

# 2

## Molecular Systematics\*

SERGEI A. SUBBOTIN,<sup>1\*\*</sup> LIEVEN WAEYENBERGE<sup>2</sup>  
AND MAURICE MOENS<sup>3</sup>

<sup>1</sup>*California Department of Food and Agriculture, USA and Biology Department, Ghent University, Belgium;* <sup>2</sup>*Institute for Agricultural and Fisheries Research (ILVO), Plant Sciences Unit, Crop Protection, Belgium;* <sup>3</sup>*Institute for Agricultural and Fisheries Research (ILVO), Belgium and Laboratory for Agrozoology, Ghent University, Belgium*

2.1.	Phylogenetics and Phylogenomics	41
2.2.	Species Concepts and Delimiting Species in Nematology	42
2.3.	Phylogenetics and Classification	43
2.4.	Molecular Techniques	44
2.4.1.	Protein-based techniques	44
2.4.2.	DNA-based techniques	44
2.4.2.1.	DNA extraction	45
2.4.2.2.	Polymerase chain reaction (PCR)	45
2.4.2.3.	PCR–restriction fragment length polymorphism (PCR–RFLP)	46
2.4.2.4.	Multiplex PCR	47
2.4.2.5.	Random amplified polymorphic DNA (RAPD)	47
2.4.2.6.	Amplified fragment length polymorphism (AFLP)	49
2.4.2.7.	Real-time PCR	49
2.4.2.8.	Loop-mediated isothermal amplification (LAMP)	50
2.4.2.9.	DNA hybridization arrays	51
2.4.2.10.	Sequencing of DNA	52
2.5.	Genes used for Molecular Systematics	52
2.5.1.	Nuclear ribosomal RNA genes	54
2.5.2.	Nuclear protein-coding genes	55
2.5.3.	Mitochondrial DNA	55
2.6.	Microsatellites	57
2.7.	DNA Bar Coding	57
2.8.	Phylogenetic Inference	57
2.8.1.	Alignment	58

\* A revision of Subbotin, S.A. and Moens, M. (2006) Molecular taxonomy and phylogeny. In: Perry, R.N. and Moens, M. (eds) *Plant Nematology*, 1st edn. CAB International, Wallingford, UK.

\*\* Corresponding author: sergei.subbotin@ucr.edu

2.8.2.	Methods for inferring phylogenetic trees	58
2.8.2.1.	Minimum evolution method	59
2.8.2.2.	Maximum parsimony	59
2.8.2.3.	Maximum likelihood	59
2.8.2.4.	Bayesian inference	60
2.8.2.5.	Evolutionary models	60
2.8.3.	Phylogenetic tree and tree terminology	60
2.8.4.	Evaluation of the reliability of inferred trees	61
2.8.5.	Testing of hypotheses	62
2.9.	Reconstruction of Historical Associations	62
2.10.	Databases	63
2.11.	Examples of Molecular Phylogenies	64
2.11.1.	Position of Nematoda within metazoans	64
2.11.2.	The phylum Nematoda	65
2.11.3.	The infraorder Tylenchomorpha	65
2.11.4.	Root-knot nematodes of the family Meloidogynidae	67
2.11.5.	Cyst nematodes of the family Heteroderidae	68
2.11.6.	Stem and gall-forming nematodes of the family Anguinidae	68
2.11.7.	Needle nematodes of the family Longidoridae	68
2.11.8.	Root-lesion nematodes of the family Pratylenchidae	70
2.11.9.	Pinewood nematode and other <i>Bursaphelenchus</i> species	71

## 2.1. Phylogenetics and Phylogenomics

The tasks of systematics are: (i) to name, identify and catalogue organisms (taxonomy); (ii) to discover the ancestral relationships among organisms (phylogenetics); and (iii) to organize information about the diversity of organisms into a hierarchical system (classification). Molecular systematics is the application of knowledge of genome information, especially sequence and structure of DNA, RNA molecules and amino acid chains, for addressing questions regarding the phylogeny and taxonomy of organisms.

There are several reasons why molecular data are more suitable for phylogenetic studies than morphological ones. First, DNA and protein sequences are strictly heritable entities, whereas morphological characters can be influenced by various environmental factors. Second, the interpretation of molecular characters, such as the assessment of homologies, is generally easier than that of morphological characters. Third, molecular characteristics generally evolve much more regularly than morphological ones and, therefore, can provide a clearer picture of relationships. Fourth, molecular characters are much more abundant than morphological features, and many can be generated in a relatively short period of time. Various preserved, deformed and partly degraded materials can often be used for molecular studies. Using standard protocols and commercial kits, sequence information of certain genes or DNA fragments can be obtained from a single nematode or even only a part of one. Using specific primers, nematode DNA can be amplified from soil or plant extracts. Moreover, specialized methods enable the extraction of short DNA fragments from long-time preserved, formalin-fixed and glycerine-embedded specimens. Recent achievements in molecular biology and the wide application of molecular techniques have revolutionized our knowledge in taxonomy and phylogeny of nematodes.

The use of such techniques is becoming routine in nematology (Jones *et al.*, 1997; Powers, 2004; Blok, 2005; Perry *et al.*, 2007; Subbotin *et al.*, 2010a,b).

Phylogenetics compares and analyses single or a few genes. However, molecular systematics enters a new era in which many thousands of nucleotides and whole genomes can be obtained inexpensively and in a relatively short period of time. The approach that involves genome data in evolutionary reconstruction is called **phylogenomics**.

## 2.2. Species Concepts and Delimiting Species in Nematology

There has been considerable debate concerning the definition of a species. Species were at first merely taxonomic units, i.e. the named categories to which Linnaeus and other taxonomists of the 18th century assigned organisms, largely on the basis of appearance. According to the **typological species concept**, the species is considered a community of specimens described by characteristic features of its type specimen. In the early 20th century, taxonomists had accumulated a great deal of evidence leading to the widely used modern concept of species. This species concept was based on two observations: (i) species are composed of populations; and (ii) populations are co-specific if they successfully interbreed with each other. This idea was articulated by E. Mayr (1942) in the **biological species concept**: ‘Species are groups of interbreeding natural populations that are reproductively isolated from other such groups’. In the last 50 years several additional species concepts have been proposed. The most popular one in systematics is the so-called **phylogenetic species concept**. This concept does not emphasize the present properties of organisms or their hypothetical future, but rather points at their phylogenetic history. However, the applicability of this concept is debatable, for it proposes operational criteria of how to delimit species only as phylogenetic taxa, rather than describing the role that species play in the living world.

As with any concept, the biological species concept has its limitations. Application of the biological species concept is restricted to sexual, outcrossing populations over a short period of evolutionary time, excluding parthenogenetic organisms. Furthermore, in practice, the diagnosis of biological species is seldom done by testing their propensity to interbreed and produce fertile offspring but is often made by examining the difference in morphology. This should not be a contradiction, because phenotypic characters often, although not always, serve as markers for reproductive isolation. Morphological distinctiveness is a good, but not infallible, criterion for separating species. Sibling species may remain undetected, even after careful morphological examination, unless allozymes or other genetic markers are studied. The genetic difference among related species appears to vary substantially but generally increases with the time elapsed since their reproductive isolation. The degree of the genetic distance among populations, estimated from allozyme frequency, nucleotide divergence, amplified fragment length polymorphism (AFLP), random amplified polymorphic DNA (RAPD), microsatellites or other markers, can be evidence to determine whether or not they should be assigned the status of species. These approaches to delimiting species are based on indirect inference of the presence or absence of gene flow.

Another approach for analysis of genetic data deals with tree-based methods. In this approach species are delimited on the properties of a phylogenetic tree, which hypothesizes the relationships of groups bound by monophyly, or the shared presence

of apomorphies. These methods can be used to detect asexual species. However, in practice, all of these methods can fail occasionally or be discordant with each other because in nature the process of speciation seems to create diffuse boundaries, or a hybrid zone, between diversifying species. As a result, groups of populations form, sometimes named as subspecies, that are not fully reproductively isolated from each other. Moreover, these methods have different sensitivities and can reflect different properties of speciation. The solution to the problem of delimiting species can be found in the concept of **polyphasic taxonomy and classification** or an integrated approach in systematics.

### 2.3. Phylogenetics and Classification

Before the 1950s, taxonomists attempted to construct classifications on the dual criteria of common ancestry and their similarities. In the 1950s, the principles of numerical taxonomy were introduced, basing classification not on a few important features but on multiple character data. Numerical methods of analysis were used to create diagrams of overall similarity among species. Such a diagram, called a phenogram, was intended to give an objective basis for classification. This approach was called **phenetics**, giving rise to **phenetic classification**. This approach does not take into account the effects of parallel or convergent evolution in taxonomic interpretations. Another system was based on the argument that classification should rigorously reflect only phylogenetic relationships, not the degree of adaptive divergence or overall similarity. Classifications based on phylogenetic principles are named phylogenetic classifications; only shared, unique character states of similarity provide evidence for phylogenetic relationships. This approach to phylogenetic inference is known as **cladistics**. Branching diagrams constructed by cladistic methods are sometimes called cladograms, and monophyletic groups are called clades. A taxon should be a monophyletic group, originating from a single common ancestor, as opposed to a paraphyletic taxon, which includes only some of the descendants of a common ancestor, or a polyphyletic taxon, whose members share only a distant common ancestor and are usually circumscribed by other characteristics (i.e. **homoplasies**). Several terms are used to describe different character states for taxa under investigation: **plesiomorphy** (ancestral character state), **symplesiomorphy** (shared ancestral character state), **apomorphy** (derived character state), **synapomorphy** (shared derived character state) and **autapomorphy** (derived character state possessed by a single taxon). Within such a framework, the concept of **parsimony** is now widely applied to the reconstruction of phylogenetic relationships. It points out that among the various phylogenetic trees hypothesized for a group of taxa, the best one requires the fewest evolutionary changes, including the fewest homoplasies. Phylogenetic classification must always rely on an inferred phylogenetic tree, which is only an estimated part of a true history of the divergence of a species. In practice, creating a phylogenetic tree to resolve the phylogenetic relation between organisms is not a simple task.

The polyphasic taxonomy, or integrated approach, refers to classifications based on a consensus of all available data and information (phenotypic, genotypic and phylogenetic) used for delimiting taxa at all levels. Such analysis leads to a transition type of taxonomy in which a compromise can be formulated on the basis of results presently at hand.

## 2.4. Molecular Techniques

Almost all information from the genome and proteome at all levels, including the sequence of fragments of DNA, RNA or amino acids, the structure of molecules, the gene arrangement and presence versus absence of proteins or genes, can be applied to molecular systematics. Various biochemical and molecular techniques have been introduced to nematology for diagnostics, the estimation of genetic diversity of populations and the inference of phylogenetic relationships between taxa. The choice of technique depends on the research question.

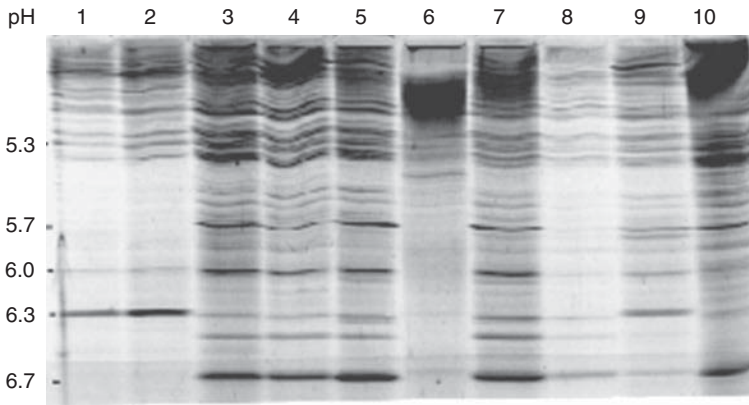
### 2.4.1. Protein-based techniques

These were the first of the molecular techniques to be applied in nematology. Soluble proteins extracted from nematodes are separated on polyacrylamide, starch, and cellulose acetate or agarose gels under an electric field on the basis of different molecular masses. Extracts from nematodes comprise thousands of different proteins but after total staining specific band patterns can be found for each sample. Differences in banding patterns between species or populations may be used as taxonomic markers. Isoelectric focusing (IEF), separating proteins on the basis of their charge in pH gradient, enables more stable profiles to be achieved and resolves proteins into sharp bands. The application of enzyme-staining techniques for characterization of a single protein or small subset of proteins on gels provides another diagnostic method. Extensive characterization of isozymes has been carried out for *Globodera*, *Heterodera*, *Radopholus*, *Meloidogyne*, *Pratylenchus* and other nematode groups. For many groups these studies revealed a wide variation between populations of the same species; however, limited interspecific variation was detected for species of root-knot nematode. The introduction of miniaturized electrophoretic systems, such as the PhastSystem (Pharmacia), has made it possible to study small amounts of soluble protein from a single sedentary female in a fully automatized process. IEF is used as a routine diagnostic technique for *Globodera pallida* and *G. rostochiensis* (Karszen *et al.*, 1995) as well as for the separation of other cyst nematode species (Fig. 2.1). Isozyme phenotypes of adult females, especially of esterase and malate dehydrogenase, are considered to be very useful as reliable markers for identification of root-knot nematodes (see Chapter 3). Because IEF differentiates root-knot nematode species only by specific isozyme patterns of young females, this technique can only be used to separate root-knot nematodes at this life stage.

Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) provides a better protein separation and fingerprint for any particular sample. In the first dimension, proteins are separated according to their charge; in the second dimension, they are separated on their mass. After staining, the position of individual proteins appears as spots of various size, shape and intensity. This technique has been applied to separate species and populations of *Globodera* and *Meloidogyne*.

### 2.4.2. DNA-based techniques

Compared with the above approaches, analysis of DNA has several advantages. DNA profiles can be obtained rapidly from a few or even single nematodes and



**Fig. 2.1.** Isoelectric focusing of proteins for species of the *Avenae* group. 1: *Heterodera avenae* (Rinkam, Germany). 2: *H. avenae* (Taaken, Germany). 3: *H. filipjevi* (Chabany, Ukraine). 4: *H. filipjevi* (Chernobyl, Ukraine). 5: *H. filipjevi* (Pushkin, Russia) 6: *H. pratensis* (Putilovo, Russia). 7, 8: *H. filipjevi* (Gorodets, Russia). 9: *H. filipjevi* (Vad, Russia). 10: *H. filipjevi* (Baimak, Russia). (From Subbotin *et al.*, 1996.)

the clarity of the results enables species to be identified very easily without the effects of environmental and developmental variation.

#### 2.4.2.1. DNA extraction

Extraction of DNA is the first step of molecular analysis. Using proteinase K is the most useful, cheap and rapid approach to extracting DNA from nematodes (Waeyenberge *et al.*, 2000). It consists of two steps: (i) mechanical destruction of the nematode body and tissues in a tube using an ultrasonic homogenizer or other tools, or repeatedly freezing samples in liquid nitrogen; and (ii) enzymatic lyses with proteinase K in a buffer for 1 h or several hours with subsequent brief inactivation of this enzyme at a high temperature. Various chemical treatments are also applied to remove cell components and purify the DNA. Phenol or phenol with chloroform extractions is often employed to remove proteins and ethanol is then used to precipitate and concentrate the DNA. Stanton *et al.* (1998) described an efficient method of DNA extraction from nematodes using chemical lyses in alkali solution without prior mechanical breaking of nematode bodies. Effective DNA extraction can also be achieved by using commercial kits developed by various companies.

#### 2.4.2.2. Polymerase chain reaction (PCR)

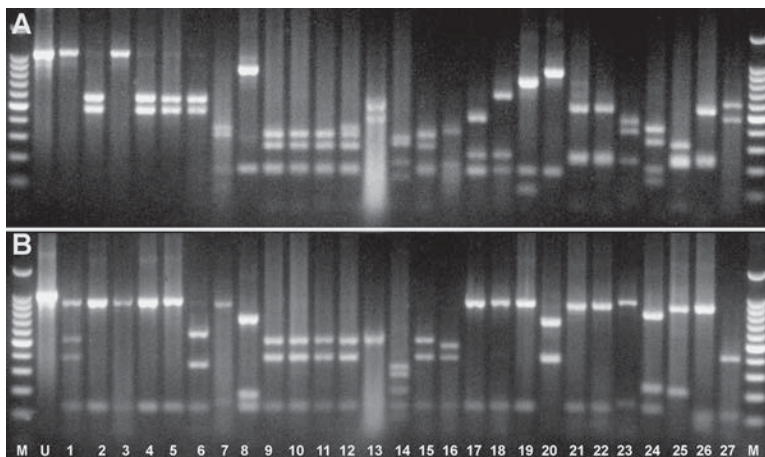
The polymerase chain reaction (PCR) technique has become one of the most widely used techniques for studying the genetic diversity of nematodes and their identification. PCR is a rapid, inexpensive and simple means of producing relatively large numbers of copies of DNA molecules via an enzyme catalyst. Any DNA fragment can be amplified and detected by PCR. The PCR method requires a DNA template (starting material) containing the region to be amplified, two oligonucleotide primers flanking this target region, DNA polymerase and four deoxynucleotide triphosphates

(dATP, dCTP, dGTP, dTTP) mixed in a buffer containing magnesium ions ( $Mg^{2+}$ ). A primer is a short, approximately 20-mer, oligonucleotide, which is complementary to the 3' end of each strand of the fragment that should be amplified. Primers anneal to the denatured DNA template and provide an initiation site for the elongation of the new DNA molecule. Universal primers are those complementary to a particular set of DNA for a wide range of organisms; primers matching only to certain species are called species-specific primers. When sequences of the flanking regions of the amplified fragment are unknown, PCR with degenerate primers, containing a number of options at several positions in the sequence that allows annealing and amplification of a variety of related sequences, can be applied.

PCR is performed in a tube in a thermocycler with programmed heating and cooling. The procedure consists of a succession of three steps determined by temperature conditions: (i) template denaturation (95°C for 3–4 min); (ii) primer annealing (55–60°C for 30 s to 2 min), and (iii) extension of the DNA chain (72°C for 30 s to 2 min). PCR is carried out for 30–40 cycles. As the result of PCR, a single target molecule of DNA is amplified into more than a billion copies. The resulting amplified products are electrophoretically separated according to their size on agarose or polyacrylamide gels and visualized using ethidium bromide, which interacts with double-stranded DNA and causes it to fluoresce under UV radiation. Once identified, nematode target DNA generated by PCR amplification can further be characterized by various analyses including restriction fragment length polymorphism (RFLP), dot blotting or sequencing. In some cases, the size of the PCR amplicon may serve as a diagnostic marker for a nematode group or species. It has been shown that primers amplifying the control region of mitochondrial DNA (mtDNA) generate different amplicon sizes for different species of root-knot nematodes; primers amplifying nuclear ribosomal intergenic spacer generated species-specific size polymorphisms for *Meloidogyne chitwoodi*, *M. hapla* and *M. fallax*.

#### **2.4.2.3. PCR–restriction fragment length polymorphism (PCR–RFLP)**

Variation in sequences in PCR products can be revealed by restriction endonuclease digestion. The PCR product obtained from different species or populations can be digested by a restriction enzyme, after which the resulting fragments are separated by electrophoresis. If differences in fragment length occur within restriction sites, the digestion of the PCR products will yield restriction fragment length polymorphism (RFLP), i.e. different RFLP profiles. PCR–RFLP of the internal transcribed spacer (ITS) regions of the ribosomal DNA is a very reliable method of identifying many plant-parasitic nematode groups including cyst (Fig. 2.2), root-knot, lesion and gall-forming nematodes, as well as nematodes from the genera *Bursaphelenchus* and *Aphelenchoides*. Using six to nine restriction enzymes enables most of the economically important species of cyst nematodes to be distinguished from each other as well as from their sibling species. RFLP of the ITS-rDNA obtained after restriction with several enzymes and their combination identifies important root-knot nematode species; however, it fails to separate species from the tropical group, including *M. javanica*, *M. incognita* and *M. arenaria*. PCR–RFLP of the mtDNA fragment between cytochrome oxidase subunit II gene and large subunit (LSU) has been applied successfully for diagnostics of these nematodes (Powers and Harris, 1993).



**Fig. 2.2.** RFLP profile of PCR-ITS rDNA generated by *AluI* (A) and *Bsh1236I* (B) for cyst-forming nematodes. M: 100 bp DNA ladder. U: unrestricted PCR product. 1 and 2: *Heterodera avenae*. 3: *H. arenaria*. 4: *H. filipjevi*. 5: *H. auklandica*. 6: *H. ustynovi*. 7: *H. latipons*. 8: *H. hordecalis*. 9: *H. schachtii*. 10: *H. trifolii*. 11: *H. medicaginis*. 12: *H. ciceri*. 13: *H. salixophila*. 14: *H. oryzicola*. 15: *H. glycines*. 16: *H. cajani*. 17: *H. humuli*. 18: *H. ripae*. 19: *H. fici*. 20: *H. litoralis*. 21: *H. carotae*. 22: *H. cruciferae*. 23: *H. cardiolata*. 24: *H. cyperi*. 25: *H. goettingiana*. 26: *H. urticae*. 27: *Meloidoderaalni*. (From Subbotin *et al.*, 2000.)

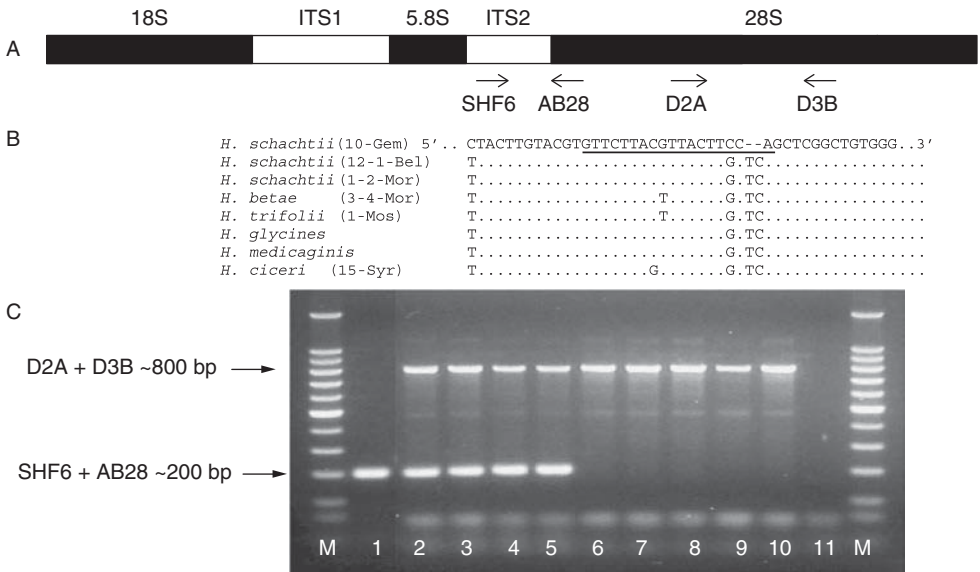
#### 2.4.2.4. Multiplex PCR

This type of PCR constitutes a major development in DNA diagnostics and enables the detection of one or several species in a nematode mixture by a single PCR test, decreasing diagnostic time and costs. In multiplex PCR, two or more unique targets of DNA sequences in the same sample are amplified by different primer pairs in the same amplification reaction. Multiplex PCR for detection of a single nematode species uses two sets of primers. One set is to amplify an internal control (e.g. universal primers for D2–D3 expansion regions of the 28S rRNA gene) confirming the presence of DNA in the sample and the success of PCR; the second set, including at least one species-specific primer, is targeted to nematode DNA sequences of interest (Fig. 2.3). Diagnostics using multiplex PCR with species-specific primers have been developed for a wide range of plant-parasitic nematodes.

#### 2.4.2.5. Random amplified polymorphic DNA (RAPD)

This method uses a single random primer of about ten nucleotides long for creating genomic fingerprints. This technique is often used for estimating genetic diversity between individuals, populations or closely related species. In this PCR approach, the short primer anneals to numerous similar sequences within the genome during the annealing step of the PCR cycle, which occurs at a lower temperature than does ‘classical’ PCR. If two complementary sequences are present on

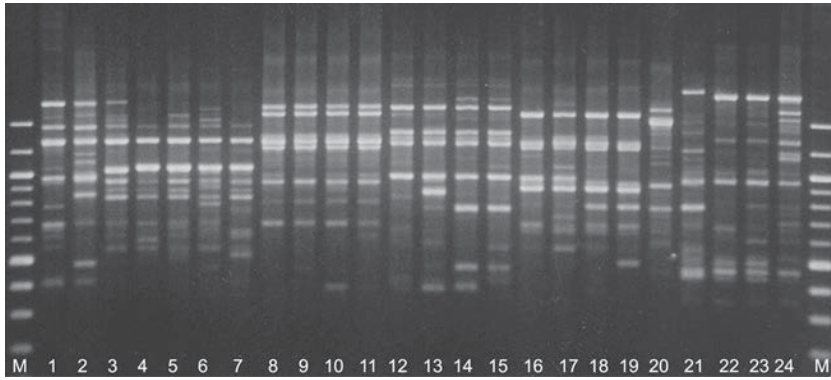




**Fig. 2.3.** PCR with a species-specific primer for the sugar beet cyst nematode *Heterodera schachtii*. **A:** Positions of species specific (SHF6) and universal primers in the rRNA gene. **B:** Sequence alignment for *H. schachtii* and closely related species with the sequence of the specific primer underlined. **C:** Agarose gel with PCR products generated with specific and universal primers (SHF6 + AB28) and universal (D2A + D3B) primers, or control band. 1–5: samples with *H. schachtii*. 6–10: nematode samples without *H. schachtii*. 11: a sample without nematode DNA. M: 100 bp DNA ladder. 1: resultant amplicon obtained with SHF6 + AB28 primer sets; resultant amplicons obtained with SHF6 + AB28 and D2A + D3B primer sets. (Modified from Amiri *et al.*, 2002.)

opposite strands of a genomic region in the correct orientation and close enough to one another, the DNA fragment between them can be amplified by PCR. Amplified DNA fragments obtained using different random primers from different samples are separated on gels and compared. RAPD polymorphisms result from the fact that if a primer hybridization site in a genome differs by even a single nucleotide, the change can lead to elimination of a specific amplification product. The resulting individual bands are considered as equivalent independent characters (Fig. 2.4). The band polymorphism can be binary scored and the data matrix is used for calculating the genetic distance between the samples under study and then presented as a dendrogram. Reproducibility of results is the most critical point for application of this technique.

The RAPD technique has been widely applied for separation of closely related species and studies of intraspecific variability of *G. pallida*, *Heterodera glycines*, *Radopholus similis*, *Ditylenchus dipsaci* and many other species (Powers, 2004; Blok, 2005). Specific sequences for certain species or races, called sequence-characterized amplified regions (SCAR) can be derived from RAPD fragments. Specific pairs of SCAR primers have been designed for identification of *M. chitwoodi*, *M. fallax*, *M. hapla* and other root-knot nematode species (Zijlstra *et al.*, 2000), as well as identification of stem nematodes *D. dipsaci* and *D. gigas*.



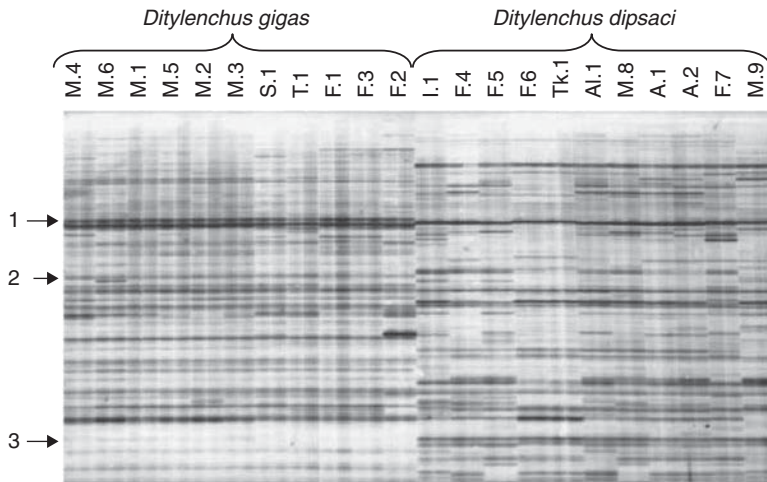
**Fig. 2.4.** Random amplified polymorphic DNA patterns for 26 populations of the *Heterodera avenae* complex. Primer G-10: 5'-AGGGCCGTCT-3'. 1: *H. avenae* (Taaken, Germany). 2: *H. avenae* (Santa Olalla, Spain). 3: *H. avenae* (Çukurova plain, Turkey). 4: *H. avenae* (Saudi Arabia). 5: *H. avenae* (Ha-hoola, Israel). 6: *H. avenae* (Israel). 7: *H. avenae* (near Delhi). 8: *H. australis* (South Australia, sample 3). 9: *H. australis* (Beulah, Australia). 10: *H. australis* (Victoria, Australia). 11: *H. australis* (Yorke Peninsula, Australia). 12: *H. mani* (Bavaria, Germany). 13: *H. mani* (Heinsberg, Germany). 14: *H. mani* (Andernach, Germany). 15: *H. mani* (Germany). 16: *H. pratensis* (Missunde, Germany). 17: *H. pratensis* (Östergaard, Germany). 18: *H. pratensis* (Lindhöft, Germany). 19: *H. pratensis* (Lenggries, Germany). 20: *H. aucklandica* (Auckland, New Zealand). 21: *H. filipjevi* (Saratov, Russia). 22: *H. filipjevi* (Akenham, UK). 23: *H. filipjevi* (Torralba de Calatrava, Spain). 24: *H. filipjevi* (Selçuklu, Turkey). M: 100 bp DNA ladder (Biolab). (After Subbotin *et al.*, 2003.)

#### 2.4.2.6. Amplified fragment length polymorphism (AFLP)

One of the most popular fingerprinting techniques, AFLP is also a random amplification technique, which does not require prior sequence information. AFLP produces a higher number of bands than is obtained by RAPD. It is a much more reliable and robust technique, unaffected by small variations in amplification parameters; however, it is more expensive. The AFLP technique represents a conceptual and practical advance in DNA fingerprinting. It comprises the following steps: (i) restriction of the total DNA with two restriction enzymes; (ii) ligation of double-stranded adapters to the ends of the restriction fragments; (iii) amplification of a subset of the restriction fragments using two 17–21 nucleotide primers complementary to the adapter and one that is 1–3 nucleotides adjacent to the restriction sites; (iv) separation and visualization of the AFLP-PCR fragments with a variety of techniques, usually on denaturing polyacrylamide gels with further staining. A comparative study of *Globodera* species and populations using AFLP revealed greater inter- and intraspecific variability than obtained by RAPD, and enabled subspecies of *G. tabacum* to be distinguished. AFLP analysis also showed a clear distinction between species of the *D. dipsaci* complex (Fig. 2.5) (Esquibet *et al.*, 2003).

#### 2.4.2.7. Real-time PCR

DNA technology also provides several methods for quantification of nematodes in samples. Real-time PCR requires an instrumentation platform that consists of a thermal



**Fig. 2.5.** Silver-stained 6% polyacrylamide gel showing AFLP amplification products generated using E-AA and M-CTG primers from 22 populations of *Ditylenchus dipsaci* and *D. gigas*. Two replicates were done for each population. Some polymorphic bands among races or populations are indicated by arrows. 1: *D. gigas*. 2: Population-specific band. 3: *D. dipsaci*. (From Esquibet *et al.*, 2003.)

cycler, optics for fluorescence excitation and emission collection, and computerized data acquisition and analysis software. The PCR quantification technique measures the number of nematodes indirectly by assuming that the number of target DNA copies in the sample is proportional to the number of targeted nematodes. Most of the difficulties with the PCR technique arise because only a very small number of the cycles (4–5 out of 40) contain useful information. The early cycles have an undetectable amount of DNA product; the final cycles, or the so-called plateau phase, are almost as uninformative. Quantitative information in a PCR comes from those few cycles where the amount of DNA grows exponentially from just above background to the plateau. The real-time technique allows continuous monitoring of the sample during PCR using hybridization probes (TaqMan, Molecular Beacons and Scorpions), allowing simultaneous quantification of several nematode species in one sample, or double-stranded dyes, such as SYBR Green, providing the simplest and most economical format for detection and quantification of PCR products in real-time reactions. Compared with traditional PCR methods, real-time PCR has advantages. It allows for faster, simultaneous detection and quantification of target DNA. The automated system overcomes the laborious process of estimating the quantity of the PCR product after gel electrophoresis. Real-time PCR has been used for detection and quantification of *Paratrichodorus pachydermus*, *H. schachtii*, *G. pallida* and *D. dipsaci*, as well as for estimating the number of virus vectoring trichodoridae nematodes.

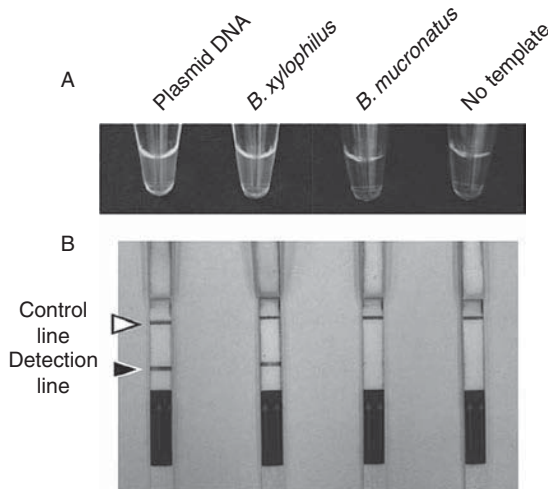
#### 2.4.2.8. Loop-mediated isothermal amplification (LAMP)

LAMP is a novel approach to nucleic acid amplification. The LAMP reaction requires a DNA polymerase with strand displacement activity (*Bst* polymerase) and a set of 4–6

specially designed primers based on distinct regions of the target DNA. Due to the specific nature of the action of these primers, the amount of DNA produced in LAMP is considerably higher than PCR-based amplification. The reaction occurs under isothermal conditions (60–65°C) and yields large amounts of product in a short time (30–60 min). LAMP products can be visualized either by gel electrophoresis or the naked eye by adding DNA intercalating dyes such as ethidium bromide or SYBR Green I in a tube. DNA concentration can also be detected by real-time detection methods. Because LAMP does not require an expensive thermal cycler and optical detection equipment and all LAMP steps are conducted within one reaction tube, this method clearly holds potential for testing in the field or in under-equipped laboratories. LAMP assays have been developed for detection of the pinewood nematode (Fig. 2.6) and common root-knot nematode species.

#### 2.4.2.9. DNA hybridization arrays

DNA arrays provide a powerful method for the next generation of diagnostics. The distinct advantage of this approach is that it combines DNA amplification with subsequent hybridization to oligonucleotide probes specific for multiple target sequences. DNA arrays can be used to detect many nematode species based on differences in the rRNA gene. In general, arrays are described as macroarrays or micro-arrays, the difference being the size and density of the sample spots, the substrate of hybridization and the type of production. Although the potential of DNA array methods for nematological diagnostics has been recognized, little progress had been made in their use,



**Fig. 2.6.** Visual inspection and lateral-flow strips used for the detection of loop-mediated isothermal amplification (LAMP) products. A: LAMP products visualized by fluorescent detection reagent and UV light. B: FAM-labelled probe detected by lateral-flow strips. LAMP reaction was carried out using internal transcribed spacer primers in the presence of plasmid DNA containing the target sequence (positive control) and genomic DNA of *Bursaphelenchus xylophilus* and *B. mucronatus*, and in the absence of template DNA (negative control). (After Kikuchi *et al.*, 2009.)

and only few research papers have been published on this technique (Blok, 2005). A reverse dot-blot assay has been developed for identification of several *Pratylenchus* species using oligonucleotides designed from the sequences of the ITS region of rRNA (Uehara *et al.*, 1999).

#### **2.4.2.10. Sequencing of DNA**

The process of determining the order of the nucleotide bases along a DNA strand is called sequencing. Several different procedures have been developed for DNA sequencing, i.e. the chemical degradation (Maxam–Gilbert) method and the chain termination (Sanger dideoxy) method, the latter being more commonly used. The chain termination sequencing method is similar to PCR in that it involves the synthesis of new strands of DNA complementary to a single-stranded template. The sequencing reaction components are template DNA, DNA polymerase with reaction buffer, one primer and the mixture of all four deoxynucleotides (dNTP) and four dideoxynucleotides (ddNTP) labels, each with a different colour fluorescent dye. As all four deoxynucleotides are present, chain elongation proceeds until, by chance, DNA polymerase inserts a dideoxynucleotide. As the dideoxy sugar lacks a 3'-hydroxyl group, continued lengthening of the nucleotide chain cannot occur. Thus, the dideoxynucleotide acts analogously to a specific chain-terminator reagent. Therefore, the result is a set of new chains with different lengths. These fluorescently labelled fragments are then separated by size using capillary electrophoresis. As each label fragment migrates through the gel pass, a laser excites the fluorescent molecule, which sends out light of a distinct colour. The detection system records the chromatogram output on a computer. A computer program (Chromas) then presents the sequencing result as chromatogram sequence files (Box 2.1).

The high demand for low-cost sequencing has driven the development of several high-throughput sequencing technologies that parallelize the sequencing process, producing thousands or millions of sequences at once. One of these is the 454 pyrosequencing technology developed by 454 Life Sciences. Pyrosequencing is based on the 'sequencing by synthesis' principle. The method amplifies DNA inside water droplets in an oil solution. A single DNA template attached to a single primer-coated bead then forms a clonal colony. The sequencing machine contains many picolitre-sized wells each containing a single bead and sequencing enzymes. Pyrosequencing differs from Sanger sequencing in that it relies on the detection of pyrophosphate release on nucleotide incorporation, rather than chain termination with dideoxynucleotides.

## **2.5. Genes used for Molecular Systematics**

A gene is usually defined as a DNA segment that codes for a polypeptide or specifies a functional RNA molecule. Eukaryotic protein-coding genes consist of transcribed and untranscribed parts, called flanking regions. Flanking regions are necessary for controlling transcription and processing pre-messenger RNA. A pre-mRNA consists of coding regions (exons), which encode amino acids, and non-coding regions containing information necessary for regulation of polypeptide production. Some segments of the non-coding regions (introns) are spliced out in the process of production of a mature mRNA.

### Box 2.1. Computer programs

A large amount of software is available for performing phylogenetic analysis. The most comprehensive list of software is given at: <http://evolution.genetics.washington.edu/phylip/software.html>

#### Packages for manipulation and align of sequences

**Chromas** (<http://www.technelysium.com.au/chromas.html>) is a program for displaying, editing and exporting chromatogram sequence files.

**Clustal** (<http://www.clustal.org>) is a package of multiple sequence alignment programs for DNA and proteins. It provides an integrated environment for performing multiple sequence and profile alignments and analysing the results.

**BioEdit** (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>) is a biological sequence alignment editor. An intuitive, multiple document interface with convenient features makes alignment and manipulation of sequences relatively easy. Several sequence manipulation and analysis options and links to external analysis programs facilitate a working environment that allows one to view and manipulate sequences with simple point-and-click operations.

#### General phylogenetic packages

**PAUP\*** (<http://paup.csit.fsu.edu>) is the most sophisticated and user-friendly program for phylogenetic analysis, with many options. It includes parsimony, distance matrix and maximum likelihood (ML) methods.

**PHYLIP** (<http://evolution.genetics.washington.edu/phylip/>) includes programs to carry out parsimony, distance matrix methods and ML, including bootstrapping and consensus trees. There are programs for data types including DNA and RNA, protein sequences, gene frequencies, restriction sites and restriction fragments, and discrete and continuous characters.

**MrBayes** (<http://mrbayes.sourceforge.net/>) is a program for the Bayesian estimation of phylogeny. The program uses a Markov chain Monte Carlo (MCMC) technique to approximate the posterior probabilities of trees. One of the program features is the ability to analyse nucleotide, amino acid and morphological data under different models in a single analysis.

**MacClade** (<http://macclade.org/>) is a program for phylogenetic analysis with analytical strength in studies of character evolution. It also provides many tools for entering and editing data and many descriptive statistics as well as for producing tree diagrams and charts.

#### Package for selecting models of evolution

**ModelTest** (<https://code.google.com/p/jmodeltest2/>) is a program for selecting the model of nucleotide substitution that best fits the data.

#### Packages for tree visualization and tree analysis

**TreeView** (<http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>) is a simple program for displaying and manipulating phylogenetic trees.

**TreeMap** (<http://taxonomy.zoology.gla.ac.uk/rod/treemap.html>) is an experimental program for comparing host and parasite trees.

**Component** (<http://taxonomy.zoology.gla.ac.uk/rod/cpw.html>) is a program for analysing evolutionary trees and is intended for use in studies of phylogeny, tree shape distribution, gene trees/species trees, host–parasite co-speciation and biogeography.

The eukaryotic cell contains two different genomes: that of the nucleus and that of mitochondria. Molecular systematics use data from both genomes. Mitochondria are inherited maternally, whereas the nucleus is biparental. The genome of *Caenorhabditis elegans* is organized as five pairs of autosomal chromosomes (coded I, II, III, IV and V) and a pair of sex chromosomes (X). The nuclear genome of this species is about 100 million base pairs in length, and encodes approximately 20,000 protein-coding genes. The mitochondrial genome of nematodes varies and in *C. elegans* consists of 13,794 base pairs.

To determine true evolutionary relationships between organisms, it is essential that the correct gene fragment or molecule is chosen for sequence studies. This is important for several reasons: (i) the molecule should be universally distributed across the group chosen for study; (ii) it must be functionally homologous in each organism, i.e. the phylogenetic comparisons must start with molecules of identical function; and (iii) it is critical in sequence comparisons to be able to align the molecules properly in order to identify regions of sequence homology and sequence heterogeneity.

### 2.5.1. Nuclear ribosomal RNA genes

Historically, the only nuclear genes with a high enough copy number for easy study were ribosomal genes. These genes code rRNAs, which are nearly two-thirds of the mass of the ribosome. The genes encoding rRNA are arranged in tandem, in several hundred copies, and are organized in a cluster that includes a small subunit (SSU or 18S) and a large subunit (LSU or 26–28S) gene, which are themselves separated by a small 5.8S gene. The whole set of genes is transcribed as a single unit. Another ribosomal gene, a ubiquitous component of large ribosomal subunits in the eukaryotic cell, is 5S rRNA. The gene is found at different localizations in different organisms. The 5S rRNA gene linked to the intergenic spacers (IGS) regions of rRNA repeated units has been described for several root-knot nematode species.

There are 55 copies of the rRNA genes on chromosome I and 110 5S rRNA genes on chromosome V in *C. elegans*. In addition to these coding sequences, the rDNA array also contains spacer sequences, which contain the signals needed to process the rRNA transcript: an external transcribed spacer (ETS) and two internal transcribed spacers, ITS1 and ITS2. A group of genes and spacer sequences together make up an rRNA transcript unit. These units are separated from each other by an IGS region, also known as a non-transcribed spacer (NTS).

The rRNA (18S and 28S) genes evolve slowly and can be used to compare distant taxa that diverged a long time ago, whereas external and intergenic spacers have higher evolution rates and so have been used for reconstructing relatively recent evolutionary events and for the comparison of closely related species and subspecies. The IGS region contains many repeats and is more variable than the ITS region.

Although rRNA genes are present in many copies, their sequences are almost identical, because the highly repetitive sequences undergo homogenization processes known as concerted evolution. If a mutation occurs in one copy of a sequence, it is generally corrected to match the other copies, but sometimes the non-mutated copies are corrected to match the mutated one, so that nucleotide

changes propagate throughout the arrays. However, this process may be disrupted, so that several different copies of this gene may be present in the genome. The risk of incorporating ITS paralogues into phylogenetic studies should be considered with caution. Inspection of some basic features of the sequence, including the integrity of the conserved motifs and the thermodynamic stability of the secondary structures of the RNA transcripts, enables rRNA pseudogenes to be excluded from the dataset.

### 2.5.2. Nuclear protein-coding genes

Protein-coding genes have some advantages over rRNA genes and their spacers in that the alignment of sequences is less problematic. Protein sequences also lend themselves to different phylogenetic weighting of bases by codon position. The intron position patterns may also serve as decisive markers for phylogenetic analysis. Heat shock proteins, RNA polymerase II, actin, major sperm protein and other genes have been used for phylogenetic studies of cyst, root-lesion and other nematodes.

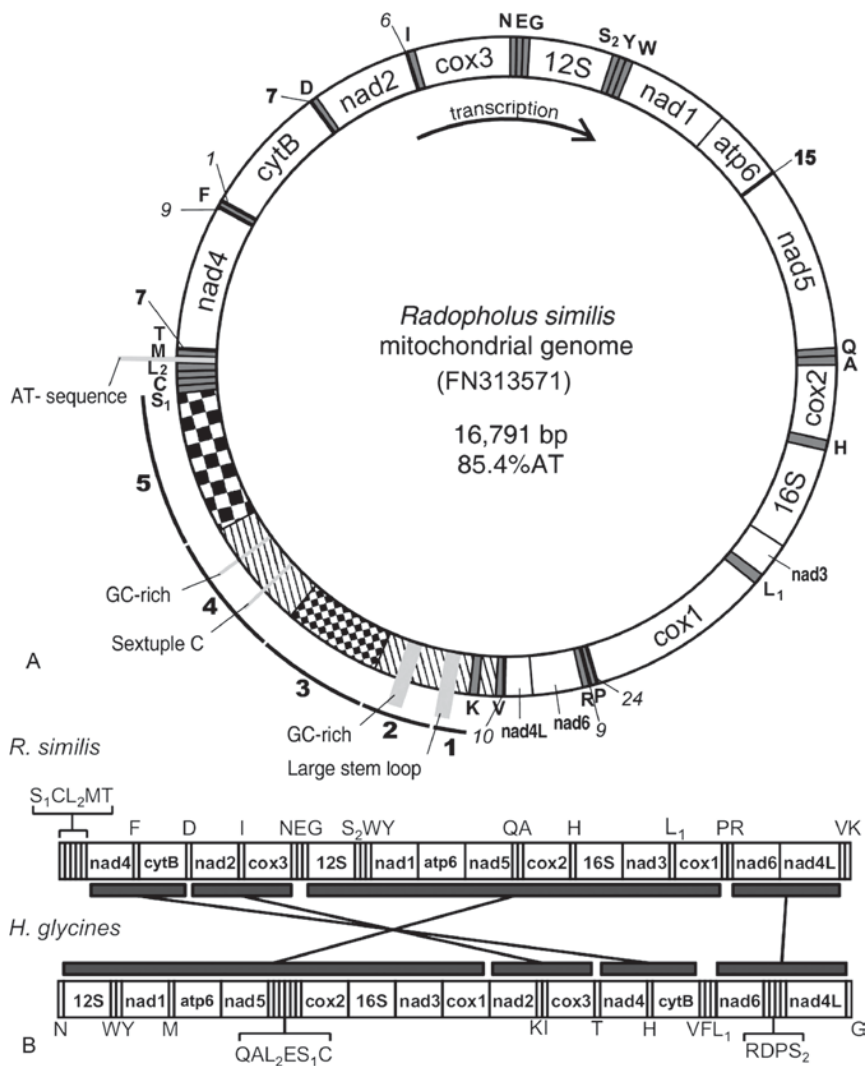
### 2.5.3. Mitochondrial DNA

Mitochondrial DNA (mtDNA) has been used to examine population structure and evolutionary relationships between different nematode groups. All nematode mtDNAs are circular, double-stranded DNA molecules. The mitochondrial genome of the majority of nematodes includes: (i) 12 protein-coding genes, all components of the oxidative phosphorylation system including subunits of cytochrome *c* oxidase (COI–COIII); (ii) 22 transfer RNA genes; and (iii) rRNA genes encoding SSU and LSU rRNAs (Fig. 2.7). In addition, there is usually a non-coding AT-rich region, or a region with high levels of the nucleotides adenine and thymine in the mitochondrial genome containing an initiation site for replication and transcription. The remainder of the approximately 1000 mitochondrial proteins is nuclear-encoded and is imported in the organelle. The arrangement of genes in the mitochondrial genome is not consistent within Nematoda. Nematodes are characterized by a surprising variation in gene order. A unique feature of the mitochondrial genome organization of nematodes is that some of them, e.g. *G. pallida*, contain at least six small circular mtDNA molecules varying in size from 6.3 to 9.5 kb.

Nematode mtDNA sequences accumulate substitution changes much more quickly than the ITS sequences and tend to be extremely A+T rich, with typical values as high as 75 and 80%. The T content seems to be greater at the third codon position, compared with the first and second positions. Although the high rate of substitution makes mtDNA very useful for low-level phylogenetic applications, failure to correct for this severe substitution bias can potentially lead to phylogenetic error.

The relatively rapid rate of evolution and rearrangements that occur in mtDNA has limited the design of universal primers and, thus, mtDNA has not been as widely used as other markers for nematode phylogenetic or diagnostic purposes, except for





**Fig. 2.7.** A. Overview of the organization of the circular mitochondrial DNA of *Radopholus similis*. Genes and non-coding regions are indicated: in white, the protein-coding and rRNA genes, in grey, the tRNA genes called by their amino acid symbol (S<sub>1</sub>: Ser-AGN, S<sub>2</sub>: Ser-UCN, L<sub>1</sub>: Leu-CUN, L<sub>2</sub>: Leu-UUR). Bold and italic numbers indicate non-coding and overlapping nucleotides between neighbouring genes, respectively. The pattern-filled part represents the large non-coding region. The repeat region of 302 bp is filled with large checkers and the 26 bp repeat region is filled with small checkers. (After Jacob *et al.*, 2009.) B. Partial mitochondrial genome organization of *R. similis* and *Heterodera glycines*. Bars joined by lines indicate regions of conserved genome organization. (After Gibson *et al.*, 2011.)

root-knot nematodes. The region between the COII and the LSU RNA gene containing an intergenic region with unique size and nucleotide polymorphism might be utilized for distinguishing different species and host races of *Meloidogyne* (Powers and Harris, 1993).

## 2.6. Microsatellites

Microsatellites or simple sequence repeats (SSRs) are short, 1–6 base nucleotide sequences (e.g. AAG) that are repeated many times in tandem (...AAGAAGAAG...). Microsatellites are found in all eukaryotic genomes. Overall, (AT)<sub>n</sub>, (AG)<sub>n</sub>, (CT)<sub>n</sub>, (AAT)<sub>n</sub> and (ATT)<sub>n</sub> are the most frequent microsatellite motifs presently known in nematode genomes. They are present in both coding and non-coding regions, and found covering from 0.09 to 1.20% of the nematode genomes (Castagnone-Sereno *et al.*, 2010). A very high mutation rate, from 10<sup>-4</sup> to 10<sup>-3</sup> mutations per microsatellite and per generation is usually associated with microsatellite loci, resulting in high heterozygosity and the presence of multiple alleles at a given locus. They are present in both coding and non-coding regions and are usually characterized by a high level of length polymorphism. Despite the fact that the mechanism of microsatellite evolution is still unclear, SSRs have been widely used as powerful markers for studies in population genetics. Microsatellites mutate over time, their alleles diverging in the number of sequence repeats. The flanking regions surrounding the microsatellites can be conserved across genera or even higher taxonomic levels and, therefore, are used as designs for PCR primers to amplify microsatellite loci. Based on analysis of the microsatellite variation in populations, inferences can be made about population structures and differences, genetic drift and the date of a last common ancestor (Jarne and Lagoda, 1996).

## 2.7. DNA Bar Coding

DNA bar coding is a taxonomic method that uses a fragment containing the first half of the COI gene of mtDNA to identify it as belonging to a particular species. It is based on a relatively simple concept: most eukaryote cells contain mitochondria and mtDNA has a relatively fast mutation rate and more differences than the ribosomal genes, which results in significant variance in mtDNA sequences between species and, in principle, a comparatively small variance within species. Molecular bar coding involves isolation of the nematodes (as individuals or in bulk), amplification of the target gene, cloning, sequencing and phylogenetic analysis leading to the assessments of species content, abundance and diversity. However, DNA bar coding is only as good as the reference database: it cannot be used to identify species not already catalogued. Bar coding will also be most reliable for identification of potential putative new species, but only for species groups whose genetic diversity has been well surveyed. DNA bar coding cannot replace the traditional methods of species description. Currently, there is insufficient information in databases for extensive nematode species identification based on the COI gene. However, the increasing deposition of DNA sequences in GenBank and NematOL databases will be beneficial for diagnostics.

## 2.8. Phylogenetic Inference

Phylogenetic analysis is a complex field of study that embraces a variety of techniques that can be applied to a wide range of evolutionary questions. However, a

complete understanding of all assumptions involved in analysis is essential for a correct interpretation of the results. A possible work flow of a molecular phylogenetic project could be presented as a flow diagram: (i) selection and sampling of a group of organisms; (ii) choice of molecular markers; (iii) sequencing and assembling of sequence data; (iv) alignment, or establishment of homology between molecules; (v) construction of phylogenetic tree using distance or discrete methods and making an assessment of the reliability of its branches; and (vi) testing of different alternative hypotheses.

### 2.8.1. Alignment

The first step of any phylogenetic study is the construction of alignment or establishment of positional homology between nucleotides or amino acid bases that have descended from a common ancestral base. Errors incurred in this step can lead to an incorrect phylogeny. The best way to compare the homologous residues is to align sequences one on top of another in a visual display, so that, ideally, each homologous base from different sequences line up in the same column. Three types of aligned pairs are distinguished: (i) **matches** (same nucleotide appears for all sequences); (ii) **mismatches** (different nucleotides were found in the same position); and (iii) **gaps** (no base in a particular position for at least one of the sequences). A gap indicates that a deletion has occurred in one sequence or an insertion has occurred in another sequence. However, the alignment itself does not enable these mutational events to be distinguished. Therefore, insertions and deletions are sometimes collectively referred to as **indels**. For closely related species the optimal alignment of sequences having the same length can easily be done manually; for distantly related organisms, where many deletion or insertion mutations have occurred, alignments are usually constructed using computer programs with particular algorithms.

The optimal automatic alignment is considered to be that in which the numbers of mismatches and gaps are minimized according to the desired criteria. The program Clustal (Box 2.1) is one of the most commonly used computerized alignment programs using a progressive alignment approach. Sequences are aligned in pairs to generate a distance matrix, which then is used for calculating a neighbour-joining guide tree. This tree gives the order in which progressive alignment should be carried out. Progressive alignment is a mathematical process that is completely independent of biological reality. The use of structural components of the given molecule can significantly improve estimations of homology, thus generating a better alignment.

### 2.8.2. Methods for inferring phylogenetic trees

The methods for constructing phylogenetic trees from molecular data can be categorized into two major groups: (i) distance methods; and (ii) discrete methods. In distance methods such as analysis by minimum evolution, sequences are converted into a distance matrix that represents an estimate of the evolutionary distances between sequences, from which a phylogenetic tree is constructed, by considering the relationships among these distance values, which are supposed to represent distances between

taxa. In discrete methods, maximum parsimony, maximum likelihood (ML), Bayesian inference methods map the history of characters onto a tree. Each method requires some assumptions about evolution.

### **2.8.2.1. Minimum evolution method**

The minimum evolution method is very useful for analysing sequences. In this method, the sum of all branch lengths is computed for all plausible trees and the tree that has the smallest sum value is chosen as the best tree. The neighbour joining (NJ) method applies the minimum evolution principle and estimates the tree based on data transformed into a pairwise distance matrix. This method does not examine all possible topologies but at each stage of taxon clustering a minimum evolution principle is used. The NJ algorithm is extremely popular because it is relatively fast and performs well when the divergence between sequences is low.

### **2.8.2.2. Maximum parsimony**

Maximum parsimony is an important method of phylogenetic inference. The goal of parsimony is to find the tree with the minimum total tree length, or the minimum amount of evolutionary changes, i.e. the transformation of one character state to another. The better a tree fits the data, the fewer homoplasies will be required and the fewer number of character state changes will be required. Several different parsimony methods have been developed for treating datasets. The problems of finding optimal trees under the maximum parsimony criterion are twofold: (i) determining the tree length; and (ii) searching over all possible trees with the minimum length. When the number of taxa is small, it is possible to evaluate each of the possible trees, or to conduct an exhaustive search. An exhaustive search is carried out by finding each of the possible trees by a branch-additional algorithm. However, if the number of trees is large, the application of this approach is near impossible, and a heuristic strategy is used.

As with any method, maximum parsimony has its pitfalls. If some sequences have evolved much faster than others, homoplasies have probably occurred more often among the branches leading to these sequences than in others, so that parsimony tends to cluster these highly divergent branches together. This effect, called long-branch attraction, can be reduced by sampling additional taxa related to those terminating the long branches, so that the branches may be broken up into smaller ones.

### **2.8.2.3. Maximum likelihood**

Maximum likelihood (ML) is the method that is generally considered to make the most efficient use of the data to provide the most accurate estimates of phylogeny. The likelihood is not the probability that the tree is the true tree; rather it is the probability that the tree has given rise to the data that were collected. The basic idea of the likelihood approach is to compute the probability of the observed data assuming it has evolved under a particular evolutionary tree and a given probabilistic model of

substitution. The likelihood is often expressed as a natural logarithm and referred to as the log-likelihood. The tree with the highest likelihood is the best estimate of the true phylogeny. The main obstacle to the widespread use of ML methods is computing time, because algorithms that find the ML score must search through a multidimensional space of parameters to find a tree. ML requires three elements: a model of sequence evolution, a tree and the observed data.

#### **2.8.2.4. Bayesian inference**

Bayesian inference of phylogeny is based on a quantity called the posterior probability of a tree. The posterior probability of a tree can be interpreted as the probability that the tree is correct. The posterior probability involves a summation over all trees and, for each tree, integration over all possible combinations of branch length and substitution model parameter values. This method is almost impossible to complete by exhaustive analysis, and so the Markov chain Monte Carlo (MCMC) is a search method used to approximate the posterior probabilities of trees. ML and Bayesian analysis are both based upon the likelihood function, although there are fundamental differences in how the two methods treat parameters.

#### **2.8.2.5. Evolutionary models**

In order to reconstruct an evolutionary tree some assumptions about the evolutionary process for the studied molecules should be made. Evolutionary substitution models for DNA are implemented in a different way in distance, ML and Bayesian analysis. The substitution model is a description of the way sequences evolved in time by nucleotide replacements. The nucleotide substitution process of a DNA sequence can be described by a so-called homogeneous Markov process that uses the Q matrix, which specifies the relative rates of change of each nucleotide along the sequences. The Jukes–Cantor model (JC69) was one of the first proposed and is perhaps the simplest model of sequence evolution. It assumes that the four bases have equal frequencies, and that all substitutions are equally likely. The general time-reversible model (GTR) is the most general model, where all eight free parameters of reversible nucleotide rate Q matrix are specified. The best-fit model of evolution for a dataset can be selected through statistical testing. The fit to the data of different models can be compared through likelihood ratio tests (LRTs) or information criteria to select the best-fit model within a set of possible ones.

### **2.8.3. Phylogenetic tree and tree terminology**

The result of a molecular phylogenetic analysis is expressed in a phylogenetic tree. A tree consists of nodes, which are connected by branches. The branch length usually represents the evolutionary distance between two consecutive nodes. Terminal nodes (leaves) are often called operational taxonomic units (OTUs). Internal nodes represent hypothetical ancestors and may be called hypothetical taxonomic units (HTUs). The ancestor of all the taxa that comprise the tree is the root of the tree. An outgroup

is a terminal taxon whose most recent common ancestor with any taxon within a given clade occurs at a node outside that clade. The OTUs within a given clade are called ingroup taxa. A group of taxa that belong to the same branch is called a cluster. Sister groups or sister taxa refer to two groups on a tree with the same immediate common ancestor, or are more closely related to each other than either is to any other taxon. The branching patterns, or the order and arrangement of nodes, are collectively called the topology of the tree. If three branches connect to an internal node, then the node represents a bifurcation, or dichotomy. If more than three branches connect to an internal node, then the node represents polytomy. A tree implicitly assumes that once two lineages appear, they subsequently never interact with each other. However, in reality such interactions might have occurred, such as through hybridization (rare in animals) or reticulate evolution, and such relationships can be presented as a network.

Parsimony analyses often arrive at multiple trees with the same length but with different branch order. Rather than choosing among these trees, systematists may simply want to determine what groups can be found in all the shortest trees. There are approaches to summarizing information which are common to two or more trees in a single tree. The resulting tree is called a **consensus tree**. A **strict consensus tree** shows only those relationships that were hypothesized in all the equally parsimonious trees, whereas a **majority consensus rule tree** shows those relationships hypothesized in more than half the trees being considered.

#### 2.8.4. Evaluation of the reliability of inferred trees

The estimation of phylogeny should be accompanied by an indication of its confidence limit. Phylogenetic trees should always be evaluated for reliability, which could be measured as the probability that the taxa of a given clade are always members of that clade. Bootstrap and jack-knife analyses are the techniques used most often for this purpose. Bootstrapping and jack-knifing are so-called re-sampling techniques, because they estimate the sampling distribution by repeatedly re-sampling data from the original dataset. These methods differ in their methods of re-sampling. Bootstrapping is the more commonly used approach for phylogenetic reconstruction. To estimate the confidence level by bootstrapping, or the bootstrap value of a clade, a series of pseudo samples or pseudo alignments is first generated by randomly re-sampling the sites in the original alignment with replacement. In such pseudo alignments some characters are not included at all, while others may be included twice or more. Secondly, for each pseudo alignment, a tree is constructed, and the proportion of each clade among all the bootstrap replicates is computed in a majority-rule consensus tree. If the value of support of the clade obtained as a result of these analyses is greater than 95%, the branch is considered to be statistically significant. Branch support less than 70% should be treated with caution.

Confidence in maximum parsimonious trees can also be evaluated by calculating the decay index, or Bremer support, which expresses the number of extra steps required for each node not to appear in the tree, i.e. the length difference between the shortest trees including the group and the shortest trees that exclude this group. The higher the decay index, the better the support for the group.

### 2.8.5. Testing of hypotheses

Once phylogenetic trees are obtained from a molecular dataset using different methods, they should be compared with each other or with trees generated from other, non-molecular datasets. There are several tests that allow the evaluation of alternative hypotheses and determine whether one tree is statistically worse than another: the Templeton test, Shimodaira–Hasegawa test and parametric bootstrapping.

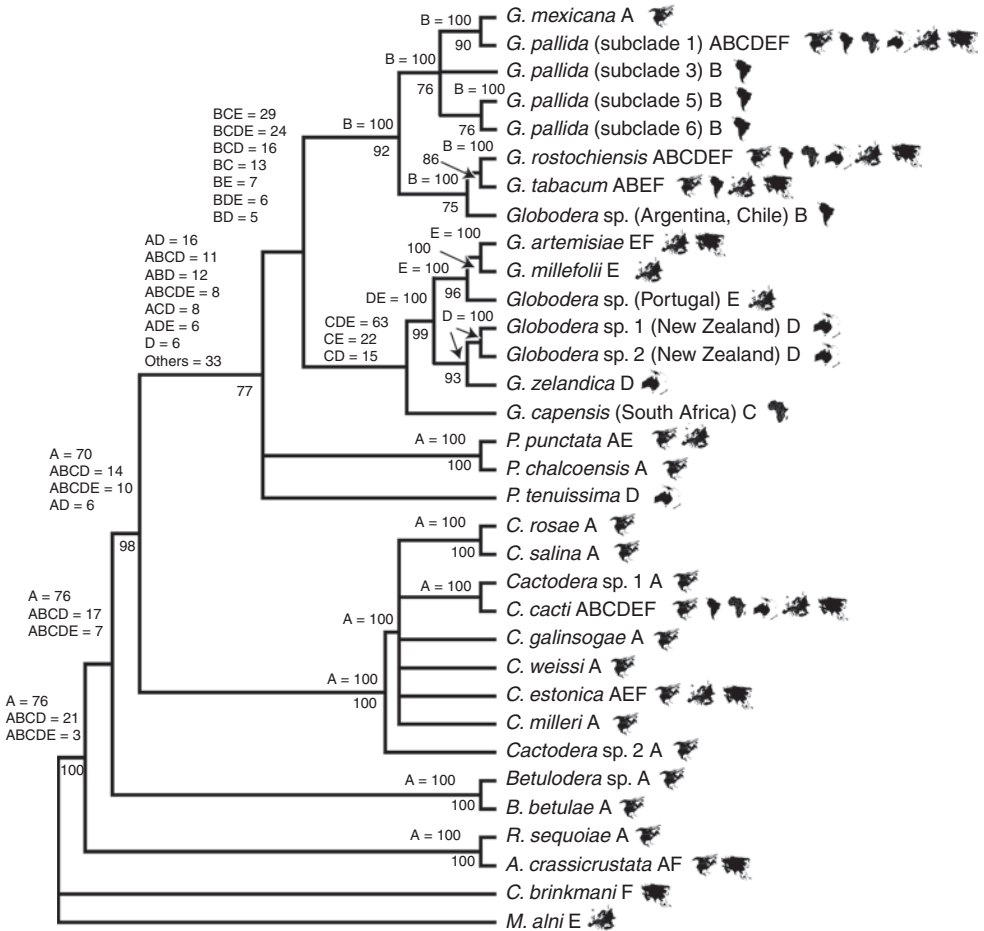
## 2.9. Reconstruction of Historical Associations







Historical associations, when a lineage tracks another lineage, can be divided into three basic categories: (i) genes and organisms; (ii) organisms and organisms; and (iii) organisms and areas (Page and Charleston, 1998). In each association, one entity tracks the other with a degree of fidelity that depends on the relative frequency of four events: co-divergence, duplication, horizontal transfer and sorting event. The testing of hypotheses for co-divergence could be made using tree topology or data-based methods. Both methods are not without problems, and should rely on the estimation of inferred phylogenies or adequacy of models of sequence evolution.

**Genes and organisms.** Each gene has a phylogenetic history that is intimately connected with, but not necessarily identical to, the history of the organisms in which the gene resides. Processes such as gene duplication, horizontal gene transfer, gene loss and lineage sorting can produce complex gene trees that differ from organismal trees.

**Organisms and organisms.** Several associations between nematodes and other organisms are known, e.g. nematodes of the Heteroderidae or Anguinidae and their host plants; nematodes from the *Xiphinema americanum* group and symbiotic bacteria of the genus *Xiphinematobacter*; seed gall forming nematodes of the genus *Anguina* and host plants or bacteria from the genus *Rathayibacter*; and soil-inhabiting nematodes and parasitic bacteria of the genus *Pasteuria*. The phylogenetic tests for co-divergence have been made for associations: anguinid nematodes and their host plants and soil-inhabiting nematodes and parasitic *Pasteuria*. Molecular data suggest that anguinid groups are generally associated with host plants from the same or related systematic groups. Although the strict co-speciation hypothesis for seed-gall nematodes and grasses was rejected, the analysis showed a high level of co-speciation events, which cannot be explained as a result of random establishments of host–parasitic association. Analysis of phylogenies of *Pasteuria* and their nematode hosts suggest that horizontal host switching is the most common event in this association (Sturhan *et al.*, 2005).

**Organisms and areas.** Organisms can track geological history such that sequences of geological events are directly reflected in the phylogenies of those organisms. Phylogeography is the study of the historical processes (vicariance, sympatry, dispersal and extinction) that take place in this association (Page and Charleston, 1998). Studies of the geographical distributions of genealogical lineages within and between species of some genera give interesting views on origin and dispersal of some nematode groups (Fig. 2.8).



**Fig. 2.8.** Phylogeny of *Globodera* and Punctoderinae sensu Krall & Krall, 1978. Strict consensus of 28 maximum parsimony trees obtained from analysis of the ITS rRNA gene sequences. Letters at nodes indicate putative ancestral areas with frequency of occurrence of the node: A: , North America. B: , South America. C: , Africa. D: , New Zealand (Oceania). E: , Europe. F: , Asia. Letters to the right of taxon name indicate present distribution and thus coding for the ancestral area analysis. Phylogeographic analysis suggested a North American origin of Punctoderinae with possible further long-distance dispersal to South America, Africa and other regions. (From Subbotin *et al.*, 2011.)

## 2.10. Databases

Phylogenetic analyses are often based on data accumulated by many investigators in different databases. All novel sequences have to be submitted in a public database. Databases are essential sources for modern bioinformatics, as they serve as information storage equipped with powerful query tools and a well-developed system of cross-references (Box. 2.2).



### Box 2.2. Databases.

Numerous genetic databases are spread out all over the world. The biggest public databases containing nucleotide sequence information are as follows: **GenBank** (National Center for Biotechnology Information, USA) (<http://www.ncbi.nlm.nih.gov>), **EMBL** (European Molecular Biology Laboratory) (<http://www.ebi.ac.uk/embl>) and **DDBJ** (DNA Data Bank of Japan) (<http://www.ddbj.nig.ac.jp>). Exchange of data between these international collaborating databases occurs on a daily basis.

**TreeBASE** (<http://www.treebase.org>) is a relational database designed to manage and explore information on phylogenetic relationships. Its main function is to store published phylogenetic trees and data matrices. It also includes bibliographic information on phylogenetic studies, and some details on taxa, characters, algorithms used and analyses performed.

There are several specialized nematode databases:

**WormBase** (<http://www.wormbase.org>) is the central data repository for information about *Caenorhabditis elegans* and related nematodes. As a model organism database, WormBase extends beyond the genomic sequence, integrating experimental results with an extensively annotated view of the genome. WormBase also provides a large array of research and analysis tools.

**NemaGene** (<http://www.nematode.net>) is a web-accessible resource for investigating gene sequences from nematode genomes. The database is an outgrowth of the parasitic nematode expressed sequence tags (EST) project. ESTs are usually shorter than the full-length mRNA from which they are derived and are prone to sequencing errors. The database provides EST cluster consensus sequence, enhanced online BLAST search tools and functional classification of cluster sequences.

**NEMBASE4** (<http://www.nematodes.org/nembase4/>) is a resource for nematode transcriptome analysis, and a research tool for nematode biology, drug discovery and vaccine design. Users may query the database on the basis of BLAST annotation, sequence similarity or expression profiles.

**NemAToL** (<http://nematol.unh.edu/>) is an open database dedicated to collecting, archiving and organizing video images of other morphological information, DNA sequences, alignments, and other reference materials for study of the phylogeny and diversity, and taxonomy, systematics and ecology of nematodes.

## 2.11. Examples of Molecular Phylogenies

### 2.11.1. Position of Nematoda within metazoans

The relative position of nematodes in animal phylogeny remains uncertain. In the traditional, morphologically based view, bilateral organisms are subdivided according to their internal organization and emerged in a universal phylogenetic tree with the following order: (i) the Acoelomata, lacking a body cavity (mainly the platyhelminths and nemertines); (ii) the Pseudocoelomata (nematodes and some other minor phyla), with an internal cavity outside the mesoderm; and (iii) the Coelomata,

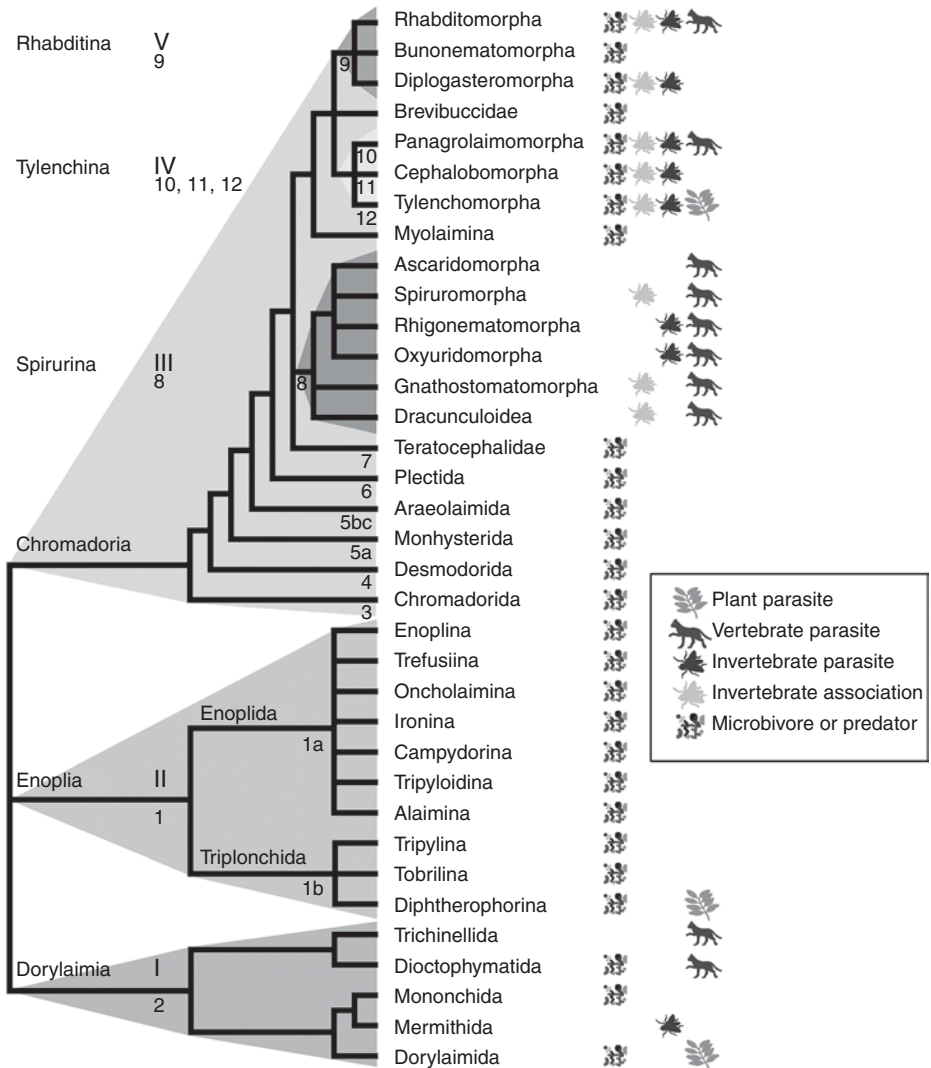
which have true coelomic cavities splitting the mesoderm. Another vision of the phylogenetic relationships between metazoan phyla was obtained after analysis of molecular data. The concept is known as the Ecdysozoa hypothesis, grouping moulting organisms, including arthropods and nematodes, in a single clade based on analysis of 18S rRNA gene sequences and recently also supported by studies of the 28S rRNA. By contrast, the phylogenetic analyses using many protein-coding genes showed the strongest and most consistent support for the coelomate topology.

### 2.11.2. The phylum Nematoda

Since the first publications of general phylogenies of nematodes based on 18S rRNA in 1998, the massive nematode rRNA database has been accumulated. The analysis of the phylum highlights a number of paraphyletic taxa and indicates new relationships between previously unconnected taxa. The Nematoda seems to have arisen from adenophorean ancestry; the classic split into Adenophorea and Secernentea is not supported. Holterman *et al.* (2006) presented a subdivision of the phylum Nematoda into 12 clades. It was suggested that animal parasitism arose independently at least five times, and plant parasitism three times: Tylenchomorpha, Dorylaimida and Diphtherophorina (Fig. 2.9) (Blaxter *et al.*, 1998; De Ley and Blaxter, 2004; van Megen *et al.*, 2009). However, the correctness of the phylogenetic reconstruction for nematodes using SSU might be influenced by two main factors: (i) grouping of long branches occurring as a result of abnormally high evolution rate and (ii) a total deficit of informative characters. Because the SSU tree is reconstructed based on a single gene, efforts are continuously made to sequence other genes. Multigene phylogeny will become available when sequences of genes from nematode genome projects are obtained.

### 2.11.3. The infraorder Tylenchomorpha

The evolutionary relationships of tylenchid and aphelenchid nematodes have been evaluated using sequence data of the 18S and the 28S rRNA genes. The order Tylenchida *sensu* Siddiqi containing plant-parasitic nematodes appears to be clearly monophyletic. The order Aphelenchida *sensu* Siddiqi comprising fungal-feeding Aphelenchidae and Aphelenchoididae is polyphyletic in all molecular analyses. Several studies have confirmed the sister relationship of tylenchids with the bacteriovorous Cephalobidae (Blaxter *et al.* 1998). The molecular datasets showed that the order Tylenchida *sensu* Siddiqi comprises lineages that largely correspond to two suborders, Hoplolaimina and Criconematina, and other taxonomic divisions by Siddiqi (2000). Molecular analysis supported the classical hypothesis of the gradual evolution of feeding types from simple forms of plant parasitism, such as root hair and epidermal feeding and ectoparasitism towards more complex forms of endoparasitism. Sedentary endoparasitism has also evolved several times

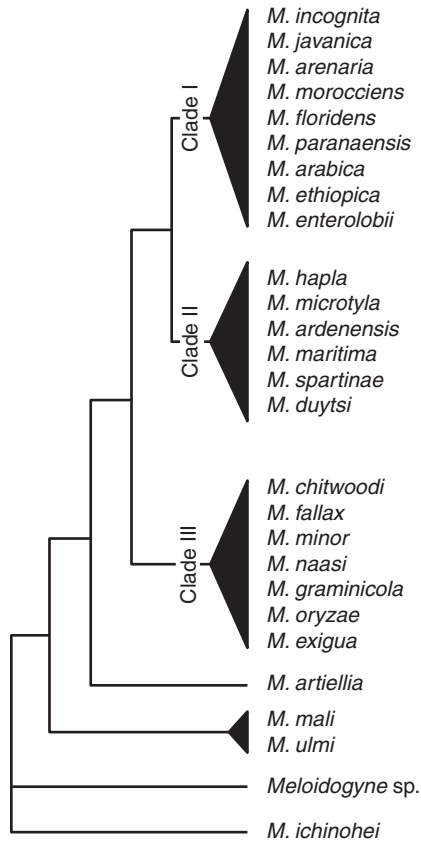


**Fig. 2.9.** An overview of phylogeny of the phylum Nematoda based on the small subunit ribosomal RNA gene. The systematic names given by De Ley and Blaxter (2004) are given, as is the 'clade' naming convention introduced by Blaxter *et al.* (1998). More recently, Helder and colleagues (Holterman *et al.*, 2006; van Megen *et al.*, 2009) have introduced a numerical clade name scheme; this is given in outlined letters below the relevant branches. Feeding mode, and animal- and plant-parasitic and vector associations, are indicated by small icons (Blaxter, 2011).

independently: (i) cyst and non-cyst nematodes of Heteroderidae probably evolved from migratory ectoparasitic nematodes; (ii) root-knot nematodes appear to be related to the false root-knot nematode *Nacobbus* and have evolved from migratory endoparasitic nematodes; and (iii) sedentary nematodes from Tylenchulidae and Sphaeronematidae (Criconematida).

#### 2.11.4. Root-knot nematodes of the family Meloidogynidae

The genus *Meloidogyne* contains more than 90 nominal species. The evolutionary relationships of root-knot nematodes have been inferred from several types of data: isozymes, DNA hybridization, DNA amplification fingerprinting, RAPD-PCR, sequencing of SSU rRNA, D2-D3 expansion segments of LSU rRNA, ITS rRNA and mtDNA. The SSU rRNA gene sequence data resolve deep relationships within *Meloidogyne*. There are several supported clades on trees generated from the SSU sequence datasets: (i) mitotic parthenogenetic species from the tropical group (*M. incognita*, *M. arenaria* and *M. javanica*) and the meiotic parthenogenetic *M. floridensis*; (ii) obligatory amphimictic *M. microtyla*, meiotic parthenogenetic *M. hapla* and two species with unknown reproductive strategy (*M. duytsi* and *M. maritima*); and (iii) meiotic parthenogenetic *M. exigua*, *M. graminophila*, *M. chitwoodi*, *M. fallax*, *M. minor* and mitotic parthenogenetic *M. oryzae*. The basal *Meloidogyne* species consist of the polyphagous *M. artiellia* and the oligophagous *M. mali*, *M. ulmi* and *M. ichinohei* (De Ley *et al.*, 2002; Holterman *et al.*, 2009) (Fig. 2.10).



**Fig. 2.10.** Schematic overview of the phylogeny of the Meloidogynidae derived from SSU rDNA sequence data (Holterman *et al.*, 2009; Bert *et al.*, 2011).

### 2.11.5. Cyst nematodes of the family Heteroderidae

Cyst nematodes are highly evolved sedentary plant parasites. Phylogenetic analysis of the ITS rRNA and D2–D3 expansion segment of the 28S rRNA gene sequences confirmed the monophyly of the subfamilies Punctoderinae and Heteroderinae with the genus *Heterodera*. The combination of molecular data with morphology of the vulval structure and the number of incisures in the lateral field of second-stage juveniles (J2) enabled six main groups within *Heterodera* to be recognized: *Avenae*, *Cyperi*, *Goettingiana*, *Humuli*, *Sacchari* and *Schachtii* groups. Close relationships were revealed between the *Avenae* and *Sacchari* groups and between the *Humuli* group and the species *H. turcomanica* and *H. salixophila*. Some inconsistencies between molecular phylogeny and earlier proposed morphological groupings may be attributed to homoplastic evolution, e.g. a bifurcated vulval cone developed independently at least three times during the evolution of cyst nematodes. Likewise, the presence of three incisures in the lateral field of J2 seems to have arisen twice independently (Subbotin *et al.*, 2001). Molecular data suggested an early divergence of tropical and temperate heteroderid species and often revealed association of nematodes with their host plants from related families (Fig. 2.11).

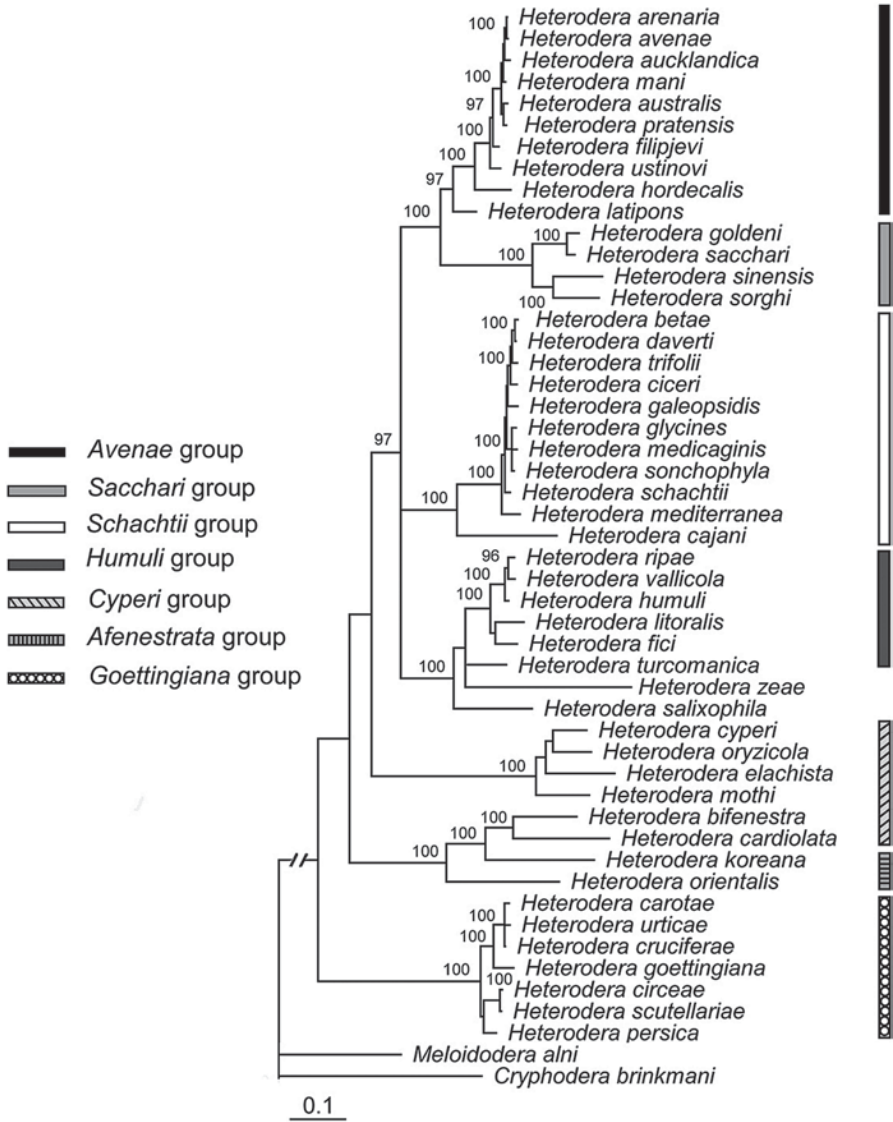
### 2.11.6. Stem and gall-forming nematodes of the family Anguinidae

The stem nematode, *D. dipsaci*, occurs as more than 20 biological races. Molecular approaches using RAPD–PCR, AFLP, PCR–RFLP and sequences of the ITS rRNA confirmed that *D. dipsaci* constitutes a complex sibling species. The phylogenetic analysis of the ITS sequences of plant-parasitic species of *Ditylenchus* revealed two main clades: (i) *D. dipsaci sensu stricto* with diploid chromosome numbers and comprising most isolates from agricultural, ornamental and several wild plants; and (ii) a complex of *Ditylenchus* species with polyploid chromosome numbers, including *D. gigas* from *Vicia faba*, *D. weischeri* and several species parasitizing various Asteraceae and a species from *Plantago maritima*. Molecular methods failed to distinguish biological races within *D. dipsaci sensu stricto*.

Over 40 nominal species of gall-forming nematodes have been described. Testing of recognized anguinid classifications using the ITS sequences strongly supported monophyly of the genus *Anguina* and paraphyly of the genera *Mesoanguina*, *Heteroanguina sensu* Chizhov & Subbotin and *Subanguina sensu* Brzeski. Molecular data demonstrate that the main anguinid groups are generally associated with host plants belonging to the same or related systematic groups. The molecular analysis supports the concept of narrow host-plant specialization of seed-gall nematodes, shows that *Anguina agrostis* causing elongate galls occurs only in one host, *Agrostis capillaries*, and reveals several undescribed species infecting other species of grass (Fig. 2.12) (Subbotin *et al.*, 2004).

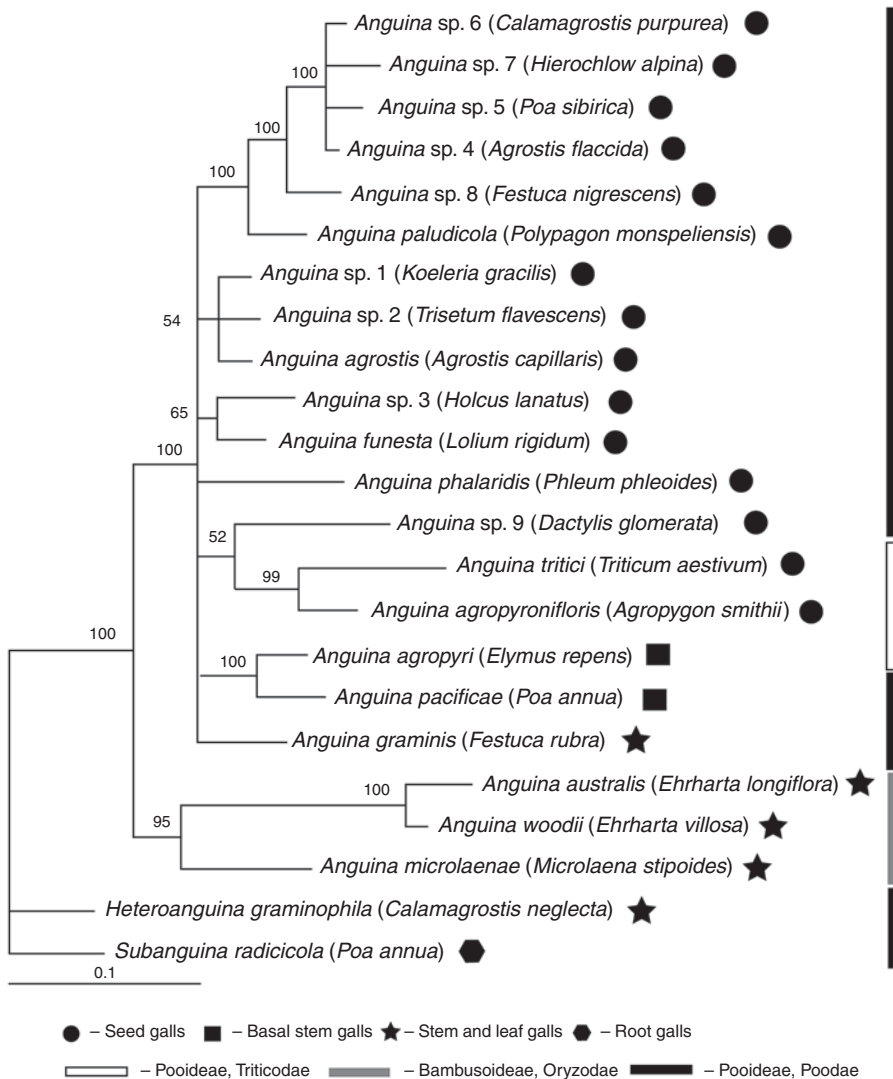
### 2.11.7. Needle nematodes of the family Longidoridae

Longidorids are a group of ectoparasitic nematodes with hundreds of species. The analysis of 18S and partial 28S rRNA gene sequences revealed several major



**Fig. 2.11.** Phylogenetic relationships among *Heterodera*: Bayesian 50% majority rule consensus tree from two runs as inferred from ITS1-5.8S-ITS2 sequences of rRNA gene alignment. Posterior probabilities more than 70% are given for appropriate clades. (Modified from Subbotin *et al.*, 2010b.)

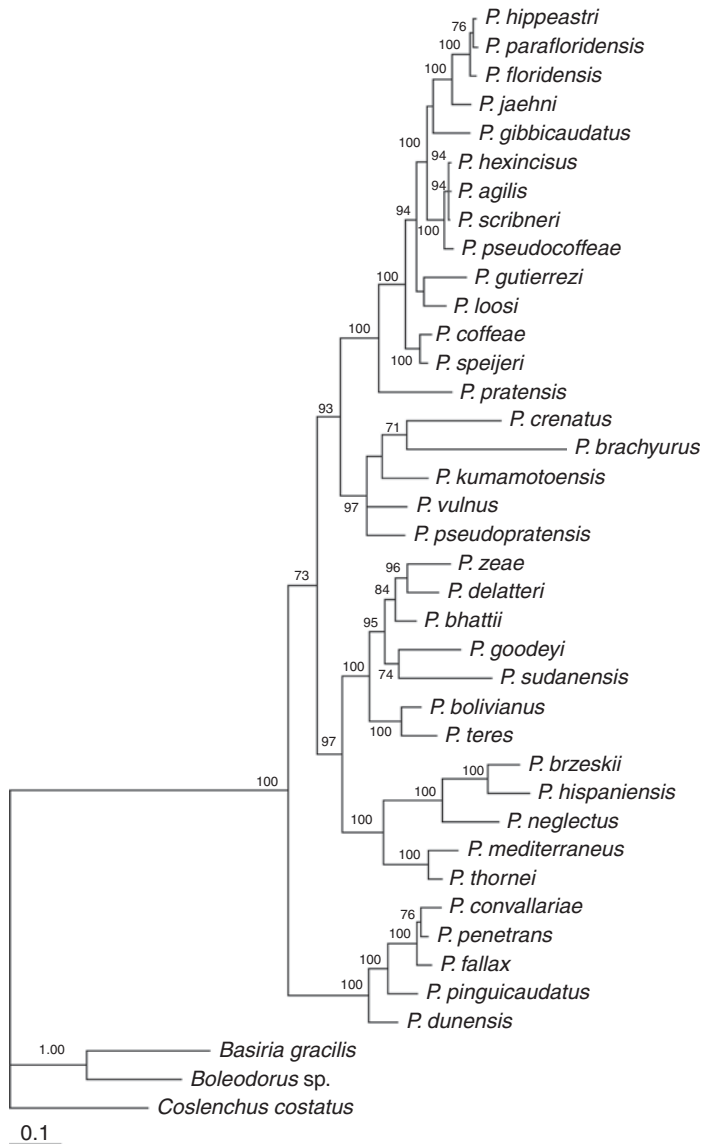
groups within Longidoridae: *Longidorus*, *Paralongidorus*, the *X. americanum* group, other *Xiphinema* species and *Xiphidorus*. The grouping of *Longidorus* species on the tree is well correlated with amphidial shapes. Although the result of the alternative phylogenetic hypotheses tests did not refute the monophyly of the genus *Xiphinema*, the species of this genus were split into two distinct clades in all trees.



**Fig. 2.12.** Phylogenetic relationships within the genus *Anguina* as inferred from the Bayesian analysis of the ITS1 rRNA gene sequences with mapping gall types and systematic position of host plants (subfamily, tribe). Posterior probabilities are given on appropriate clades. (From Subbotin and Riley, 2012.)

### 2.11.8. Root-lesion nematodes of the family Pratylenchidae

Phylogenetic analyses of the D2–D3 of 28S and 18S rRNA gene sequences of large numbers of geographically diverse isolates of genus *Pratylenchus* species confirmed that most classical morphospecies are monophyletic and revealed several species complexes. Analyses revealed at least six distinct major clades of examined *Pratylenchus* species and these clades are generally congruent with those defined



**Fig. 2.13.** Phylogeny of the genus *Pratylenchus*. Bayesian majority rule consensus tree as inferred from analysis of the D2–D3 expansion segments of 28S rRNA gene sequence alignment. Posterior probabilities are given on appropriate clades.

by characters derived from lip patterns, numbers of lip annules and shape of the spermatheca (Fig. 2.13).

### 2.11.9. Pinewood nematode and other *Bursaphelenchus* species

A phylogeny of *Bursaphelenchus* species from Europe, North America, Central America and Asia representing much of the known biological diversity in this genus



has been reconstructed using sequences of the 18S, 28S and ITS of rRNA and COI genes. Phylogenetic analyses using several methods of inference were congruent, with the greatest resolution obtained with combined datasets. Phylogenetic analysis revealed *B. abruptus* as the basal taxon among all investigated *Bursaphelenchus* species and a large number of significantly supported monophyletic groups that are largely consistent with morphological and life history variation in the genus. The genus is divided into seven groups with four incisures, four groups with three incisures and two groups with two incisures. MtDNA data were limited by non-stationary base composition and apparent saturation above the species level (Ye *et al.*, 2005).