TEACHING TOOLS IN PLANT BIOLOGY™: LECTURE NOTES

Genomic Analysis of Botanical Collections: opportunities and challenges

ABSTRACT

Botanical collections, comprising living or preserved specimens, are invaluable repositories of plant biodiversity. Genomic analysis of these collections can help answer critical questions about species evolution and extinction and contribute to conservation of plant resources. However, these assets are often underutilized due to limited resources, technical difficulties or lack of awareness. In this lesson, we provide an overview of the factors influencing genomic studies of botanical collections, we introduce the challenges and solutions of plant genomics for botanical collections and stimulate reflection on innovations in establishing botanical collections for future research.

BOTANICAL COLLECTIONS AS A RESOURCE FOR THE STUDY OF PLANT GENETIC VARIATION

Botanical collections are organized assemblages of plant materials and their associated data, that can be used for scientific research, education, conservation, reference. There are many types of botanical collections, each serving different purposes in scientific research, conservation, education, and public outreach. These collections are stored in institutions like botanical gardens, museums or research and education institutions where they are maintained and catalogued to ensure their accessibility and preservation for future uses. Botanical collections include:

1. **Herbaria**: Dried and pressed plant specimens, stored systematically, are primarily used for taxonomic studies, plant identification, and historical documentation of biodiversity. They are also valuable for generating reference genomes. Examples: The New York Botanical Garden Herbarium, the Royal Botanic Gardens Kew Herbarium, the Missouri Botanical Garden Herbarium
2. **Living Collections**: Plants maintained in botanical gardens and arboreta, used for research, conservation of endangered species, and public education. Examples: The New York Botanical Garden (NYBG), The Royal Botanic Gardens Kew, The Arnold Arboretum of Harvard University.
3. **Seed, Pollen, and Spore Banks**: These collections preserve genetic diversity by storing seeds, pollen or spores under controlled conditions, supporting conservation and restoration efforts. Examples: The Millennium Seed Bank, Svalbard Global Seed Vault, Chicago Botanic Garden's Pollen Bank.
4. **DNA and Tissue Banks**: Repositories of DNA and plant tissues that enable genomics and conservation genetics research. Examples: The Global Genome Biodiversity Network (GGBN), DNA Bank Network, The DNA Bank at the Missouri Botanical Garden, The Center for Comparative Genomics CryoCollection at the California Academy of Sciences or the DNA bank and the Langenheim Resin Collection at the NYBG.
5. **Ethnobotanical Collections**: These focus on plants used by indigenous people and document traditional knowledge and plant-based remedies. Examples: National Museum of Natural History Ethnobotany Collections, University of Michigan Herbarium Ethnobotanical Collection.
6. **Digital Collections**: Digitized plant specimens, including genetic data, make biodiversity accessible globally for virtual studies and collaborative research. Examples: Integrated Digitized Biocollections (iDigBio), Global Biodiversity Information Facility (GBIF), Global Plants on JSTOR.
7. **Palynological Collections**: Pollen and spore collections, often used for studies in morphology, paleobotany, and climate change. Examples: The John P. Smol Paleolimnology and Environmental Change Laboratory, University of Arizona Palynology Collection.
8. **Carpological and Spirit Collections**: These preserve fruits, seeds, and other plant parts, important for taxonomic studies and the historical record of plant diversity. Examples: Royal Botanic Garden Edinburgh Carpological Collection, the Royal Botanic Gardens Kew Spirit Collection.

Large collections of samples gathered for specific studies are also considered as botanical collections in this text.

Botanical collections are vital for documenting plant diversity, supporting taxonomic classification, conserving rare and endangered species, and providing reference material for various scientific fields. Big institutions, like the New York Botanical Garden, the Royal Botanic Gardens Kew, or the Svalbard Global Seed Vault, are well-known for the impact of their projects on plant diversity studies and conservation. But regardless of the size, collections from smaller institutions all-around the world, collections of samples gathered for scientific studies and even personal herbaria are a treasury of key-resources and unique specimens for studying and preserving biodiversity. An increasingly important application of botanical collections is the study of genetic diversity. DNA is a remarkably stable molecule that –under the right conditions– can persist in biological specimens for extended periods (sometimes thousands of years). Therefore, even old, preserved specimens can be a source of DNA for genetic investigations. DNA preservation typically depends on factors like the specimen's age and the environmental conditions in which it has been stored.

Technological advancements have given renewed relevance to botanical collections and especially to herbaria. On one hand, the advance of molecular methods and the decrease in their costs has enabled the genetic study of increasingly older preserved specimens. Simultaneously, the progress of sequencing technologies has increased the number of samples and markers that can be studied, facilitating the use of botanical collections for the study of plant genetic diversity. Finally, digitization of collections has made a growing number of herbaria and other collections accessible to scientists worldwide, facilitating their consultation and providing important information about where specimens of interest are stored. However, despite these advancements, there are still many bottlenecks constraining the full exploitation of the resources present in botanical collections. In this paper, we will first introduce the concept of genetic variation and the processes that drive it. Next, we will provide an overview of Next-Generation Sequencing (NGS) methods used to detect these variants. Lastly, we will address the opportunities and practical challenges of using botanical collections for the study of genetic diversity in plants through -omics technologies.

What is genetic diversity and why do we need to study it in plants?

Genetic diversity represents the genetic differences among individuals within or between populations. Genetic variation manifests as differences in DNA sequence ('polymorphisms'). While polymorphisms are primarily caused by mutations, sexual reproduction facilitates the formation of new combinations through meiotic recombination, secondarily contributing to genetic diversity. On the other hand, sometimes polymorphisms do not affect the phenotype but even then, they are useful to measure the differences between individuals and populations.

These polymorphisms can manifest at several scales, ranging from a single nucleotide base to larger modifications involving entire chromosome segments. The potential effect of polymorphisms is related to its scale. A **Single Nucleotide Variant (SNV)**, or **Single Nucleotide Polymorphism (SNP)**, is the simplest type of polymorphism and occurs when a single base in the DNA is substituted, deleted, or inserted. Being a small change, SNPs often have no effect on the phenotype as they may occur in non-coding regions or lead to small changes that do not affect gene and/or protein function. However, occasionally, SNPs can affect gene expression or protein functions. **Multiple Nucleotide Polymorphisms (MNP)**, involve changes across a longer sequence of adjacent nucleotides. Consequently, they are more likely to cause changes in gene/protein function and may be associated with more complex traits, such as disease resistance or tolerance to specific environmental conditions. **Structural Variants (SV)** are larger variations, encompassing nucleotide sequences longer than 50 base pairs. These changes can include **deletions**, **insertions**, **translocations** (where a DNA segment changes its position), **inversions** (where the sequence is reversed), and **copy number variations** (changes in the number of copies of a gene or sequence). Structural variants can have significant effects on gene functionality and the adaptive response of plants. Finally, **chromosomal rearrangements** represent even larger-scale modifications that involve entire portions of chromosomes. These rearrangements can result in inversions, translocations, losses, or duplications of entire chromosomes. Such changes can lead to post-zygotic reproductive barriers, which can trigger reproductive isolation between different populations, thereby influencing the evolution and diversification of plant species.

Genetic variation is acquired through different processes. Gene flow via interbreeding with a different population is a potential source of new allelic variants. Genetic diversity can also be gained through genetic mutations like point mutations or massive genomic modifications, such as transposable element (TEs) proliferation, polyploidization or hybridization. Plants are more tolerant to genomic modifications than animals, so these events are common in plant evolution.

On the other hand, several factors can lead to genetic diversity loss in populations. One of the most significant is genetic drift, which can cause the random elimination or fixation of genetic variants. Due to its stochastic nature, its effects are particularly relevant in small, isolated populations, where the contribution of each individual to allele frequencies is higher. Small populations are also more susceptible to inbreeding, which increases population homozygosity; this can result in a higher incidence of deleterious traits and non-viable individuals (inbreeding depression). Reduced gene flow can contribute to genetic diversity reduction. Additionally, natural selection can fix certain alleles and remove others depending on their fitness effects.

Genetic variation can influence the physical, physiological, and ecological characteristics of organisms. Populations with low genetic diversity are less capable of adapting to changing environments and face a higher risk of extinction than those with higher diversity. Studying genetic diversity patterns in plants helps address key botanical questions like identifying vulnerable species or populations, understanding how plants respond to climate change or anthropic impact and species delimitation. Therefore, population genetics and evolutionary studies are essential for biodiversity conservation. Plant genetics is also crucial in agronomy, aiding to select disease- and climate-resilient crop varieties. In summary, analyzing genetic diversity offers insights into the ecological, evolutionary, and economic dynamics of the plant world.

SEQUENCING TECHNOLOGIES AND HOW THEY HAVE ENABLED THE STUDY OF GENETIC VARIATION

Genetic variation is analyzed through any method that detects genetic variants. Today this definition drives us to think about changes in DNA sequences, but before the development of DNA sequencing technologies in the 1970s, genetic diversity was studied through morphological, biochemical, or cytological markers related to DNA variation. DNA sequencing was a breakthrough enabling the study of DNA sequences regardless of their phenotypic effect. The first sequencing methods were developed in the 1970s (first-generation). Sanger sequencing is the only first-generation method that remains in use today for specific applications. Sanger sequencing has a very low error rate, but only a single, relatively short fragment (around 1kb) can

be sequenced per reaction. Still, it enabled the first DNA variation studies and the production of the first reference genomes from eukaryotic organisms (i.e., human or the model plant *Arabidopsis thaliana*). A **reference genome** is a representation, as complete as possible, of the genomic DNA of an organism or a set of organisms belonging to a species. It is annotated with the position of gene sequences and other functional elements, such as repeats, untranslated regions, or transcript isoforms. They provide a representative framework for comparing sequences and identifying variants among individuals or species.

Second-generation sequencing, also known as high throughput sequencing, parallel sequencing or next generation sequencing (NGS), drastically increased the outputs and reduced the costs and time required for sequencing. NGS can sequence hundreds of samples simultaneously (hence, "in parallel") even without previous genetic information on the target organism. Several second-generation methods were developed in the 2000's, but only the Illumina platform is still used nowadays. Illumina provides high amounts of sequences ("reads") with low error rate, but it is limited by read length (50-300 bp). Nevertheless, second-generation methods enabled the first wave of plant genomic studies, greatly increasing the amount of genetic information available and the number of plant genomes sequenced. Its high throughput required new bioinformatic and statistical tools to analyze the large quantity of sequences produced. The genetic diversity detected with this approach is commonly referred to as "genomic diversity" in place of "genetic diversity" to underscore its genome-wide scope.

Presently, third-generation sequencing technologies like Oxford Nanopore and PacBio, developed in the late 2010's, can sequence DNA molecules without amplifying them, which allows to obtain much longer sequences (up to hundreds of kilobases). Long reads facilitate the reconstruction of big and complex genomes and repetitive regions and are especially relevant for plants because phenomena like hybridization, polyploidization, and introgression that complicate genome assembly are common in plant genomes. Although third-generation sequences were initially less accurate, at present their precision is close to that of Illumina sequences, offering significant advantages for studying structural variants (SVs) such as inversions, deletions and copy number variations. Third-generation sequencing has enabled the creation of large-scale sequencing initiatives, such as the 10KP Initiative, the Earth Biogenome Project, and the Darwin Tree of Life, which aim to sequence and catalog global genomic diversity.

NGS APPROACHES FOR THE STUDY OF GENETIC VARIATION

Whole genome sequencing

Whole genome sequencing and resequencing (WGS and WGR) is the process of determining the sequence of the entire genome of interest (or as complete as possible). Resequencing refers to the complete sequencing of additional individuals when a reference genome is already available. The reads are compared to the reference to identify their location on the genome ("alignment") and accurately determine the genetic variants present in each individual ("variant calling"). This (mostly) complete and unbiased set of variants can be employed for any genetic diversity study, from marker assisted selection of crops, to population genomics, to estimation of inbreeding in threatened species. While WGR would be the ideal choice for assessing the genotype of an individual, it presents some significant challenges, including high sequencing costs (dependent on genome size) and substantial computational resource requirements due to the large amount of data generated. These costs can make WGR prohibitive for many studies (i.e., non-model species with limited genomic resources).

Reduced representation sequencing techniques

Reduced representation sequencing (RRS) techniques address the challenges mentioned above by sequencing only a fraction of the genome (about 1% to 5% of the total size), reducing sequencing costs, computational resources, and data storage requirements compared to WGS strategies. RRS can still detect hundreds to thousands of SNPs across the genome, sufficient for population genetic analyses. Another advantage of RRS is that a reference genome is not mandatory to identify SNPs: since the number of sequenced DNA fragments is lower than for WGR, it is computationally feasible to align and cluster reads of the same locus and identify variants within each sequenced fragment. This aspect has contributed to the spread of RRS techniques in non-model organisms and in species with complex genomes. However, when a reference genome of the target species (or at least of a close relative) is available, bioinformatic analysis is greatly facilitated and a greater number of markers can be analyzed. Additionally, with a reference genome, it is also possible to infer the position of the markers in the genome and investigate whether some variants may affect fitness (i.e., by altering a protein sequence) or be linked to genes coding for specific traits.

RRS methods rely on different approaches to select genome fractions. Many of them, like RAD-seq (Restriction Site-Associated DNA sequencing), or GBS (Genotyping-By-Sequencing), use restriction enzymes. Others, like K-seq, rely on annealing of short oligonucleotides designed to target single-copy regions randomly distributed in the genome.

A common issue of RRS methods is missing data or allele dropout, which is the lack of information on the genotype of some individuals at some loci. This might be due to real variation among individuals, poor DNA quality or technical variability in library preparation. For example, in methods based on restriction enzymes, mutations in the restriction sites or DNA modifications like methylation can prevent digestion and, thus, sequencing. Despite these limitations, RRS remains a cost-effective and widely used genotyping approach, particularly for large studies involving hundreds of individuals.

Other methods available for genome-wide genetic analysis are, for example, transcriptome sequencing (RNA-Seq), low coverage WGS, Pool-seq, targeted amplification and sequence capture. Transcriptome sequencing, or RNA-seq, is the sequencing of RNA molecules present in a sample (plant organ, tissue or even single cells). Comparison of samples can help to associate genes to specific plant traits or functions. Low coverage WGS, also known as genome skimming, consists of sequencing genomes at a very low depth; this allows recovery and genotyping at regions that are naturally present in many

copies, such as organelle DNA. Pool-seq, instead, consists of the simultaneous sequencing of the DNA of multiple individuals mixed in a pool of samples. Pools can include individuals with similar phenotype, or same origin, habitat or species. The pooling can be useful to identify genetic variants associated with the trait shared by the pooled individuals when compared with a pool where the trait is absent. The pooling reduces the sequencing effort and cost. Targeted amplification and sequence capture act similarly, in that the regions to be sequenced are pre-determined and are either selectively amplified with specific primers (i.e. TruSeq custom amplicons) or enriched in a sample via hybridization to single-stranded probes (sequence capture).

HOW CAN BOTANICAL COLLECTIONS HELP UNDERSTAND AND CONSERVE GENETIC VARIATION OF PLANTS?

Preserved collections

Preserved specimens, like herbaria, provide information about the genetic variants occurring in a locality and at the time in which the individual was collected, capturing a "snapshot" of the genetic make-up of a species in that moment. Genetic data from historical preserved specimens offer a unique opportunity to reconstruct past genetic diversity patterns and species evolution. Comparing individuals collected at different times allows researchers to track changes in population genetic structure, allelic frequencies, and evolutionary processes like selection or bottlenecks. Even pathogens, preserved within plant tissues in herbaria — such as fungi and viruses — can be studied, providing insight also into co-evolutionary processes.

The study of herbarium samples with NGS methods has been termed herbariomics. Several approaches have been successfully applied, from WGR to SNP assays to targeted capture, allowing sequencing of specific nuclear and plastid DNA fragments (or markers). However, they present some additional challenges when compared to fresh samples (see the paragraph *DNA quality and quantity*).

Herbaria are also repositories of the plants used to generate a reference genome, or for any kind of genetic study. It is especially important in the case of wild species that a sample from the individual used to generate the reference genome is vouchered and stored and several international sequencing initiatives like the Earth Biogenome Project require that specimens used to generate reference genomes are vouchered. This ensures the preservation of the sample, data reproducibility, and allows for the association of a genotype with a specific phenotype.

Ethnobotanical collections, which include plants of cultural and medicinal interest, share many characteristics with herbarium specimens but often with fewer botanical details and more information about their use. Nonetheless, they are a valuable resource for understanding human use of plants across history and the evolution of domesticated plants and local varieties. The digitization of herbarium specimens is promoting the use of herbaria for genetic studies because they facilitate the dissemination and consultation of collections and allow the preservation of the plant's appearance if part of the physical specimen is used for DNA extraction. Additionally, digital specimens can be linked to other datasets, creating a more cohesive relationship between different research fields.

Living and germplasm collections

Botanical gardens were created in the Renaissance period to grow medicinal plants so medical students could learn to recognize them *in vivo*. Today, botanical gardens and arboreta maintain their educational purposes, but also function as centers for research and conservation.

The cultivation of plants in botanical gardens offers scientists many opportunities to study plant genetic diversity. For instance, reproduction capacity of threatened species can be studied in botanical gardens without stressing natural populations. Additionally, botanical gardens are excellent settings to perform common garden experiments (i.e. where plants of different origins are grown together in common conditions) that can help in disentangling genetics and environmental effects in trait expression or stress response.

Living collections also provide abundant fresh material that is indispensable for cytological and cytogenetic investigation methods, like genome size estimation with flow cytometry, or for NGS-based studies that require good amounts of high-quality DNA (i.e. for reference genome generation). Because of that, botanical garden collections play a key role in confirming species identification and phylogenetic investigations. Botanical garden collections were essential to produce the most recent angiosperm tree of life providing almost 8000 samples from all over the world.

Additionally, individuals from living collections can be used to reintroduce or reinforce wild populations. This led living collections, including plant nurseries of native plants, to acquire new value for *ex-situ* conservation practices and restoration projects. Moreover, in some cases botanical gardens represent the last chance for species before complete extinction.

Seed banks are facilities specialized in preserving seeds, which are stored dehydrated in a controlled environment to maintain their viability over long periods. Initially created for crops, with time, they became repositories also of wild germplasm. For example, the International Rice Genebank (IRRI) is the largest collection of rice genetic diversity in the world, whereas the seed collection at the Millenium Seed Bank of Kew Gardens (MSB) is the most diverse wild plant species seed collection in the world, serving as a global genetic resource. Similar to seed banks, pollen and spore banks cryopreserve pollen and spores for breeding programs or for germplasm conservation and exchange. Thanks to their small dimensions, they are easier to store than other propagules (i.e., tubers) but require adequate equipment to maintain the proper temperatures for long-term viability. Seed and spore viability are tested periodically and maintained through regeneration, replacing old material with new. These facilities are invaluable resources for studying genetic diversity. Seeds and spores stored in these facilities not only preserve a stock for conservation purposes, but are useful also for comparative genetic studies within and among populations over time, allowing to examine their evolution and their response to climate change. Pollen banks play a crucial role in studying hybridization potential and associated evolutionary processes, as pollen can be used for controlled breeding and gene flow analysis. Thus, seed, pollen and spore banks are also living collections that can provide fresh, high-quality plant material for genetic studies.

CHALLENGES FOR THE STUDY OF BOTANICAL COLLECTIONS USING NGS

DNA quality and quantity

Botanical collections hold valuable genetic information that can be studied thanks to NGS technologies. However, obtaining high-quality DNA directly from a preserved collection is a frequent limitation that strongly impacts downstream steps. The quality of starting DNA affects all stages of NGS samples preparation, impacting the final set of molecular markers revealed and, consequently, research outcomes.

The first factor influencing DNA extraction is the starting material. Best results are always obtained with fresh or liquid nitrogen (LN)-frozen material. Rapid freezing in LN (-196°C), helps preserve cell structure reducing the risk of DNA degradation. However, LN's practical limitations, such as the need for specialized containers, permits for transport or its rapid evaporation, makes it difficult to use it in fieldwork. This makes living botanical collections ideal for genomic diversity studies, especially if they are close to a molecular laboratory. When LN or a rapid DNA extraction can't be applied (like in the fieldwork), sample preservation is critical for the success of the project. Silica gel is commonly used to dry freshly collected plant samples. If the drying rate is fast enough (within 24 hours) it effectively prevents DNA degradation. Other methods such as preservation or pretreatment of samples in ethanol are also used, although not universally applicable. In fact, ethanol inhibits hydrolytic enzymes and facilitates the homogenization of cell walls, but it does not ensure the preservation of DNA integrity. Silica gel drying is also unsuitable for RNA extraction, which requires tissue to be flash-frozen or stabilized in specific buffers. Different methods can be tested in the early stages of a research project to determine those most suitable for the taxa under study.

Another factor that impairs DNA extraction is the presence of specialized metabolites in the starting material, like terpenes or phenolic compounds. Moreover, the type and concentration of these metabolites vary between plant species and are influenced by environmental conditions, making it challenging to develop a universally effective DNA extraction protocol. Two common problems related with specialized metabolites are DNA oxidation by phenolic compounds and co-precipitation of polysaccharides. For example, adding polyvinylpyrrolidone (PVP) to the extraction buffer helps remove phenols and prevents oxidation by suppressing quinone formation. Adding sorbitol also reduces oxidation. When starting from fresh/LN-frozen material, isolating nuclei prior to DNA extraction is also helpful as most specialized metabolites localize into the vacuole. In this way, specialized metabolites are eliminated before starting the extraction and the DNA never gets in contact with them. Additionally, this technique yields very long DNA fragments (required for third generation sequencing) because it minimizes mechanical damage to DNA during extraction.

Extracting DNA from herbaria specimens presents a range of challenges. Most notably, the DNA is often damaged and fragmented. While DNA damage and fragmentation are generally age-related, the correlation is not straightforward, as several factors can influence this process. These include storage conditions and chemical treatments commonly used in the past to 'disinfect' plants before and during storage. Additionally, specialized metabolites may have bound to the DNA over time, complicating the extraction. To address these issues, dedicated protocols have been developed for working with historical samples. While some DNA can usually be extracted, the fragments are often quite small (<100 bp), limiting the use of herbarium specimens to short-read sequencing or sequence capture-based methods. Despite technical advances are enabling the retrieval of higher-quality and longer DNA fragments opening the door to long-read approaches, some herbarium specimens remain unsuitable for techniques requiring high-quality, unfragmented DNA or large amounts of plant tissue. Therefore, since herbarium sampling is destructive, it is crucial to carefully assess the potential risks and benefits before extracting DNA. This helps prevent irreparable damage to the specimen, which could limit its future use in genomic studies and as a phenotypic reference. In this regard, the use of living collections grown in garden conditions is advantageous as it can provide more homogenous DNA extracts, improving DNA performance in downstream reactions. Another option to circumvent these problems is using young leaves for DNA extraction, as they are less likely to have suffered environmental stress (leading to the accumulation of specialized metabolites) and physical damage that could compromise DNA integrity.

On the other hand, the drawbacks of herbaria samples sometimes can be turned into advantages as they can also provide unique information. For example, DNA damage can be used to estimate DNA methylation of historical samples, providing a glimpse into the epigenome of the samples.

Genome size and sequencing depth

Sequencing technologies might be very accurate (i.e., Illumina) but none is completely error free. Therefore, for unequivocal identification of variants, each nucleotide position needs to be sequenced several times, to avoid mistaking real genetic variants with sequencing errors; a concept known as sequencing coverage or depth. For Illumina, a 30x coverage is recommended, meaning that the amount of sequencing required is 30 times the species' genome size. In other words, each nucleotide will be sequenced 30 times on average. On occasions lower coverage (e.g., 5x–10x) is enough to achieve reliable variant calling, i.e., particularly when working with species with small and well-annotated reference genomes or when not aiming to detect SNPs. Nevertheless, whenever possible, 30x coverage is considered optimal for ensuring accurate detection and avoiding false positives or missing variants. Therefore, genome size greatly impacts the amount of data needed and the cost. For example, in *Arabidopsis thaliana*, with a genome size of 120 megabases, a 30X coverage represents 3.6 Gigabytes of sequencing data, but in *Zea mays*, with a genome of 2.4 Gigabases, it represents 72 Gigabytes of data. RRS and target enrichment address this by selecting a small portion of the genome, allowing sequencing at high coverage with a lower amount of total data. Genome size also influences the choice of enzyme and protocol for digestion-based RRS approaches as, depending on genome size, enzymes with higher- or lower-cutting frequencies will be preferred to modulate the number of fragments generated. Thus, knowing the genome size of at least a close relative is helpful for planning NGS experiments. Flow cytometry is the most

common method for genome size estimation, but most protocols require fresh material. Herbarium or silica-gel dried specimens were considered unsuitable for genome size estimation via flow cytometry, but ongoing technical advancements have led to the development of specialized protocols that increasingly overcome these limitations. Nevertheless, living collections such as those in botanical gardens remain the first choice for genome size estimation, providing abundant and high-quality plant material. Estimates of genome size for many plant species are available in the Kew Plant DNA C-values database.

Polyploidy

Polyploidy is the heritable condition of possessing more than two complete sets of chromosomes. Polyploidization is a frequent event in plant evolution. Stable polyploid populations or species can either derive from whole-genome duplication (WGD) events (which originate autopolyploids) or from hybridization between close-related species (originating allopolyploids). Polyploids, and especially allopolyploids, pose challenges for genomic studies as reads of related genes coming from each subgenome (homeolog genes) are easily confounded, complicating the alignment. On the other hand, WGD is often followed by partial re-diploidization events adding further complexity. Most tools for genomic analysis are designed for diploids so, not accounting for ploidy level can result in biased conclusions, especially in populations where individuals of different ploidy coexist. Producing reference genomes of species with big polyploid genomes is still a major challenge, so species with big genomes were often excluded from genomic studies, but long-read sequencing is helping bridge this gap. An example of the use of botanical collections for the study of polyploid species is the common reed (*Phragmites australis)* at Aarhus University, in Denmark. This collection, of over 200 genotypes collected worldwide, provided insight into the phylogeographic pattern of this cosmopolitan species and the role of cytological variation (i.e., chromosome number, ploidy, genome size) in its evolution and invasion success. Such living collections allow evolutionary and invasion biologists to study the taxonomic complexity of a species across its distribution range.

Challenges for the preparation of collections for genetic diversity studies

Most botanical collections were not established with the objective of studying genetic variation, so this can limit the genetic diversity contained in them. Therefore, when establishing new collections, it is crucial to adopt some precautions to collect the most genetic variation possible and to maintain it for future genetic studies. Moreover, understanding the challenges and the genetic dynamics within living plant collections can be useful to assess the suitability of older collections for genetic studies.

When new material is collected, obtaining a good representation of the genetic diversity present in the field is not straightforward. Phenotypic differences do not always correspond with genetic variation, and genetic screening is not usually performed in the field, so sampling strategy has to be carefully designed, and factors like biological characteristics that affect genetic diversity (i.e., sexual/asexual reproduction, autogamy/allogamy, seed dispersal strategy) or the evolutionary forces acting on the population (i.e., gene flow, genetic drift, selection) have to be considered. Therefore, phylogeographic studies of the genetic variation pattern are necessary to guide the establishment of new collections. Sampling theory and spatial statistics can also aid in drawing a strategy. Several guidelines developed for or by seed banks for collecting a good representation of genetic diversity are available online.

Moreover, sampling should comply with international agreements like the Nagoya protocol (SCBD, 2015), or IUCN (International Union for the Conservation of Nature) for protected and invasive species. Searching information about the target species is as crucial as being informed about the legal issues and conservation laws on species (i.e. plants in CITES).

Once the material is collected, it is crucial to maintain genetic diversity inside the collection. Living collections can be considered as isolated populations and are subjected to the same evolutionary forces that affect genetic diversity in nature (i.e., selection, genetic drift, inbreeding, mutation events). Since the collection aims to mirror the genetic diversity of the natural populations, these dynamics should be carefully monitored. An important issue to consider is avoiding the involuntary selection of certain genotypes. The captivity environment may differ from the original habitat of the species, posing less competition for resources (light, water, nutrients) and biotic stresses (pathogens, parasites, or herbivores). This can lead to proliferation of more susceptible individuals and lead to an alteration of the original genetic diversity of the species. This phenomenon is most relevant for annual plants that are renovated every year with seeds produced *ex situ*. Growing plants in conditions resembling their wild environment should minimize these effects. However, this is not always feasible or even desirable (if a species is extremely rare, it is risky to cultivate it in conditions that challenge its survival). Therefore, this factor can be mitigated but not eliminated. Another issue is accidental hybridization, due to cultivation of closely related allopatric species in proximity. This can create hybrids that will alter the genetic diversity pool and might even be harmful if the collection is meant for species reintroduction in nature or restoration.

Exchanging material between institutions facilitates backing up the collections, reduces the likelihood of losing genotypes and promotes the maintenance of genetic diversity.

Take-home message

With the decreasing costs of sequencing technologies and the vast number of approaches available, there has never been a better time to apply genome-wide analyses of genetic diversity to plant species for conservation and other botanical purposes. Genomic data can inform conservation strategies on a resolution previously unmatched and botanical collections represent both a rich reservoir of diversity and a foundation for establishing future collections guided by genomic principles. Nevertheless, studying plant genetic diversity presents many challenges. Genome sequencing and downstream analyses can be challenging for botanists and naturalists with no molecular background; moreover, translating genomic data into conservation practice is challenging *per se*.

Martina Deglialberti
Instituto Universitario para la Conservación y Mejora de la Agrodiversidad Valenciana (COMAV), Universidad Politécnica de Valencia (UPV) Camino de Vera s/n, 46022 Valencia, Spain

Instituto de Biología Molecular y Celular de Plantas (IBMCP) Primo-Yúfera, Consejo Superior de Investigaciones Científicas- Universidad Politécnica de Valencia, Avenida fausto Elio s/n, 46022 Valencia, Spain

Università degli Studi di Milano, Department of Biosciences Via Giovanni Celoria 26, 20133 Milan, Italy

Chiara Paleni
Università degli Studi di Milano, Department of Biosciences Via Giovanni Celoria 26, 20133 Milan, Italy

Federico Fainelli
Università degli Studi di Pavia, Department of Earth and Environmental Science, Via S. Epifanio 14, 27100 Pavia, Italy

Carla Lambertini
Università degli Studi di Milano, Department of Biosciences Via Giovanni Celoria 26, 20133 Milan, Italy

Silvia Manrique
Instituto Universitario para la Conservación y Mejora de la Agrodiversidad Valenciana (COMAV), Universidad Politécnica de Valencia (UPV) Camino de Vera s/n, 46022 Valencia, Spain

Università degli Studi di Milano, Department of Biosciences Via Giovanni Celoria 26, 20133 Milan, Italy

Instituto de Recursos Naturales y Agrobiología de Salamanca (IRNASA), Consejo Superior de Investigaciones Científicas, Cordel de Merinas 40 Salamanca - Spain

ACKNOWLEDGMENTS

Thanks to Martin Kater (University of Milan), Simone Orsenigo (University of Pavia) and Aureliano Bombarely (CSIC, Spain) for support and helpful feedback on the manuscript. MD is grateful for funding from Erasmus+ and the Catalan Biogenome Project (IEC-BG1-2023-2, Institut d'Estudis Catalans). CP and FF are funded by Programma Operativo Nazionale Ricerca e Innovazione 2014-2020, FSE REACT-EU Azione IV.4 "Dottorati e contratti di ricerca su tematiche dell'innovazione", and Azione IV.5 "Dottorati su tematiche Green" resources. CL and SMU are grateful for funding from the University of Milan (Linea 2, 2021-2022, Piano di sviluppo di ateneo). SMU's work has been funded by the University of Milan (linea 2 to CL), UPV and the Spanish Ministry of University (María Zambrano program funded by the European Union Next Generation-EU), the Junta de Castilla y León (Andrés Laguna program) and CBP (IEC-BG1-2023-1, Institut d'Esudis Catalans).

RECOMMENDED READING

Botanical collections as a resource for the study of plant genetic variation

Delves J, Albán-Castillo J, Cano A, Fernández Aviles C, Gagnon E, Gonzáles P, Knapp S, Leon B, Marcelo-Peña JL, Reynel C, et al. Small and in-country herbaria are vital for accurate plant threat assessments: A case study from Peru. *Plants People Planet*. 2024:6(1):174–185. https://doi.org/10.1002/ppp3.10425.

González-Toral C, Cires E. Relevance of DNA preservation for future botany and ecology. *Mol Ecol*. 2022:31(20):5125–5131. https://doi.org/10.1111/mec.16652.

JSTOR. JSTOR Global Plants. Accessed 10 July 2024. https://plants.jstor.org.

Marsico TD, Krimmel ER, Carter JR, Gillespie EL, Lowe PD, McCauley R, Morris AB, Nelson G, Smith M, Soteropoulos DL, et al. Small herbaria contribute unique biogeographic records to county, locality, and temporal scales. *Am J Bot*. 2022:107(11):1577–1587. https://doi.org/10.1002/ajb2.1563.

Primack RB, Ellwood ER, Gallinat AS, Miller-Rushing AJ. The growing and vital role of botanical gardens in climate change research. *New Phytol*. 2021:231(3):917–932. https://doi.org/10.1111/nph.17410.

What is genetic diversity and why do we need to study it in plants?

Bernatchez L, Ferchaud AL, Berger CS, Venney CJ, Xuereb A. Genomics for monitoring and understanding species responses to global climate change. *Nat Rev Genet*. 2024:25(3):165–183. https://doi.org/10.1038/s41576-023-00657-y.

Chung MY, Merilä J, Li J, Mao K, López-Pujol J, Tsumura Y, Chung MG. Neutral and adaptive genetic diversity in plants: An overview. *Front Ecol Evol*. 2023:11:1116814. https://doi.org/10.3389/fevo.2023.1116814.

Ennos RA. The contribution of population genetic studies to plant conservation. *Bot J Scotl*. 2003:55(1):89–100. https://doi.org/10.1080/03746600308685051.

Frankham R, Ballou JD, Briscoe DA. Introduction to conservation genetics (2nd ed). Cambridge University Press. 2010. https://doi.org/10.1017/CBO9780511809002.

Rieseberg LH, Willis JK. Plant speciation. *Science*. 2010:317(5840):910–914. https://doi.org/10.1126/science.1137729.

Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. *Genome Biol*. 2016:17:1–14. https://doi.org/10.1186/s13059-016-0908-1.

Willi Y, Kristensen TN, Sgrò CM, Weeks AR, Ørsted M, Hoffmann AA. Conservation genetics as a management tool: The five best-supported paradigms to assist the management of threatened species. *Proc Natl Acad Sci USA*. 2022:119(1):e2105076119. https://doi.org/10.1073/pnas.2105076119.

Sequencing technologies and how they have enabled the study of genetic variation

Akintunde O, Tucker T, Carabetta VJ. The evolution of next-generation sequencing technologies (arXiv:2305.08724). *arXiv*. 2023. https://doi.org/10.48550/arXiv.2305.08724.

Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux P-M, Li F-W, Melkonian B, Mavrodiev EV, Sun W, et al. 10KP: A phylodiverse genome sequencing plan. *GigaScience*. 2018:7(3):giy013. https://doi.org/10.17863/CAM.33002.

CNGBdb. 10KP: The 10,000 Plant Genomes Project Database. *China National GeneBank Database*. Accessed 20 August 2024. https://db.cngb.org/10kp/.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*. 2011:12(7):499–510. https://doi.org/10.1038/nrg3012.

Earth BioGenome Project. Earth BioGenome Project. Accessed 5 August 2024. https://earthbiogenome.org.

ERGA Consortium. European Reference Genome Atlas (ERGA). Accessed 10 September 2024. https://erga-biodiversity.eu.

Frankham R, Ballou JD, Briscoe DA. Introduction to conservation genetics (2nd ed). *Cambridge University Press*. 2010. https://doi.org/10.1017/CBO9780511809002.

Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Mol Ecol*. 2017:26(20):5369–5406. https://doi.org/10.1111/mec.14264.

Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J*. 2020:18:9–19. https://doi.org/10.1016/j.csbj.2019.11.002.

Hamilton JP, Buell CR. Advances in plant genome sequencing. *Plant J*. 2012:70(1):177–190. https://doi.org/10.1111/j.1365-313X.2012.04894.x.

Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016:107(1):1–8. https://doi.org/10.1016/j.ygeno.2015.11.003.

Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. *Hum Immunol*. 2021:82(11):801–811. https://doi.org/10.1016/j.humimm.2021.02.012.

Idrees M, Irshad M. Molecular markers in plants for analysis of genetic diversity: A review. *Eur Acad Res*. 2014:2(1):1513–1540. https://euacademic.org/UploadArticle/514.pdf.

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, et al. Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci USA*. 2018:115(17):4325–4333. https://doi.org/10.1073/pnas.172011.

Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*. 2008:9(1):387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005:437(7057):376–380. https://doi.org/10.1038/nature03959.

Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci USA*. 1977:74(2):560–564. https://doi.org/10.1073/pnas.74.2.560.

McCartney AM, Formenti G, Mouton A, De Panis D, Marins LS, Leitão HG, Pellicer J. The European Reference Genome Atlas: Piloting a decentralised approach to equitable biodiversity genomics. *npj Biodiversity*. 2024:3(1):28. https://doi.org/10.1038/s44185-024-00054-6.

Onda Y, Mochida K. Exploring genetic diversity in plants using high-throughput sequencing techniques. *Curr Genomics*. 2016:17(4):358–367. https://doi.org/10.2174/1389202917666160331202742.

Padmanabhan R, Jay E, Wu R. Chemical synthesis of a primer and its use in the sequence analysis of the lysozyme gene of bacteriophage T4. *Proc Natl Acad Sci USA*. 1974:71(6):2510–2514. https://doi.org/10.1073/pnas.71.6.2510.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021:592(7856):737–746. https://doi.org/10.1038/s41586-021-03451-0.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*. 1977:74(12):5463–5467. https://doi.org/10.1073/pnas.74.12.5463.

Stadler T, Magnus C, Vaughan TG, Barido-Sottani J, Bošková V, Huisman J, Pečerska J. Decoding Genomes: From Sequences to Phylodynamics. *ETH Zurich*. 2024. https://doi.org/10.3929/ethz-b-000664449.

The Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. *Proc Natl Acad Sci USA*. 2022:119(4):e2115642118. Accessed 2 September 2024. https://www.darwintreeoflife.org/.

NGS approaches for the study of genetic variation

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet*. 2016:17(2):81–92. https://doi.org/10.1038/nrg.2015.28.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*. 2008:3(10):e3376. https://doi.org/10.1371/journal.pone.0003376.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*. 2011:12(7):499–510. https://doi.org/10.1038/nrg3012.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*. 2011:6(5):e19379. https://doi.org/10.1371/journal.pone.0019379.

Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Mol Ecol*. 2017:26(20):5369–5406. https://doi.org/10.1111/mec.14264.

Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals—Mining genome-wide polymorphism data without big funding. *Nat Rev Genet*. 2014:15(11):749–763. https://doi.org/10.1038/nrg3803.

Theissinger K, Fernandes C, Formenti G, Bista I, Berg PR, Bleidorn C, Bombarely A, Crottini A, Gallo GR, Godoy JA, et al. How genomics can help biodiversity conservation. *Trends Genet*. 2023:39(7):545–559. https://doi.org/10.1016/j.tig.2023.01.005.

Ziarsolo P, Hasing T, Hilario R, Garcia-Carpintero V, Blanca J, Bombarely A, Cañizares J. K-seq, an affordable, reliable, and open Klenow NGS-based genotyping technology. *Plant Methods*. 2021:17(1):30. https://doi.org/10.1186/s13007-021-00733-6.

How can botanical collections help understand and conserve genetic variation of plants?

Preserved collections

Böhne A, Fernández R, Leonard JA, et al. Contextualising samples: Supporting reference genomes of European biodiversity through sample and associated metadata collection. *npj Biodivers*. 2024:3:26. https://doi.org/10.1038/s44185-024-00053-7.

Brown TA, Cappellini E, Kistler L, Lister DL, Oliveira HR, Wales N, Schlumbaum A. Recent advances in ancient DNA research and their implications for archaeobotany. *Veg Hist Archaeobot*. 2015:24(1):207–214. https://doi.org/10.1007/s00334-014-0489-4.

Buckner JC, Sanders RC, Faircloth BC, Chakrabarty P. The critical importance of vouchers in genomics. *eLife*. 2021:10:e68264. https://doi.org/10.7554/eLife.68264.

Burbano HA, Gutaker RM. Ancient DNA genomics and the renaissance of herbaria. *Science*. 2023:382(6666):59–63. https://doi.org/10.1126/science.adi1180.

De Albuquerque UP, Hurrell JA. Ethnobotany: One concept and many interpretations. In: Recent Developments and Case Studies in Ethnobotany. 2010:87–99.

Dodsworth S, Guignard MS, Christenhusz MJ, Cowan RS, Knapp S, Maurin O, Forest F, et al. Potential of herbariomics for studying repetitive DNA in angiosperms. *Front Ecol Evol*. 2018:6:174. https://doi.org/10.3389/fevo.2018.00174.

James SA, Soltis PS, Belbin L, Chapman AD, Nelson G, Paul DL, Collins M. Herbarium data: Global biodiversity and societal botanical needs for novel research. *Appl Plant Sci*. 2018:6(2):e1024. https://doi.org/10.1002/aps3.1024.

Kistler L, Bieker VC, Martin MD, Pedersen MW, Madrigal JR, Wales N. Ancient plant genomics in archaeology, herbaria, and the environment. *Annu Rev Plant Biol*. 2020:71:605–629. https://doi.org/10.1146/annurev-arplant-081519-035837.

Lanaud C, Vignes H, Utge J, Valette G, Rhoné B, Garcia Caputi M, Angarita Nieto NS, Fouet O, Gaikwad N, Zarrillo S, et al. A revisited history of cacao domestication in pre-Columbian times revealed by archaeogenomic approaches. *Sci Rep*. 2024:14(1):2972. https://doi.org/10.1038/s41598-024-53010-6.

Lawniczak MK, Durbin R, Flicek P, Lindblad-Toh K, Wei X, Archibald JM, Richards S, et al. Standards recommendations for the Earth BioGenome Project. *Proc Natl Acad Sci USA*. 2022:119(4):e2115639118. https://doi.org/10.1073/pnas.2115639118.

Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M. Genetic analyses from ancient DNA. *Annu Rev Genet*. 2004:38(1):645–679. https://doi.org/10.1146/annurev.genet.37.110801.143214.

Yeates DK, Zwick A, Mikheyev AS. Museums are biobanks: Unlocking the genetic potential of the three billion specimens in the world's biological collections. *Curr Opin Insect Sci*. 2016:18:83–88. https://doi.org/10.1016/j.cois.2016.09.009.

Living and germplasm collections

BGCI 2024a. Seed Conservation Hub and Training Resources. *Botanic Gardens Conservation International*. Richmond, U.K. Accessed 17 August 2024. https://www.bgci.org/resources/bgci-tools-and-resources/seed-conservation-hub-and-training-resources/.

Chen G, Sun W. The role of botanical gardens in scientific research, conservation, and citizen science. *Plant Divers*. 2018:40(4):181–188. https://doi.org/10.1016/j.pld.2018.07.006.

Dinato NB, Imaculada Santos IR, Zanotto Vigna BB, de Paula AF, Fávero AP. Pollen cryopreservation for plant breeding and genetic resources conservation. *CryoLetters*. 2020:41(3):115–127. https://www.cryoletters.org/documents/perspectives/perspective-41-3-115-127-dinato.pdf.

Forgiarini C, Parzefall F, Reisch C. The impact of ex situ cultivation on the genetic variation of endangered plant species–Implications for restoration. *Biol Conserv*. 2023:284:110221. https://doi.org/10.1016/j.biocon.2023.110221.

FAO. Genebank Standards for Plant Genetic Resources for Food and Agriculture. Rev. ed. *Food and Agriculture Organization of the United Nations*. 2014. Rome. https://www.fao.org/4/i3704e/i3704e.pdf.

Gratzfeld J, ed. From Idea to Realisation – BGCI's Manual on Planning, Developing and Managing Botanic Gardens. *Botanic Gardens Conservation International*. 2016. Richmond, U.K. https://www.bgci.org/wp/wp-content/uploads/2019/04/BGCI%20Botanic%20Garden%20Manual.pdf.

Griffith MP, Beckman E, Calicrate T, Clark JR, Clase T, Deans S, Dosmann M, Fant J, Gratacos X, Havens K, et al. Toward the metacollection: Safeguarding plant diversity and coordinating conservation collections. *Botanic Gardens Conservation International-US*. 2019. https://arbnet.org/sites/arbnet/files/Toward-the-Metacollection-Coordinating-conservation-collections-to-safeguard-plant-diversity.pdf.

International Rice Research Institute. International Rice Genebank. *IRRI*. Accessed 10 August 2024. https://www.irri.org/international-rice-genebank.

Royal Botanic Gardens, Kew. Seed Collection. *Kew*. Accessed 6 August 2024. https://www.kew.org/science/collections-and-resources/collections/seed-collection.

Spencer R, Cross R. The origins of botanic gardens and their relation to plant science, with special reference to horticultural botany and cultivated plant taxonomy. *Muelleria*. 2016:35:43–93. https://doi.org/10.5962/p.291985.

Zuntini AR, Carruthers T, Maurin O, Bailey PC, Leempoel K, Brewer GE, Epitawalage N, Françoso E, Gallego-Paramo B, McGinnie C, et al. Phylogenomics and the rise of the angiosperms. *Nature*. 2024:629(8013):843–850. https://doi.org/10.1038/s41586-024-07324-0.

Challenges for the study of botanical collections using NGS

DNA quality and quantity

Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res*. 2010:38(6):e87. https://doi.org/10.1093/nar/gkp1163.

Carey SJ, Becklund LE, Fabre PP, Schenk JJ. Optimizing the lysis step in CTAB DNA extractions of silica-dried and herbarium leaf tissues. *Appl Plant Sci*. 2023:11(3):e11522. https://doi.org/10.1002/aps3.11522.

Chase MW, Hills HH. Silica gel: An ideal material for field preservation of leaf samples for DNA studies. *Taxon*. 1991:40(2):215–220. https://doi.org/10.2307/1222975.

Gutaker RM, Reiter E, Furtwängler A, Schuenemann VJ, Burbano HA. Extraction of ultrashort DNA molecules from herbarium specimens. *Biotechniques*. 2017:62(2):76–79. https://doi.org/10.2144/000114517.

Johnson G, Canty SWJ, Lichter-Marck IH, Wagner W, Wen J. Ethanol preservation and pretreatments facilitate quality DNA extractions in recalcitrant plant species. *Appl Plant Sci*. 2023:11(3):e11519. https://doi.org/10.1002/aps3.11519.

Latorre SM, Lang PLM, Burbano HA, Gutaker RM. Isolation, library preparation, and bioinformatic analysis of historical and ancient plant DNA. *Curr Protoc Plant Biol*. 2020:5(4):e20121. https://doi.org/10.1002/cppb.20121.

Marinček P, Wagner ND, Tomasello S. Ancient DNA extraction methods for herbarium specimens: When is it worth the effort? *Appl Plant Sci*. 2022:10(3):e11477. https://doi.org/10.1002/aps3.11477.

Moreira PA, Oliveira DA. Leaf age affects the quality of DNA extracted from Dimorphandra mollis (Fabaceae), a tropical tree species from the Cerrado region of Brazil. *Genet Mol Res*. 2011:10(1):353–358. https://doi.org/10.4238/vol10-1gmr1030.

Murray MG, Thompson WF. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res*. 1980:8(19):4321–4326. https://doi.org/10.1093/nar/8.19.4321.

Quatela A, Cangrén P, Jafari F, Michel T, De Boer H, Oxelman B. Retrieval of long DNA reads from herbarium specimens. *AoB Plants*. 2023:15. https://doi.org/10.1093/aobpla/plad074.

Rogers SO, Bendich AL. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol Biol*. 1985:5(2):69–76. https://doi.org/10.1007/BF00020088.

Staats M, Cuenca A, Richardson JE, Vrielink-van Ginkel R, Petersen G, Seberg O, Bakker FT. DNA damage in plant herbarium tissue. *PLoS One*. 2011:6(12):e28448. https://doi.org/10.1371/journal.pone.0028448.

Wagner S, Plomion C, Orlando L. Uncovering signatures of DNA methylation in ancient plant remains from patterns of post-mortem DNA damage. *Front Ecol Evol*. 2020:8:11. https://doi.org/10.3389/fevo.2020.00011.

Genome size and sequencing depth

Bennett MD, Leitch IJ. Nuclear DNA amounts in angiosperms: Targets, trends and tomorrow. *Ann Bot*. 2011:107(3):467–590. https://doi.org/10.1093/aob/mcq258.

Dolezel J. Plant DNA flow cytometry and estimation of nuclear genome size. *Ann Bot*. 2005:95(1):99–110. https://doi.org/10.1093/aob/mci005.

Hesse U. K-mer-based genome size estimation in theory and practice. In: Heitkam T, Garcia S, eds. *Plant Cytogenetics and Cytogenomics*. Vol. 2672. Springer US; 2023:79–113. https://doi.org/10.1007/978-1-0716-3226-0_4.

Pan J, Wang B, Pei ZY, Zhao W, Gao J, Mao JF, Wang XR. Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. *Mol Ecol Resour*. 2015:15(4):711–722. https://doi.org/10.1111/1755-0998.12342.

Pellicer J, Leitch IJ. The Plant DNA C-values database (release 7.1): An updated online repository of plant genome size data for comparative studies. *New Phytol*. 2020:226(2):301–305. https://doi.org/10.1111/nph.16261.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: Key considerations in genomic analyses. *Nat Rev Genet*. 2014:15(2):121–132. https://doi.org/10.1038/nrg3642.

Polyploidy

Heslop-Harrison JS (Pat), Schwarzacher T, Liu Q. Polyploidy: Its consequences and enabling role in plant diversification and evolution. *Ann Bot*. 2023:131(1):1–10. https://doi.org/10.1093/aob/mcac132.

Lambertini C, Gustafsson MHG, Frydenberg J, Lissner J, Speranza M, Brix H. A phylogeographic study of the cosmopolitan genus *Phragmites* (Poaceae) based on AFLPs. *Plant Syst Evol*. 2006:258(3):161–182. https://doi.org/10.1007/s00606-006-0412-2.

Lambertini C, Sorrell BK, Riis T, Olesen B, Brix H. Exploring the borders of European Phragmites within a cosmopolitan genus. *AoB Plants*. 2012. https://doi.org/10.1093/aobpla/pls020.

Liu LL, Yin MQ, Guo X, Wang JW, Cai YF, Wang C, Guo WH, et al. Cryptic lineages and potential introgression in a mixed-ploidy species (*Phragmites australis*) across temperate China. *J Syst Evol*. 2022:60(2):398–410. https://doi.org/10.1111/jse.12672.

Meirmans PG, Liu S, van Tienderen PH. The analysis of polyploid genetic data. *J Hered*. 2018:109(3):283–296. https://doi.org/10.1093/jhered/esy006.

Meyerson LA, Cronin JT, Bhattarai GP, Brix H, Lambertini C, Lučanová M, Rinehart S, Suda J, Pyšek P. Do ploidy level and nuclear genome size and latitude of origin modify the expression of *Phragmites australis* traits and interactions with herbivores? *Biol Invasions*. 2016:18(9):2531–2549. https://doi.org/10.1007/s10530-016-1200-8.

Woodhouse M, Burkart-Waco D, Comai L. Polyploidy. *Nat Educ*. 2009:2(1):1.

Challenges for the preparation of collections for genetic diversity studies

BGCI. Garden Search. *Botanic Gardens Conservation International*. 2024b. Richmond, U.K. Retrieved 6 June 2024. http://gardensearch.bgci.org/.

BGCI. PlantConnect. *Botanic Gardens Conservation International*. 2024c. Richmond, U.K. Accessed 5 June 2024. https://plantconnect.bgci.org/.

BGCI. Seed Conservation Hub and Training Resources. *Botanic Gardens Conservation International*. 2024a. Richmond, U.K. Accessed 17 August 2024. https://www.bgci.org/resources/bgci-tools-and-resources/seed-conservation-hub-and-training-resources/.

Convention on Biological Diversity. About the Nagoya Protocol. *Secretariat of the Convention on Biological Diversity*. 2015, June 9. Accessed 23 August 2024. https://www.cbd.int/abs/about/default.shtml.

ENSCONET. ENSCONET Seed Collecting Manual for Wild Species. Royal Botanic Gardens, Kew & Universidad Politécnica de Madrid (Eds.). *Royal Botanic Gardens, Kew*. 2009. Accessed 17 August 2024. https://brahmsonline.kew.org/Content/Projects/msbp/resources/Training/ENSCONET_Collecting_protocol_English.pdf.

Enßlin A, Sandner TM, Matthies D. Consequences of ex situ cultivation of plants: Genetic diversity, fitness and adaptation of the monocarpic Cynoglossum officinale L. in botanic gardens. *Biol Conserv*. 2011:144(1):272–278. https://doi.org/10.1016/j.biocon.2010.09.001.

Forgiarini C, Parzefall F, Reisch C. The impact of ex situ cultivation on the genetic variation of endangered plant species–Implications for restoration. *Biol Conserv*. 2023:284:110221. https://doi.org/10.1016/j.biocon.2023.110221.

Griffith MP, Beckman E, Calicrate T, Clark JR, Clase T, Deans S, Dosmann M, Fant J, Gratacos X, Havens K, et al. Toward the metacollection: Safeguarding plant diversity and coordinating conservation collections. *Botanic Gardens Conservation International-US*. 2019. https://arbnet.org/sites/arbnet/files/Toward-the-Metacollection-Coordinating-conservation-collections-to-safeguard-plant-diversity.pdf.

Guerrant EO Jr, Havens K, Vitt P. Sampling for effective ex situ plant conservation. *Int J Plant Sci*. 2024:175(1):11–20. https://doi.org/10.1086/674131.

Lockwood DR, Richards CM, Volk GM. Probabilistic models for collecting genetic diversity: Comparisons, caveats, and limitations. *Crop Sci*. 2007:47(2):861–866. https://doi.org/10.2135/cropsci2006.04.0262.

Murrell OG, Diaz-Martin Z, Havens K, Hughes M, Meyer A, Tutt J, Zerega N, Fant JB. Using pedigree tracking of the ex situ metacollection of Amorphophallus titanum (Araceae) to identify challenges to maintaining genetic diversity in the botanical community. *Ann Bot*. 2025:mcaf038. https://doi.org/10.1093/aob/mcaf038.

Neel MC, Cummings MP. Genetic consequences of ecological reserve design guidelines: An empirical investigation. *Conserv Genet*. 2003:17(4):427–439. https://doi.org/10.1023/A:1024758929728.

Ward SM, Jasieniuk M. Sampling weedy and invasive plant populations for genetic diversity analysis. *Weed Sci*. 2009:57(6):593–602. https://doi.org/10.1614/WS-09-082.1.

Take-home message

Hogg CJ. Translating genomic advances into biodiversity conservation. *Nat Rev Genet*. 2016:25(5):362–373. https://doi.org/10.1038/s41576-023-00671-0.

Rossetto M, Yap JYS, Lemmon J, Bain D, Bragg J, Hogbin P, Gallagher R, Rutherford S, Summerell B, Wilson TC. Conservation genomics workflow to guide practical management actions. *Glob Ecol Conserv*. 2016:26:e01492. https://doi.org/10.1016/j.gecclo.2021.e01492.

Genomic Analysis of Botanical Collections: Opportunities and Challenges – Teaching guide

Overview

Botanical collections, comprising living or preserved specimens, are invaluable repositories of plant biodiversity. Genomic analysis of these collections can help answer critical questions about species evolution and extinction and contribute to conservation of plant resources. However, these biobanks are often underutilized due to limited material, technical difficulties or lack of awareness. This lecture, designed for university students or researchers with little experience in genomics, provides an overview of the factors influencing genomic studies of botanical collections and introduces the challenges and solutions of studying plant genomics from this kind of collections.

Learning Objectives

By the end of this lecture the student should be able to:

- Source material from different types of botanical collections
- Understand the main drivers of genetic diversity
- Describe how genome sequencing can be used to study genetic diversity
- Describe the workflow of a sequencing experiment
- Identify the main challenges for applying NGS to botanical collections
- Describe different NGS approaches for studying botanical collection, depending on the purpose of the study and the economic resources available
- Understand why genetic diversity is important for developing and maintaining a collection

Study/exam questions

- Why might small populations have higher incidence for recessive deleterious traits?
- What are the differences between Next Generation Sequencing (NGS) and Sanger sequencing?
- What are the differences between Illumina and Oxford Nanopore sequencing technologies?
- What constraints can limit the use of whole genome resequencing of botanical collections? What are the limitations of using herbarium samples for genomic studies?
- What are potential drawbacks of living collections concerning their potential to conserve genetic diversity?
- Name three plants that you can find in a supermarket that you think will be difficult to sequence because they produce large amounts of secondary metabolites.
- How can chromosomal rearrangement lead to reproductive isolation?
- A project wants to focus on variants located on promoter regions of different genes, which genotyping methods out of the ones described could be used?
- What genotyping method would you use for a population genomic study?

Discussion questions

Project 1: Useful databases for genomic resources

Choose a species of your interest and find all genomic information available (i.e., genome size, chromosome number and whether a reference genome is available). Here are some resources you can use:

https://www.ipni.org/
https://goat.genomehubs.org/
https://cvalues.science.kew.org/
https://ccdb.tau.ac.il/browse/

Project 2: Preparation of a botanical collection for ex-situ conservation of genetic diversity in a wild plant species

Choose any wild plant species. Your goal is to develop a project to prepare an *ex-situ* collection of this plant. You can use the following questions as a guide, which address the main challenges that you can encounter in such a project. Prepare a 10-minute presentation explaining your project.

1. The species
    1. Conservation status (IUCN): is it under conservation laws?
    2. Life history traits: plant life form, breeding system, dispersal, distribution range
    3. What is its genetic variation pattern at the population scale like?
    4. Where do you expect to find the most diverse genetic variants?
    5. Are there domesticated populations of this plant?
2. The collection
    1. What kind of collection would you like to make? (i.e., Living, preserved or other types of collection)
    2. What do you need to harvest from the field? (Is it compatible with the conservation status of your species? Are there any conservation policies you must follow during the harvest?)
    3. What is your sampling strategy to maximize genetic variation with minimum disturbance in the populations?
    4. Can it be relevant to collect also other taxa (i.e. local subspecies or hybrids)
3. Genetic diversity analysis
    1. Which technique would you choose?
    2. Is the genome size known?
    3. Is a reference genome available?
    4. Which DNA extraction protocols have been used for this plant (or close relatives) in other studies?
4. Management of the collection
    1. How will you preserve the material of the collection?
    2. What is your strategy for maintaining genetic diversity over time? (in living collections)
    3. How can this collection be useful for the conservation of the species?
    4. How would you promote your collection?

Slide concepts

Slide	***Concepts***
1	Title
PART 1	**WHAT ARE BOTANICAL COLLECTIONS?**
2	**Botanical collections are organized assemblages of plant materials and metadata**
3-10	**Types of botanical collections and their uses**
4	Herbaria
5	Living collections
6	Seed, pollen and spore banks
7	DNA and tissue banks
8	Ethnobotanical collections
9	Digital collections
10	Palynological, carpological & spirit collections
PART 2	**WHAT IS GENETIC DIVERSITY AND WHY DO WE NEED TO STUDY IT IN PLANTS?**
11-16	**What is genetic variation?**
11-12	Genetic diversity refers to differences in DNA sequences between individuals
13-14	Genetic diversity is an important part of biodiversity and influences many aspects of plant life
15-17	Processes that cause gain or loss of genetic diversity
PART 3	**HOW DID DNA SEQUENCING METHODS CHANGE THE STUDY OF PLANT BIODIVERSITY?**
18-27	**How to study genetic variation with NGS**
18	Before DNA sequencing, other methods were used to study plant genetic diversity.
19-23	Sequencing methods are divided into three families or “generations”. Second generation sequencing or NGS introduced “parallel” sequencing which drastically increased sequencing output and decreased cost.
24-28	Regardless of the sequencing method, the preparation of a sequencing experiment follows the same workflow: isolation of genetic material, preparation of the sequencing library, sequencing, data analysis.
PART 4	**WHAT SEQUENCING STRATEGIES CAN BE USED FOR GENOTYPING A COLLECTION?**
29-41	**Sequencing methods to genotype a collection**
29	Genotyping means determining the DNA sequence of an individual at one or more genomic positions. Generally, data generated for an individual is compared to a reference genome.
30-32	Whole genome sequencing can be used to study a collection, but it is expensive because each genomic position needs to be targeted by many sequencing reads to accurately determine its sequence.

33-34	LcWGS and Pool-seq reduce the “amount” of sequencing (depth) per individual while still targeting the whole genome, and so they are more cost-effective, but less information can be gained.
35-40	Reduced representation sequencing methods
35	RRS methods are a family of cost-effective genotyping methods that focus on sequencing fragments of the genome instead of the whole sequence.
36	RRS methods select a fraction of the genome in several ways and sequence that fraction at high depth.
37	Some RRS approaches are targeted amplification, hybridization, enzymatic digestion + size selection, transcriptome sequencing.
PART 5	**WHAT UNIQUE OPPORTUNITIES DO BOTANICAL COLLECTIONS PROVIDE FOR THE STUDY OF GENETIC DIVERSITY?**
38-47	**Botanical collections provide unique opportunities to study genetic variation, but have unique challenges**
38	Botanical collections allow studying genetic variation across time and space, and they can be used to provide historical preserved specimens or large amounts of fresh material. Living collections can also be used for unique studies on plant reproduction and gene expression.
39-40	It is important to be aware of genetic diversity when developing, maintaining, and using a collection. For example, it is important to sample adequately to represent the diversity of the wild populations, (sampling pressure should not endanger the populations); it is also important to monitor evolutionary processes, such as genetic drift, that may cause a living collection to diverge from its wild relatives.
411-43	Plant genomics has some unique challenges when compared to human and animal genomes: DNA extraction can be more challenging; genome size may be much larger; polyploidy is very common.
44-45	Opportunities and challenges of herbaria for herbariomics
446-47	Opportunities and challenges of living and seed collections
PART 6	**TAKE-HOME MESSAGE**
48	**Synthesis and take-home message**

Lecture synopsis

PART 1: WHAT ARE BOTANICAL COLLECTIONS? (slides 1- 10)

Botanical collections are organized assemblages of plant materials and their associated data. They can be used for scientific research, education, conservation, reference, and public outreach.

These collections are stored in institutions like botanical gardens, museums, or research and education institutions where they are maintained and catalogued to ensure their accessibility and preservation for future uses. Large collections of samples gathered for specific studies are also considered as botanical collections in this text. Links to online collections are provided as examples, for teaching/research activities and for further detail.

There are several types of botanical collections, including: herbaria, living collections, seed, pollen, and spore banks, DNA and tissue banks, palynological, carpological, and spirit collections, digital collections, and ethnobotanical collections.

In herbaria, dried and pressed plant specimens are collected, mounted on sheets of paper and stored in a systematic manner. These collections play a key role in taxonomic studies, in identification of plant species, and as a historical record of plant biodiversity. Herbaria are also used to deposit plants used for reference genome generation.

Living collections are formed by living plants maintained in dedicated facilities, like botanical gardens, arboreta, experimental outdoor or indoor areas, plant nurseries, greenhouses and conservatories. They serve as a resource for research, education, restoration ecology, and public display. They are also important for the conservation of rare and endangered species.

Seed, Pollen, and Spore Banks are facilities that store spores, seeds or pollen under controlled environmental conditions to preserve genetic diversity and ensure the survival of plant species. They are used for conservation, restoration projects, and research on plant genetics and breeding.

DNA and Tissue Banks are repositories of DNA samples, tissues, and other genetic materials from plants. They provide resources for genomics, molecular biology, and conservation genetics research. Ethnobotanical collections are focused on plants used by indigenous communities and in the traditional culture of a community, for medicinal, nutritional, and cultural purposes. They document traditional knowledge, support conservation efforts, and facilitate research on plant-based remedies and sustainable use.

Digital Collections contain digitized records of plant specimens, including high-resolution images and its associated metadata. They enhance accessibility of biodiversity data, support global research collaborations, and allow for virtual studies of plant diversity. Other botanical repositories in which botanical specimens are collected are Palynological, Carpological and Spirit Collections. Palynological collections are assemblages of pollen and spores, often prepared as slides for microscopic examination. Carpological collections store fruits, seeds, flowers or other plant parts (usually, associated with herbarium collections) that cannot be mounted on herbarium sheets because of their tridimensional form. These collections are used for studies in taxonomy and species identification, paleobotany, palynology, climate change research, and forensic science.

Spirit Collections contain specimens preserved in fluid stored in glass jars. These collections are particularly useful for preserving organs that are not suitable for drying (i.e., fleshy flowers and fruits). They can also allow more accurate measurements of plant organs as drying may lead to shrinkage and better observation of three-dimensional arrangement of organs (i.e., flower parts).

PART 2: WHAT IS GENETIC DIVERSITY AND WHY DO WE NEED TO STUDY IT IN PLANTS? (Slides 11-17)

Genetic diversity refers to the genetic differences among individuals within or between populations and species. It manifests as differences in DNA sequence ('polymorphisms') caused by sexual reproduction, meiotic recombination, and mutations. Many DNA polymorphisms do not affect the phenotype but even then, they are useful to measure the differences between individuals and populations and understand evolutionary dynamics.

At the sequence level, these polymorphisms can have various levels of complexity, ranging from changes in a single nucleotide among individuals or populations to larger modifications involving entire chromosome segments. A Single Nucleotide Variant (SNV), or Single Nucleotide Polymorphism (SNP), is the simplest type of polymorphism and occurs when a single base in the DNA is replaced by a different base, deleted, or inserted. Multiple Nucleotide Polymorphisms (MNP), on the other hand, involve changes of adjacent nucleotides up to 50 base pairs. Structural Variants (SV) encompass nucleotide sequences longer than 50 base pairs. These changes can include deletions, insertions, translocations (where a DNA segment changes its position), inversions (where the sequence is reversed), and copy number variations (changes in the number of copies of a gene or of a specific sequence). Finally, chromosomal rearrangements represent even larger-scale modifications that involve entire portions of chromosomes. These rearrangements can result in inversions, translocations, losses, or duplications of entire chromosomes.

Depending on the site where the modification occurs (non-coding regions, exons, highly conserved portions of genes) and on the extent of the modification, DNA changes can have different effects on the phenotype, the survival, and the fitness of the plants and, therefore, on their evolution. Short DNA modifications might have no effect on the fitness of individuals or in gene expression or protein functions. Changes in long DNA fragments have generally more chances to involve coding genes or regulative parts of the genome, leading to effects on gene functionality and even to post-zygotic reproductive barriers (especially for chromosome rearrangements), which can trigger reproductive isolation between individuals or populations, thereby influencing the evolution and diversification of plant species.

Genetic diversity is an important component of biodiversity because it helps defining species and delimiting groups of related individuals, like subspecies or ecotypes. It also affects the phenotypic variation of organisms and their ecology and evolution. Populations with higher genetic variation have more chances to overcome environmental challenges. Studying genetic diversity is also crucial for the conservation of endangered plant species or for crop improvement because it helps researchers to understand the adaptive potential of populations, develop strategies for preserving genetic resources or improving traits related to yield, quality, and resistance in crops.

Genetic variation can change over time due to factors that alter allele frequencies in populations, providing selective advantages or disadvantages, a process known as natural selection. New variants can also arise through mutations or by seed or pollen dispersal, increasing genetic diversity. However, phenomena like genetic drift and inbreeding tend to reduce genetic variability. Sexual reproduction plays a key role in enhancing the genetic variation of populations, promoting the recombination of existing alleles within individuals. Therefore, the number of individuals capable of sexual reproduction significantly affects the diversity produced. Population size and isolation greatly impact genetic diversity and its conservation. Small populations are more affected by genetic drift and inbreeding, leading to a higher occurrence of homozygotes for deleterious recessive traits. Studying genetic variation patterns within and between populations reveals genetic dynamics, identifies isolated populations, and highlights diversifying populations, offering valuable insights into evolutionary processes, fitness, and conservation status.

Since botanical collections store individuals collected from different areas and from different times, they provide an invaluable resource for studying genetic diversity changes across space and time. However, there are also challenges that must be considered when setting up a collection that can be mitigated by following some precautions during the collection phase. One crucial issue to consider is that the collection must capture as much of the genetic diversity as possible because it must ensure the conservation of the evolutionary capacity of a species and the genetic resources necessary for applied breeding research.

PART 3: HOW DID DNA SEQUENCING METHODS CHANGE THE STUDY OF PLANT BIODIVERSITY? (SLIDES 18-28)

Before the development of DNA sequencing technologies in the 1970s, genetic diversity analysis was studied through morphological, biochemical, or cytological variants. The possibility of sequencing DNA was an important breakthrough in biological research as it enabled the study of DNA sequences regardless of the phenotype. The introduction of DNA sequencing has made it possible to obtain direct genetic information, overcoming the limitations of previous methods.

The first sequencing methods were developed in the 1970s (first-generation sequencing). Sanger sequencing is the only first-generation method that remains in use today for specific applications. Sanger sequencing has a very low error rate, but only a single, relatively short fragment (around 1kb) can be sequenced per reaction. Nowadays, it is used to target specific short sequences in the genome as covering large parts of the genome is costly and labor intensive. In Sanger sequencing, the fragment to be sequenced is amplified with specific PCR primers. Then, a polymerase reaction is set up with a mix of normal nucleotides (dNTPs) and fluorescently labeled chain-terminating inhibitors (ddNTPs). The ddNTPs are randomly incorporated and, whenever this happens, the DNA chain is terminated, resulting in a collection of fragments of varying lengths, all terminating with a fluorescent ddNTP. These fragments are then separated by capillary electrophoresis. Then, a laser is used to excite the fluorescent labels, identifying the ddNTP at the end of each fragment. By analyzing the sequence of fluorescent signals, the DNA sequence is determined.

Newer sequencing methods, also known as *high throughput sequencing*, *parallel sequencing* or *next generation sequencing* (NGS), drastically increased the sequencing output because they can sequence many fragments at once.

Two generations of NGS methods are currently available. Second-generation methods granted the possibility of sequencing many fragments simultaneously for the first time, but the length of the fragments was limited (up to 300 bp). Of the different methods developed, only Illumina is still in use. In Illumina sequencing, the fragments are bound to a support known as *flow cell* thanks to hybridization of predefined complementary sequences that are added to both the fragments and the flow cell. The DNA is then amplified while still bound to the flow cell, in a process called *bridge amplification*, forming clusters of identical molecules. Next, fluorescently labeled nucleotides are incorporated one at a time into the complementary strand in multiple sequencing cycles. In each cycle, only one type of nucleotide is added, and the fluorescence emitted by each cluster is detected by a laser. At the end, the fluorescence emitted by each cluster in each cycle is analyzed, allowing the sequences to be reconstructed.

Third-generation sequencing technologies like Oxford Nanopore and PacBio, developed in the late 2010's, provide much longer sequences (up to hundreds of kilobases). Although third-generation sequences were initially less accurate, at present their precision is close to that of Illumina sequences, offering significant advantages for studying structural variants (SVs) such as inversions, deletions and copy number variations. However, they are more expensive. The third generation was kickstarted by Oxford Nanopore Technologies (ONT). In this method, a difference in potential is applied between two sides of a membrane. Current flows between the two sides of a membrane through a set of pores made of a specially engineered protein. An enzyme "pulls" DNA molecules from one side of the membrane to the other through the pore. This alters the current flowing through the pore with a specific pattern depending on the nucleotide sequence of the DNA that is passing through the pore. The sequencing method developed by PacBio is more similar to Illumina as it is based on polymerase reactions and fluorescent nucleotides, but the reaction is miniaturized and occurs in nano wells; each nano well contains one molecule of polymerase and one molecule of DNA. Additionally, the DNA molecule is circularized with special adapters, so that the polymerase can keep catalyzing the reaction in a loop and the same molecule is sequenced multiple times which allows for error correction.

All NGS methods share a common workflow, consisting of several phases:

1. The nucleic acid (DNA or RNA) is isolated from different samples. Samples shall be accurately labelled and appropriately preserved, and extraction of good quality nucleic acid is imperative.
2. A sequencing library is prepared (the protocol depends on the NGS method). Generally, the DNA must be either fragmented or digested by enzymes to reduce fragment size. Additionally, specific sequences are added to the fragment ends with a ligation enzyme according to the different sequencing technologies. Then, many copies of all fragments are obtained with PCR amplification.
3. The libraries are loaded in the chosen sequencing platform and the signal (fluorescence, current, etc.) is converted into nucleotides.

4. The quality of the data is assessed and several bioinformatic analyses can be carried out.

PART 4: WHAT SEQUENCING STRATEGIES CAN BE USED FOR GENOTYPING A COLLECTION? (SLIDES 29-37)

Genotyping is the process of determining the DNA sequence at positions within the genome of an individual. Therefore, having a reference genome can simplify the process. A reference genome is the sequence, as complete as possible, of the genomic DNA of an organism belonging to a species, annotated with the position of gene sequences and other functional elements, such as regulatory regions, repeats or transposable elements. The reference genome provides a framework for comparing sequences and identifying variants among individuals of the same or different species.

Genotyping can be approached in various ways; generally, the decision on the approach to follow is a trade-off between the amount of information that can be gained and the cost of sequencing. Whole genome sequencing and resequencing (WGS and WGR) is the process of determining the sequence of the entire genome of interest. The reads obtained (after a reference genome is available), are compared to the reference genome to determine their location on the genome (this process is called “alignment”) and accurately determine the genetic variants present in the studied individual/individuals compared to the reference genome (“variant calling”). This (mostly) complete and unbiased set of variants can be employed for any genetic diversity study, but it has significant downsides as it is costly and requires large amounts of computing power and storage. Moreover, because NGS is not error-free, many independent reads “covering” the same genomic position are needed to make sure that the inferred sequence is accurate (the number of reads is referred to as “sequencing depth”). This makes WGS very expensive for species with large genomes. As such, different more cost-effective approaches were developed.

Low coverage WGS, also known as genome skimming, consists of sequencing genomes at a very low depth; this allows recovery and genotyping at regions that are naturally present in many copies, such as organelle DNA. This has a very low cost but is limited in that only a few regions can be analyzed.

Pool-seq is a technique in which DNA from multiple individuals is combined into a single sample, or "pool," and then sequenced together. Individuals in a pool share a specific characteristic (such as a phenotype, geographic origin, or species). By comparing the genetic data of different pools—for example, one with individuals showing a trait and another without—it is possible to identify genetic variants associated with that trait as, despite the variability derived from sequencing multiple individuals together, the region responsible for the characteristic of interest will stand out as neatly different. Pool-seq can also be used to estimate allele frequencies across populations. However, it’s important to ensure equal DNA contributions from each individual, as uneven amounts can introduce bias.

Finally, there is a family of methods, known as Reduced Representation Sequencing (RRS), that relies on selecting a portion of the genome for sequencing. Therefore, the regions selected can be sequenced at a high depth for a lower cost. Additionally, a reference genome is not mandatory for these approaches.

RRS methods rely on different approaches to select genome fractions. Many of them, like RAD-seq (Restriction Site-Associated DNA sequencing), or GBS (Genotyping-

By-Sequencing), use restriction enzymes. Others, like K-seq, rely on short random oligonucleotides designed to target single-copy regions randomly distributed in the genome. In Targeted Amplification and Sequence Capture the regions to be sequenced are pre-selected and either selectively amplified with specific primers (i.e. TruSeq custom amplicons) or enriched via hybridization to single-stranded probes (Sequence Capture).

PART 5: WHAT UNIQUE OPPORTUNITIES DO BOTANICAL COLLECTIONS PROVIDE FOR THE STUDY OF GENETIC DIVERSITY? (SLIDES 38-47)

Botanical collections can provide a picture of plant variation patterns in their range and at the time of collection, but they can also be used to obtain large amounts of fresh plant material (for example, by growing new plants in a nursery or with seeds from a seed bank). We can now perform genetic studies on both precious, old, preserved specimens, and large amounts of fresh samples, at an affordable price.

However, most botanical collections were not established with the objective of studying genetic variation, so this can limit the genetic diversity contained in them and bias the results. Therefore, when establishing new collections, it is crucial to adopt some precautions to collect as much genetic variation as possible and maintain it for future genetic studies. Moreover, understanding the challenges and the genetic dynamics within living plant collections can be useful to assess the suitability of older collections for genetic studies. Once the material is collected, it is crucial to maintain genetic diversity inside the collection. Living collections can be considered as isolated populations and are subject to the same evolutionary forces that affect genetic diversity in nature (i.e., selection, genetic drift, inbreeding, mutation events). These effects are stronger for species that reproduce mainly by seed. Since the collection aims to mirror the genetic diversity of the natural populations, these dynamics should be carefully monitored.

In general, plant genetic and genomic research has some additional challenges when compared to human/animal research.

First, obtaining high-quality DNA from plant tissues can be challenging, as plants can produce large amounts of specialized metabolites, such as polysaccharides or aromatic compounds that are difficult to remove from the DNA. The type and concentration of these metabolites vary between plant species and are influenced by environmental conditions, making it challenging to develop a universally effective DNA extraction protocol. Some adjustments can be made to common protocols, such as adding polyvinylpyrrolidone or sorbitol to the extraction buffer, or isolating nuclei priori to DNA extraction. Additionally, young leaves tend to accumulate less of these interfering compounds and thus perform better in DNA extraction.

Secondly, genome size is highly variable in plants, and it can reach up to 150 Gbp (human genome has only 3 Gbp). Due to the need of sequencing depth, genome size may increase the amount of sequencing data required for accurate genotyping and WGS may become extremely costly.

Third, polyploidy is very common in plants. Polyploidy is the heritable condition of possessing more than two complete sets of chromosomes. Stable polyploid populations or species can either derive from whole-genome duplication events (which originate autopolyploids) or from hybridization between close-related species

(originating allopolyploids). Polyploids, and especially allopolyploids, pose challenges for genomic studies as reads of related genes coming from each subgenome (homeolog genes) are easily confounded, complicating alignment and annotation at the individual level.

The study of herbarium samples with NGS methods is called herbariomics. Herbaria provide information about the genetic variants occurring at the time and place where the individual was collected, capturing a “snapshot” of the genetic make-up of a species in that moment. Genetic data from historical preserved specimens offer a unique opportunity to reconstruct past genetic diversity patterns and species evolution. Even pathogens, preserved within plant tissues in herbaria, such as fungi or viruses, can be studied, providing insight also into co-evolutionary processes. However, genotyping herbarium material is challenging. DNA from old samples tends to be somewhat degraded, so it is impossible to extract high molecular weight DNA. Specialized metabolites can bind DNA hindering extraction, or the specimen may be contaminated by the DNA of molds, fungi, or other samples. Herbarium samples provide also a very limited quantity of tissue that can be used for DNA extraction without irreparably damaging the herbarium sample. The digitization of herbarium specimens is promoting the use of herbaria for genetic studies because it facilitates the dissemination and consultation of collections and allow the preservation of the plant's morphology after a part of the physical specimen is used for DNA extraction.

The use of living collections grown in garden conditions and plants derived from seeds preserved in seed banks is advantageous as it can provide large amounts of young, fresh material and more homogenous DNA extracts, improving DNA performance in downstream reactions. Living collections can provide plants and environmental conditions to study the reproduction capacity of threatened species without stressing natural populations. Common garden experiments (i.e. where plants of different origins are grown together in common conditions) can help disentangle genetics and environmental effects in trait expression or stress response. Additionally, individuals from living collections can be used to produce germplasms to reintroduce or reinforce wild populations. However, unintentional divergence of the collection’s genetic makeup from that of the wild population needs to be considered when using such collections.

TAKE-HOME MESSAGE (SLIDE 48)

With the decreasing costs of sequencing technologies and the vast number of genotyping approaches available, there has never been a better time to apply genome-wide analyses of genetic diversity to plant species for conservation and other botanical purposes. Genomic data can inform conservation strategies on a previously unmatched resolution and botanical collections represent both a rich reservoir of diversity and a foundation for plant research and conservation especially if the establishment of future collections is guided by genetic principles. Nevertheless, studying plant genetic diversity presents many challenges. In addition to the several challenges associated with the genomes of the botanical collections, genome sequencing and downstream analyses can be challenging for botanists and naturalists with no molecular background and translating genomic data into conservation practice is challenging *per se.*

Genomic Analysis of Botanical Collections: Opportunities and Challenges

Martina Deglialberti, Chiara Paleni, Federico Fainelli, Carla Lambertini, Silvia Manrique

1- What are botanical collections?

Herbarium specimen of *H. umbellatum*, herbarium, Milan herbarium, (https://erbario.lim.unimi.it)

Salvia desoleana plant, Orto Botanico di Brera, Milan (https://ortibotanici.unimi.it)

- Organized assemblages of plant materials
- Associated with **metadata** (i.e., GPS coordinates, collection date)
- Useful for:
  - Scientific research
  - Education
  - Conservation
  - Reference

- Important for documenting **plant diversity**, for studying the **variation within and between species**, their **current** and **past distribution**, for better understanding **plant evolution and ecology**.

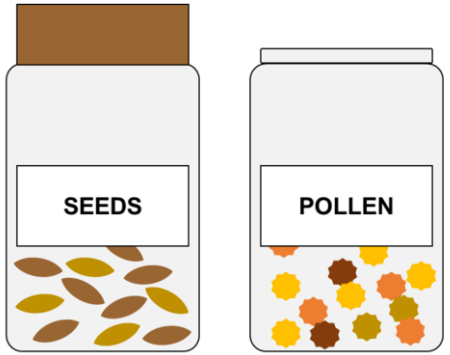
Types of collections

Herbaria

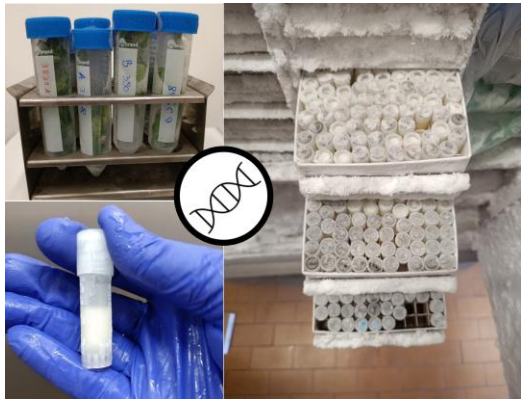
Living Collections

Designed by Freepik

Seed, Pollen, and Spore Banks



DNA and Tissue Banks



Palynological, Carpological & Spirit Collections

(pollen, fruit and other of specimens preserved in alcohol for morphology analysis)

Designed by Seksak Kerdkanno from Vecteezy.com

Digital Collection

Designed by redgreystock (PC) and pikisuperstar (herbs) from Freepik.

Ethnobotanical Collections

Adapted with permission from Cámara-Leret & Bascompte (2021). Language extinction triggers the loss of unique medicinal knowledge. *PNAS*, *118*(24), e2103683118.

Types of collections

Herbaria

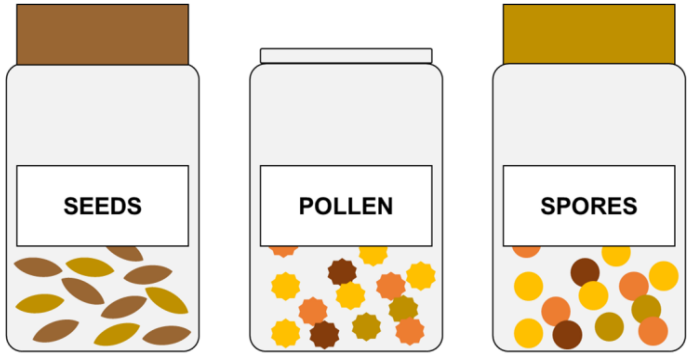
What	Dried and pressed plant specimens, stored systematically
Uses	Taxonomic and floristic studies, plant identification, and historical documentation of biodiversity. They are also repositories to store voucher specimens of individuals used in reference genome sequencing.
Examples	New York Botanical Garden Herbarium (https://sweetgum.nybg.org/science/) Royal Botanic Gardens Kew Herbarium (https://www.kew.org/science/collections-and-resources/collections/herbarium) Missouri Botanical Garden Herbarium (https://www.missouribotanicalgarden.org/plant-science/plant-science/about-science-conservation/departments-centers-of-excellence/herbarium)

Types of collections

Living collections

What	Plants maintained in botanical gardens, arboreta and nurseries
Uses	Research, conservation of endangered species, and public education
Examples	The New York Botanical Garden (https://www.nybg.org/) The Royal Botanic Gardens Kew (https://www.kew.org/) Missouri Botanical Garden (https://www.missouribotanicalgarden.org/)

Types of collections



Seed, pollen & spore banks

What	Stored seeds, pollen, or spores under controlled conditions
Uses	Conservation and restoration efforts, study of past climate
Examples	Millennium Seed Bank (https://www.kew.org/wakehurst/whats-at-wakehurst/millennium-seed-bank) Svalbard Global Seed Vault (https://seedvault.no/) Chicago Botanic Garden's Pollen Bank (https://www.chicagobotanic.org/research/pollen-bank)

Types of collections

DNA & tissue banks

What	Repositories of DNA and plant tissues
Uses	Genomics and conservation genetics research
Examples	Global Genome Biodiversity Network (GGBN) (https://www.ggbn.org/) The DNA Bank at the New York Botanical Garden (https://sweetgum.nybg.org/science/digital-collections/dna-bank/) The Center for Comparative Genomics CryoCollection at the California Academy of Sciences (https://www.calacademy.org/scientists/ccg/ccg-cryocollection)

Types of collections

Adapted from Cámara-Leret & Bascompte (2021). Language extinction triggers the loss of unique medicinal knowledge. *PNAS*, *118*(24), e2103683118.

Ethnobotanical collections

What	Collections with a focus on plants used by Indigenous peoples
Uses	Document traditional knowledge and plant-based remedies
Examples	Missouri Botanical Garden Ethnobotany Search (https://tropicos.org/ethnobotany/Search) National Museum of Natural History Paris Ethnobotany Collections (https://www.mnhn.fr/en/ethnobotany-collection) University of Michigan Herbarium Ethnobotanical Collection (https://museumcollab.anthro.lsa.umich.edu/s/Anishinaabe/page/ebot) Ecomuseo delle Erbe Palustri (Ecomuseum of marsh plants) (Bagnacavello, Ravenna, Italy) https://ecomuseoerbepalustri.it/

Types of collections

Digital collections

What	Digitized plant specimens, including genetic data
Uses	Global access to herbarium data for virtual studies and collaborative research, genetic and genomic information on species
Examples	Integrated Digitized Biocollections (iDigBio) (https://www.idigbio.org/) Global Biodiversity Information Facility (GBIF) (https://www.gbif.org/) Global Plants on JSTOR (https://plants.jstor.org/)

Types of collections

Designed by Seksak Kerdkanno from Vecteezy.com

Palynological, carpological, & spirit collections

What	Collections of pollen, spores, fruits, seeds, or other parts
Uses	Morphology, paleobotany, climate change research
Examples	Personal and department collections, like: ● The John P. Smol Paleolimnology and Environmental Change Laboratory (https://www.queensu.ca/pearl/) ● The University of Arizona Palynology Laboratory (https://www.geo.arizona.edu/palynology) The Royal Botanic Garden Edinburgh Spirit Collection (https://www.rbge.org.uk/science-and-conservation/preserved-collections/herbarium/our-collections/spirit/) The Royal Botanic Gardens Kew Spirit Collection (https://www.kew.org/science/collections-and-resources/collections/spirit-collection)

2- What is genetic diversity and why is it important?

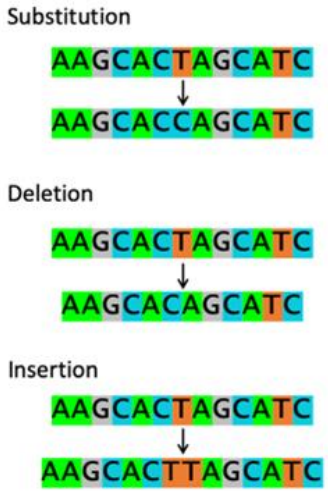
Differences or changes in the DNA sequence can be observed…

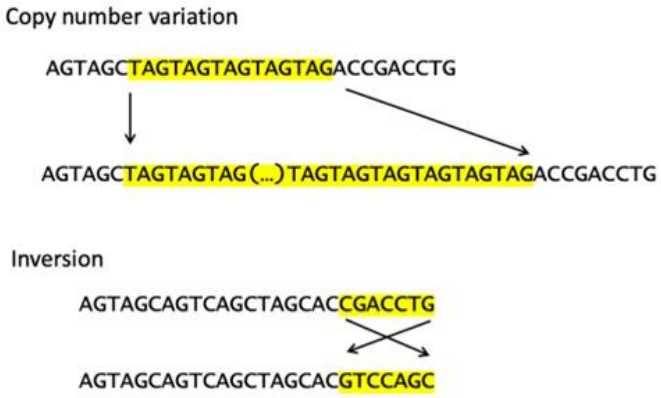
among individuals within one population

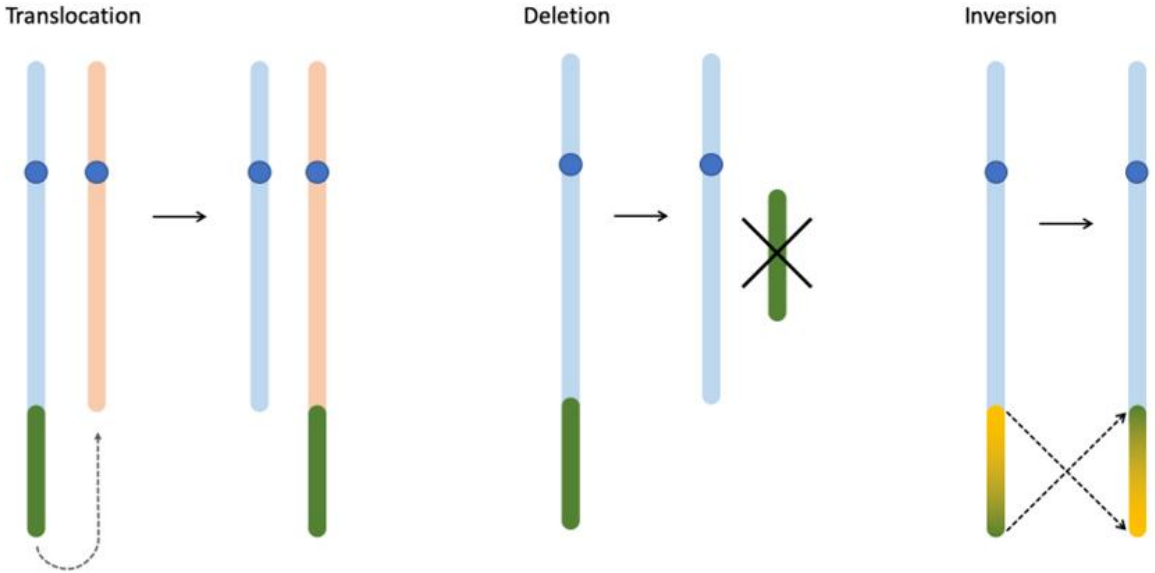
between populations of the same species

between species

Designed with icons by Flaticon







Genetic variation can have different levels of complexity

- Differences on a single base (SNPs*)
- Differences on multiple bases (MNPs*)
- Large-scale differences (SVs*)

- Bases can be replaced, deleted, or their order can be inverted

- Chromosome-level rearrangements can lead to lethal defects or reproductive isolation

*SNPs (Single Nucleotide Polymorphisms), MNPs (Multi-Nucleotide Polymorphisms), SVs (Structural Variants)

Why study genetic variation?

Solberg, S. Ø., et al. (2022). Photo of Luis M. Salazar/Crop Trust.

- It influences **phenotypic** variation
- It influences plant **ecology**
- It helps **delimit species**, subspecies and ecotypes
- It helps identify vulnerable populations
- It can help make crops more resilient
- Genetic diversity is an important component of **biodiversity!**

Genetic diversity can be selected and provide adaptive potential

(Example of environmental changes due to the introduction in the environment of a caterpillar that feeds only on flowers with a short stem)

Population with high diversity

The environment changes

Selection of beneficial alleles

Population adapted to new conditions

Population with low diversity

The environment changes

The population cannot adapt

Genetic diversity can change over time

Designed by Freepik

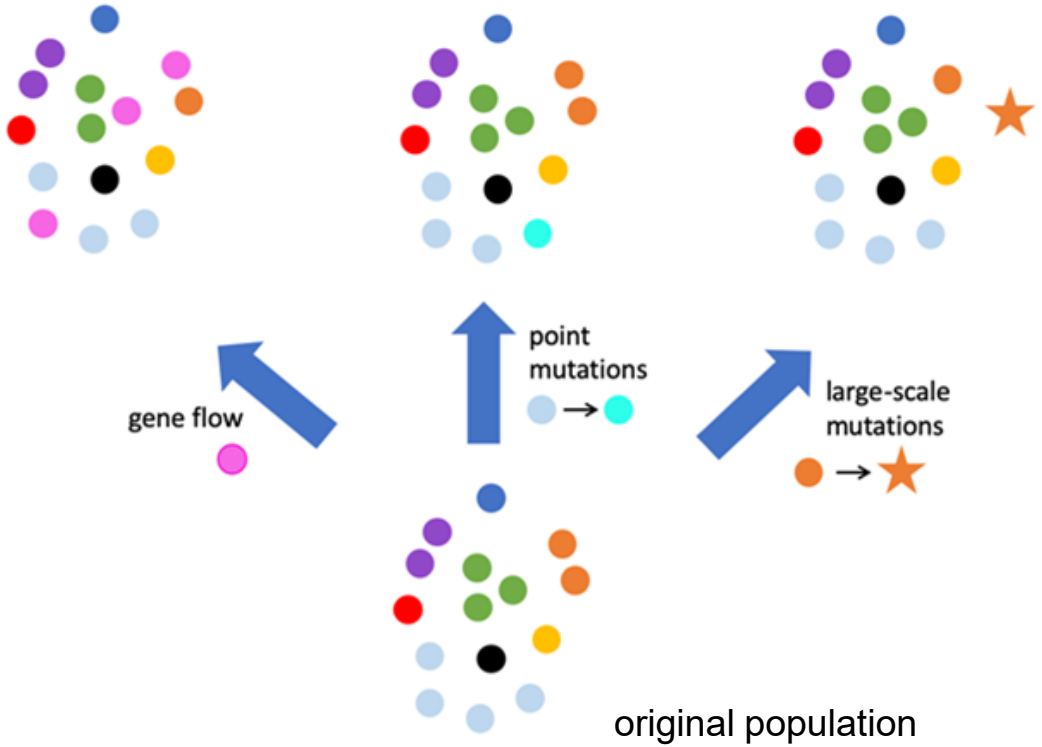
Genetic diversity

Can increase:

- mutations
- gene flow with other populations (dispersal, migration)
- sexual reproduction (new combinations of alleles)

Can decrease:

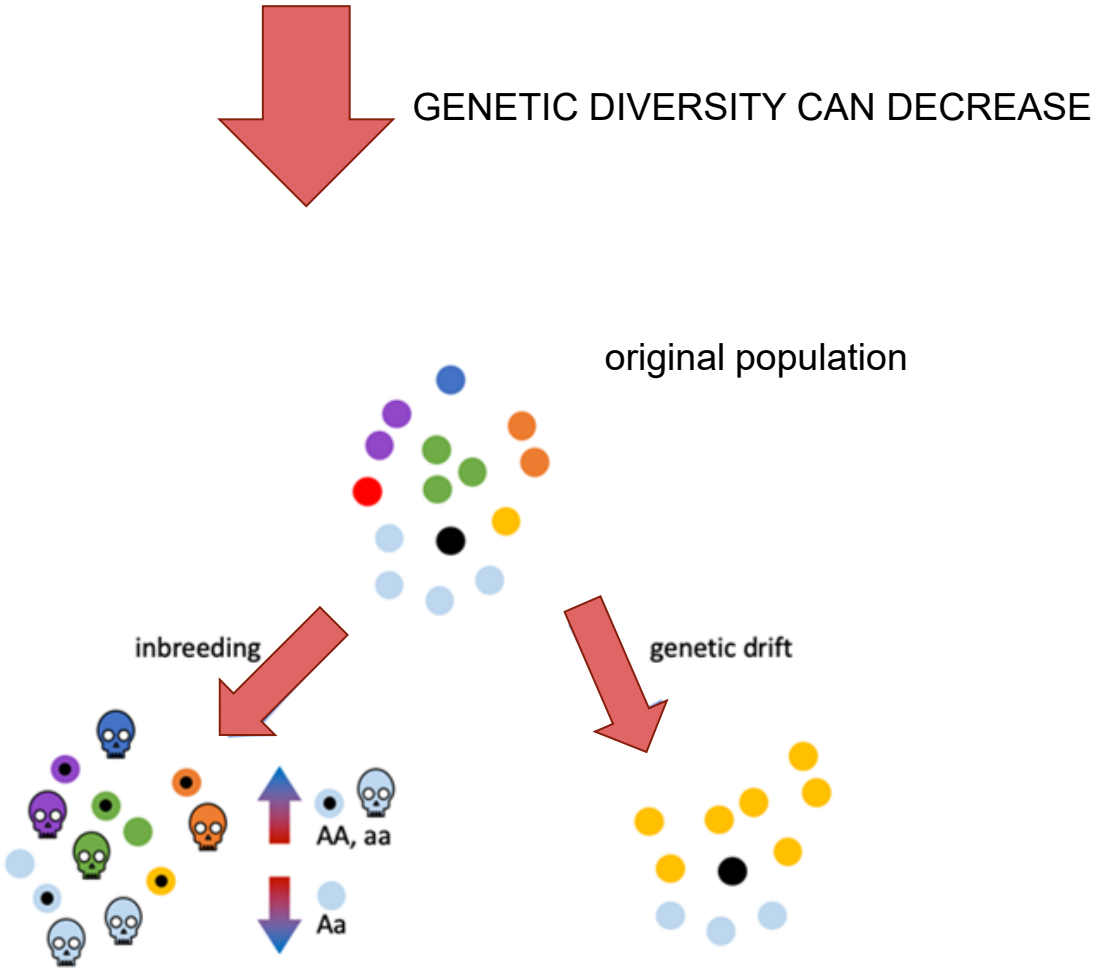
- selection
- genetic drift
- inbreeding



GENETIC DIVERSITY CAN INCREASE

Sources of genetic variation

- **Novel alleles** can be introduced from a separate population by gene flow.
- **Genetic mutations** also introduce new alleles.
- **Large-scale mutations** (at chromosome level) can cause reproductive isolation.
  - **Plant genomes** are tolerant to those events and so they are not always lethal.



Genetic diversity is easily lost in small populations

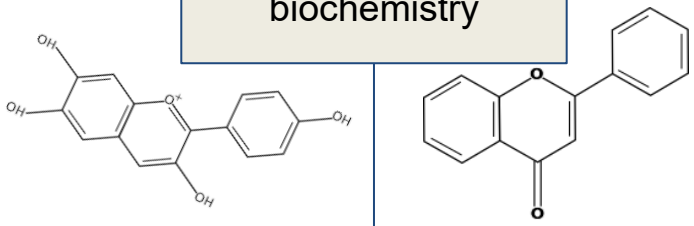
- **Genetic drift** causes genetic variants to be lost or fixed in a population at random.
- **Natural selection** causes alleles to be lost or fixed depending on their effect on the **fitness** of the individual.
- **Inbreeding** increases the rate of homozygosity.
- Isolated populations will experience all these effects independently.

3- How did DNA sequencing methods change the study of plant biodiversity?

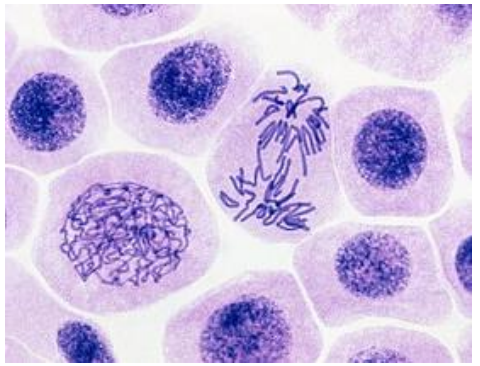
Methods used to study plant diversity <u>before</u> DNA sequencing:

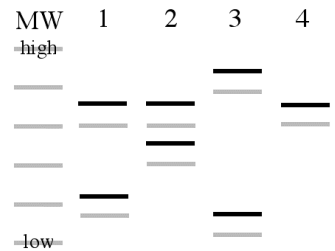
phenetics

biochemistry



cytology and cytogenetics





influenced by

Some methods can be **influenced by external factors**, leading to limitations in accuracy and consistency.

<u>After</u> DNA sequencing the study of plant genetic variability can be conducted by directly analyzing DNA.

DNA (genetic diversity)

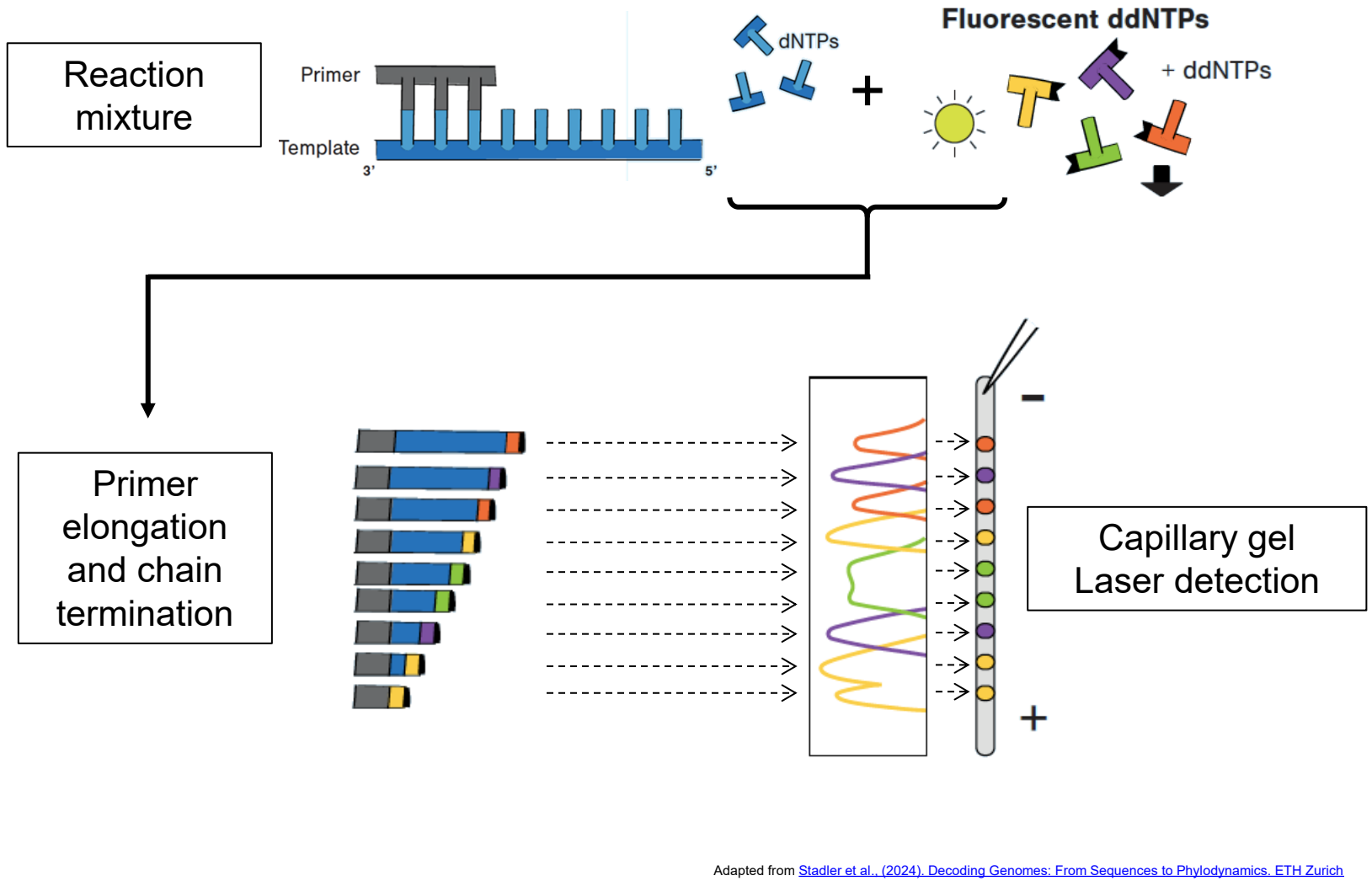
These technologies can detect diversity that does not affect the phenotype of the organisms, overcoming the limitations of previous methods.

Onion cell image from Doc. RNDr. Josef Reischig, CSc., CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0>, via Wikimedia Commons

DNA sequencing

	First Generation Sequencing	Next Generation Sequencing	Third generation sequencing
Development	1970s	2000s	2000s-2010s
Main technology	Sanger sequencing	Illumina	Oxford Nanopore and PacBio
Reads length	500-1000 bp	50-300 bp	10.000-100.000 bp
Pros	Very high accuracy, still used as a golden standard method	Sequence hundreds of samples in parallel (high throughput and lower costs per sequence)	Generate longer sequenced fragments (useful for genome assembly or diagnostic)
Cons	Costly and labor intensive Only one sequence at run	Lower accuracy than Sanger → Use the large amount of data to correct	Accuracy comparable to Illumina More expensive

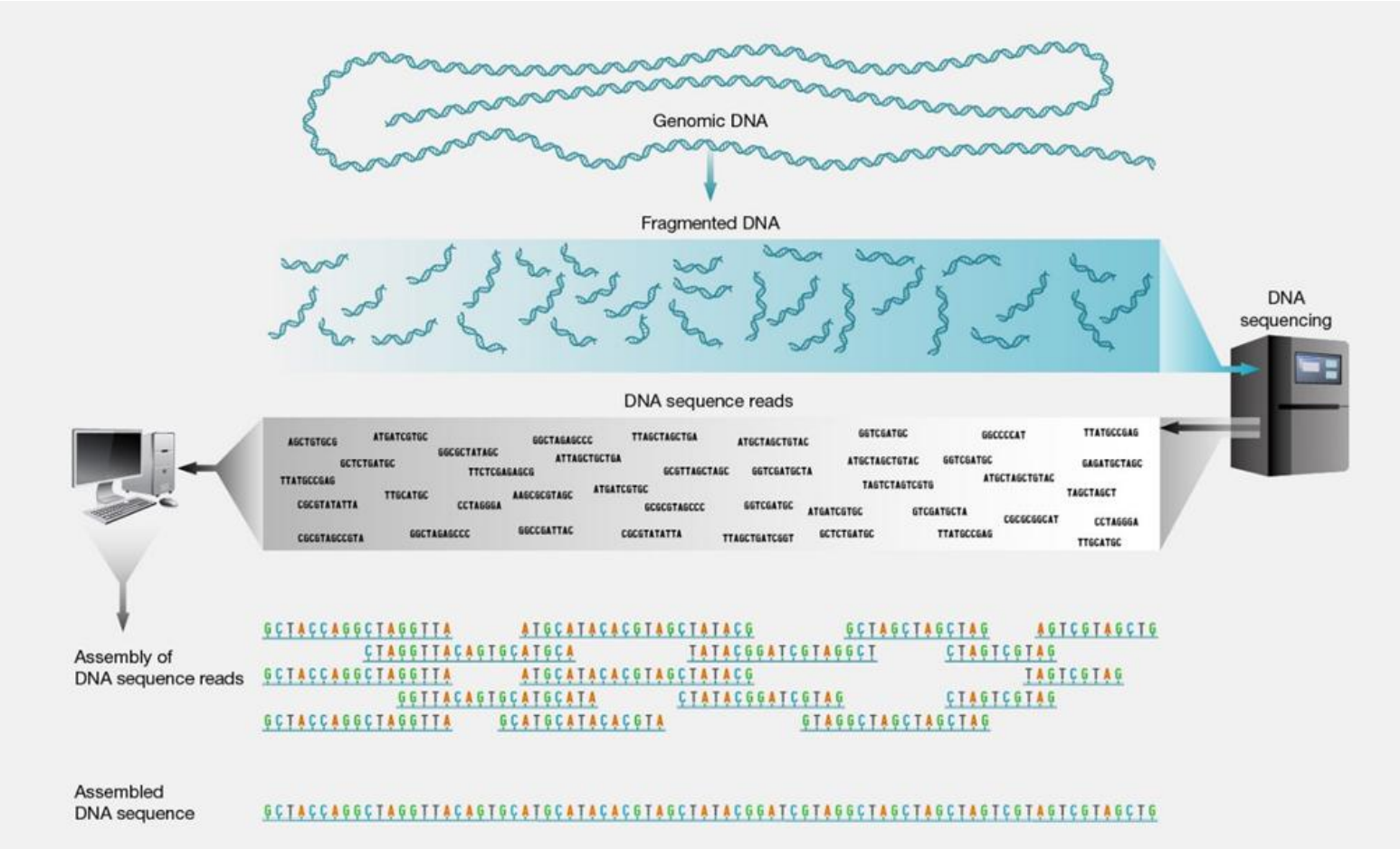
The First Generation: Sanger sequencing



Adapted from Stadler et al., (2024). Decoding Genomes: From Sequences to Phylodynamics. ETH Zurich

- Previously, target DNA is PCR-amplified with specific primers
- Nucleotides (dNTPs) and fluorescently labelled chain-terminating inhibitors (ddNTPs) are incorporated into the complementary strand by a DNA polymerase. ddNTPs stop the polymerase reaction
- This produces a mix of fragments with various lengths all terminating with one type of ddNTP
- Electrophoresis and a laser are used to determine the length of the fragment and the terminating ddNTP according to the fluorescence
- The full “target" sequence is reconstructed

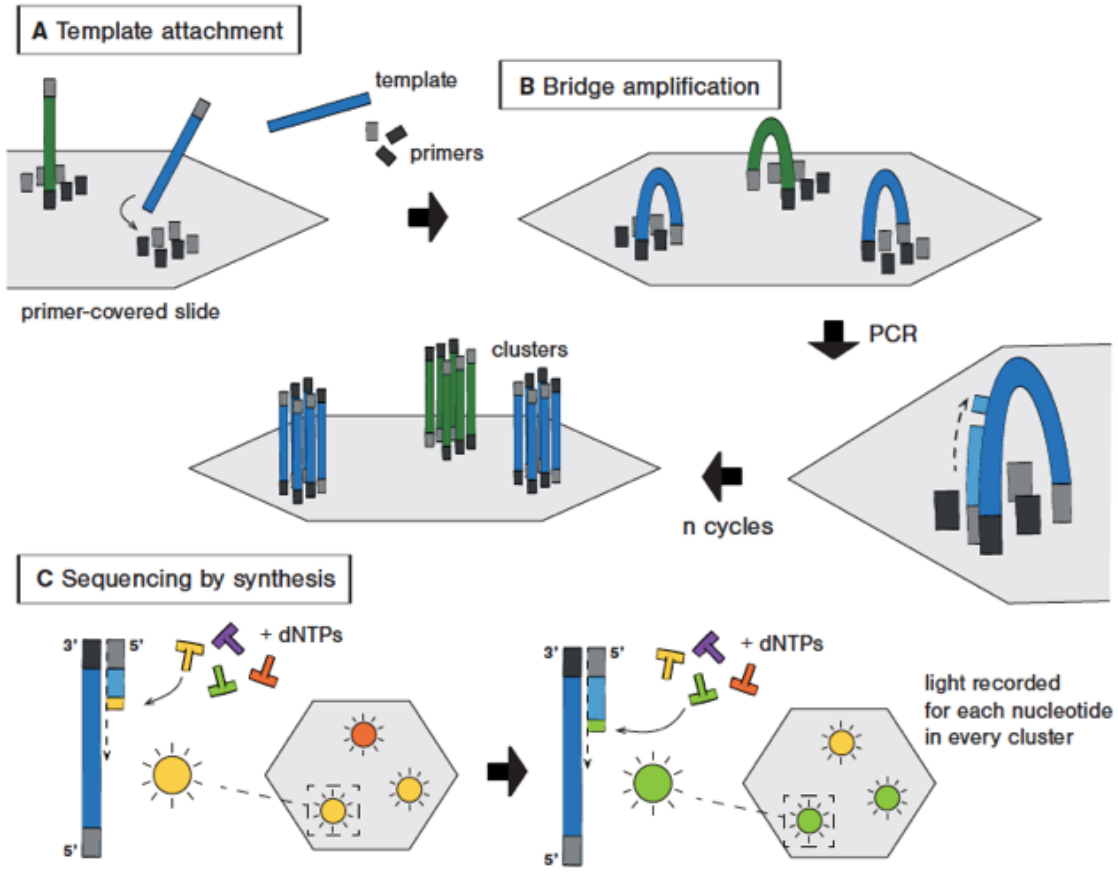
Next Generation Sequencing and high throughput technologies



Courtesy: National Human Genome Research Institute

- Sequence in parallel massive amounts of previously fragmented DNA
- Short length of the sequenced fragments (=reads)
- Increased output, reduced cost
- Also known as Second Generation Sequencing

NGS: Illumina sequencing by synthesis



. Stadler et al., (2024). Decoding Genomes: From Sequences to Phylodynamics. ETH Zurich

Library Preparation

- The DNA is fragmented, and specific **adapters** are added to the sequence ends to bind universal Illumina sequencing **primers** during sequencing.

Cluster Formation

- The DNA is bound to a flow cell and amplified to create **clusters** of identical sequence copies (**bridge amplification**).

Sequencing by Synthesis (SBS)

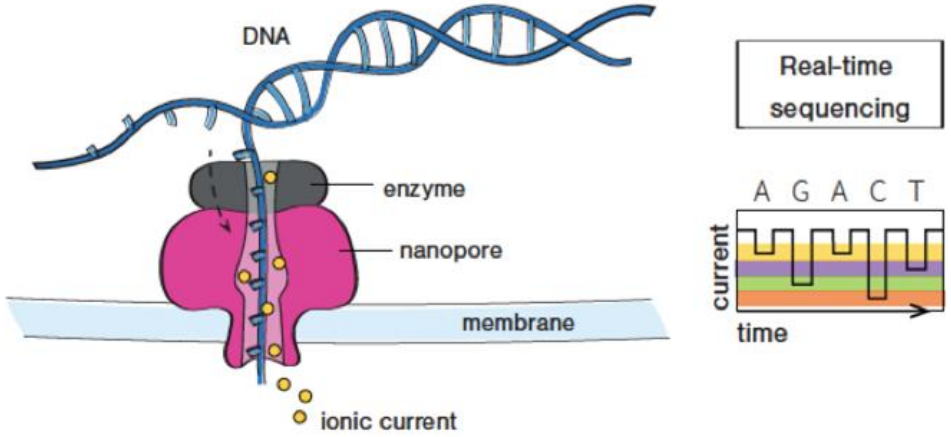
- Fluorescently labeled nucleotides (ATP, TTP, CTP, GTP) are incorporated one at a time into the complementary strand.
- Each nucleotide emits a **fluorescent** signal, which is detected by a laser.

Sequence Identification

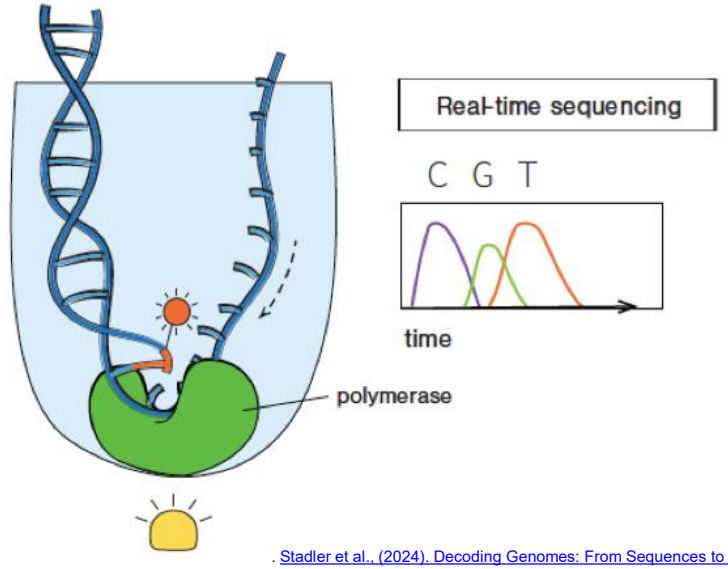
- The fluorescent signals are recorded and processed to determine the DNA nucleotide sequence.

Third Generation Sequencing - Long reads

ONT



PacBio HiFi



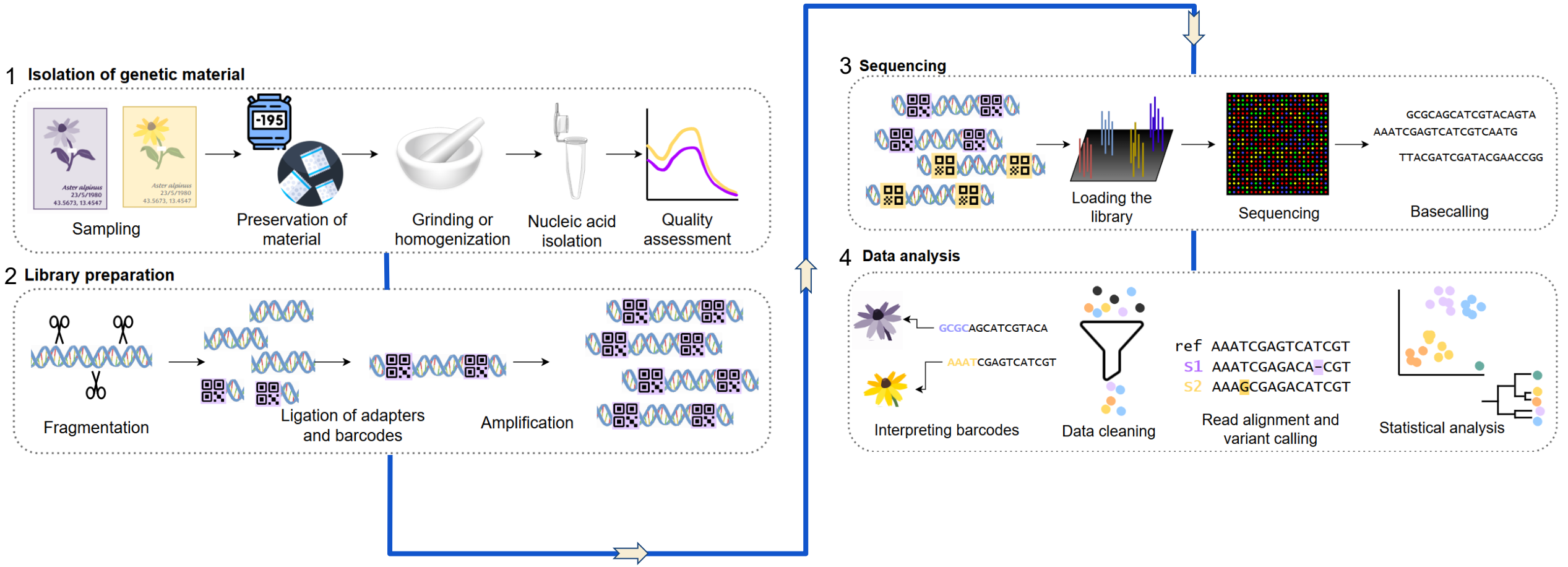
- Third gen methods allow to sequence DNA without amplification
- The sequenced fragments are up to hundreds of kilobases long
- More expensive

- **Oxford Nanopore Technologies (ONT)** is based on a protein nanopore. Different nucleotides alter the current signal while flowing through the pore with specific patterns
- **PacBio CLR/HiFi** is based on miniaturized polymerase reactions and fluorescence

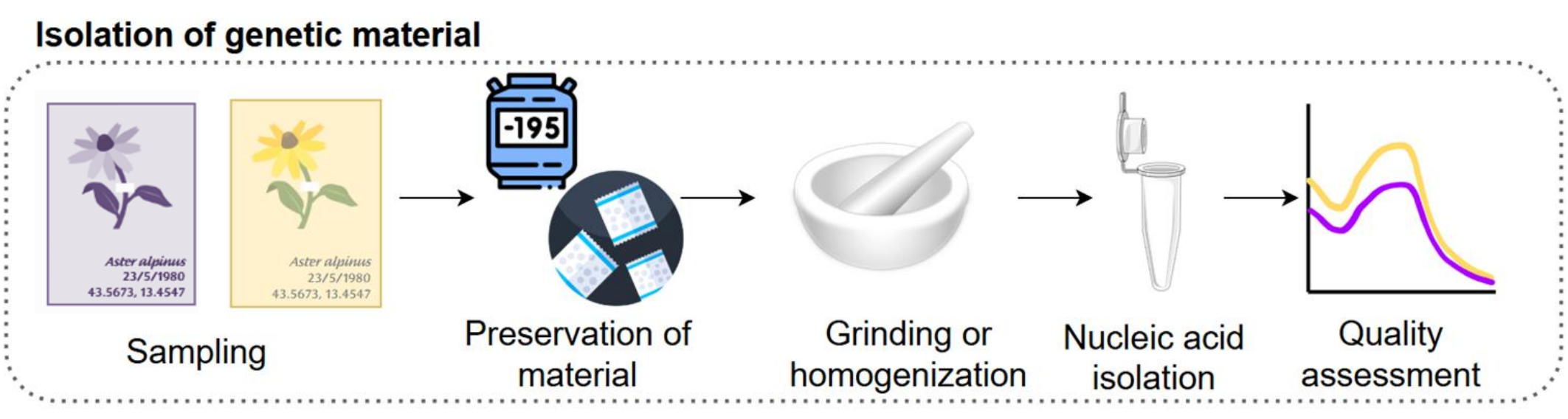
. Stadler et al., (2024). Decoding Genomes: From Sequences to Phylodynamics. ETH Zurich

Common workflow in next generation sequencing technologies

An example for a population genomics study with Illumina sequencing

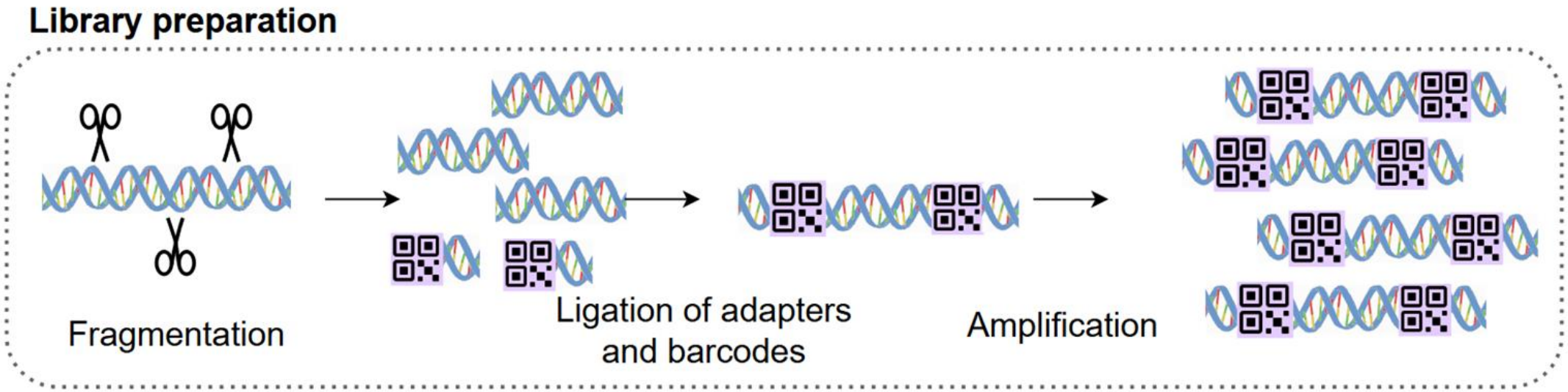


How to prepare Next Generation Sequencing experiments - 1



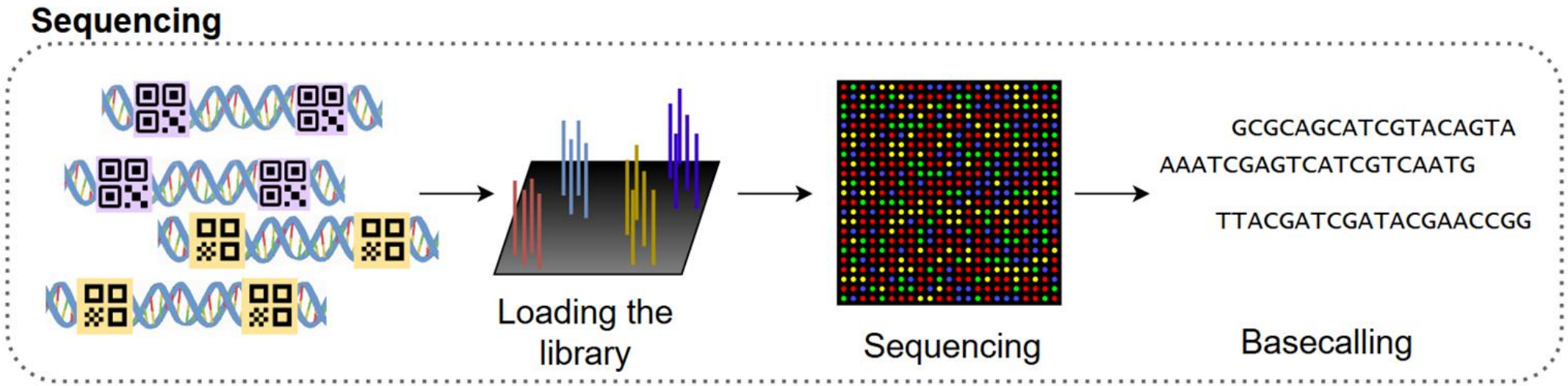
- Samples with appropriate **metadata** (coordinates, collection date etc.)
- Appropriately **preserved** (no mold, no parasites, no chemical treatments)
- Extraction of **good quality** DNA (long DNA fragments)

How to prepare Next Generation Sequencing experiments - 2



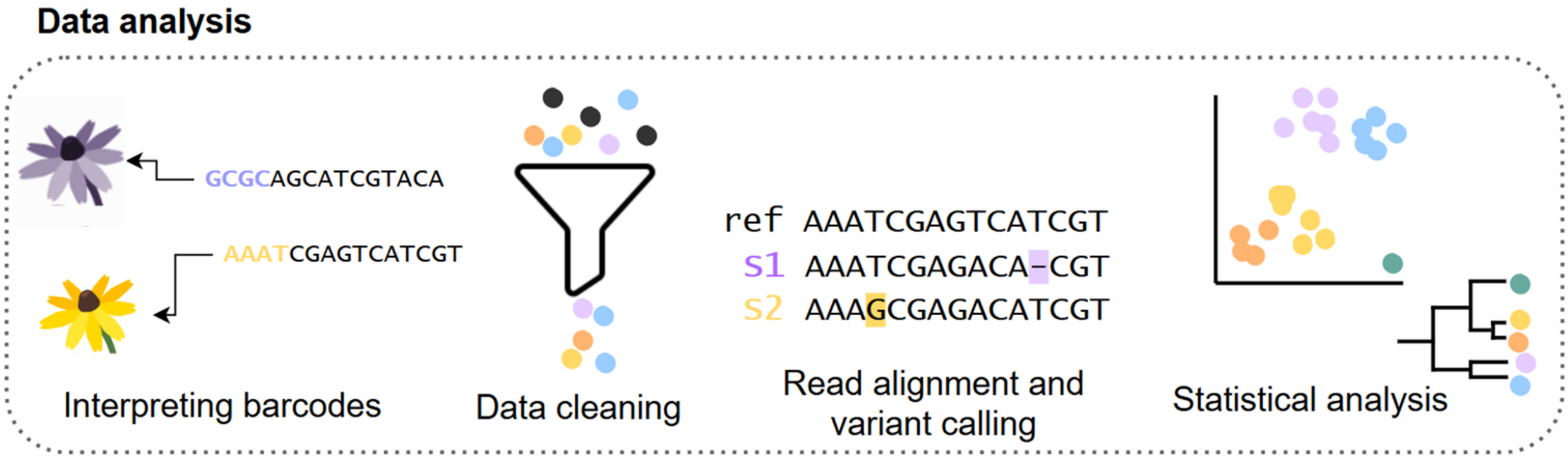
- DNA is **fragmented** or **digested** with restriction enzymes
- Specific oligonucleotides (**adapters**) are added to fragment ends for sequencing
- Optionally, **barcodes** can be added to identify samples (**multiplexing**)
- Fragments are amplified in a PCR

How to prepare Next Generation Sequencing experiments - 3



- The library is loaded in the chosen platform (i.e. Illumina)
- The signal (i.e. fluorescence) is converted into nucleotides

How to prepare Next Generation Sequencing experiments - 4



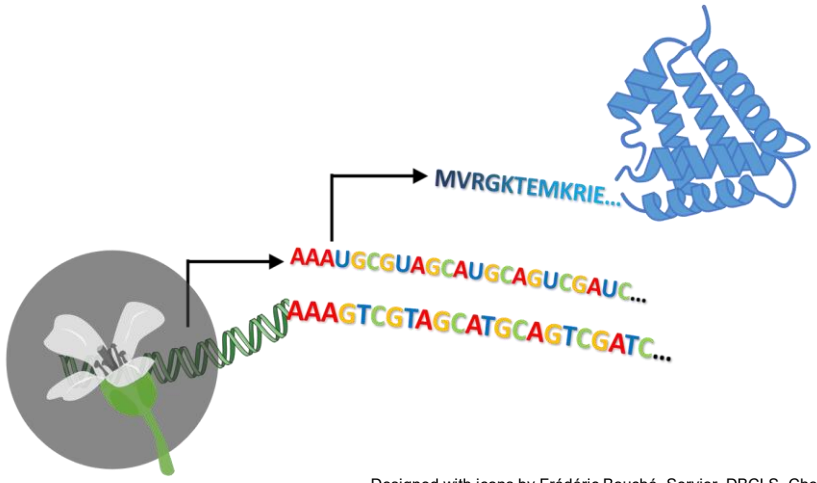
- Interpreting barcodes (**demultiplexing** → separating reads from different samples)
- Sequence **quality** control (filtering low quality sequences)
- Analysis of genetic data (phylogeny, population structure, genetic diversity…)

4- What sequencing strategies can be used for genotyping a collection?

Genotyping: It is the process of determining the **DNA sequence** profile of an individual

A reference genome can simplify genotyping process:

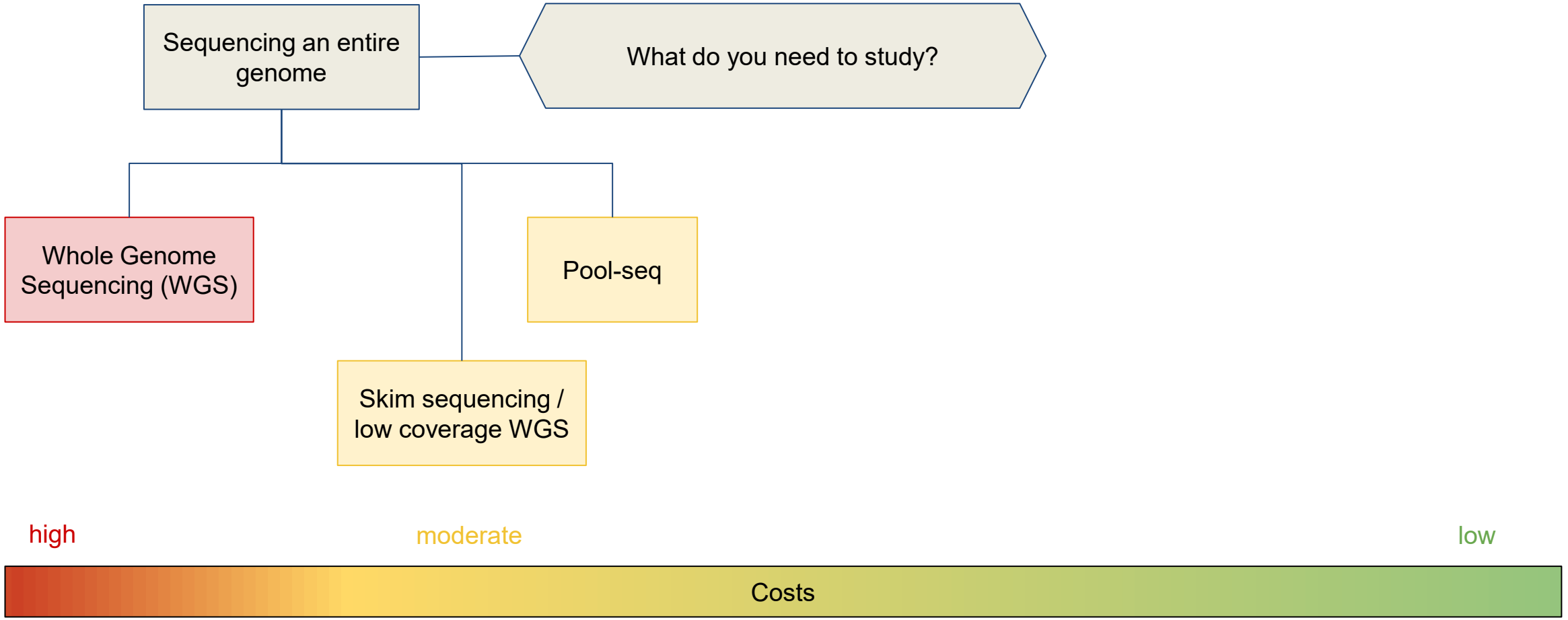
- It is the sequence of the genomic DNA of an organism or a species.
- It is annotated with the position of gene sequences and other functional elements.
- It provides a reference framework for comparing sequences of different individuals and identifying variants among individuals or species.




```
ref  AAATCGAGTCATCGT
s1   AAATCGAGACA-CGT
s2   AAAGCGAGACATCGT
```

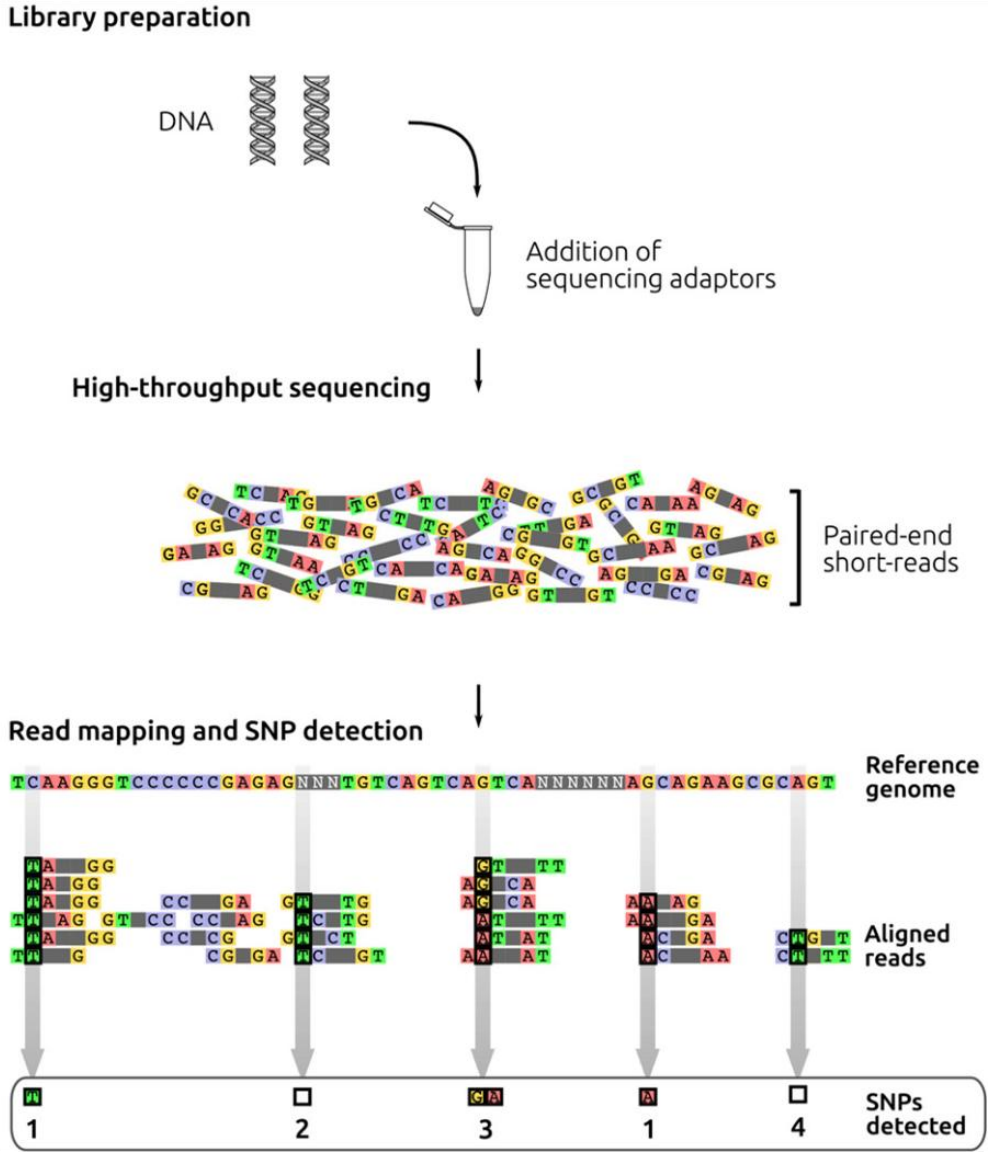
Designed with icons by Frédéric Bouché, Servier, DBCLS, Chenxin-Li (Bioicons)

NGS methodologies used for genotyping



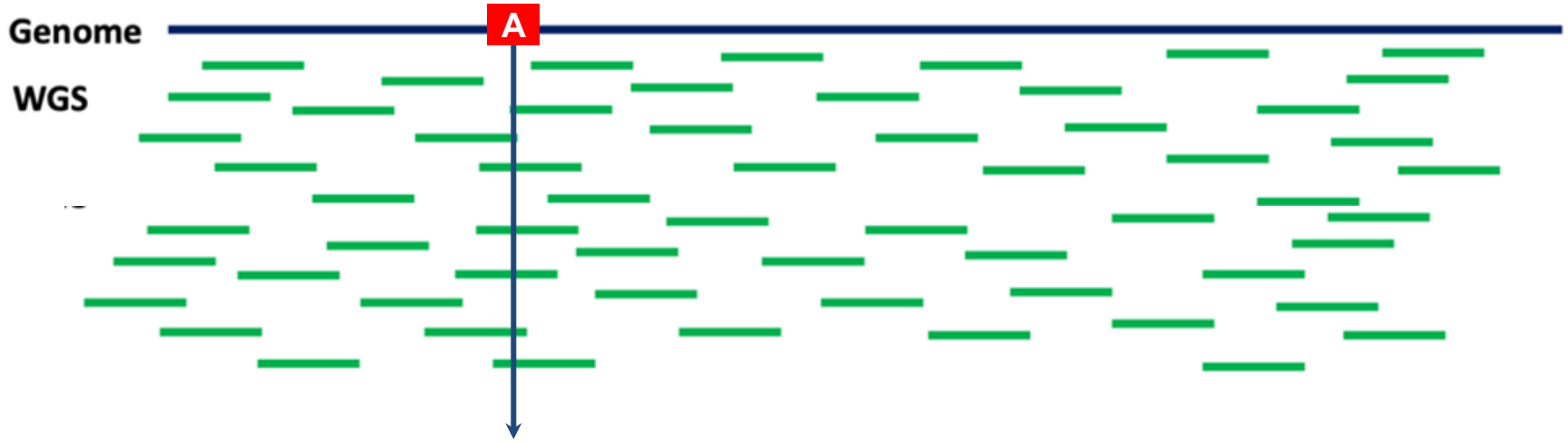
Whole genome sequencing

- Known as WGS or WGR (re-sequencing).
- The entire genome of all samples is sequenced (with short or long reads).
- Reads are compared to reference genome.
- Individual differences between samples are detected.
- This provides the most complete genetic information on the studied samples.
- Highest cost.



Reused with permission from Fuentes-Pardo, A. P., and Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Molecular Ecology *26:5369–5406.*

Whole genome sequencing: the “depth” problem

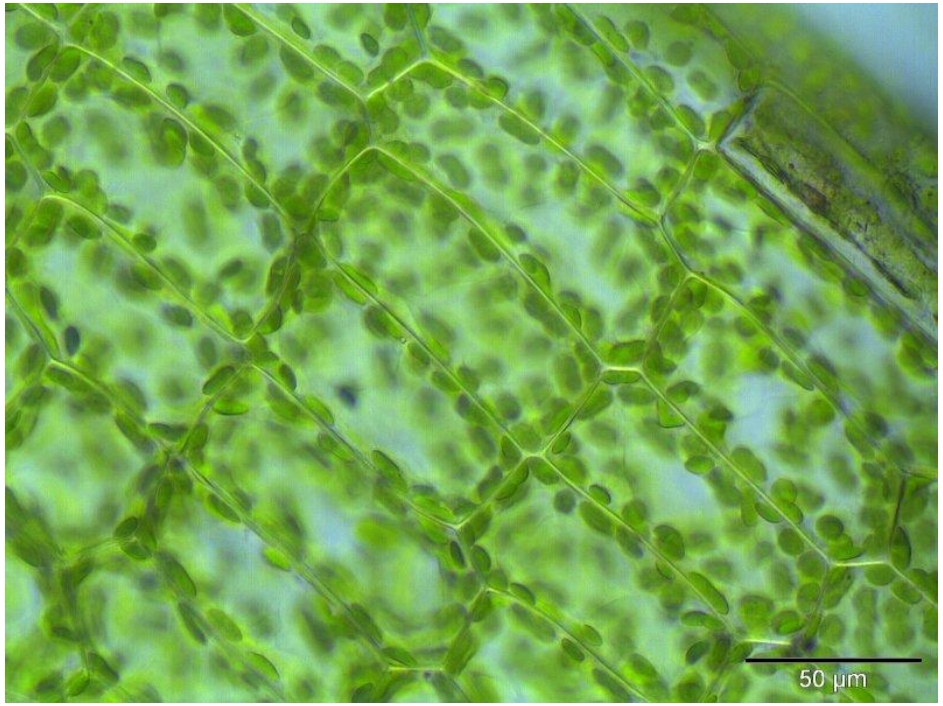


99% of reads have a “A” (adenine)
1% of the reads have a “G” (guanine)
→ A is the correct genotype

A high **depth** of sequencing (=avg. read number per genomic position) is needed for **accurate** genotyping.
This increases the **costs and computational resources** needed for WGS.

Skim sequencing / low coverage Whole Genome Sequencing (WGS)

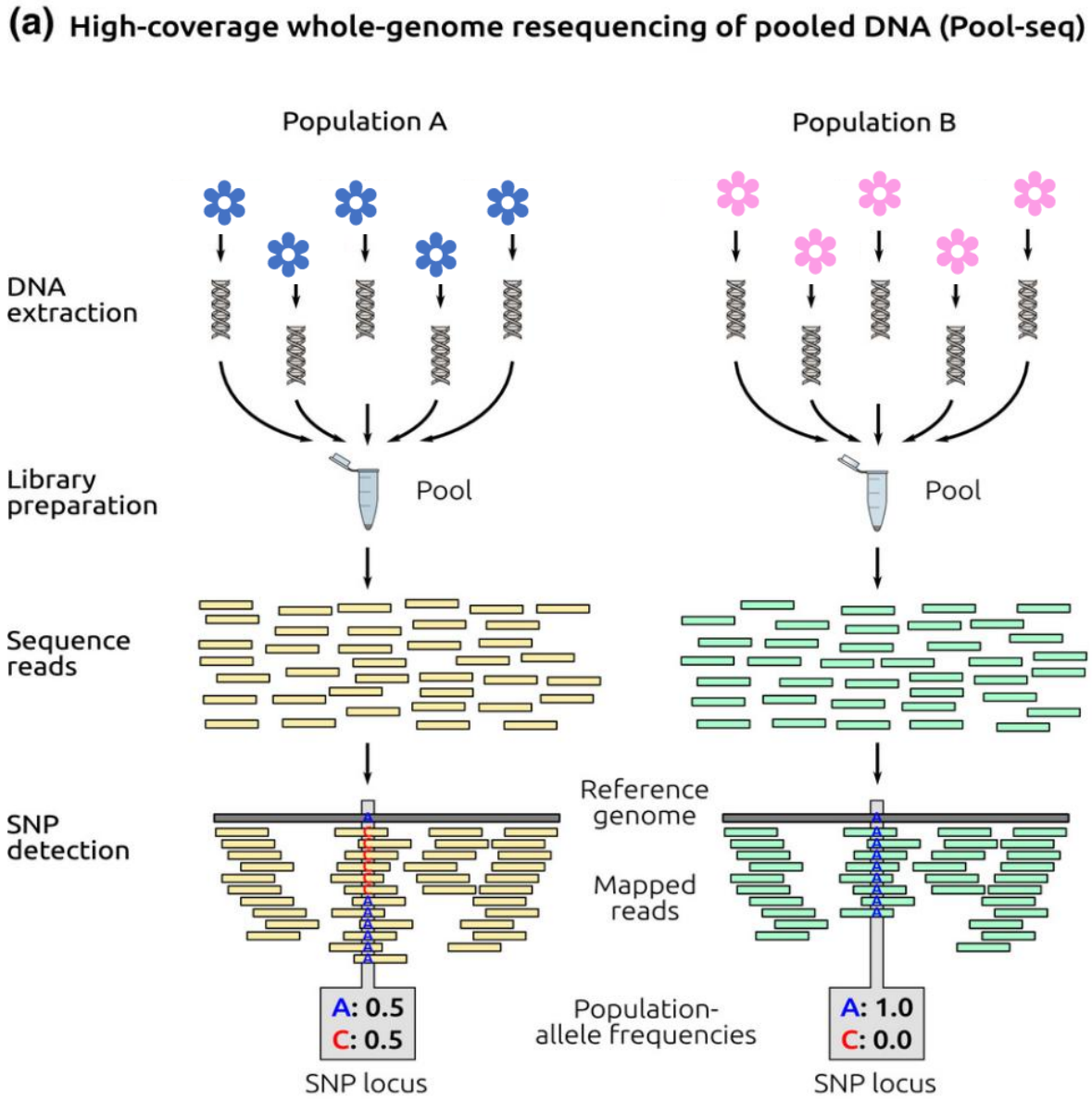
- Some DNA sequences are naturally present in many copies in the cell:
  - Mitochondria
  - Chloroplasts
  - High copy number sequences
- By sequencing at a low depth repeated sequences have still a high coverage.
- Low cost and easy to analyze without prior information.
- Limited to mitochondrial and chloroplast markers and a few other genes.



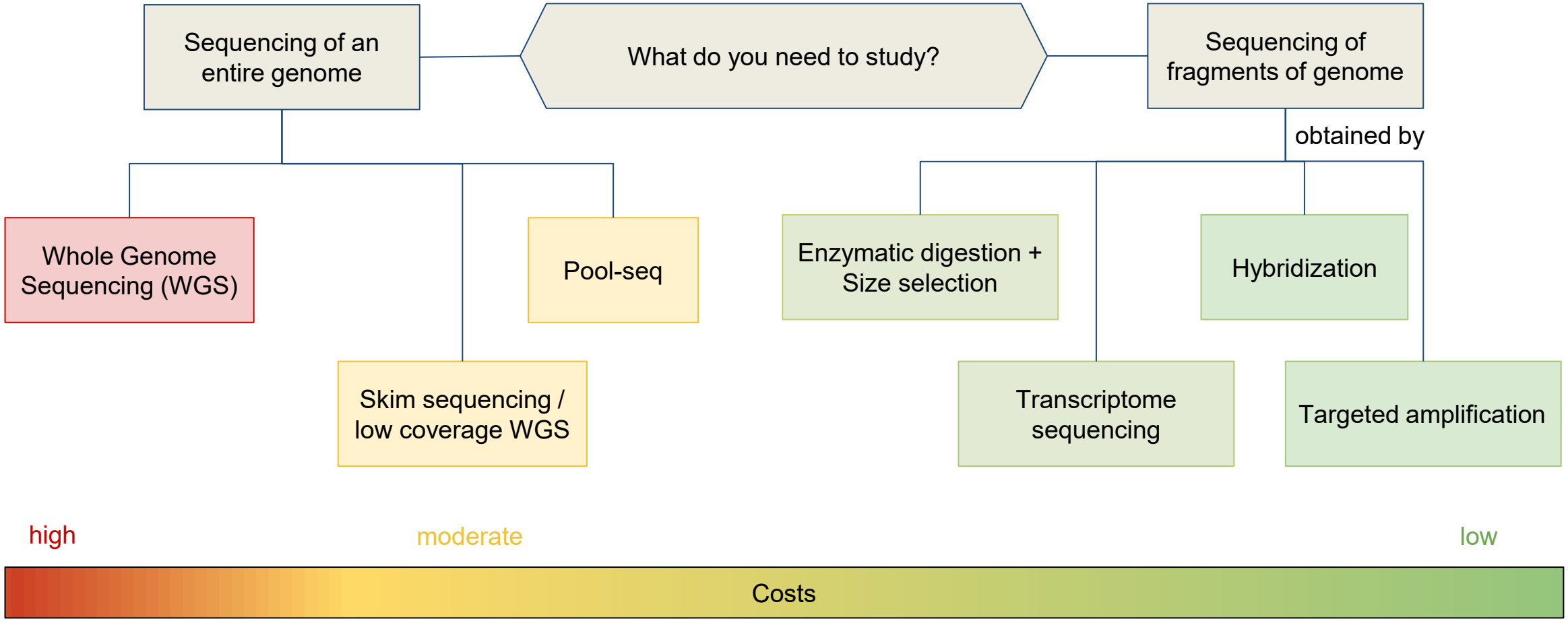
Kristian Peters -- Fabelfroh 09:12, 28 February 2007 (UTC), CC BY-SA 3.0 <http://creativecommons.org/licenses/by-sa/3.0/>, via Wikimedia Commons

Pool-seq

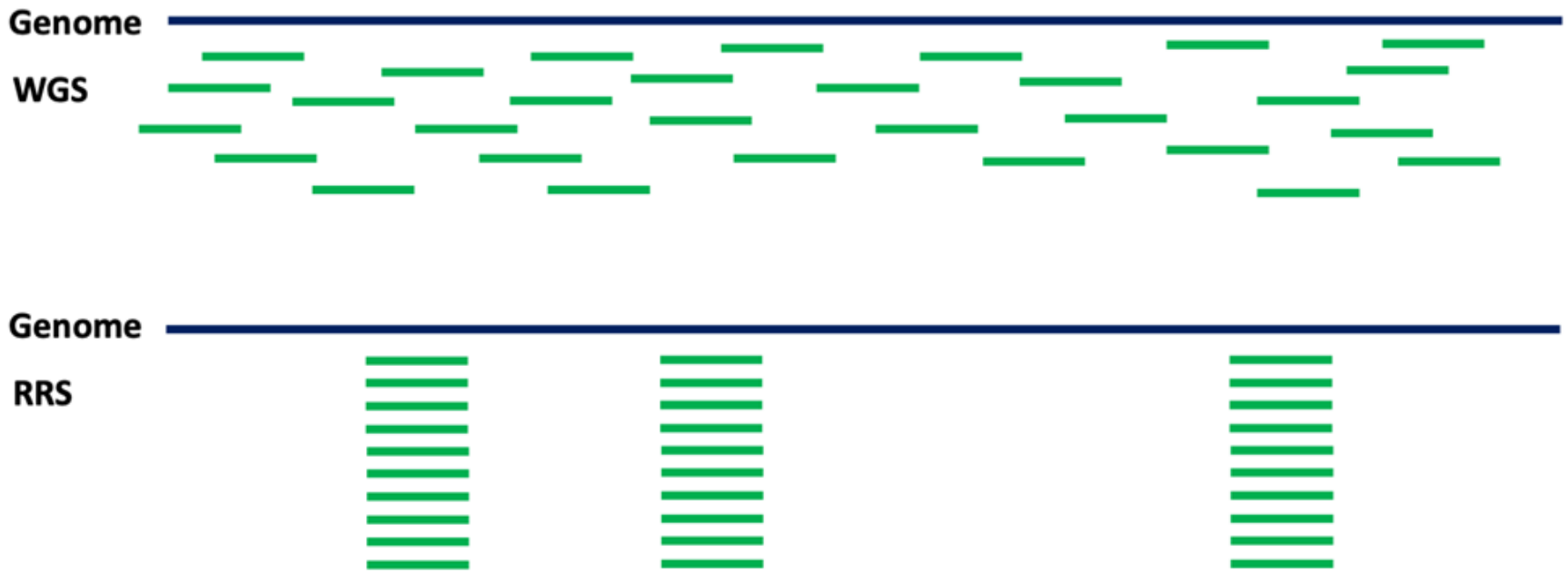
- Sequence a pool of individuals with high depth.
- Lose information on the genotype of the single individual.
- Accurate information on **allelic frequencies** in the pools.
- Can be biased by uneven amounts of DNA in the pools.
- Rare variants are lost.



Next Generation Sequencing methodologies used for genotyping



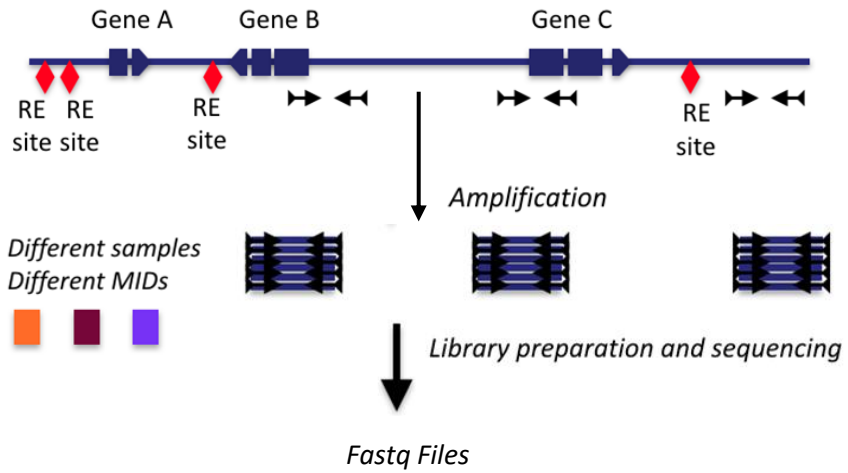
Cost-effective genotyping with Reduced Representation Sequencing (RRS)



A high depth of sequencing is needed for accurate genotyping.
With RRS the cost is reduced by selecting DNA fragments that represent a fraction (1-5%) of the genome for sequencing.
Some RRS approaches can be applied even **without** a reference genome.

GENOTYPING APPROACHES: Reduced Representation Approaches

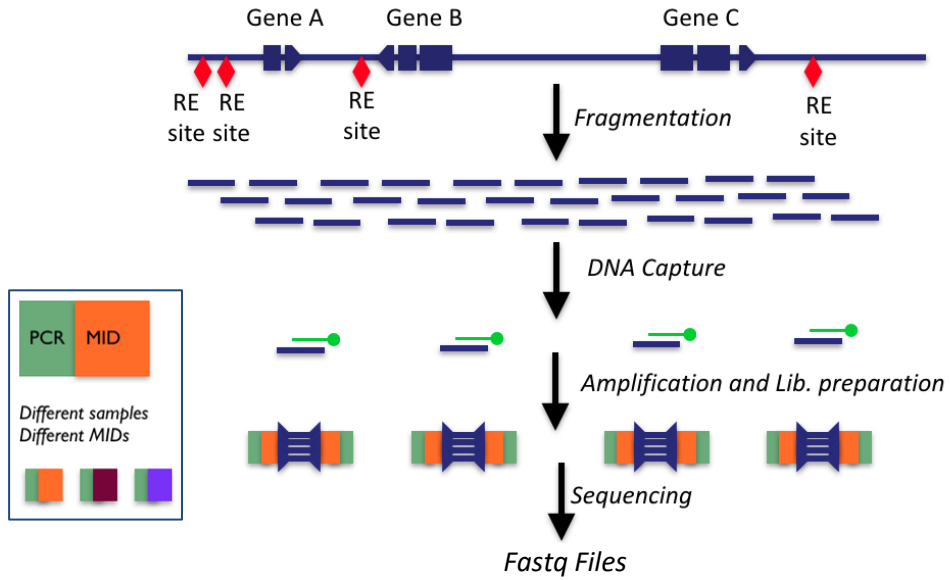
1. Targeted amplification



Specific DNA fragments are selected and amplified with a high number of primers targeting different genomic regions.
Examples:

- TrueSeq Custom Amplicon
- K-Seq

2. Hybridization



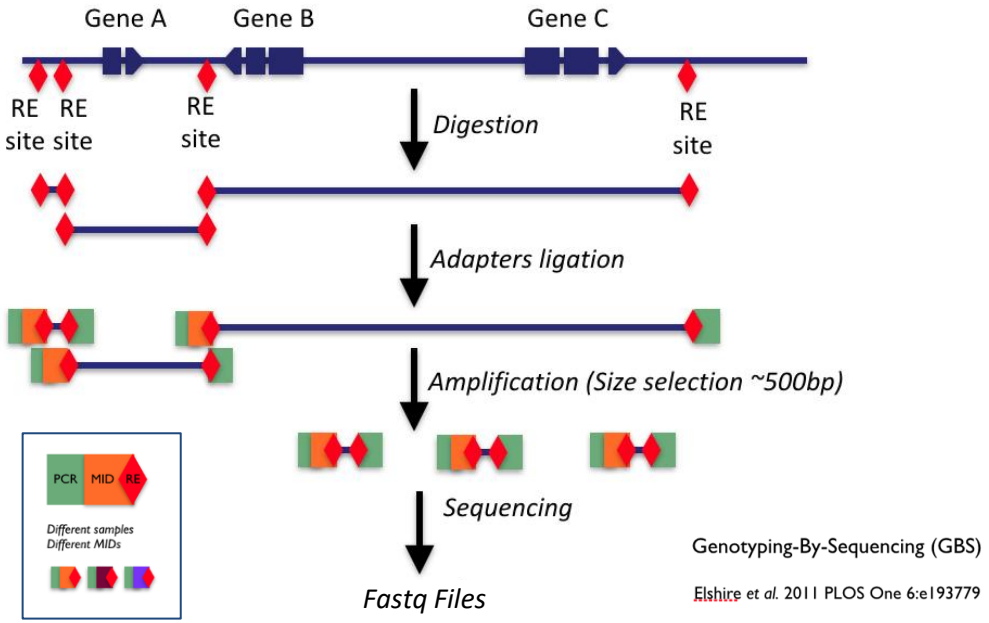
Specific DNA fragments are selected with the use of probes that hybridize to the region of interest. Fragments bound to the probes are recovered and amplified.
Examples:

- Sequence capture

Adapted from lecture slides by Aureliano Bombarely

GENOTYPING APPROACHES: Reduced Representation Approaches

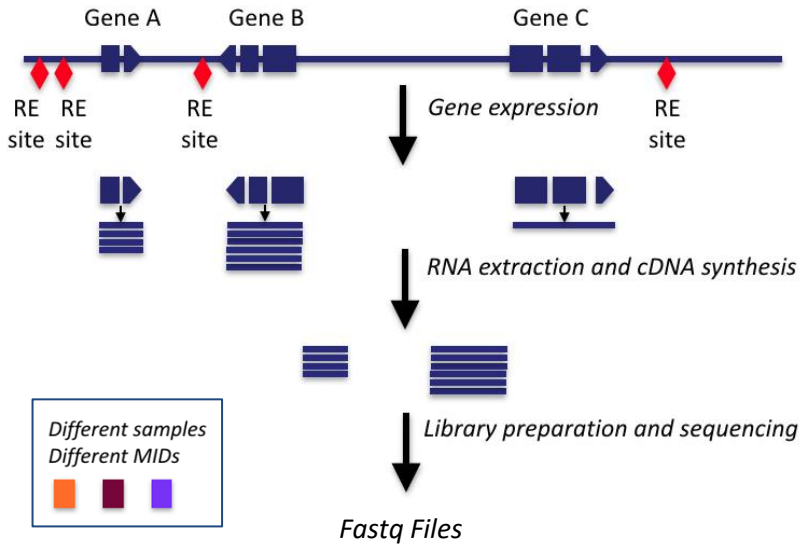
3. Enzymatic Digestion + Size selection



Specific DNA primers are selected by digesting the genomic DNA with one or more restriction enzymes and selecting fragments flanked by 2 restriction sites within a certain length range.
Examples:

- Genotyping by Sequencing (GBS)
- Rad-Seq

4. Transcriptome sequencing



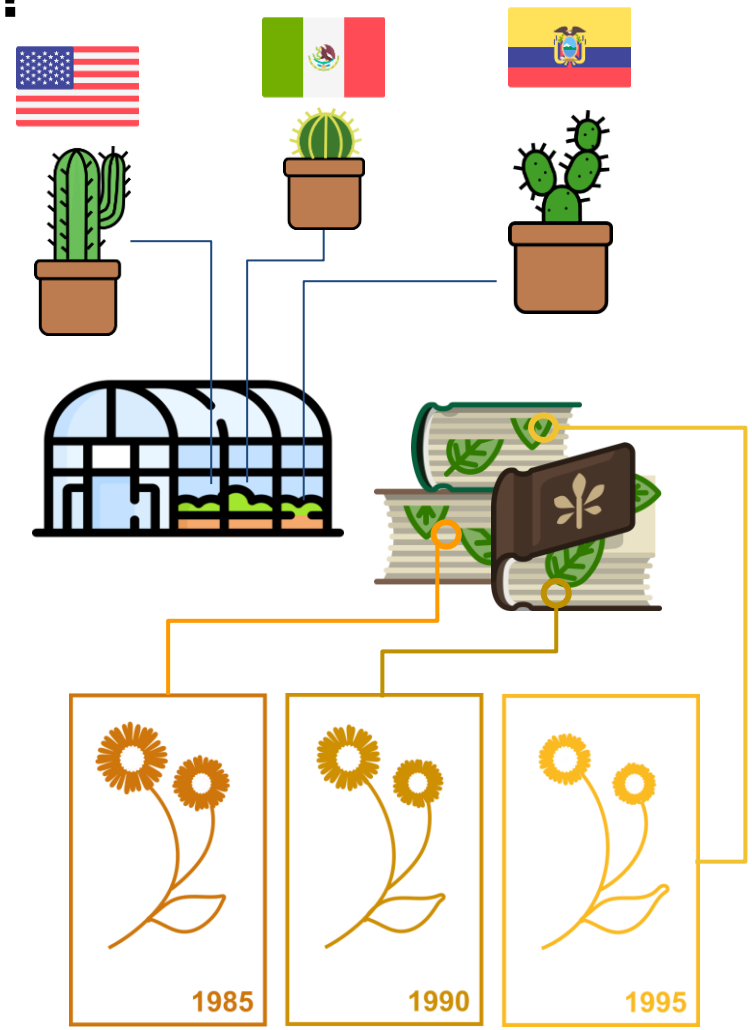
The RNA from one or more tissue types is extracted and sequenced. Only expressed genes are sequenced.
Examples:

- RNA-Seq
- Iso-Seq

Adapted from lecture slides by Aureliano Bombarely

5- What unique opportunities do botanical collections provide for the study of genetic diversity?

- Provide a picture of plant variation patterns in specific ranges at the time of collection (herbaria)
- Can also be used to obtain large amounts of fresh plant material (for example, by growing new plants from propagules in a nursery or from seeds from a seed bank)
- Can be prepared for specific studies to study gene expression under manipulated environmental conditions (living collections for research)
- Can be used to perform hybridization experiments by hand pollination (germplasms for breeding)
- Can be used to study sexual reproduction (research living collections)



Designed with icons by Freepik , Yumminky (Flaticon)

Key aspects for developing a collection

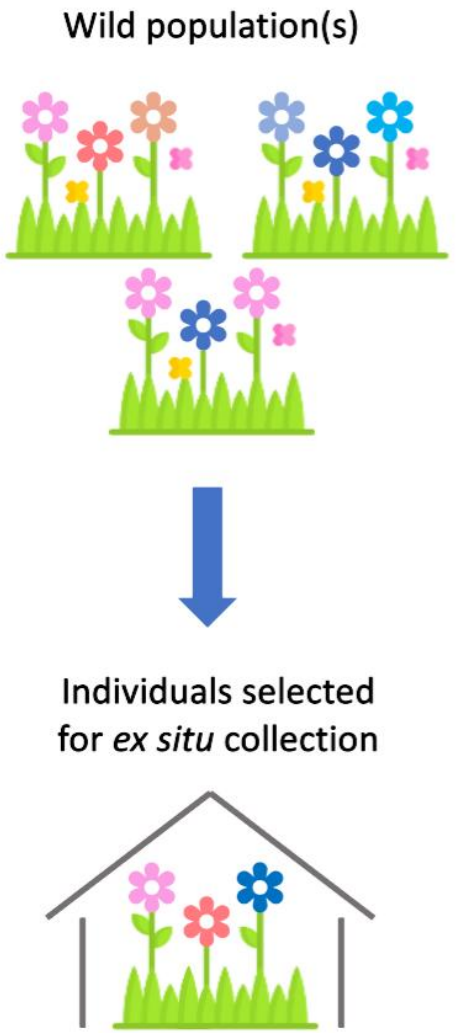
1. Collecting natural genetic variability

→ Sampling must reflect the genetic diversity of wild populations.

→ Sampling strategy should consider biological traits and evolutionary processes.

→ Use phylogeographic studies, spatial statistics, and existing guidelines.

→ Always comply with international legal and conservation frameworks.



Designed with icons by Freepik (Flaticon)

Key aspects for maintaining a living collection

1. Collecting natural genetic variability

→ Sampling must reflect the genetic diversity of wild populations.

→ Sampling strategy should consider biological traits and evolutionary processes.

→ Use phylogeographic studies, spatial statistics, and existing guidelines.

→ Always comply with international legal and conservation frameworks.

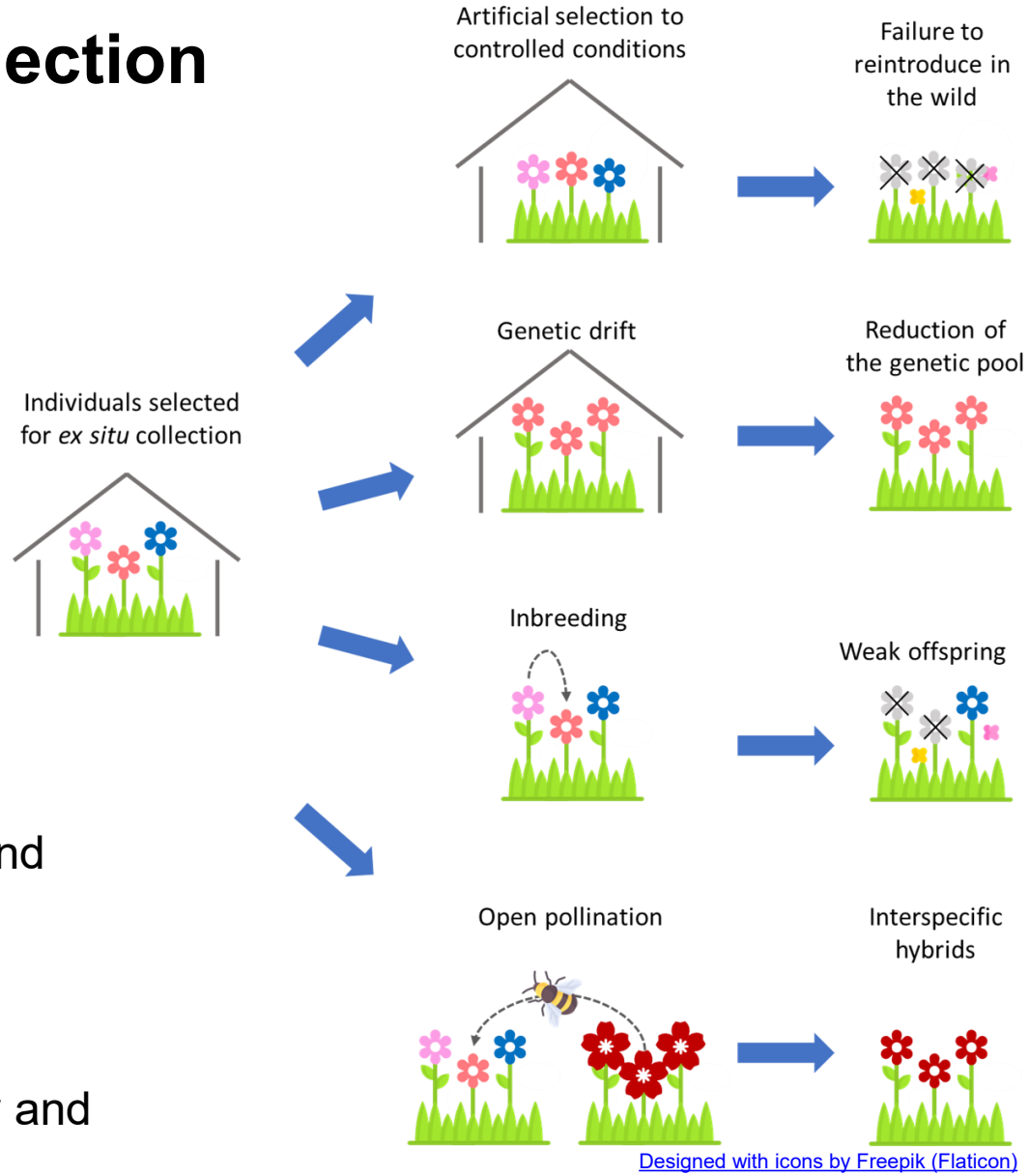
2. Maintaining genetic diversity over time

→ Living collections evolve: monitor genetic drift, inbreeding, and selection.

→ Avoid unintentional selection due to *ex situ* conditions.

→ Prevent hybridization with related species grown nearby.

→ Exchange material between institutions to preserve diversity and prevent inbreeding



Designed with icons by Freepik (Flaticon)

Challenges in plant genomics

Plant genomes present unique challenges when compared to human or animal genomes.

DNA extraction

Pinus sylvestris, by Georgi Kunev (Wikimedia Commons)

Difficult in tissues that produce specialized metabolites. Can be solved by:

- using young leaves
- optimizing protocols

Challenges in plant genomics

Plant genomes present unique challenges when compared to human or animal genomes.

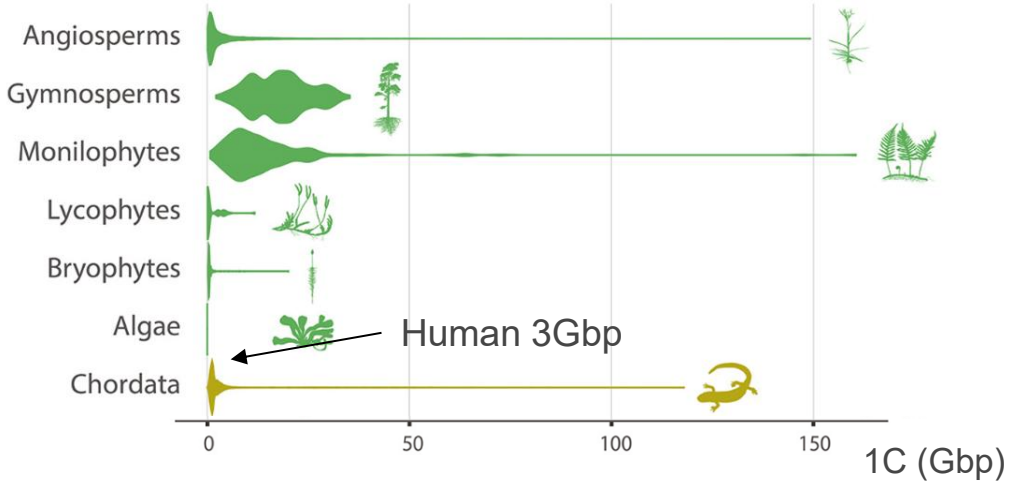
DNA extraction

Pinus sylvestris, by Georgi Kunev (Wikimedia Commons)

Difficult in tissues that produce specialized metabolites. Can be solved by:

- using young leaves
- optimizing protocols

Genome size



Can be very large, increasing the cost of WGS. Examples:

- *Arabidopsis thaliana* 120 Mbp
- *Zea mays* 2.4 Gbp
- *Picea abies* 20 Gbp
- *Tmesipteris oblanceolata* 160 Gbp

Reprinted from iScience, *27:109889*, Fernández et al., A 160 Gbp fork fern genome shatters size record for eukaryotes. Copyright 2024, with permission from Elsevier

Challenges in plant genomics

Plant genomes present unique challenges when compared to human or animal genomes.

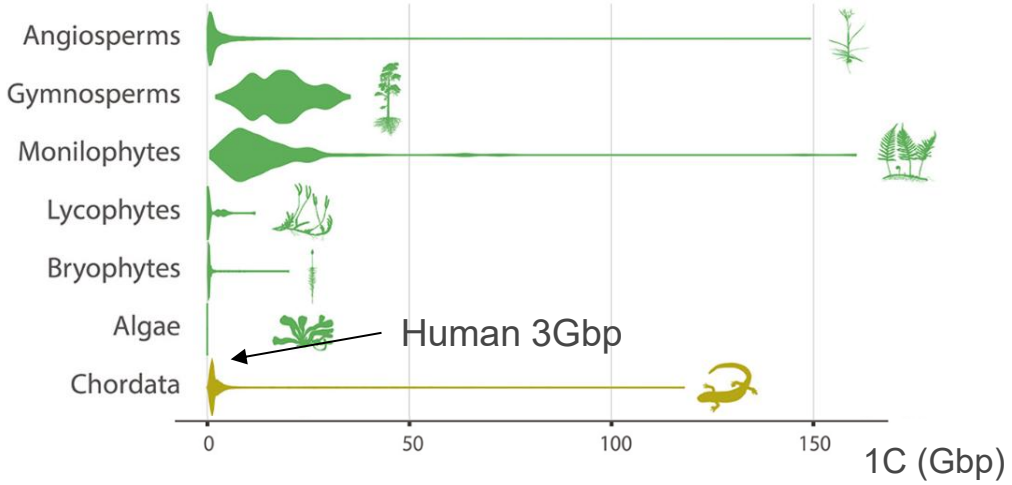
DNA extraction

Pinus sylvestris, by Georgi Kunev (Wikimedia Commons)

Difficult in tissues that produce specialized metabolites. Can be solved by:

- using young leaves
- optimizing protocols

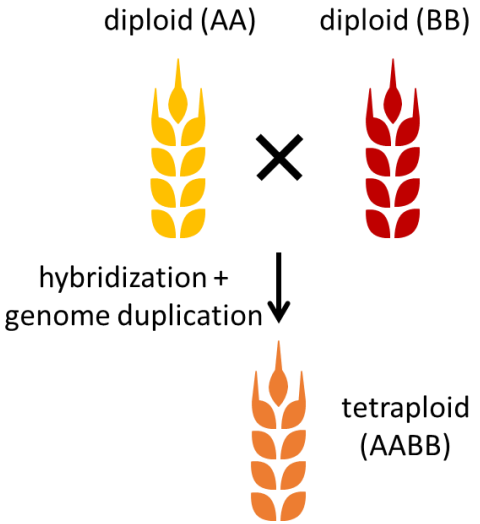
Genome size



Can be very large, increasing the cost of WGS. Examples:

- *Arabidopsis thaliana* 120 Mbp
- *Zea mays* 2.4 Gbp
- *Picea abies* 20 Gbp
- *Tmesipteris oblanceolata* 160 Gbp

Polyploidy



> 2 sets of homologous chromosomes, originating from:

- the same species
- hybridization

Complicates data analysis.

Herbariomics: opportunities & challenges

✓ Opportunities
Snapshot of the species at the moment of collection
Investigate past genetic diversity and trends throughout time
Digitization can preserve the appearance and make accession data available online
Record can be linked to data from different databases (taxonomic, genetic, …)
Data on pathogen evolution

Herbarium specimen of *H. umbellatum*, herbarium, Milan herbarium, (https://erbario.lim.unimi.it)

Herbariomics: opportunities & challenges

✓ Opportunities
Snapshot of the species at the moment of collection
Investigate past genetic diversity and trends throughout time
Digitization can preserve the appearance and make accession data available online
Record can be linked to data from different databases (taxonomic, genetic, …)
Data on pathogen evolution

Herbarium specimen of *H. umbellatum*, herbarium, Milan herbarium, (https://erbario.lim.unimi.it)

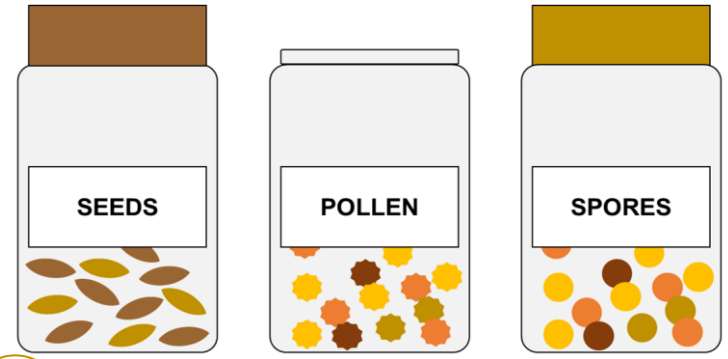
⚠ Challenges
DNA from historical samples is degraded and damaged
Specialized metabolites might be bound tightly to DNA
Cannot be used for RNA extraction
Limited and precious amount of material
Possible contaminations with fungi, molds, or material from other specimens

Living and seed collections: opportunities & challenges

✓ Opportunities
Preserve threatened species
Grow different genotypes in a uniform environment
Prepare plants for reintroduction
Provide large amount of fresh material
Seeds, pollen, and spores are preserved for breeding and germplasm conservation
Include cultivars, crops, and key resources for domestication
Collections safeguard genetic diversity and evolutionary history

Designed by Freepik

Living collections



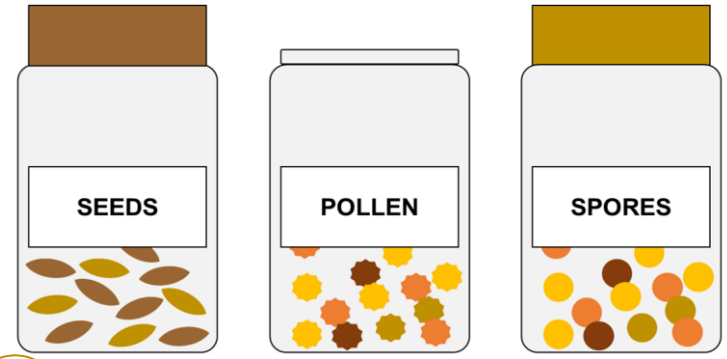
Seed, Pollen, and spores collections

Living and seed collections: opportunities & challenges

✓ Opportunities

- Preserve threatened species
- Grow different genotypes in a uniform environment
- Prepare plants for reintroduction
- Provide large amount of fresh material
- Seeds, pollen, and spores are preserved for breeding and germplasm conservation
- Include cultivars, crops, and key resources for domestication
- Collections safeguard genetic diversity and evolutionary history

Living collections



Seed, Pollen, and spores collections

⚠ Challenges

- Not originally designed for genetic studies
- Limited representation of genetic diversity
- Risk of unintentional selection in artificial environments
- Loss of seed viability over time
- Need for regular regeneration and viability testing

Genomic analysis of botanical collections

✓ Opportunities	⚠ Challenges
Decreasing sequencing costs	DNA extraction / material availability
Genomic data for targeted strategies	Need for bioinformatics training
Genetic resources for past, present and future	Must adequately represent the natural genetic diversity
High resolution phylogenetic and population analyses	Translating data into concrete conservation action

Freepik (Flaticon)