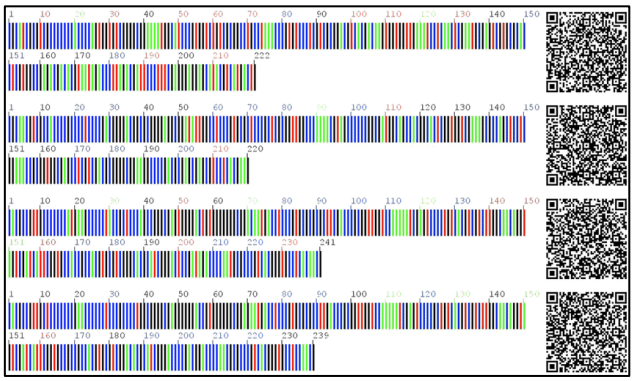
DNA Barcoding

DNA Barcoding started with the seminal work of Hebert *et al*., who demonstrated that individual species from a collection of 200 closely allied species of lepidopterans could be identified with 100% accuracy using the mitochondrial gene cytochrome *c* oxidase subunit I (COI). Barcoding is now a well-established technique for species identification in animals. In DNA barcoding, the unique nucleotide sequence patterns of small DNA fragments (400–800 bp) are used as specific reference collections to identify specimens and to discover overlooked species. Thus, the initial goal of the DNA barcoding process is to construct on-line libraries of barcode sequences for all known species that can serve as a **standard to which DNA barcodes of any identified or unidentified specimens can be matched**. DNA barcoding, can provide the taxonomists, conservationists and others who need the identification of species, a cost-effective and efficient tool, much as a barcode that identifies supermarket products.

In order to promote use of DNA barcoding for all eukaryotic life in this planet, a Consortium for the Barcode of Life (CBOL) was established in May 2004, which currently includes more than 120 organizations from 45 nations. With the support of CBOL, the effort of DNA barcoding has been slowly progressing with controversies and intense debates.

[A brief history of the Barcode of Life initiative is available at http://www.dnabarcodes.org/pa/ge/history_of_boli. Other useful information on DNA barcoding is also available at www.barcodinglife.org, http://barcoding.si.edu/, http://www.dnabarcoding.ca, http://www.kew.org/barcoding and http://www.ibolproject.org.]

BASIC FEATURES OF BARCODING SEQUENCES



The most important characteristic features of a DNA barcode are its universality, specificity on variation and easiness on employment. This means that the gene segment used as a barcode should be **suitable for a wide range of taxa**, should have **high variation between species** but should be **conserved within the species**, so that the intra-specific variation will be insignificant. Consequently, **an ideal DNA barcode should also be routinely retrievable with a single primer pair**.

I. NUCLEAR GENOME SEQUENCES

Till date internal transcribed spacer (ITS) regions of the ribosomal DNA (rDNA) are the only nuclear DNA that have been tested for suitability as barcodes in plants.

Internal transcribed spacer regions of nuclear ribosomal cistron

The rDNA cistron is a multigene family encoding the nucleic acid core of the ribosome. Within the cell, the rDNA is arranged as tandemly repeated units containing *18S*, *5.8S*, *28S* coding regions and two internal transcribed spacers (*ITS1* and *ITS2*) present on either side of *5.8S* region.

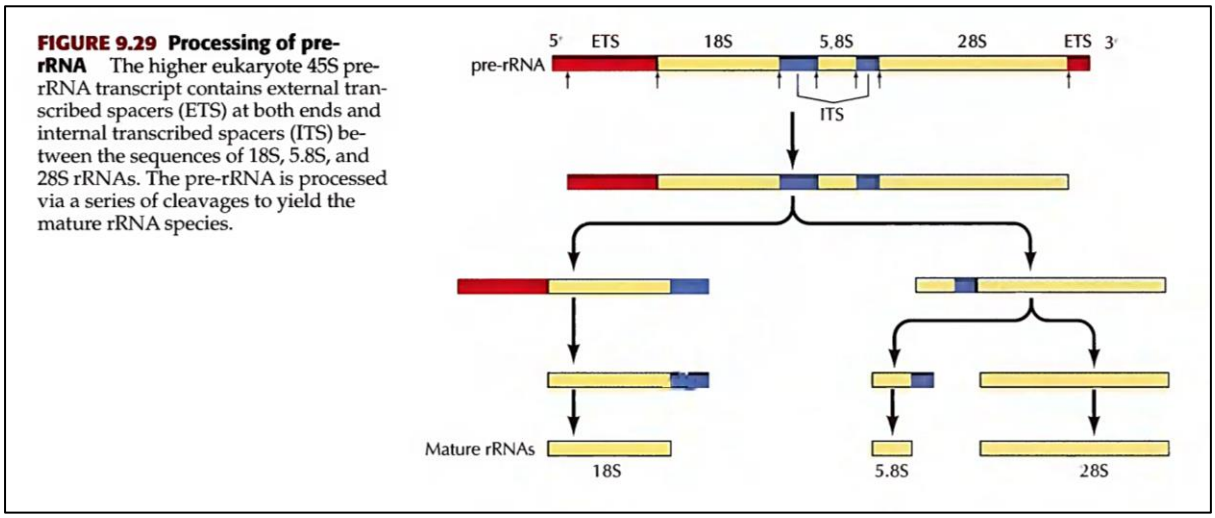
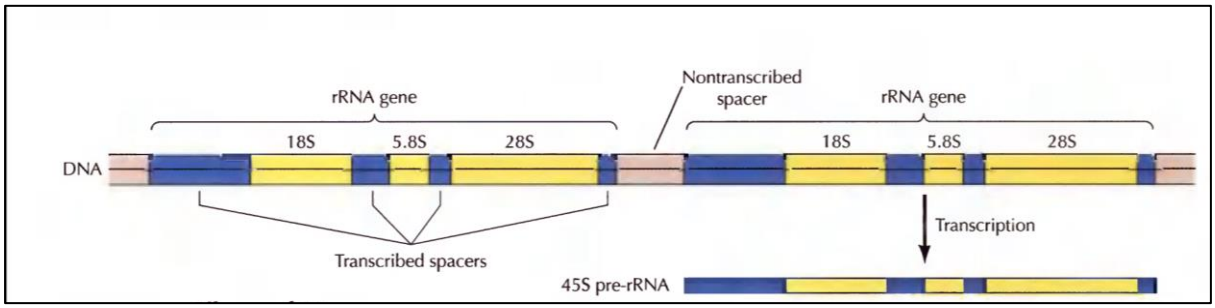


FIGURE 9.29 Processing of pre-rRNA The higher eukaryote 45S pre-rRNA transcript contains external transcribed spacers (ETS) at both ends and internal transcribed spacers (ITS) between the sequences of 18S, 5.8S, and 28S rRNAs. The pre-rRNA is processed via a series of cleavages to yield the mature rRNA species.



nrITS (nuclear region ITS) is considered as one of the most useful phylogenetic markers for both plants and animals, because of its ubiquitous nature, biparental inheritance, and comparatively higher evolutionary changes due to less functional constrains. Likewise, species-level discrimination and technical ease have also contributed to its wider acceptability as a powerful phylogenetic marker.

Another advantage is that the ITS1 and ITS2 regions can be PCR-amplified separately in the conserved coding genes. This facilitates easy amplification of ITS even from poor quality or degraded DNA. Universal primers are also available for amplification of ITS1 and ITS2 regions.

Some of the recent reports in tree plants and asexually propagated plants revealed the presence of some degree of intra-individual variations among the copies of ITS1 and ITS2 sequences. Various reasons, such as, hybridization, recombination among copies, high mutation rate are considered to be the reasons for such variations.

CBOL-Plant Working Group has not regarded *nrITS* suitable for a universal plant DNA barcode, but as a **supplementary locus** for taxonomic groups.

II. CHLOROPLAST GENES

Regions of chloroplast genes, including rbcL—RuBisCo (Ribulose-1,5-bisphosphate carboxylase oxygenase) large subunit—and matK—maturase K—are used for barcoding plants. The most abundant protein on earth, RuBisCo catalyzes the first step of carbon fixation, while maturase K encodes for a protein that assists with RNA editing.

Chloroplast genes could be considered as analogous to the mitochondrial gene that has been used for DNA barcoding in animals. However, compared to mtDNA genes in animals, chloroplast genes in plants have slower rate of evolution; therefore, finding suitable gene sequences with sufficient species discriminatory power is a great challenge.

Chloroplast genome of higher plants is a circular structure with a size of 120–160 k bp. Architecture of the chloroplast genomes is represented by a large and a small single-copy region (LSC and SSC) intervened by two copies of a large inverted repeat (Ira and Irb).

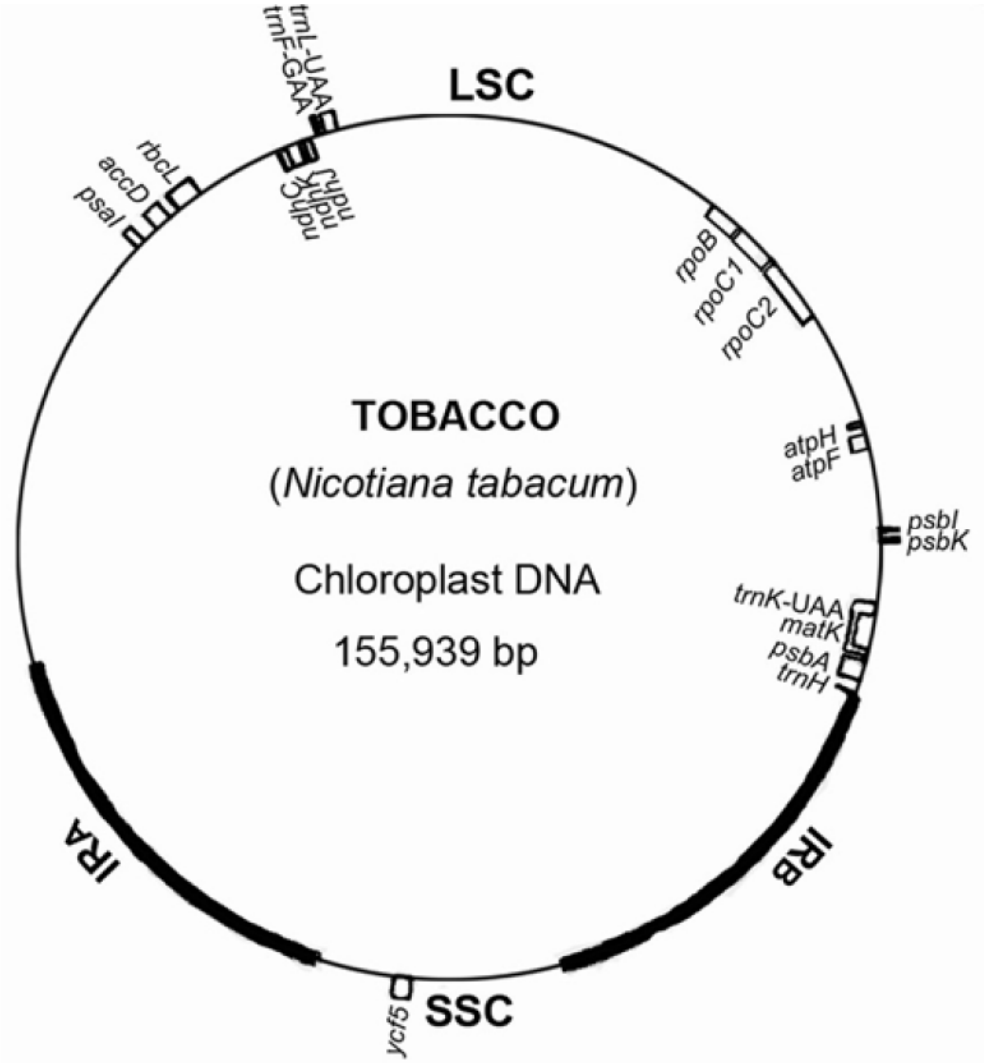


Figure 1. Structural organization of the chloroplast genome of *Nicotiana tabacum*. Locations of the seven loci proposed by different investigating group as potential candidates for barcoding in plants have been shown as described by Shinozaki *et al.*[36]. Genes shown outside the circle are encoded on the A strand and transcribed counter-clockwise. Genes shown inside are encoded on the B strand and transcribed clockwise. LSC, Large single copy region; IR, Inverted repeat; SSC, Small single copy region.

rbcL gene sequence

Among the plastid genes, *rbcL* is the best characterized gene sequence. Therefore, most of the investigating groups tested its suitability in barcoding. It encodes the large subunit of rubilose-1,5-bisphosphate carboxylase/oxygenase (RUBISCO). As RUBISCO is a critical photosynthetic enzyme, *rbcL* was the first gene that was sequenced from the plants.

Most of the phylogenetic studies suggest that *rbcL* is best suited to reconstruct the relationships down to the generic levels, but is not useful for specific levels. In order to obtain enough species discrimination, the entire ~ 1430 bp needs to be sequenced, which

acts as a limiting factor for its use as a barcoding sequence because an ideal DNA barcoding region should be short enough to amplify from degraded DNA and analysed via single-pass sequencing.

Primers for PCR amplification and sequencing for such short sequence within the *rbcL* gene have been developed accordingly for most of the taxa. The CBOL–Plant Working Group has recently recognized *rbcL* as one of the most potential gene sequences for DNA barcoding in plants. *rbcL* should be used in conjunction with other markers. It has been used in combination with *matK*.

matK gene sequence

Among the chloroplast genes, *matK* is one of the most rapidly evolving genes. It has a length of about 1550 bp and encodes the enzyme maturase. Since *matK* is embedded in the lysine gene *trnK,* it can be easily PCR-amplified with a primer set designed from the conserved regions of the genes *trnK*, *rps16* and *psbA*. Phylogenetically, the rate of evolution of *matK* was found suitable for resolving intergeneric as well as interspecies relationships in many angiosperms.

Ford *et al*., after testing *matK* along with 11 other cpDNA loci in 98 land plant taxa, **proposed a combination of *rpoC1* + *rpoB* + *matK* as the most promising combination for barcoding of land plants**. Starr *et al*., advocated the use of *matK* alone as a universal barcode for land plants.

rpoB and rpoC1 gene sequences

Genes *rpoB*, *rpoCl* and *rpoC2* encode three out the four subunits of the chloroplast RNA polymerase. *rpoB* has been considered as the core gene for phylogenetic analyses and identification of bacteria, especially when studying closely related isolates. Together with the 16S rRNA gene, *rpoB* helps delineate new bacterial species and refine bacterial community analysis.

Logacheva *et al.,* also found that *rpoA*, *rpoB*, *rpoC1* and *rpoC2* as a group, are ideal for phylogenetic studies, but *rpoB* and *rpoC1* alone may not generate good results. These genes have been proposed for barcoding either individually or in combination by various groups. PCR amplification of *rpoB* failed in *Araucaria*, *Ephedra*, *Equisetum*, *Isoetes*, *Lycopodium* and *Mannia. rpoC1* has been found **highly useful for barcoding the bryophytes** (mosses).

accD gene sequence

The plastid *accD* gene encodes the β -carboxyl transferase subunit of **acetyl-CoA carboxylase** and is present in most flowering plants, except in grasses. The *accD* gene sequence has been used for several phylogenetic studies in plants.

ycf5 gene sequence

ycf5 is the only gene from the small single-copy region being seriously studied for its suitability in DNA barcoding. *ycf 5* encodes a protein containing 313 amino acids. This gene is conserved across all land plants and has been tested for its suitability for DNA barcoding by several groups.

ndhJ gene sequence

The plastid *ndh* gene complex, identified originally from the tobacco plastid genomes and liverwort, codes for subunits of a functional respiratory protein complex of size ~ 550 kDa within the mature chloroplast. *ndhJ* encodes NADH dehydrogenase 30 kDa subunit.

ndhJ is **absent** in *Pinus* and *Cuscuta*. Further, it was reported that all *ndh* genes are absent across Gnetales and Pinaceae.

atpF–atpH intergenic sequence

The genes *atpF* and *atpH* encode ATP synthase subunits CFO I and CFO III respectively. Testing of the intergeneric spacer between these two genes as barcode in the flora of the Kruger National Park, South Africa, revealed that PCR amplification was easier but alignment of sequences was considerably difficult due to significant length variations. Thus, it was found useful only as a supplementary locus in combination with *matK* for barcoding in plants.

The CBOL–Plant Working Group also observed its high universality but less species discriminatory power.

trnH–psbA intergenic sequence

Because of the high species discriminatory power exhibited by this small segment of DNA, Kress *et al.* proposed it along with *nrITS* for DNA barcoding in plants. The *trnH–psbA* locus has been successfully PCR amplified from a wide range of angiosperms and gymnosperms, ferns, mosses, and wild liverworts.

psbK–psbI intergenic sequences

psbK and *psbI* genes encode two low molecular mass polypeptides, **K and I respectively, for the photosystem II** and are conserved from algae to land plants. The potential of the *psbK–psbI* intergenic region as a barcode for plants was tested in the flora of the Kruger National Park, South Africa.

psbK–psbI was proposed in combination with other loci such as *matK*, *trnH–psbA* and *atpF–atpH* for barcoding of plants. The species discriminatory power of this locus was better than that of *matK*.

trnL (UAA)–trnF (GAA): genic, intron and intergenic sequences

The *trnL* (UAA)–*trnF* (GAA) locus contains the *trnL* (UAA) gene, its intron and the intergenic region between *trnL* (UAA) and *trnF* (GAA). Taberlet *et al.* employed *trnL* intron for the first time in plant systematic studies. Despite can be used as a barcode for plants, as universal primers are available.