the selection process), or if the SNPs belong to regulatory regions.

Also, for these candidate genes, you can analyze if the sequence changes detected by the SNPs are related to modifications in the amino acids of the codified protein, or if they are apparently "neutral", *i.e.*, the change does not modify the protein; however, the genomic changes may also be related to the expression of the gene. If you have complete genomes, you could explore the complete gene in different individuals, and conduct more detailed analyses, in particular comparing dN/dS (the amount of change in non-synonymous and in synonymous sites) (Hedrick 2011, page 330 and following), for instance. Also, phenotypic plasticity or differential gene expression may relate to epigenetic changes and the differences in methylation patterns between populations can now be relatively easily obtained, as mentioned above (see Steps 1, 2, and 3).

These natural selection/adaptation analyses are indeed very powerful and can be later incorporated in detail into a Genome-Wide Association Study (GWAS) analyses (see a review in Korte & Farlow 2013), to explore the genetic basis of adaptation or of a trait, where using crosses you can associate phenotypic and/or adaptive traits to particular loci or regions of the genome. For a recent GWAS study in a Solanaceae, see the analyses for *Capsicum* in Wu *et al.* (2019). A related study was carried out in a series of papers using another Solanaceae, *Datura stramonium* (known in Mexico as *toloache*) by Juan Nuñez-Farfán's group, that interested readers may want to explore (De-la-Cruz *et al.* 2020a, b, 2021).

Step 10. The next frontier: detailed coalescent analyses, in particular to estimate effective population size and evolutionary forces

The coalescent theory has proved to be revolutionary for population genetics thinking and analyses (Hedrick 2011, page 347 and following), by describing how the allelic frequencies change under different evolutionary forces and scenarios, not from the present to the future —as the classic models of Sewall Wright, R.A. Fisher and J.B.S Haldane and textbooks teach us (Eguiarte 1986, Hedrick 2011)—, but from the present to the past, by analyzing the standing variation that you sample and tracing it to the past to infer different scenarios (see for instance Hahn 2019, page 111 and following). This method allows to model different possible evolutionary histories, and to infer the more probable critical parameters, like the time of coalescence (origin) of the alleles, the effective population sizes in the present and in the past, and different patterns of gene flow and fragmentation.

One problem with this approach is that coalescence simulations of genomic data can be very computer time consuming, so usually only some evolutionary scenarios can be explored. Also, it is possible for the analyses to be very sensitive to missing populations in the sampling, and only provide you with an estimate, given all the analyzed scenarios, of which one is the most probable of the considered scenarios, while in reality all the analyzed scenarios could be wrong because of the missing populations (Beaumont 2010). That is why it is useful to have all the descriptive statistics of genetic variation and differentiation mentioned in earlier steps to inform on the more plausible scenarios that are worth analyzing in the coalescence simulations.

Also, conducting parallel analyses of the paleoclimate of the studied populations has proven very useful to ponder if the scenarios and the results are realistic (Alvarado-Serrano & Knowles 2014), as we have done in several population genomics studies conducted in Mexico that we know well, for example with teosinte (Aguirre-Liguori *et al.* 2019a, b), pumpkins (Barrera-Redondo *et al.* 2021), and yuccas (Arteaga *et al*. 2020).

A couple of programs for conducting coalescent analyses using genomic data that we can mention here are DI-YABC (Cornuet *et al*. 2014) and Fastsimcoal (Excoffier & Foll 2011). For a recent review on the use of these methods see Barrera-Redondo *et al.* (2020).

Step 11. (Optional) Using the genomic data for conservation genetics and studies of genetic resources

For some years, an important concern of modern conservation biology has been the study of the genetic aspects involved in conservation of alleles, populations, and species for population genetics (*i.e.*, conservation genetics, see Frankel & Soulé 1981, Eguiarte & Piñero 1990). For example, population genetics analyses are necessary to decide the minimal population sizes that we want to maintain in future managed populations. They are also critical to guide conservation efforts, for instance to decide which population would be more interesting or relevant for conservation (Eguiarte *et al.* 1999, Delgado *et al.* 2008, Castellanos-Morales *et al*. 2016), both for *in situ* and *ex situ* conservation, to maintain most of the genetic variation, and for future reintroductions.

The possibilities of using genomic data in conservation were very clear from the beginning, as exemplified by the now classic review of Allendorf *et al*. (2010). Genomic data allow for detailed and less biased estimates of the levels of genetic variation, which is one of the most important parameters of conservation genetics. They also permit to obtain a more reliable evaluation of the number and relationships of different groups within a species. These genetic groups within a species can define the stocks, relevant for fisheries management, and help define subspecies, varieties, lineages, or just groups of related populations that are relevant for conservation. Genomic information also allows us to infer the patterns of historical gene flow and past demographic histories, in particular the effective population sizes, which is considered a critical parameter to know (and to maintain as large as possible) for conservation biology.

Populations that have passed through bottlenecks in the recent past but are now "healthy" (*i.e.*, the populations have recovered their large sizes and the individuals have high fitness) in many cases have purged (lost) their deleterious recessive alleles, either by random genetic drift processes or by natural selection, and thus they should be very easy to be further preserved in small populations, as can be the case of botanical gardens and in ecological preserves. Interestingly, the opposite would happen in species that have had until recently very large population sizes, where these deleterious alleles accumulate if there are no selective purges, as these alleles are usually recessive and thus are seldom expressed in large populations (see for instance Morin *et al.* 2021 and reference therein). We can speculate that this is what happened in formerly very common species, as the passenger pigeon *Ectopistes migratorius*, which was a very common species in North America, perhaps in numbers of billions of individuals, but due to anthropogenic pressure (habitat change and hunting) its populations started to drastically dwindle until becoming completely extinct in 1914 (Arita 2016, page 161). Perhaps this also happened in the near-extinction of the American bison (*Bison bison*).

Genomic techniques can also allow understanding and perhaps even reducing the effect of inbreeding depression (Allendorff *et al*. 2010), which can be one of the most important risks for the long-term conservation of (some) small populations. Using genomic tools, we can explore if the inbreeding depression is caused by true overdominance (*i.e.*, advantage of the heterozygote in a case of one locus and two alleles), also called "balancing selection" (Eguiarte 1986). For instance, by looking at the age of the alleles, they should be very old in cases of balancing selection, and the levels of genetic variation along chromosomes should be very high in regions maintained by balancing selection (Hedrick 2011, pages 324-327). In addition, we could directly use GWAS and similar methods to analyze if the heterozygous individuals have indeed higher fitness. Furthermore, inbreeding depression may be caused if some alleles are defective, or even there can be missing sections of the genes in some chromosomes of the population, or they may even lack the complete gene. An inbred organism may be homozygous to the condition and thus completely lack an important protein or function. This was found in the early analyses of the genome of the potato, *Solanum tuberosum*, where some copies of a given chromosome completely lacked some genes, and thus the homozygous condition for these chromosomes was lethal (The Potato Genome Sequencing Consortium 2011).

Also, genomic data, along with selection and local adaptation studies, can illuminate the relevant adaptations of a given population that we want to preserve and that can be significant for future survival of these populations, as we will detail in the next Step (12) in relation to climate change.

Clearly, genomic analysis can help define the sampling strategies for designing germplasm collections, and for analyzing the diversity included in these collections. Also, these analyses can help design field sampling expeditions and decide which samples (accession, in the genetic resources terminology) are more relevant for preservation (*i.e.*, the most different samples, or the ones that include interesting genes for adaptation for plant improvement, or to survive global climate change). For instance, population genomic analyses can help define how many individuals are needed to guarantee a given percentage of the total gene pool represented in a germplasm collection, and to evaluate the minimum sample sizes needed to reach a collected level of the pangenome, as well as to describe the available germplasm collections in a formal population genetics way.

Conservation genomics principles and tools are very relevant for the preservation of genetic resources, for instance to analyze the descendants of the wild populations where the plants were originally selected from (that can have interesting adaptations, as disease resistances, etc.), or to find out if the samples represent a large part of the genetic pool of a cultivated species, like local landraces. We can also study the populations that, although not directly the ancestral ones, may represent populations or related species that can have relevant adaptations and genetic variation that later can be mobilized into the genome of our target species. Moving those genes to the target species can be accomplished by traditional crosses, or by assisted crosses using biotechnological methods than can allow breaking incompatibility barriers, and even using molecular engineering encompassing classic or modern methods including CRISPR-Cas9 strategies (see a recent review of these ideas in Barrera-Redondo *et al.* 2020).

Step 12. (Optional) Adaptability to climate change analyses

One related and exciting research possibility is to use genomic data to predict if a given population can adapt to different climate change scenarios. If we already have the SNPs, and if we have conducted an analysis of local adaptation, we can also use the present distribution data (geographic coordinates of the presence of the species). This distribution information can be retrieved from herbarium labels and databases, and also from specialized databases as GBIF (www.gbif.org). Then we can simulate the geographic distribution of the populations not only in the past, but also in the future (Gotelli & Stanton-Geddes 2015). Using all this information, we can further analyze if the alleles (present in the current populations) that could be potentially adapted to future climatic conditions are available in a given population (*i.e.*, drought resistance genes), or if not, if they can reach these populations (by dispersal and/or gene flow) where they are "needed", *i.e.*, where the alleles could help the species adapt and survive, and thus to estimate the probability that a given population may follow the change and adapt to it (Fitzpatrick & Keller 2015, Capblancq *et al.* 2020). Analyses of this kind were conducted for both cultivated maize and for teosintes by Aguirre-Liguori *et al.* (2019b, 2021).

Obviously, these analyses depend on how reliable the different models and scenarios are, and in the quality of the data, including the genomic information, present distribution, and the environmental variables. But it is also clear that their potential is enormous, allowing to move from knowing just the geographic distribution of the species, to assessing the probabilities of adapting to climate change.

These analyses obviously can be done for plants of economic or ecological value and also for animals that are important to these plants, as is the case of their pollinators. An example is provided by the joint analysis of species in the *Agave* genus and its main pollinators, the nectar feeding bat *Leptonycteris*, as explored in the doctoral thesis of Trejo Salazar (2022), or for the pumpkins, the genus *Cucurbita* and their specialized pollinators, the bees *Xenoglossa* spp. and *Peponapis* spp. (Giannini *et al.* 2011, Castellanos-Morales *et al.* 2018).

Step 13. (Optional) Domestication studies

Closely related to the conservation genetics and genetic resources studies is the field of domestication (Eguiarte *et al.* 2018, Barrera-Redondo *et al.* 2020, 2021). This has been a very dear field for biologists since the publication of *The Origin of Species* by Charles Darwin (1859) and his subsequent books, in particular his book on the domestication of plants and animals (Darwin 1868), and in the classic studies of Alphonse de Candolle (1883) and of Nikolai Vavilov in the first part of the last century (Vavilov 1922, 1992, Jardón Barbolla 2015).

Genomic data and the tools described in the previous steps fit perfectly well to study human-mediated evolution of cultivated plants: when, how, where and for how long have the plants been under human management. For instance, recently, we analyzed the domestication of the pipiana pumpkin, *Cucurbita argyrosperma*, using the variety of described tools; first, we sequenced and analyzed in detail the genome of the domesticated plant (*C. argyrosperma* ssp. *argyrosperma*), also using transcriptomic data to assist in the assembly and annotation, along with available genomes of other species in the genus (Barrera-Redondo *et al.* 2019). Then we sequenced the wild relative (*C. argyrosperma* ssp. *sororia)* and analyzed SNPs from many wild and cultivated populations, with SNPs obtained using GBS (Barrera-Redondo *et al.* 2021). We found that the most likely area for its domestication was western Mexico centered on the coast of Jalisco, and that the domestication process took a long time, involving constant gene flow between early domesticated plants and the wild gene pools for a long period. The coalescent estimated time of origin was very early, more than 13 thousand generations ago (since it is an annual plant, each generation is a year), compared with the accepted archeological dating of less than 10 thousand years (J. Barrera-Redondo pers. comm.). This inconsistency may be either an artifact of our used coalescence methods, or a consequence of the fragmentary and incomplete nature of the archeological record, as can be expected given the preservation complexities of the original small populations of plants during the domestication process.

We also conducted a similar analysis on the domestication of maize (Moreno-Letelier *et al*. 2020), with similar results: we found that the most probable initial domestication of maize took place in the lowlands of Jalisco in *Zea mays parviglumis* populations, also under the presence of gene flow for a long period. A later adaptation to the highlands was mediated by gene flow and introgression of adapted genes from the other wild subspecies, *Z. mays mexicana* from the highlands of central Mexico. In this case, we were unable to estimate the time of origin given the intrinsic ascertainment bias of the used Illumina 50 K chip (as the analyzed genes and SNPs were a biased sample of all the possible SNPs, given that only very variable SNPs of possible adaptive relevance were included in the original design of the chip).

Similar studies, including carefully designed and inclusive sampling as well as analyses not only of the wild populations but also of as many traditional landraces as possible, will be very important for countries like Mexico, where plant diversity is very high and where we still have many landraces being locally preserved and cultivated; and also and very importantly, where we still have the wild ancestral populations of the same species from which they were domesticated initially, along with a rich archeological record from which DNA can sometimes be extracted and analyzed, even if it is degraded (Barrera-Redondo *et al.* 2020). For analyses using genetic data from archeological samples of maize, see the papers by Ramos-Madrigal *et al.* (2016), Vallebueno-Estrada *et al.* (2016) and Swarts *et al.* (2017).

We expect to see in the near future a plethora of similar domestication studies using genomic data, and powerful coalescent analyses in Mexico and all of the Americas.

Step 14. (Optional) From the gene to the form, function, and improvement

The methods reviewed and discussed so far can be used for inferring the genetic basis of adaptation through comparative analyses of different populations. Also, as mentioned before, the genomic data can be used to infer the genealogical relationships and the degree of relatedness of individuals within populations, and if we have the phenotypes, we can now estimate the heritability of the traits of interest (Stanton-Geddes *et al.* 2013, Perrier *et al.* 2018). For instance, using GWAS and related methods, if on one hand we have the morphometric, ecological or chemical/metabolomic profiles, etc. of different individuals, and if on the other hand we know their genealogical relationships (of data derived from crosses, pedigrees, or if we can infer relatedness among the individuals in a population, as mentioned in Step 7), we can infer the genetic basis (*i.e.*, the heritability) of the studied phenotypes, and with these heritability estimates, we can predict how many generations of selection would be needed for a desired change in the population.

It is important to stress that these traits can be of ecological and evolutionary relevance for the fitness of the population, as shown in the above-mentioned *Datura stramonium* study by Juan Núñez-Farfán's laboratory. But the traits can also bear an important economic value, as for instance in *Datura*, whose secondary compounds are used in medicine, in particular the scopolamine alkaloid (De-la-Cruz *et al.* 2020a, b, 2021), and relevant for other applied reasons, like increasing the plants resistance to pests, diseases, viruses, or even to allow adaptations to new climate conditions (temperature, rain patterns and water availability) or soil types.

Once you have identified the genes or regions that contain the relevant SNPs (and again that is why it is critical to have a reference genome and transcriptome), as mentioned above (see a review in Barrera-Redondo *et al.* 2020), the plants with the desired genes and phenotypes, for example, can be later used to improve the yield of cultivated plants.

Also, we can use molecular markers to improve the artificial selection of a specific trait. The molecular marker-based selection strategy can be particularly useful for long lived-plants like trees, as instead of waiting 20 or more years to see the phenotype of the adults grown from the crosses, you can just screen the seedlings for the genotype with the marker gene related to the desired phenotype. This idea has been pursued for many years in forestry trees, for instance, in Canada in trees like *Picea*, by choosing the seedlings with the adequate genetic markers that can have a candidate gene for drought resistance. These trees will be able to live in future warmer climates (due to global change; Namroud *et al.* 2008, Holliday *et al.* 2010).

The agronomic and forestry potential of population genomics along modern genetic methods are endless and exciting, and this may allow humans to survive the difficult contaminated and extreme climates we will face in the future.

Step 15. Considering conducting phylogenomic studies

Even though this review has concentrated on the perspectives of using genomic data in a populational setting, genomic data are of obvious immense value to phylogenetic studies. It has been a long way since the earlier phylogenetic studies using only one gene in plants. Initially the chloroplast sequence of *rbcL* was used in these phylogenetic studies, some of them published in the predecessor of this journal (see for instance our phylogenetic studies of monocotyledon and agave related plants: Chase *et al.* 1993, Eguiarte *et al.* 1994, Eguiarte 1995). Later, botanists started also using nuclear regions, usually ITS and other different chloroplast regions, sometimes using both types of markers at the same time, also illustrated by our *Agave* studies (Eguiarte *et al.* 2000, Jiménez-Barron *et al.* 2020), and more recently, botanists started analyzing complete chloroplasts sequences (see for instance McKain *et al.* 2016).

The idea of using genomic data to obtain better resolved and more robust phylogenies has been in the air for a while (see for instance Conte *et al.* 2008, Cibrián-Jaramillo *et al*. 2010, Lee *et al.* 2011), but the recent massive parallel sequencing, along with different reduced representation methods has boosted their use (see for instance the recent studies of McKain *et al.* 2018, Hipp *et al.* 2019, Leebens-Mack *et al.* 2019, Kapli *et al.* 2020, Cruz-Nicolás *et al.* 2021, and Lara-Cabrera *et al.* 2021).

Nevertheless, it is very easy to encounter difficulties while attempting these phylogenomic studies, as reviewed by McKain *et al.* (2018). Some of these problems stem from different sources. One of the inherent problems of the phylogenomic studies stems directly from comparing complete genomes because the management and analysis of these large databases is daunting, coupled with increased probabilities of assembling errors in some of the used genomes, and possible sequencing artifacts. Another potential problem is that different genes will have different coalescent times, and there can be incomplete lineage sorting (less related species may have more related versions of a gene: *i.e.*, gene trees are not the same as species trees; Pamilo & Nei 1988, Maddison 1997), and thus different genes may result in slightly or very different phylogenies. These phylogenomic complexities are illustrated in the saguaro genome paper (Copetti *et al*. 2017, see their Figure 2). When comparing the genomes of different cactus species, we experienced a phenomenon called *hemiplasy*: given the complexities of the genomes, it is easy to find discordant phylogenies for different genes. In particular in plants, besides all the problematic scenarios mentioned above, both hybridization and polyploidy events are common, and this can easily result in a confusion between true homologous (called orthologous genes) and duplicated genes (paralogous) that can be lost in different lineages. These incongruent phylogenies may in some cases be also the result of hybridization processes in the past.

As the number of genes, and in particular of DNA bases, are not only in the thousands but often in the millions, it is impossible to check everything by hand. So perhaps it is a good idea not to use the complete genomes, but rather to pick some thousands of SNPs or hundreds of genes that we can possibly curate by hand for conducting phylogenomic studies.

Perspectives of plant genomic studies

We predict that the following decades will become the golden ages of genomics and omics. While the number of living angiosperms is huge, perhaps more than 350 thousand species (Ollerton *et al.* 2011), and as each species is formed by many populations, and the pangenomes of each species will be very large, we can start working now with the more interesting, charismatic, or important species, or just with our favorite taxon, using the methods briefly mentioned above, to disentangle their genetic variation, their evolutionary history and their adaptation and ecology. This paramount goal will become faster and easier to reach as many reference genomes are being assembled and annotated as initiatives such as the Earth Biogenome project and affiliated project networks (www.earthbiogenome.org) advance in their objectives of obtaining the genomes of all or most of the Earth's eukaryotic biodiversity (Lewin *et al*. 2022).

This genomic information will be also critical for conservation studies and implementations, to describe the genetic resources of the plants relevant for future agriculture and for ecology, and to preserve their populations in the future. For important species according to economic, agronomic or ecological reasons, using these data and results can help or facilitate the dispersal of the adapted population or adaptive genes, or at least to evaluate their possibilities of surviving while facing the global climatic change, new diseases and future different human related disturbances.

These future efforts will be informed by the previous work of all the botanists that collected, described and preserved plants in herbaria, from which we can now extract DNA and in some cases even RNA. These preserved plants also provide the occurrence data we need to infer the previous distributions, demographic patterns, and future possible adaptations of the species.

We recognize the impressive contributions of these earlier botanists and their insight in defining the clades, families, species and subspecies/varieties, etc., that will be critical for the development of evolutionary and genomic studies in the future, along with better and more efficient ways to extract DNA, RNA, and to sequence and analyze them. Surely, we will find many ways to use their economic and ecological potential in order to conserve their biodiversity.

Acknowledgements

We thank the organizers and editors of this special number of *Botanical Sciences* for their invitation and all the effort to edit the chapters, in particular Drs. Jorge A. Meave and Arturo de Nova for all their patience and advice and Dr. Meave for his detailed technical review. We also thank the two anonymous reviewers for their detailed comments and observations. This manuscript was written during the stay of VS and LEE in Punta Arenas, Chile, in the Centro de Estudios del Cuaternario Fuego-Patagonia Antártica (CEQUA); we deeply acknowledge all the support of Dr. Paola Acuña and the staff at the Centro. Our studies have been supported by different grants by CONACyT, CONABIO and PAPIIT, DGAPA, UNAM, in particular by the recent PAPIIT IG200122 and also by the project R20F0009 CEQUA - ANID, Chile, as well as by the Operative Budget of the Instituto de Ecología, UNAM. We also want to acknowledge all of our collaborators, students, and friends who have discussed and helped us develop the ideas presented in this paper, in particular our teacher Dr. Daniel Piñero, and Drs. Rafael Lira, Jaime Gasca-Pineda, Josué Barrera-Redondo, Alejandra Moreno-Letelier, Yocelyn Gutiérrez-Guerrero, Jonás Aguirre-Liguori, Rosalinda Tapia, Peter Tiffin, Juan Pablo Jaramillo Correa, Maud Tenaillon and Brandon S. Gaut.

Literature cited

Aguirre-Liguori JA, Gaut BS, Jaramillo-Correa JP, Tenaillon MI, Montes-Hernández S, García-Oliva F, Hearne SJ, Eguiarte LE. 2019a. Divergence with gene flow is driven by local adaptation to temperature and soil phosphorus concentration in teosinte subspecies (*Zea mays parviglumis* and *Zea mays mexicana*). *Molecular Ecology* **28**: 2814-2830. DOI: https://doi.org/10.1111/mec.15098

Aguirre-Liguori JA, Luna-Sánchez JA, Gasca-Pineda J, Eguiarte LE. 2020. Evaluation of the minimum sampling design for population genomic and microsatellite studies: an analysis based on wild maize. *Frontiers in Genetics* **11**: 870. DOI: https://doi.org/10.3389/fgene.2020.00870

Aguirre-Liguori JA, Ramírez-Barahona S, Gaut BS. 2021. The evolutionary genomics of species' responses to climate change. *Nature Ecology & Evolution* **5**: 1350-1360. DOI: https://doi.org/10.1038/s41559-021-01526-9

Aguirre-Liguori JA, Ramírez-Barahona S, Tiffin P, Eguiarte LE. 2019b. Climate change is predicted to disrupt patterns of local adaptation in wild and cultivated maize. *Proceedings of the Royal Society B* **286**: 20190486. DOI: https://doi.org/10.1098/rspb.2019.0486

Aguirre-Liguori JA, Tenaillon MI, Vázquez-Lobo A, Gaut BS, Jaramillo-Correa JP, Montes-Hernandez S, Souza V, Eguiarte LE. 2017. Connecting genomic patterns of local adaptation and niche suitability in teosintes. *Molecular Ecology* **26**: 4226-4240. DOI: https://doi.org/10.1111/mec.14203

Aguirre-von-Wobeser E, Rocha-Estrada J, Shapiro LR, de la Torre M. 2018. Enrichment of Verrucomicrobia, Actinobacteria and Burkholderiales drives selection of bacterial community from soil by maize roots in a traditional milpa agroecosystem. *PloS One* **13**: e0208852. DOI: https://doi.org/10.1371/journal.pone.0208852

Ahrens CW, Rymer PD, Stow A, Bragg J, Dillon S, Umbers KDL, Dudaniec RY. 2018. The search for loci under selection: trends, biases and progress. *Molecular Ecology* **27**: 1342-1356. DOI: https://doi.org/10.1111/mec.14549

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**: 1655-1664. DOI: https://doi.org/10.1101/gr.094052.109

Allendorf FW, Hohenlohe PA, Luikart G. 2010. Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**: 697-709. DOI: https://doi.org/10.1038/nrg2844

Alvarado-Serrano DF, Knowles LL. 2014. Ecological niche models in phylogeographic studies: applications, advances and precautions. *Molecular Ecology* **14**: 233-248. DOI: https://doi.org/10.1111/1755-0998.12184

Anderson EC, Thompson E. 2002. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160:** 1217-1229. DOI: https://doi.org/10.1093/genetics/160.3.1217

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe P. 2016. Harnessing the power of RAD-seq for ecological and evolutionary genomics. *Nature Reviews Genetics* **17**: 81-92. DOI: https://doi.org/10.1038/nrg.2015.28

Andrews KR, Luikart G. 2014. Recent novel approaches for population genomics data analysis. *Molecular Ecology* **23**: 1661-1667. DOI: https://doi.org/10.1111/mec.12686

Arita HT. 2016. *Crónicas de la Extinción: La Vida y la Muerte de las Especies Animales.* Mexico City: Fondo de Cultura Económica. ISBN: 978-6071646125

Arteaga MC, Bello-Bedoy R, Gasca-Pineda J. 2020. Hybridization between yuccas from Baja California: Genomic and environmental patterns. *Frontiers in Plant Science* **11:** 685. DOI: https://doi.org/10.3389/fpls.2020.00685

Barrera-Redondo J, Ibarra-Laclette E, Vázquez-Lobo A, Gutiérrez-Guerrero YT, Sánchez de la Vega G, Piñero D, Montes-Hernández S, Lira-Saade R, Eguiarte LE. 2019. The genome of *Cucurbita argyrosperma* (silver-seed gourd) reveals faster rates of protein-coding gene and long noncoding RNA turnover and neofunctionalization within *Cucurbita*. *Molecular Plant* **12:** 506-520. DOI: https://doi.org/10.1016/j.molp.2018.12.023

Barrera-Redondo J, Piñero D, Eguiarte LE. 2020. Genomic, transcriptomic and epigenomic tools to study the domestication of plants and animals: A field guide for beginners. *Frontiers in Genetics* **11:** 742. DOI: https://doi.org/10.3389/fgene.2020.00742

Barrera-Redondo J, Sánchez-de la Vega G, Aguirre-Liguori JA, Castellanos-Morales G, Gutiérrez-Guerrero YT, Aguirre-Dugua X, Aguirre-Planter E, Tenaillon MI, Lira-Saade R, Eguiarte LE. 2021. The domestication of *Cucurbita argyrosperma* as revealed by the genome of its wild relative. *Horticulture Research* **8:** 109. DOI: https://doi.org/10.1038/s41438-021-00544-9

Beaumont MA. 2010. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics* **41:** 379-406. DOI: https://doi.org/10.1146/annurev-ecolsys-102209-144621

Beerli P, Mashayekhi S, Sadeghi M, Khodaei M, Shaw K. 2019. Population genetic inference with MIGRATE. *Current Protocols in Bioinformatics* **68:** e87. DOI: https://doi.org/10.1002/cpbi.87

Borges RM. 2017. Co-niche construction between hosts and symbionts: ideas and evidence. *Journal of Genetics* **96:** 483-489. DOI: https://doi.org/10.1007/s12041-017-0792-9

Bourge M, Brown SC, Siljak-Yakovlev S. 2018. Flow cytometry as tool in plant sciences, with emphasis on genome size and ploidy level assessment. *Genetics & Applications* **2**: 1-12.

Capblancq T, Fitzpatrick MC, Bay RA, Exposito-Alonso M, Keller SR. 2020. Genomic prediction of (Mal)adaptation across current and future climatic landscapes. *Annual Review of Ecology, Evolution and Systematics* **51:** 245-269. DOI: https://doi.org/10.1146/annurev-ecolsys-020720-042553

Castellanos-Morales G, Gutiérrez-Guerrero YT, Gámez N, Eguiarte LE. 2016. Use of molecular and environmental analyses for integrated *in situ* and *ex situ* conservation: The case of the Mexican prairie dog. *Biological Conservation* **204**: 284-95. DOI: https://doi.org/10.1016/j.biocon.2016.10.036

Castellanos-Morales G, Paredes-Torres LM, Gámez N, Hernández-Rosales HS, Sánchez-de la Vega G, Barrera-Redondo J, Aguirre-Planter E, Vázquez-Lobo A, Montes-Hernández S, Lira-Saade R, Eguiarte LE. 2018. Historical biogeography and phylogeny of *Cucurbita*: insights from ancestral area reconstruction and niche evolution. *Molecular Phylogenetics and Evolution* **128:** 38-54. DOI: https://doi.org/10.1016/j.ympev.2018.07.016

Caye K, Jumentier B, Lepeule J, François O. 2019. LFMM 2: Fast and accurate inference of gene-environment associations in genome-wide studies. *Molecular Biology and Evolution* **36:** 852-860. DOI: https://doi.org/10.1093/molbev/msz008

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu Y-L, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim K-J, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang Q-Y, Plunkett

GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn Jr GH, Graham SW, Barrett SCH, Dayanandan S, Albert VA. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. *Annals of the Missouri Botanical Garden* **80:** 528-580. DOI: https://doi.org/10.2307/2399846

Cibrián-Jaramillo A, De la Torre-Bárcena JE, Lee EK, Katari MS, Little DP, Stevenson DW, Martienssen R, Coruzzi GM, DeSalle R, 2010. Using phylogenomic patterns and gene ontology to identify proteins of importance in plant evolution. *Genome Biology and Evolution* **2:** 225-239. DOI: https://doi.org/10.1093/gbe/evq012

Compant S, Samad A, Faist H, Sessitsch A. 2019. A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application. *Journal of Advanced Research* **19:** 29-37. DOI: https://doi.org/10.1016/j.jare.2019.03.004

Conte MG, Gaillard S, Droc G, Perin C. 2008. Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. *BMC Genomics* **9:** 183. DOI: https://doi.org/10.1186/1471-2164-9-183

Coop G, Witonsky D, Di Rienzo A, Pritchard JK. 2010. Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185:** 1411-1423. DOI: https://doi.org/10.1534/genetics.110.114819

Copetti D, Búrquez A, Bustamante E, Charboneau JLM, Childs KL, Eguiarte LE, Lee S, Liu TL, McMahon MM, Whiteman NK, Wing RA, Wojciechowski MF, Sanderson MJ. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. *Proceedings of the National Academy of Sciences of the United States of America* **114:** 12003-12008. DOI: https://doi.org/10.1073/pnas.1706367114

Cornejo OE, Yee M-C, Dominguez V, Andrews M, Sockell A, Strandberg E, Livingstone III D, Stack C, Romero A, Umaharan P, Royaert S. Tawari NR, Ng Pauline, Gutierrez O, Philips W, Mockaitis K, Bustamante CD, Motamayor JC. 2018. Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. *Communications Biology* **1**: 167. DOI: https://doi.org/10.1038/s42003-018-0168-6

Cornuet J-M, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, Marin J-M, Estoup A. 2014. DIYABC V2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* **30:** 1187-1189. DOI: https://doi.org/10.1093/bioinformatics/btt763

Cosgrove A. 2021. Probing nuclear architecture with Hi-C. *Nature Milestones, Genomic Sequencing*, **2021:** S14.

Cruz-Nicolás J, Villarruel-Arroyo A, Gernandt DS, Fonseca RM, Aguirre-Planter E, Eguiarte LE, Jaramillo-Correa JP. 2021. Non-adaptive evolutionary processes governed the diversification of a temperate conifer lineage after its migration into the tropics. *Molecular Phylogenetics and Evolution* **160**: 107125. DOI: https://doi.org/10.1016/j.ympev.2021.107125

Darwin CR. 1859. *On the Origin of Species.* London: John Murray. ISBN: 978-1546622499

Darwin CR. 1868. *The Variation of Animals and Plants Under Domestication.* London: John Murray. ISBN: 978-0814720639

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker. Discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12:** 499-510. DOI: https://doi.org/10.1038/nrg3012

David P, Pujol B, Viard F, Castella V, Goudet J. 2007. Reliable selfing rate estimates from imperfect population genetic data. *Molecular Ecology* **16:** 2474-2487. DOI: https://doi.org/10.1111/j.1365-294X.2007.03330.x

de Bruijn FJ. 2015. Biological nitrogen fixation. *In*: Lugtenberg B, ed. *Principles of Plant-Microbe Interactions.* Cham: Springer, pp. 215-224. DOI: https://doi.org/10.1007/978-3-319-08575-3_23

De Candolle A. 1883. *Origine des Plantes Cultivées.* Paris: Librairie Germer Baillière et Cie. ISBN: 978-1514228043

De-la-Cruz IM, Cruz LL, Martínez-García L, Valverde PL, Flores-Ortiz CM, Hernández-Portilla LB, Núñez-Farfán J. 2020a. Evolutionary response to herbivory: population differentiation in microsatellite loci, tropane alkaloids and leaf trichome density in *Datura stramonium*. *Arthropod-Plant Interactions* **14:** 21-30. DOI: https://doi.org/10.1007/s11829-019-09735-7

De-la-Cruz IM, Hallab A, Olivares-Pinto U, Tapia-López R, Velázquez-Márquez S, Piñero D, Oyama K, Usadel B, Núñez-Farfán J. 2021. Genomic signatures of the evolution of defence against its natural enemies in the poisonous and medicinal plant *Datura stramonium* (Solanaceae). *Scientific Reports* **11:** 882. DOI: https://doi.org/10.1038/s41598-020-79194-1

De-la-Cruz IM, Merilä J, Valverde PL, Flores-Ortiz CM, Núñez-Farfán J. 2020b. Genomic and chemical evidence for local adaptation in resistance to different herbivores in *Datura stramonium*. *Evolution* **74**: 2629-2643. DOI: https://doi.org/10.1111/evo.14097

De Mita S, Thuillet A-C, Gay L, Ahmadi N, Manel S, Ronfort J, Vigouroux Y. 2013. Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology* **22:** 1383-1399. DOI: https://doi.org/10.1111/mec.12182

de Villemereuil P, Gaggiotti OE. 2015. A new F_{ST}-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution* **6:** 1248-1258. DOI: https://doi.org/10.1111/2041-210X.12418

Delgado P, Eguiarte LE, Molina-Freaner F, Alvarez-Buylla ER, Piñero D. 2008. Using phylogenetic, genetic and demographic evidence for setting conservation priorities for Mexican rare pines. *Biodiversity and Conservation* **17:** 121-137. DOI: https://doi.org/10.1007/s10531-007-9234-y

Díez CM, Gaut BS, Meca E, Scheinvar E, Montes-Hernandez S, Eguiarte LE, Tenaillon MI. 2013. Genome size variation in wild and cultivated maize along altitudinal gradients. *New Phytologist* **199:** 264-276. DOI: https://doi.org/10.1111/nph.12247

Earl DA, vonHoldt BM. 2012. Structure Harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**: 359-361. DOI: https://doi.org/10.1007/s12686-011-9548-7

Eguiarte LE. 1986. Una guía para principiantes a la genética de poblaciones. *Ciencias Número Especial* **1:** 30-39.

Eguiarte LE. 1995. Hutchinson (Agavales) vs. Huber y Dahlgren (Asparagales): análisis moleculares sobre la filogenia y evolución de la familia Agavaceae *sensu* Hutchinson dentro de las monocotiledóneas. *Botanical Sciences* **56**: 45-56. DOI: https://doi.org/10.17129/botsci.1463

Eguiarte LE, Duvall MRR, Learn Jr GH, Clegg MT. 1994. The systematic status of the Agavaceae and Nolinaceae and related Asparagales in the monocotyledons: An analysis based on the rbcL gene sequence. *Botanical Sciences* **54:** 35-56. DOI: https://doi.org/10.17129/botsci.1427

Eguiarte LE, Hernández-Rosales HS, Barrera-Redondo J, Castellanos-Morales G, Paredes-Torres LM, Sánchez-de la Vega G, Ruiz-Mondragón KY, Vázquez-Lobo A, Montes-Hernández S, Aguirre-Planter E, Souza V. 2018. Domesticación, diversidad y recursos genéticos y genómicos de México: El caso de las calabazas. *TIP. Revista Especializada en Ciencias Químico-Biológicas* **21:** 85-101. DOI: https://doi.org/10.22201/fesz.23958723e.2018.0.159

Eguiarte LE, Larson-Guerra J, Nuñez-Farfán J, Martinez-Palacios A, Santos Del Prado K, Arita HT. 1999. Diversidad filogenética y conservación: ejemplos a diferentes escalas y una propuesta a nivel poblacional para *Agave victoriae-reginae* en el desierto de Chihuahua, México. *Revista Chilena de Historia Natural* **74:** 475-92.

Eguiarte LE. Piñero D. 1990. Genética de conservación: leones vemos, genes no sabemos. *Ciencias Número Especial* **4:** 34-97.

Eguiarte LE, Souza V, Aguirre X, eds. 2007. *Ecología Molecular*. Mexico City: Instituto Nacional de Ecología. ISBN: 978-9688178393

Eguiarte LE, Souza V, Silva-Montellano A. 2000. Evolución de la familia Agavaceae: filogenia, biología reproductiva y genética de poblaciones. *Botanical Sciences* **66:**131-50. DOI: https://doi.org/10.17129/botsci.1618

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14:** 2611-2620. DOI: https://doi.org/10.1111/j.1365-294X.2005.02553.x

Excoffier L, Foll M. 2011. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27:** 1332-1334. DOI: https://doi.org/10.1093/bioinformatics/btr124

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* **61:** 717-726. DOI: https://doi.org/10.1093/sysbio/sys004

Ferretti L, Ramos-Onsins SE, Pérez-Enciso M. 2013. Population genomics from pool sequencing. *Molecular Ecology* **22:** 5561-5576. DOI: https://doi.org/10.1111/mec.12522

Figueroa-Corona L, Delgado Valerio P, Wegrzyn J, Piñero D. 2021. Transcriptome of weeping pinyon pine, *Pinus pinceana*, shows differences across heterogeneous habitats. *Trees* **35:** 1351-1365. DOI: https://doi.org/10.1007/s00468-021-02125-8

Fitzpatrick MC, Keller SR. 2015. Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters* **18:** 1-16. DOI: https://doi.org/10.1111/ele.12376

Fitzpatrick CR, Salas-González I, Conway JM, Finkel OM, Gilbert S, Russ D, Teixeira PJPL, Dangl JL. 2020. The plant microbiome: From ecology to reductionism and beyond. *Annual Review of Microbiology* **74:** 81-100. DOI: https://doi.org/10.1146/annurev-micro-022620-014327

Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* **180:** 977-993. DOI: https://doi.org/10.1534/genetics.108.092221

Frankel OH, Soulé ME. 1981. *Conservation and Evolution*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/S0030605300017853

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147:** 915-925. DOI: https://doi.org/10.1093/genetics/147.2.915

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* **133:** 693-709. DOI: https://doi.org/10.1093/genetics/133.3.693

Fustier M-A, Brandenburg J-T, Boitard S, Lapeyronnie J, Eguiarte LE, Vigouroux Y, Manicacci D, Tenaillon MI. 2017. Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples. *Molecular Ecology* **26:** 2738-2756. DOI: https://doi.org/10.1111/mec.14082

Gasca-Pineda J, Gutiérrez-Guerrero YT, Aguirre-Planter E, Eguiarte LE. 2020. The role of environment, local adaptation, and past climate fluctuation on the amount and distribution of genetic diversity in two subspecies of Mexican wild *Zea mays*. *American Journal of Botany* **107:** 1542-1554. DOI: https://doi.org/10.1002/ajb2.1561

Giannini TC, Lira-Saade R, Ayala R, Saraiva AM, Alves-dos-Santos I. 2011. Ecological niche similarities of *Peponapis* bees and non-domesticated *Cucurbita* species. *Ecological Modelling* **222:** 2011-2018. DOI: https://doi.org/10.1016/j.ecolmodel.2011.03.031

Gotelli NJ, Stanton-Geddes J. 2015. Climate change, genetic markers and species distribution modelling. *Journal of Biogeography* **42:** 1577-1585. DOI: https://doi.org/10.1111/jbi.12562

Grotkopp E, Rejmánek M, Sanderson MJ, Rost TL. 2004. Evolution of genome size in pines (*Pinus*) and its life-history correlates: Supertree analyses. *Evolution* **58:** 1705-1729. DOI: https://doi.org/10.1111/j.0014-3820.2004.tb00456.x

Guerra-García A, Suárez-Atilano M, Mastretta-Yanes A, Delgado-Salinas A, Piñero D. 2017. Domestication genomics of the open-pollinated scarlet runner bean (*Phaseolus coccineus* L.). *Frontiers in Plant Science* **8**: 1891. DOI: https://doi.org/10.3389/fpls.2017.01891

Hahn MW. 2019. *Molecular Population Genetics*. New York: Oxford University Press. ISBN: 978-0878939657

Hamrick JL, Godt MW. 1990. Allozyme diversity in plant species. *In*: Brown AHD, Clegg MT, Kahler AL, Weir BS, eds. *Plant Population Genetics, Breeding, and Genetic Resources*. pp 43-63. Sunderland: Sinauer. DOI: http://dx.doi.org/10.2307/1311547

Hedrick PW. 2011. *Genetics of Populations*. 4th ed. Sudbury: Jones and Bartlett Publishers. ISBN: 978-0763757373.

Hernández-Álvarez C, García-Oliva F, Cruz-Ortega R, Romero MF, Barajas HR, Piñero D, Alcaraz LD. 2022. Squash root microbiome transplants and metagenomic inspection for *in situ* arid adaptations. *Science of The Total Environment* **805**: 150136. DOI: https://doi.org/10.1016/j.scitotenv.2021.150136

Heyduk K, McKain MR, Lalani F, Leebens-Mack J. 2016a. Evolution of a CAM anatomy predates the origins of Crassulacean acid metabolism in the Agavoideae (Asparagaceae). *Molecular Phylogenetics and Evolution* **105:** 102-113. DOI: https://doi.org/10.1016/j.ympev.2016.08.018

Heyduk K, Trapnell DW, Barrett CF, Leebens-Mack J. 2016b. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biological Journal of the Linnean Society* **117:** 106-120. DOI: https://doi.org/10.1111/bij.12551

Hime PM, Lemmon AR, Lemmon ECM, Prendini E, Brown JM, Thomson RC, Kratovil JD, Noonan BP, Pyron RA, Peloso PL, Kortyna ML. 2021. Phylogenomics reveals ancient gene tree discordance in the amphibian tree of life. *Systematic Biology* **70:** 49-66. DOI: https://doi.org/10.1093/sysbio/syaa034

Hipp AL, Manos PS, Hahn M, Avishai M, Bodénès C, Cavender-Bares J, Crowl AA, Deng M, Denk T, Fitz-Gibbon S, Galling O, González-Elizondo S, González-Rodríguez A, Grimm GW, Jiang X-L, Kremer A, Lesur I, McVay JD, Plomion C, Rodríguez-Correa H, Schulze E-D, Simeone MC, Sork VL, Valencia-Avalos S. 2019. Genomic landscape of the global oak phylogeny. *New Phytologist* **226**: 1198-1212- DOI: https://doi.org/10.1111/nph.16162

Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, Poss ML, Reed LK, Storfer A, Whitlock MC. 2016. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist* **188:** 379-397. DOI: https://doi.org/10.1086/688018

Holliday JA, Ritland K, Aitken SN. 2010. Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). *New Phytologist* **188:** 501-514. DOI: https://doi.org/10.1111/j.1469-8137.2010.03380.x

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921. DOI: https://doi.org/10.1038/35057062

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang T-H, Lan T, Welch AJ, Juárez MJA, Simpson J, Fernández-Cortés A, Arteaga-Vázquez M, Góngora-Castillo E, Acevedo-Hernández G, Schuster SC, Himmelbauer H, Minoche AE, Xu S, Lynch M, Oropeza-Aburto A, Cervantes-Pérez SA, Ortega-Estrada MJ, Cervantes-Luevano JI, Michael TP, Mockler T, Bryant D, Herrera-Estrella A, Albert VA, Herrera-Estrella L. 2013. Architecture and evolution of a minute plant genome. *Nature* **498:** 94-98. DOI: https://doi.org/10.1038/nature12132

Janes JK, Miller JM, Dupuis JR, Malefant RM, Gorrell JC, Cullingham CI, Andrew RL. 2017. The *K* = 2 conundrum. *Molecular Ecology* **26:** 3594-3602. DOI: https://doi.org/10.1111/mec.14187

Jardón Barbolla L. 2015. Orígenes y diversidad en las montañas: Nicolai Vavilov, México y las plantas domesticadas. *Oikos* **14:** 6-10.

Jiménez-Barron O, García-Sandoval R, Magallón S, García-Mendoza A, Nieto-Sotelo J, Aguirre-Planter E, Eguiarte LE. 2020. Phylogeny, diversification rate, and divergence time of *Agave sensu lato* (Asparagaceae), a group of recent origin in the process of diversification. *Frontiers in Plant Science* **11:** 536135. DOI: https://doi.org/10.3389/fpls.2020.536135

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11:** 94. DOI: https://doi.org/10.1186/1471-2156-11-94

Kapli P, Yang Z, Telford MJ. 2020. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* **21:** 428-444. DOI: https://doi.org/10.1038/s41576-020-0233-0

Korte A, Farlow A, 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9:** 29. DOI: https://doi.org/10.1186/1746-4811-9-29

LaFlamme B. 2021. Genomes go platinum. *Nature Milestones. Genomic Sequencing* **2021:** S20.

Lara-Cabrera SI, Perez-Garcia ML, Maya-Lastra CA, Montero-Castro JC, Godden GT, Cibrian-Jaramillo A, Fisher AE, Porter JM. 2021. Phylogenomics of *Salvia* L. subgenus *Calosphace* (Lamiaceae). *Frontiers in Plant Science* **12:** 725900. DOI: https://doi.org/10.3389/fpls.2021.725900

Lee EK, Cibrian-Jaramillo A, Kolokotronis S-O, Katari MS, Stamatakis A, Ott M, Chiu JC, Little DP, Stevenson DW, McCombie WR, Martienssen RA, Coruzzi G, DeSalle R. 2011. A functional phylogenomic view of the seed plants. *Plos Genetics* **7:** e1002411. DOI: https://doi.org/10.1371/journal.pgen.1002411

Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S, Porsch M. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574:** 679-685. DOI: https://doi.org/10.1038/s41586-019-1693-2

Lewin HA, Richards S, Aiden EL, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, Blaxter ML, Cai J, Caperello ND, Carlson K, Castilla-Rubio JC, Chaw S-M, Chen L, Childers AK, Coddington JA, Conde DA, Corominas M, Crandall KA, Crawford AJ, DiPalma F, Durbin R, Ebenezer TE, Edwards SV, Fedrigo O, Flicek P, Formenti G, Gibbs RA, Gilbert MTP, Goldstein MM, Graves JM, Greely HT, Grigoriev IV, Hackett KJ, Hall N, Haussler D, Helgen KM, Hogg CJ, Isobe S, Jakobsen KS, Janke A, Jarvis ED, Johnson WE, Jones SJM, Karlsson EK, Kersey PJ, Kim J-H, Kress WJ, Kuraku S, Lawniczak MKN, Leebens-Mack JH, Li X, Lindblad-Toh K, Liu X, Lopez JV, Marques-Bonet T, Mazard S, Mazet JAK, Mazzoni CJ, Myers EW, O'Neill RJ, Paez S, Park H, Robinson GE, Roquet C, Ryder OA, Sabir JSM, Shaffer HB, Shank TM, Sherkow JS, Soltis PS, Tang B, Tedersoo L, Uliano-Silva M, Wang K, Wei X, Wetzer R, Wilson JL, Xu X, Yang H, Yoder AD, Zhang G. 2022. The Earth Biogenome Project 2020: starting the clock. *Proceedings of the National Academy of Science of the United States of America* **119:** e2115635118. DOI: https://doi.org/10.1073/pnas.2115635118

Lewontin RC, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74:** 175-95. DOI: https://doi.org/10.1093/genetics/74.1.175

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475:** 493-496. DOI: https://doi.org/10.1038/nature10231

Liu S, Westbury MB, Dussex N, Mitchell KJ, Sinding M-HS, Heintzman PD, Duchêne DA, Kapp JD, von Seth J, Heiniger H, Sánchez-Barreiro F, Margaryan A, André-Olsen R, De Cahsan B, Meng G, Yang C, Chen L, der Valk T, Moodley Y, Rookmaaker K, Bruford MW, Ryder O, Steiner C, Bruins-van Sonsbeek LGR, Vartanyan S, Guo C, Cooper A, Kosintsev P, Kirillova I, Lister AM, Marques-Bonet T, Gopalakrishnan S, Dunn RR, Lorenzen ED, Shapiro B, Zhang G, Antoine P-O, Dalén L, Gilbert MTP. 2021. Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. *Cell* **184:** 4874-4885. DOI: https://doi.org/10.1016/j.cell.2021.07.032

Liu Y, Siegmund KD, Laird P, Berman BP. 2012. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biology* **13:** R61. DOI: https://doi.org/10.1186/gb-2012-13-7-r61

Lotterhos KE, Whitlock MC. 2015. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology* **24**: 1031-1046. DOI: https://doi.org/10.1111/mec.13100.

Luu K, Bazin E, Blum MG. 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources* **17**: 67-77. DOI: https://doi.org/10.1111/1755-0998.12592

Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* **46:** 523-536. DOI: https://doi.org/10.1093/sysbio/46.3.523

Martínez-González C, Castellanos-Morales G, Barrera-Redondo J, Sánchez-de la Vega G, Hernández-Rosales HS, Gasca-Pineda J, Aguirre-Planter E, Moreno-Letelier A, Escalante AE, Montes-Hernández S, Lira-Saade R. 2021. Recent and historical gene flow in cultivars, landraces, and a wild taxon of *Cucurbita pepo* in Mexico. *Frontiers in Ecology and Evolution* **9:** 656051. DOI: https://doi.org/10.3389/fevo.2021.656051

Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC. 2014. Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology* **15**: 28-41. DOI: https://doi.org/10.1111/1755-0998.12291

McKain MR, Johnson MG, Uribe-Convers S, Eaton D, Yang Y. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences* **6:** e1038. DOI: https://doi.org/10.1002/aps3.1038

McKain MR, McNeal JR, Kellar PR, Eguiarte LE, Pires JC, Leebens-Mack J. 2016. Timing of rapid diversification and convergent origins of active pollination within Agavoideae (Asparagaceae). *American Journal of Botany* **103:** 1717-1729. DOI: https://doi.org/10.3732/ajb.1600198

Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, Acuña-Alonzo V, Sandoval K, Eng C, Romero-Hidalgo S, Ortiz-Tello P. 2014. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science*. **344:** 1280-1285. DOI: https://doi.org/10.1126/science.1251688

Moreno-Letelier A, Aguirre-Liguori JA, Piñero D, Vázquez-Lobo A, Eguiarte LE. 2020. The relevance of gene flow with wild relatives in understanding the domestication process. *Royal Society Open Science* **7:** 191545. DOI: https://doi.org/10.1098/rsos.191545

Morin PA, Archer FI, Avila CD, Balacco JR, Bukhman YV, Chow W, Fedrigo O, Formenti G, Fronczek JA, Fungtammasan A, Gulland FMD, Haase B, Heide-Jorgensen MP, Houck ML, Howe K, Misuraca AC, Mountcastle J, Musser W, Paez S, Pelan S, Phillippy A, Rhie A, Robinson J, Rojas-Bracho L, Rowles TK, Ryder OA, Smith CR, Stevenson S, Taylor BL, Teilmann J, Torrance J, Wells RS, Westgate AJ, Jarvis ED. 2021. Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Molecular Ecology Resources* **21**:1008-1020. DOI: https://doi.org/10.1111/1755-0998.13284

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J. 2008. Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology* **17:** 3599-3613. DOI: https://doi.org/10.1111/j.1365-294X.2008.03840.x

Nature Milestones. *Genomic Sequencing*. 2021. **2021:** S1-S21.

Ollerton J, Winfree R, Tarrant S. 2011. How many flowering plants are pollinated by animals? *Oikos* **120:** 321-326. DOI: https://doi.org/10.1111/j.1600-0706.2010.18644.x

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5:** 568-583. DOI: https://doi.org/10.1093/oxfordjournals.molbev.a040517

Paun O, Verhoeven KJ, Richards CL. 2019. Opportunities and limitations of reduced representation bisulfite sequencing in plant ecological epigenomics. *New Phytologist* **221:** 738-742. DOI: https://doi.org/10.1111/nph.15388

Perrier C, Delahaie B, Charmantier A. 2018. Heritability estimates from genomewide relatedness matrices in wild populations: Application to a passerine, using a small sample size. *Molecular Ecology Resources* **18:** 838-853. DOI: https://doi.org/10.1111/1755-0998.12886

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One* **7:** e37135. DOI: https://doi.org/10.1371/journal.pone.0037135

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8:** e1002967 DOI: https://doi.org/10.1371/journal.pgen.1002967

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **15:** 945-959. DOI: https://doi.org/10.1093/genetics/155.2.945

Ramos-Madrigal J, Smith BD, Moreno-Mayar JV, Gopalakrishnan S, Ross-Ibarra J, Gilbert MTP, Wales N. 2016. Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. *Current Biology* **26:** 3195-31201. DOI: https://doi.org/10.1016/j.cub.2016.09.036

Rebollar EA, Sandoval-Castellanos E, Roessler K, Gaut BS, Alcaraz LD, Benítez M, Escalante AE. 2017. Seasonal changes in a maize-based polyculture of central Mexico reshape the co-occurrence networks of soil bacterial communities. *Frontiers in Microbiology* **8:** 2478. DOI: https://doi.org/10.3389/fmicb.2017.02478

Rebolleda Gómez M, Ashman TL. 2019. Floral organs act as environmental filters and interact with pollinators to structure the yellow monkeyflower (*Mimulus guttatus*) floral microbiome. *Molecular Ecology* **28**: 5155-5171. DOI: https://doi.org/10.1111/mec.15280

Rebolleda-Gómez M, Forrester NJ, Russell AL, Wei N, Fetters AM, Stephens JD, Ashman TL. 2019. Gazing into the anthosphere: considering how microbes influence floral evolution. *New Phytologist* **224:** 1012-1020. DOI: https://doi.org/10.1111/nph.16137

Schlötterer C, Tobler R, Kofler R, Nolte V. 2014. Sequencing pools of individuals –mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics* **15:** 749-763. DOI: https://doi.org/10.1038/nrg3803

Stanton-Geddes J, Yoder JB, Briskine R, Young ND, Tiffin P. 2013. Estimating heritability using genomic data. *Methods in Ecology and Evolution* **4:** 1151-1158. DOI: https://doi.org/10.1111/2041-210X.12129

Swarts K, Gutaker RM, Benz B, Blake M, Bukowski R, Holland J, Kruse-Peeples M, Lepak N, Prim L, Romay MC, Ross-Ibarra J, Sanchez-Gonzalez JJ, Schmidt C, Schuenemann VJ, Krause J, Matson RG, Weigel D, Buckler ES, Burbano HA. 2017. Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. *Science* **357:** 512-515. DOI: https://doi.org/10.1126/science.aam9425

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585-595. DOI: https://doi.org/10.1093/genetics/123.3.585

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* **102:** 13950-13955. DOI: https://doi.org/10.1073/pnas.0506758102

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796-815. DOI: https://doi.org/10.1038/35048692

The Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato *Nature* **475:** 189-195. DOI: https://doi.org/10.1038/nature10158

Trejo Salazar RE. 2022. *Filogeografía y Conservación del Murciélago Magueyero Menor* Leptonycteris yerbabuenae *(Martínez y Villa 1940)*. PhD Thesis, Universidad Nacional Autónoma de México.

Trivedi P, Leach JE, Tringe SG, Sa T, Singh BK. 2020. Plant–microbiome interactions: from community assembly to plant health. *Nature Reviews Microbiology* **18:** 607-621. DOI: https://doi.org/10.1038/s41579-020-0412-1

Vallebueno-Estrada M, Rodríguez-Arévalo I, Rougon-Cardoso A, González JM, Cook AG, Montiel R, Vielle-Calzada JP. 2016. The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. *Proceedings of the National Academy of Sciences of the United States of America* **113:** 14151-14156. DOI: https://doi.org/10.1073/pnas.1609701113

van der Heijden MGA, Martin FM, Selosse M-A, Sanders IR. 2015. Mycorrhizal ecology and evolution: the past, the present, and the future. *New Phytologist* **205:** 1406-1423. DOI: https://doi.org/10.1111/nph.13288

van Gurp TP, Wagemaker NCAM, Wouters B, Vergeer P, Ouborg JNJ, Verhoeven KJF. 2016. epiGBS: reference-free reduced representation bisulfite sequencing. *Nature Methods* **13:** 322-324. DOI: https://doi.org/10.1038/nmeth.3763

Vandenkoornhuyse P, Quaiser A, Duhamel M, Le Van A, Dufresne A. 2015. The importance of the microbiome of the plant holobiont. *New Phytologist* **206:** 1196-1206. DOI: https://doi.org/10.1111/nph.13312

Vavilov NI. 1922. The law of homologous series in variation. *Journal of Genetics* **12:**47-89. DOI: https://doi.org/10.1007/BF02983073

Vavilov NI.1992. *Origin and Geography of Cultivated Plants*. Cambridge: Cambridge University Press. ISBN 0-521-49427-4

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, 265 more authors. 2001. The sequence of the human genome. *Science* **291:** 1304-1351. DOI: https://doi.org/10.1126/science.1058040

Wu L, Wang P, Wang Y, Cheng Q, Lu Q, Liu J, Li T, Ai Y, Yang W, Sun L, Shen H. 2019. Genome-wide correlation of 36 agronomic traits in the 287 pepper (*Capsicum*) accessions obtained from the SLAF-seq-Based GWAS. *International Journal of Molecular Sciences* **20:** 5675. DOI: https://doi.org/10.3390/ijms20225675

Wilson GA, Rannala B. 2003. Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163:** 1177-1191. DOI: https://doi.org/10.1093/genetics/163.3.1177

Workman PL, Niswander JD. 1970. Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. *American Journal of Human Genetics* **22:** 24-29.

Wrighton KH. 2021. Filling in the gaps telomere to telomer. *Nature Milestones*. *Genomic Sequencing* **2021:** S21.

Xu Q, Zhu C, Fan Y, Song Z, Xing S, Liu W, Yan J, Sang T. 2016. Population transcriptomics uncovers the regulation of gene expression variation in adaptation to changing environment. *Scientific Reports* **6:** 25536. DOI: https://doi.org/10.1038/srep25536

Zaidem ML, Groen SC, Purugganan MD. 2019. Evolutionary and ecological functional genomics, from lab to the wild. *The Plant Journal* **97:** 40-55. DOI: https://doi.org/10.1111/tpj.14167

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ, Neale DB, Salzberg SL, Yorke JA, Langley CH. 2014. Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics.* **196**: 875-90. DOI: https://doi.org/10.1534/genetics.113.159715

Associate editor: Ken Oyama

Author contributions: LEE, study design, literature review, writing; EAP, study design, literature review, writing; GCM, literature review, writing; VS, literature review, writing.