eDNA metabarcoding:

A realistic perspective

Sophie Moinier, Utrecht University
Master Program Environmental Biology –
Natural Resource Management, Master Thesis
Supervisors: Arjen de Groot (Alterra-WUR) & Merel Soons (UU)
January 14, 2013, Utrecht

Dear reader,

This thesis has been written as a part of the Environmental Biology/Natural Resources Management master track of Utrecht University. The subject of this thesis has a molecular character, despite the fact that my specialization is not particularly in the molecular ecology field. Extra reading was needed to get familiar with the principles of (e)DNA (meta)barcoding and the techniques used for (e)DNA (meta)barcoding. It was a little bit of extra work, but it was worth the effort; it was very interesting and I learned many new things while writing this thesis.

I want to thank Arjen, my supervisor, for his good advice, his knowledge and for letting me get a glimpse of a research field I was hitherto unfamiliar with.

I hope you enjoy reading this thesis!

Sophie Moinier,
Utrecht, January 2013

In the glossary on page 5, some terms that have been used in this thesis are listed, together with a short explanation. When a term listed in the glossary is used for the first time in the text, it is written in *italics*.
In the Appendix, a timeline shows the references in a chronological order.
(e)DNA (meta)barcoding refers to eDNA barcoding, eDNA metabarcoding and metabarcoding.

Glossary

Cryptic taxa: different taxa that are indistinguishable based on morphology.
DNA polymerase: enzyme which replicates DNA.
DNA sequencing: the process of determining the precise order of nucleotides within a DNA molecule.
eDNA: DNA that can be extracted from environmental samples without first isolating any target organisms.
(fluorescent) capillary electrophoresis: technique used to separate DNA strands of different lengths.
Genetic variability: a measure for the tendency of individual genotypes in a population to vary from one another.
GUI: windows-screen with clickable options.
Heteroplasmy: several (mt) genomes existing within the same cell.
Indel: collective term for a group of DNA mutations; INsertions of nucleotides and DELetions of nucleotides.
Intron: non-coding DNA or RNA, which is removed through a process called splicing.
Introgression: gene flow from one species to another by backcrossing of hybrid species with one of its parents species.
Metabarcoding: the identification of multiple taxa via one single experiment.
Mitochondrium: part of the animal cell, which facilitates the cell's energy. Mitochondria are originally bacteria which entered the (animal) cell through a process called endosymbiosis. This is why mitochondria have their own (haploid) DNA.
NGS: Next-Generation Sequencing; a collective term for new, faster and cheaper methods for DNA sequencing.
Nucleotide: building block of DNA.
PCR: Polymerase Chain Reaction; a biochemical technique to amplify a single or a few copies of DNA.
Phenotypic plasticity: the capacity of a single genotype to exhibit variable phenotypes in different environments.
Primer: small strand of DNA used to start DNA replication. Also used for PCR.
Specimen: quantity of material used for examination.

Table of Contents

Chapter 1: Introduction

In 2003, Hebert and colleagues published an article on the use of the mitochondrial gene cytochrome *c* oxidase (CO1) for identifying animal species. After sequencing the CO1 gene, the gene was compared to a reference database (containing previously sequenced CO1 genes of known species) to find a match. The use of DNA for taxonomic purposes was not new, just as the use of short DNA sequences for species identification (i.e. Valentini et al 2009, Taylor & Harris 2012). However, the idea of Hebert et al (2003) was revolutionary in the sense that the researchers proposed that a single standardized DNA region (a barcode) could be used to identify all kinds of species. In the years after the famous publication, many scientists wrote articles on the promises of DNA barcoding for (taxonomic) research and the many benefits of this new, revolutionary technique (i.e. Janzen 2004, Stoeckle 2003, Hebert & Gregory 2005, Gregory 2005, Marshall 2005). Some other scientists displayed a more skeptical and negative view (i.e. Will et al. 2005, Will & Rubinoff 2004, Ebach & Holdrege 2005, Cameron et al 2006). Their main remarks on the DNA barcoding movement were that DNA barcoding was not new and revolutionary at all, it could not answer the hard questions of taxonomy, it was too expensive and by no means a replacement for taxonomy. Instead, Will et al. proposed an integrative approach which combined DNA barcoding with the standard taxonomic identification techniques without replacing them (Will et al 2005).

Even though DNA was already used as a tool for species identification, up until 2003 most of the taxonomical research was still morphology-based and relied much on morphological diagnoses (Hebert et al 2003). Determining species can be a hard task; Janzen (2004) used the example that "neither [he], nor the other millions of wild biodiversity users, [could] carry in their pocket the tens of thousands of pages of taxonomic descriptions, keys and images, and their authors". All the taxonomists together have a lot of knowledge on the world's biodiversity, but a single taxonomist is not able to use this for identifying all the organisms in a certain ecosystem because he only has knowledge on roughly 0,01% of the total biodiversity in the world (Hebert et al 2003). Furthermore, morphological identification could be impossible when morphological characteristics are poorly defined or not present at all. Hebert and colleagues listed four important limitations of taxonomical research. The morphology-based approach could lead to (Hebert et al. 2003):

1) Incorrect identification of species due to *genetic variability* and *phenotypic plasticity*.
2) Overlooking of *cryptic taxa*.
3) Non-identification of individuals due to the fact that many methods are solely applicable to a certain life stage or gender.
4) Misdiagnosis due to the high level of expertise needed.

The use of a standardized DNA region to identify species (whether plant, animal, fungus or bacteria) could be the perfect solution to the limitations mentioned above, and it could help making taxonomic research easier and more efficient; scientists with a specialization in different fields can work together on a shared project (Casiraghi et al 2010). A relatively short strand of DNA (of between 500 and 1000 bp) potentially contains enough information to identify millions of species (Stoeckle 2003, Hebert et al

2003). Combined with the fact that *DNA sequencing* becomes cheaper, faster and more efficient (due to the discovery of *NGS*; Next-Generation Sequencing), these two developments inspired the scientific world to announce the "Golden Age" of DNA barcoding in the early 2000's (Hajibabaei, 2012, Valentini et al 2009).

In the early days, DNA barcoding focused mainly on taxonomic research. Nowadays, DNA barcoding could be used for a wide variety of purposes and the research field has diversified itself. Not only DNA of intact and isolated species is used, but also DNA that is shed into the environment by organisms (e.g DNA from skin, nails, hairs, waste products, etc.). This form of DNA is called environmental DNA, or *eDNA*, and is usually highly degraded. It could be found in environmental samples like air, water or soil and it could be extracted without isolating the organisms (Taberlet et al. 2012[a]). The eDNA fragments are shorter than normal DNA fragments and thus an adjusted barcoding method is needed. Specifically, it requires the use of shorter barcodes (Hajibabaei et al 2006). A second direction of development concerns the use of environmental samples, not only to identify one single species, but to identify a wide variety of species in one experiment. This is called *metabarcoding*. The development of metabarcoding approaches was aided by the discovery of NGS, which allows parallel reading of DNA sequences from a single DNA-extract without a necessity for cloning (i.e. Taberlet et al. 2012[a,c], Hajibabaei 2012). In this respect, metabarcoding differs from normal DNA barcoding in the sense that classic DNA barcoding aims to identify intact *specimens* (e.g. with complete genomes) up to species level, and metabarcoding aims to identify degraded DNA samples (eDNA) up to family level or higher.

Due to the natures of both eDNA barcoding and metabarcoding, these methods could be very useful for ecological purposes. The methods could also be combined; this is called eDNA metabarcoding. The possibility of identifying all the species present in an environmental sample opens doors to new approaches in ecological research. Therefore, eDNA barcoding, metabarcoding as well as eDNA metabarcoding will all be addressed in this thesis.

(e)DNA (meta)barcoding could be a good addition to the already used taxonomic methods instead of a replacement, and an integrative approach is widely adopted by many scientists in recent years (i.e. Hajibabaei et al. 2008). Due to the development of NGS and other new techniques, an ever growing number of applications of (e)DNA (meta)barcoding appears:

- eDNA barcoding could be used for conservation biology; for biodiversity surveys and to find traces of nearly extinct species (Stoeckle 2003, Taberlet 2012[a], Callaway 2012, Valentini et al. 2009).
- DNA from the stomach or faeces of animals could be analyzed to determine the interactions between species and it could be used for diet analysis, this is an example of metabarcoding (Stoeckle 2003, Taberlet 2012[a], Valentini et al. 2009).
- Changes in species distribution and stability of the niche can be investigated with eDNA barcoding (Taberlet et al. 2012[a])
- eDNA barcodes could be used for biosecurity; it has the potential to accurately identify invasive species (Armstrong and Ball, 2005, Valentini et al. 2009)

- DNA from environmental samples could be analyzed for biomonitoring (i.e. Hajibabaei 2012, Taberlet et al 2012[a], Valentini et al. 2009)
- (e)DNA could be used to monitor illegal trade and by-products (Valentini et al. 2009)
- interspecies-relations of cryptic species could be analyzed with eDNA (Valentini et al. 2009)

Based on the examples written above, (e)DNA (meta)barcoding seems a very promising research field. However, there are still some gaps in the knowledge on (e)DNA (meta)barcoding. There are also some limitations concerning the different techniques involved in (e)DNA (meta)barcoding. Current limitations are the dependence of NGS on *PCR* to amplify the barcodes and the dependence on the availability of money to construct reference libraries (Taberlet et al. 2012[c]). Scientists work hard to fill the gaps, but the big question remains whether this will be enough or not. Many promises have been made concerning the future prospects on the use and application of (e)DNA (meta)barcoding, and the objective of this thesis is to investigate what the main promises are and which promises already have been achieved. These two questions will help to find a realistic perspective on the (future) uses of (e)DNA (meta)barcoding. To summarize:

Main research question:
What is a realistic perspective on the (future) uses of (e)DNA (meta)barcoding?

Subquestions:
A: what are the main promises on (e)DNA (meta)barcoding?
B: which of these promises have already been achieved?

This thesis aims to give an overview of the discoveries and developments in the DNA barcoding research field during the past decade. The articles used as references in this thesis are shown in a timeline (see Appendix 1) to visualize the development of (e)DNA (meta)barcoding through time and the development of associated techniques, like NGS. The timeline serves as a 'backbone'; the first chapters represent the first period after the discovery of DNA barcoding, and the subsequent chapters explain the development of DNA barcoding through time, ending with the current situation in chapters 4 and 5.

Chapter 2 will explain the general method of DNA barcoding, dating from before the discovery of NGS. Secondly, in chapter 3 the promises made on DNA barcoding and (e)DNA (meta)barcoding are listed, following a chronological order. Chapter 4 gives an overview of which promises have been achieved up until now. In chapter 5, still remaining challenges and gaps that have to be filled will be explained and an example of present research will be given to illustrate the current state of research, with the limitations included. In the last chapter, research questions will be answered.

Chapter 2: DNA barcoding

This chapter will give an overview of DNA barcoding in general. The first part of this chapter focuses on the basic approach of DNA barcoding. Subsequently, DNA barcodes and the techniques used for DNA barcoding will be explained.

2.1 Basic approach

A small strand of DNA – potentially – contains enough information to identify millions of species. Paul Hebert was one of the first scientists who recognized this and who used this knowledge to develop a revolutionary and unique technique for species identification. And, his method proved to be successful: numerous studies have shown that DNA barcoding can be effective in several groups; birds, fish, cowries, spiders and several arrays of Lepidoptera (Hajibabaei et al. 2008).
The main idea behind DNA barcoding is that new species can be identified by comparing (part of) their DNA to DNA from other species. This reference-DNA could be collected in for instance a reference library. If the DNA (i.e. the barcode) of the target species differs enough from the reference-DNA, the target species could be considered as a new or different species. When a target species is collected, its barcode-region first needs to be amplified and sequenced. The barcode is then compared to the reference library to determine if the barcode belongs to a "new" species (which was not yet present in the reference library) or if the barcode belongs to a species already present in the library.

2.2 The optimal barcode

In their search for a DNA region suitable for barcoding, Hebert and his colleagues discovered that using the *mitochondrial* genome gives better results than using the normal nuclear genome. This is because the mitochondrial genome does not contain *introns* (which has a higher mutation rate than used DNA), there is little recombination and mitochondria are inherited maternally (haploid inheritance) (Hebert et al 2003). Maternal inheritance is beneficial for the DNA sequencing procedure. When nuclear DNA is used, cloning is needed to separate the two potentially different alleles of the diploid DNA (one chromosome is inherited maternally, one chromosome is inherited paternally). When a potentially heterozygous fragment is sequenced, this may lead to ambiguous results as the two different codes may be mixed up (Metzker, 2005).

Hebert and his colleagues evaluated genes of the mitochondrial genome on their suitability for DNA barcoding. Ribosomal genes were excluded due to the high number of *indels* found in those genes, and the scientists focused on a gene called the cytochrome *c* oxidase I – gene (CO1). The use of this gene for DNA barcoding has a lot of benefits:

- The CO1-gene has a high rate of molecular evolution, due to the high incidence of base substitutions at 3rd position *nucleotides*. This evolution is fast enough to allow the identification (or discrimination) of closely related species.
- When the barcodes are amplified, the *primers* that are used to start amplification are universal (e.g. several primers can be used for a lot of different species) and the primers are robust (e.g.

they remain effective under variable conditions). This makes the primers applicable to a lot of different (animal) groups.

- The changes in amino acid sequence occur more slowly in the CO1-gene than in any other mitochondrial gene.

For animals (and for metazoans in particular), DNA barcoding with the CO1-gene seems to work really well; 95% of the animal species possess unique CO1 barcode sequences (Hajibabaei et al 2008). However, more research on DNA barcoding revealed that the CO1-gene could not be used for all the different taxa. In plants, for example, molecular evolution in the CO1-gene goes too slowly and therefore it is not suitable for species identification. Furthermore, the CO1-gene has slow substitution rates and intra-molecular recombination occurs (Casiraghi et al. 2010).

Criteria for the ideal barcoding system:

1. The gene region sequenced should be nearly identical among individuals of the same species, but different between species,
2. It should be standardized, with the same DNA region used for different taxonomic groups.
3. The target DNA region should contain enough phylogenetic information to easily assign unknown or not yet 'barcoded' species to their taxonomic group (genus, family, etc,).
4. It should be extremely robust, with highly conserved priming sites and highly reliable DNA amplifications and sequencing.
5. The target DNA region should be able to allow amplification of eDNA (this will be explained in chapter 4).

Adapted from Valentini et al. 2009

When DNA barcoding is concerned, there are two different definitions. The first one is DNA barcoding *sensu stricto*, and the second one is DNA barcoding *sensu lato*. With DNA barcoding *sensu stricto*, a single standardized DNA fragment is used to identify a sample to species level. With DNA barcoding *sensu lato*, any DNA fragment could be used to identify a sample, to any taxonomical level (Valentini et al. 2009).

There is still an ongoing search for new – and better – barcodes, seeing the fact that the mitochondrial CO1-gene is a good start, but it is not the universal barcode which could be used for all the different existing taxa. Two chloroplast genes, rbcL and trnLF, have for instance been elected as barcode for plants (CBOL plant working group, PNAS 2010), but augmentation with additional barcodes is necessary in certain groups (e.g. De Groot et al. 2011).

Valentini et al. (2009) summarized five criteria for an ideal barcoding system, and these criteria are shown in the box on the left. The criteria are very similar to the benefits of the CO1-gene, mentioned above, with a few extra points added.

The use of different definitions on DNA barcoding requires a focus on different criteria. Criteria 1, 2 and 3 are important when DNA barcoding *sensu stricto* is concerned, and criteria 4 and 5 are important when DNA barcoding *sensu lato* is concerned.

2.3 Data gathering

The first step in the process of DNA barcoding is data gathering. (Environmental) samples need to be collected, and afterwards a series of molecular techniques is used to extract the DNA and make it suitable for further analysis. When an environmental sample is collected and stored in such a way that its DNA is preserved, the DNA needs to be isolated. DNA isolation can be performed with two different

groups of methods. The first group consists of DNA-release methods. DNA-release methods rapidly release DNA into a solution. However, released-based methods are not very sensitive and the DNA cannot be stored for more than a year. The second group consists of DNA-extraction methods, and it works by binding target DNA to a membrane or it chemically fractionates the DNA to purify it. The isolated DNA can be stored for a longer period of time, but this method can be time consuming and sometimes toxic materials are used (Hajibabaei et al. 2005)

When the DNA of a sample is isolated, the DNA barcode has to be amplified. This is done with a technique called PCR, Polymerase Chain Reaction. First, the double-stranded DNA is denaturated to create single stranded DNA. A primer is added to the solution, which binds to the target region (in this case the barcode) (Alberts et al. 2002). *DNA polymerase* replicates the DNA. Primers are very important for a PCR; minor adjustments in primer sequences can have a large impact (Hajibabaei et al. 2005).

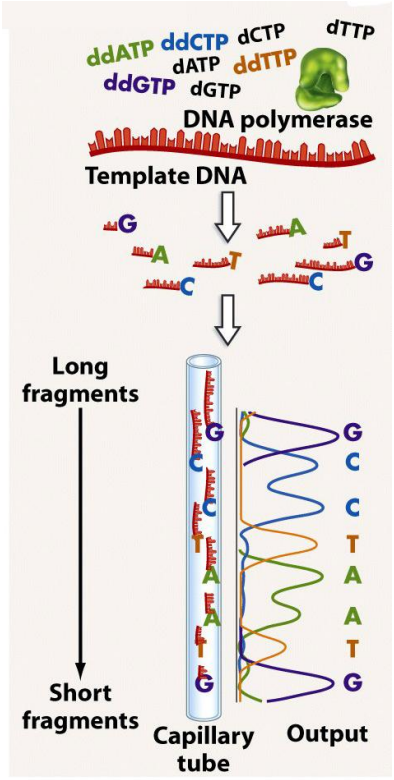

Figure 1: Sanger sequencing with Fluorescent capillary electrophoresis. Adapted from: www.uic.edu/classes/bios/bios100/lextures/techniques.htm

DNA that has been isolated and amplified can be sequenced; this is the next step in the data gathering process. One of the first and basic sequencing-methods is the Sanger-sequencing method (Sanger, 1977). For this type of sequencing, a primer is added to a solution with amplified (target) DNA. The enzyme DNA polymerase is also added, plus the nucleotides to elongate the newly created strand of DNA. The last thing that is added to the solution is a replication terminator. Four different replication terminators are added; G, T, A and C terminators. First, a primer binds to the DNA. Secondly, DNA polymerase binds to this primer and starts replicating the DNA, but stops when it incorporates a replication terminator in the newly created DNA strand. The replication terminators (ddntp's) are nucleotides which lack a hydroxyl group (-OH) and normal nucleotides cannot bind to the growing DNA strand after a replication terminator is incorporated. Furthermore, the replication terminators are tagged with a fluorescent dye (which emits fluorescent light) to automate the process. Each of the four different terminators emits a different color. When DNA replication is ended, the solution contains a lot of different DNA fragments of different lengths. The DNA fragments are separated by capillary electrophoresis, and the light emitted by each fragment is detected. The color of each fragment indicates what kind of replication terminator was incorporated, i.e. the last nucleotide (A, C, G or T) incorporated the DNA fragment. The sequence of the DNA fragment is revealed by the sequence of the detected colors (Metzker, 2005). This Sanger sequencing method with *fluorescent capillary electrophoresis* is visualized in figure 1.
The Sanger-sequencing method was invented by F. Sanger in 1977. New sequencing techniques enable faster and more efficient DNA sequencing. These so-called Next-Generation Sequencing techniques are especially suitable for the barcoding of degraded DNA or environmental DNA, and examples of these techniques will be discussed in chapter 4.

2.4 Data storage

20 million barcodes are needed for the ± 1.5 million described species, when 10 barcodes per species are desired (Hajiababaei et al. 2005). This requires a very large reference library. In 2004, the Consortium of the Barcode of Life was established. The consortium aims to develop a standard protocol for DNA barcoding and it aims to construct a DNA barcode library. The International Barcode of Life Project (iBOL) is a huge collaboration of 25 countries, and it aims to construct a reference library for all multi-cellular life. Among others 500.000 species will be barcoded to construct this library. Currently, there are two large databases available; the Barcode of Life Database (BOLD) and The International Nucleotide Sequence Database Collaborative (The Barcode of Life, www.barcodeoflife.org).

2.5 Data analysis

When the DNA is gathered and stored, the newly created DNA barcoding samples need to be analyzed in order to identify them. DNA barcoding does not define species but discriminates between them, and the question is when a molecular entity (i.e. a DNA barcode) should be considered as a separate species when it is compared to a reference library. The use of suitable bioinformatics tools is crucial for this analysis (Casiraghi et al. 2010).

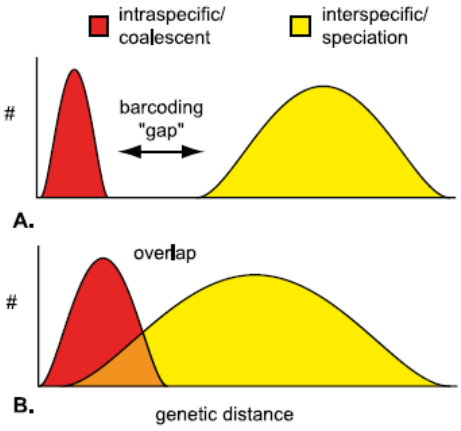

Figure 2: graphical representation of the barcoding gap. Meyer and Paulay 2005.

At the heart of the DNA barcoding method lays the basic assumption that the genetic variation of the chosen DNA region – the barcode – is bigger between species than within species (Hajibababei 2008, Elias et al 2007, Hebert et al 2004, Casiraghi et al 2010). DNA of closely related species is more similar than DNA of distant related species, and DNA of different individuals within a certain species shows so little variation that both individuals can be considered as one species instead of individuals from different species. However, there is a difficulty concerning this assumption: if there is overlap between the variations found within species and between species, it is hard to discriminate between species. In that case, DNA barcoding is less effective. Figure 2b shows what happens when there is overlap between intraspecific genetic variation (variation within species) and interspecific genetic variation (variation between species). The orange-colored area represents the overlap and, in other words, it represents the uncertainty of the DNA barcoding method for those particular species when a certain barcode is used. Figure 2a shows the desired situation. In this situation, the difference between intraspecific variation and interspecific variation is so big, that there is a gap in between; this is called the "barcoding gap" (i.e. Meyer and Paulay, 2005). Hebert et al (2004) introduced the 10-fold rule, which is (although sometimes criticized) now commonly used in the DNA barcoding movement: the barcoding gap – which is the difference between intra- and interspecific variation) – should be ten times the intraspecific variation.

The question whether two samples are different enough to assign them to two different species, is open for debate. The approach which uses a threshold value to discriminate between species is called the threshold or distance-based approach, and it searches for (pair-wise) similarities between an unknown DNA sequence and known DNA sequences stored in a reference library. The number of identical nucleotides determines whether or not the unknown DNA sequence (i.e. the barcode) belongs to a new species or not. If the similarity between the unknown DNA sequence and a species found in the reference library is high enough, both DNA sequences can be assigned to the same species (de Groot et al. 2011, Casiraghi et al. 2010). The distance-based approach is among others used by the Barcode of Life Database (BOLD) (Casiraghi et al. 2010). The benefits of this method are that it is fast and that it does not require knowledge on population structure or phylogenetic relationships. Furthermore, this method could also be useful when the reference library is limited due to for instance missing species (Casiraghi et al. 2010, de Groot et al. 2011).

Another example of a method that is used as a species discrimination tool for DNA barcoding is the method of character-based parsimony analysis. With this method, a phylogenetic tree is reconstructed, in which the unknown DNA sequence either represents a new monophyletic group or is assigned to one. The method uses the assumption that the most plausible outcome is the tree with the least evolutionary changes needed to construct it (deGroot et al. 2011) This method demands high computational power, and therefore it can be very time-consuming (Casiraghi et al. 2010).

Both of the methods have some drawbacks when used for species assignment. When the reference database is incorrect or incomplete, species could be clustered wrongly when character based parsimony analysis is used. A monophyletic group could be created while the DNA differs only slightly from other species in the database. However, also the distance based approach has some drawbacks. Species that resemble each other closely according to the distance based method might in fact be more different because they differ on crucial nucleotide positions (e.g. 3rd position nucleotides). To achieve the best results, the best approach seems to be to combine both of the methods (de Groot et al. 2011).

Chapter 3: (e)DNA (meta)barcoding, *promises*

From the first publication on DNA barcoding in 2003 onwards, (e)DNA (meta)barcoding is said to be a very promising technique for taxonomy and ecology. This chapter gives an overview of the promises that have been made in the past decade. The promises are put in a chronological order, starting with the promises made by Hebert et al. (2003) on classic DNA barcoding and ending with promises made on eDNA barcoding and metabarcoding in later years.

3.1 The early days of DNA barcoding

The article of Hebert and his colleagues (2003) is generally acknowledged to be the first article in which DNA barcoding is described the way it is used nowadays. Hebert et al. wrote in their article that it was time for a new approach for taxon recognition. The use of a mitochondrial gene for the identification of animal species was their solution for the problems related to morphological (mis)identification. The researchers compared the CO1-gene used in their experiments with a barcode which can be found on for instance groceries in the grocery store. Ideally, this barcode just needed to be "scanned" (i.e. sequenced and compared to a reference library) to identify the product (or specimen) and get access to essential information related to this product (Hebert et al. 2003).

DNA barcoding seemed to work really well and Hebert et al. concluded that, based on their research, DNA barcoding could be a very promising identification technique for the future: "microgenomic identification systems, which permit life's discrimination through the analysis of a small segment of the genome, represent the one extremely promising approach to the diagnosis of biological biodiversity". Hebert and his colleagues were convinced of the fact that the use of DNA barcoding could be the only possibility to identify species with a sustainable approach in the future. The use of the mitochondrial CO1-gene for identification of animals is – if the method is well developed – an accessible, reliable and cost-effective solution for the problems related to species identification (Hebert et al. 2003).

The presentation of DNA barcoding as a concept in 2003 has caused great uproar in the scientific world. Some researchers were so excited about the concept that they wrote science fiction-like articles which described visions in which all the biodiversity in the world could be identified in no-time, on the spot, with some sort of "tricoder" or portable barcoding device (just like in Star Trek). According to Janzen (2004) and Savolainen et al. (2005), this scenario could become reality if the DNA barcoding technique would be further developed. It would potentially take just 3 years to develop such a device, and it could be linked to the internet to link all kinds of information to a specific DNA sequence (Savolainen et al. 2005). Three developments would be needed for the realization of such a scenario: first, a "miniaturization" needed to happen in science; large devices needed to become so small that people could carry them in their pockets. Secondly, a reference-library needed to be developed (Janzen 2004). The third crucial development according to Janzen (2004) was a feedback-loop, where money is earned every time the barcoding device is used to sequence a DNA barcode.

"The time is ripe for a barcoder", is Janzen's overall conclusion. The correct use of DNA barcodes for species identification could be a tool for sharing the knowledge of species-level identification with all the people in the world. In this way, knowledge on biodiversity becomes available to everyone (Janzen, 2004).

The examples listed above are very optimistic, but may not be as unrealistic as they seem. Most of the proponents of DNA barcoding agreed on a few major advantages of DNA barcoding and these advantages made DNA barcoding a very promising identification tool for the future (e.g. Stoeckle 2003, Janzen 2004, Hebert et al 2003, Savolainen et al. 2005). DNA barcoding could be a fast and accurate approach for the discovery and identification of species. Perhaps not with a handheld barcoder, but the method could still be an improvement compared to other techniques that were used in the first half of the years 2000. DNA barcoding could be a cost effective solution for species identification; due to the ever decreasing costs of DNA sequencing analysis, it became easier and cheaper to sequence DNA and to analyze it. DNA barcoding could also help to make biodiversity and the knowledge on biodiversity more accessible to everyone (Stoeckle 2003, Hebert et al. 2003, Savolainen et al. 2005, Janzen 2004)

In the early days, applications of DNA barcoding mainly focused on the research field of taxonomy; DNA barcoding could benefit taxonomic science and most of the proponents of DNA barcoding argued that it would by no means compete with taxonomy (i.e. Savolainen et al. 2005, Janzen 2004, Hebert et al. 2003, Hebert and Gregory 2005, Gregory 2005). Some researchers proposed another, more fundamental approach, in which the use of DNA sequences for identification would form the basis of a whole new taxonomic system (Tautz et al. 2003). However, not many researchers supported this idea and therefore further discussion of this subject is not within the scope of this thesis.

3.2 The emergence of (e)DNA (meta)barcoding & the development NGS

The promises on DNA barcoding were quite general in the early days. At the time, relatively little was known about DNA barcoding because it was a new technique. The more the scientific world discovered on the applications of DNA barcoding and the techniques used for DNA barcoding, the more knowledge gaps could be filled. However, some limitations remained. One of the critical points of DNA barcoding is its inability to infer species boundaries, and it is unsure whether DNA barcoding could work well enough to identify species up to species level (Moritz & Cicero, 2004). For the application of DNA barcoding in ecology, the possibility to identify specimens up to species level is not crucial for achieving good results. When the object of the ecological research is to create an overview of the biodiversity in a certain soil sample, identifying specimens up to genus or family level could also give satisfying results. Combined with the fact that the development of NGS made it possible to sequence large quantities of short DNA fragments in a short amount of time, this led to two new branches within the DNA barcoding research field: eDNA barcoding and metabarcoding (both of the methods could also be combined; eDNA metabarcoding).

With eDNA barcoding, identification of species is possible even if only part of the species is present, and eDNA could be used for identification of species which cannot be morphologically identified (e.g. soil organisms) and which could be of crucial importance to an ecosystem (Valentini et al. 2009).

Metabarcoding emerged as a result of the development of NGS: multiple specimens found in one sample could now be sequenced simultaneously (Shokralla et al. 2012). When both of the methods are combined (eDNA metabarcoding), it should be possible to identify all the species present in an environmental sample (e.g. soil, water, faeces, etc.) and to determine in which quantities these identified species occurred (Taberlet et al. 2012[c]).

Scientists discovered that (e)DNA (meta)barcoding could be used for a wide variety of purposes; not only in the field of taxonomy but also in other research fields, such as ecology (Valentini et al. 2009). (e)DNA (meta)barcoding could be a good addition to the research field of ecology and it will probably be integrated in a growing amount of ecological studies (Taberlet et al. 2012[a]).

Chapter 4: (e)DNA (meta)barcoding, *where are we now?*

In previous chapters, both DNA barcoding and (e)DNA (meta)barcoding were introduced. However, in this chapter, focus will be just on metabarcoding, both for normal DNA and eDNA. Because of the development of NGS and other new tools, the research field of DNA barcoding has diversified itself. New opportunities opened up and a growing number of techniques are developed to improve the new (e)DNA metabarcoding method. In this chapter, an overview will be given of the present state of (e)DNA metabarcoding research field.

Microbiologists started with the use of eDNA in order to study microbial biodiversity. There are many bacterial and fungal species that cannot be cultivated, and with the use of eDNA, these researchers could perform experiments in the field and study organisms found in environmental samples (Taberlet et al. 2012[a]). Environmental samples can be divided in two groups: aquatic environmental samples and terrestrial environmental samples (Taberlet et al. 2012[a]). Because the techniques used for these groups of samples and the limitations which occur when these groups are analyzed differ between the two groups, this thesis will mainly focus on terrestrial environmental samples, i.e. soil samples.

4.1 Minibarcodes

eDNA differs from ordinary nuclear DNA in the sense that it is usually degraded. As a consequence, a different approach is needed for species identification and it involves the use of different kinds of barcodes. With classic DNA barcoding, DNA fragments of 500-1000bp are used as barcodes; the mitochondrial CO1-gene which is used for the identification of animal species, is ± 650bp long (Hebert et al, 2003). However, extracting a 500-1000bp fragment from eDNA could be difficult, as the degraded eDNA often consists of shorter fragments. Besides, there can be a high variability in the DNA that has to be sequenced, and thus it might not be possible to use universal primers (Meusnier et al. 2008). The use of a shorter barcode, the so called minibarcode, could be a good solution to overcome these problems (Hajibabaei et al. 2006, Shokralla et al. 2011, Hajibabaei et al. 2011, Meusnier et al. 2008). Researchers found suitable primers for replicating and sequencing minibarcodes and research also revealed that DNA fragments of 100-250 bp in length were for 90-95% successful for identifying species (Meusnier et al. 2008). The ideal minibarcode is, like any barcode, a variable DNA fragment, surrounded by conserved regions. The variable region of the barcode is for the identification of the specimen and the conserved region is for the binding of primers (e.g. Bienert et al. 2012, Epp et al. 2012). For classic DNA barcoding the ideal primer is a primer which is specific for the targeted taxonomic group so sequencing of other DNA can be avoided (Epp et al. 2012). Primers used for (e)DNA metabarcoding should preferably work for a wider range of species, to enhance the amplification of all the (e)DNA found in a particular sample (Coissac et al. 2012). When eDNA is concerned, there might be a high variability between the different DNA sequences found in a sample because eDNA is usually degraded. As a consequence, there can be many different DNA fragments in one sample which all belong (possibly) to one species, and the use of different primers for each taxonomic group might be necessary (Meusnier et al. 2008). For correct

species assignment, the barcoding gap (see chapter 2) for eDNA barcoding should be bigger than when classic DNA barcoding is concerned.

Examples of minibarcodes that are often used in recent research are a fragment of the mtDNA CO1-gene (Meusnier et al. 2008) and parts of the mitochondrial 12S/16S rRNA genes (Epp et al. 2012). mtDNA has a high number of copies per cell, which makes it very suitable for barcoding of degraded (e)DNA (Epp et al. 2012).

4.2 Reference library

The minibarcodes that are used for (e)DNA (meta)barcoding are different from the standardized barcodes used for classic DNA barcoding. Therefore, well-known databases such as the BOLD database cannot be used as a reference-library. Because a large amount of (different) minibarcodes is used, a different kind of library is needed. This library could still contain sequences of many different species, but different barcodes will be used to identify different groups of organisms. The reference library would in fact be a collection of different libraries, each library being the result of a different study (Coissac et al. 2012).

4.3 Next-Generation Sequencing

The development of Sanger's DNA sequencing method in 1977 (see chapter 2) caused a lot of commotion in the scientific world. For the first time in history it was possible to sequence strands of DNA; even whole genomes could be sequenced (e.g. the Human Genome Project) However, Sanger sequencing is expensive and very time consuming (Harismendy et al. 2009). Moreover, DNA strands can only be sequenced one at a time, which makes this technique not suitable for metabarcoding since the samples used for this type of barcoding usually contain (e)DNA originating from many different species (Shokralla et al. 2012).

In the past few years, new sequencing techniques have been developed which enabled parallel sequencing of short DNA fragments in a short amount of time: Next-Generation Sequencing, or NGS (Shokralla et al. 2012). Another major benefit of these revolutionary sequencing methods is that the time-consuming cloning step to separate the two templates of nuclear DNA is bypassed (Margulies et al. 2005, Ansorge 2009). Because of the fact that with NGS, only short DNA fragments can be sequenced, the use of these DNA sequencing techniques involves the use of minibarcodes instead of using conventional barcodes. As a consequence, for metabarcoding the use of minibarcodes is needed.

There are several different platforms of Next-Generation Sequencing techniques, and these platforms can be divided in two groups: the first group uses PCR to amplify the barcodes prior to DNA sequencing. The second group uses a technique called single-molecule sequencing, and this technique does not include an amplification-step (with PCR) prior to sequencing (Shokralla et al. 2012). 454-pyrosequencing was the first available NGS technology, and it was developed in 2005 by Margulies and his colleagues. Up to this date, it is still the most commonly used NGS technique, and therefore this is the only NGS technique that will be fully explained in this thesis. Other examples of PCR-based

technologies are Illumina sequencers and Applied Biosystems SOLiD sequencer (Shokralla et al. 2012, Ansorge 2009, Mardis 2008)

Next-Generation Sequencing techniques can produce enormous amounts of reads per run (one read = one sequenced nucleotide or basepair). To illustrate this: 454-pyrosequencing could produce up to 1 billion reads per run and Illumina could produce up to 6 billion reads per run (Coissac et al. 2012) Furthermore, according to Yoccoz (2012), the amount of money needed to sequence one basepair decreased by half every 5 months. Although NGS is still quite expensive, costs are rapidly declining.

4.4 454-pyrosequencing

454-pyrosequencing uses light, which is produced each time a nucleotide is incorporated by the enzyme DNA polymerase, to monitor the order of the added nucleotides. In figure 3, a schematic overview of the workflow is shown. The workflow can be divided in two phases (adapted from Margulies et al. 2005, Ansorge 2009, Shokralla et al. 2012):

Phase A (preparatory step) :

- An (e)DNA fragment is separated into single stranded DNA (see fig. 3A, I)
- The DNA fragments bind to special beads; one fragment per bead.
- The beads are placed in oil-droplets in an emulsion
- Within each droplet: PCR amplification, this type of PCR is called emulsion-PCR (see fig. 3A, II)
- After the PCR reaction, each bead contains billions of copies of the same DNA fragment
- The beads are placed into picolitre wells in a special fibre-optic slide (see fig 3A, III)
- Small beads with sulfurylase and luciferase enzymes attached are added to the wells
- DNA polymerase (for replication of the DNA) and a primer are added to the wells
- An unlabeled nucleotide is added to start the replication of the DNA strands

Phase B (the pyrosequencing reaction):

- Each step, a different labeled nucleotide (A, T, C, or G) is added to the wells
- Incorporation of this nucleotide releases a pyrophosphate group, which is first converted to ATP by the enzyme sulfurylase, and then converted to light by the enzyme luciferase (see fig. 3B).
- Between each step is a washing step to wash away the nucleotides that have not been incorporated
- The amount of light is directly proportional to the amount of nucleotides incorporated. Amplification of the light signal is needed to improve the reliability of the signal.
- The presence of light indicates incorporation of a nucleotide. Because the identity of the incorporated nucleotide is known, the sequence of the growing DNA strand becomes visible.

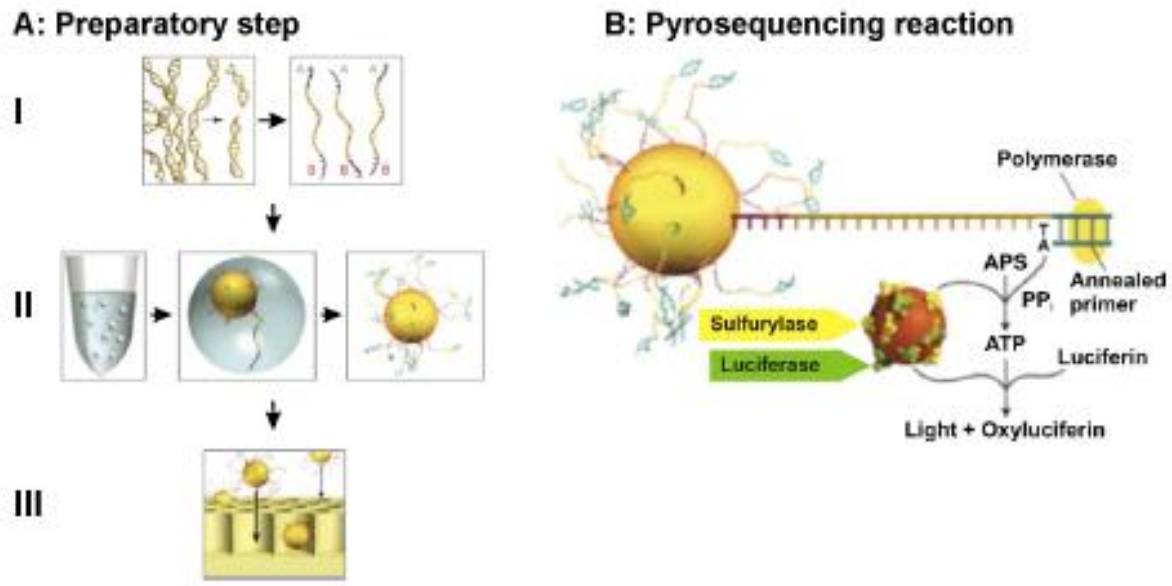

Figure 3: schematic presentation of the 454-pyrosequencing workflow. In figure 3A the creation of single-stranded DNA fragments (I), the emulsion PCR (II) and loading of the wells (III) are shown. In figure 3B the pyrosequencing reaction is shown. (adapted from Ansorge, 2009)

4.5 Bioinformatics tools

eDNA was first used for micro-organisms, and bioinformatics tools focused mainly on these organisms (Taberlet et al. 2012a). Furthermore, the bioinformatics tools used for classic DNA barcoding differ from the ones used for (e)DNA (meta)barcoding, because of the differences between the different methods. For (e)DNA (meta)barcoding, shorter barcodes are used and for metabarcoding DNA of multiple species is simultaneously amplified (compared to the classic DNA barcoding method where one DNA sample at a time is amplified and sequenced) (Coissac et al. 2012). The (e)DNA (meta)barcoding research field is relatively new, and therefore the bioinformatics tools used for these methods are new too. Many computer programs have been developed which could serve as a bioinformatics tools for (e)DNA (meta)barcoding.

Coissac and his colleagues (2012) wrote down an overview of the bioinformatics tools currently used for (e)DNA (meta)barcoding, and they divided the tools in different categories:

Suitable eDNA metabarcodes: For eDNA metabarcoding, new markers and primers are needed. Bioinformatics tools can help to select and test these markers and primers.

The multiplexing of samples for NGS: NGS platforms can produce up to 6 billion reeds per run. Because an environmental sample usually contains far less DNA than the amount of DNA which could possibly be sequenced in one run, DNA samples are sequenced multiple times, and these samples need to be tagged so the reads (i.e. the output of NGS) can be separated according to their corresponding samples. Bioinformatics tools could help with this.

Analyzing large datasets: the enormous amount of data produced by NGS, needs to be analyzed in order to use the sequencing output. Coissac and his colleagues suggested a few different analysis programs which preferably run on a unix system (e.g. MacOSx or Linux)

Errors during amplification/sequencing: during the (e)DNA (meta)barcoding process, errors could occur (e.g. PCR-generated errors or sequencing errors). Bioinformatics tools could help to filter these errors out of the results.
Species assignment: In chapter 2, some of the tools used for species assignment have been explained. During the past years, these methods have been expanded and new methods have been developed. Coissac and his colleagues divide the methods for species assignment in two different groups: one group focuses on the comparison of unknown species with a reference database, the other group works without a reference database and aims to partition the DNA sequences into distinct groups. To the first group belong the methods listed in chapter 2; the distance-based approach and the character-based parsimony analysis. Methods belonging to the second group can be used without a reference library, and an example of such a method is the ABGD-method, the Automatic Barcode Gap Discovery method (Puillandre et al. 2012)

Chapter 5: (e)DNA (meta)barcoding, *limitations*

This chapter will give an overview of gaps and the current limitations of the (e)DNA (meta)barcoding method. To conclude this chapter, an example of present research (and it's imitations) will be given.

(e)DNA (meta)barcoding is a relatively new research field and a lot of research is conducted to improve the methods. The (e)DNA (meta)barcoding approach seems very promising for the future, but in order to achieve those promises (see chapter 3), some gaps need to be filled and limitations have to be overcome. There are different limitations for metabarcoding and eDNA barcoding. Most of the issues that are limited for metabarcoding are also limited for eDNA barcoding, because both of the methods use small (degraded) DNA fragments for analysis. In the first 6 paragraphs, limitations are listed which concern both metabarcoding and eDNA. In paragraph 7, additional limitations (concerning eDNA) are listed.

5.1 Sampling problems

The first step in almost every experiment is the sampling of the material that has to be analyzed. For metabarcoding, DNA from multiple species is analyzed in one experiment, and retrieving good samples is a challenge. In chapter four, it is already explained that in this thesis, focus is on terrestrial environmental samples. There are several limitations of metabarcoding that have to be overcome in order for the metabarcoding approach to work properly when a soil sample is collected. Theoretically, for metabarcoding, all the DNA found in a sample should be analyzed. However, this is difficult and science has not found a straightforward approach to do this, yet (de Groot et al. 2012, de Groot 2012 pers. com.). When there are large differences between the sizes of the target organisms and when there is a high heterogeneity of organisms in the soil, the main challenge is to choose a sample size which is representative for the biodiversity living in the soil (de Groot et al. 2012 pers. com; Taberlet et al. 2012b). Taberlet et al. (2012b) suggests to use several kilograms of soil (retrieved from different places and depths to account for heterogeneity) and to take a subsample of this for analysis. Taberlet and his colleagues also suggest to focus on the target organisms with the coarsest heterogeneity and to adjust the sample area to this group of organisms.

Another challenge is to extract the DNA from the soil. Extracellular DNA – like eDNA – could be extracted with a saturated phosphate buffer (Taberlet et al. 2012b), but extracting the intracellular DNA, which is still inside (living) organisms, proves to be more difficult (de Groot 2012, pers. com.). Up until now, researchers first extract whole organisms and before extracting the DNA itself (de Groot et al. 2012). However, strictly spoken, this is not metabarcoding but classic DNA barcoding and often different techniques are used to extract DNA of different organisms. These techniques might conflict with each other.

5.2 Universal barcodes

Barcodes form the basis of (e)DNA (meta)barcoding and (e)DNA (meta)barcoding requires a different use of barcodes than classic DNA barcoding. The barcodes used are shorter than the classic DNA barcodes, and the use of primers is different (see chapter 4). The search for new primers and barcode regions is one of the most important remaining challenges of (e)DNA (meta)barcoding (Coissac et al. 2012). Primers are suggested for the minibarcode of the mtDNA CO1-gene (Meusnier et al 2008), but these primers do not work for vertebrates and scientists have not been able to use the same primerset in consecutive studies (Coissac et al. 2012). Furthermore, there are some remarks concerning the use of mitochondrial DNA. Because of a phenomenon called *introgression*, closely related species could possess exactly the same copies of a mt-gene. This causes errors, because the DNA sequences will be assigned to one species, instead of to two closely related species. *Heteroplasmy* could also occur, which could lead to the misidentification of species (Valentini et al. 2009). Some bioinformatics tools have been developed which could help with the search for new primers and barcodes, but still, challenges remain. Often these tools have difficulties with running large number of sequences which are available in public databases (Coissac et al. 2012).

It seems that the (e)DNA (meta)barcoding approach is currently missing some form of standardization. This standardization was one of the major benefits of the classic DNA barcoding approach, and enabled the use of a standard set of primers and the application of the same barcode region for a whole range of different target organisms. It is unsure whether it is possible to incorporate this standardization in the (e)DNA (meta)barcoding approach.

5.3 Reference libraries

The biggest limitation is that there is a need for high-quality reference databases (Yoccoz 2012, Taberlet et al. 2012c). Because of the fact that the (e)DNA (meta)barcoding method is not as standardized as classic DNA barcoding and well-known databases like the BOLD database cannot be used, a different approach is needed. Bioinformatics tools could be used to delimit species when no reference library is present; an example of such a tool is the Automatic Barcode Gap Discovery tool (Puillandre et al. 2012). Another widely used approach is the developing of group-specific reference libraries (Coissac et al. 2012). Up to this date, the number of reference libraries is still expanding and a large part of the present research on (e)DNA (meta)barcoding includes building of reference libraries (De Groot et al, 2012 unpublished data). Taberlet et al. 2012c suggests that it could also be a good idea to design the barcodes used for (e)DNA (meta)barcoding in such a way that they are mini versions of the standardized barcodes used for classic DNA barcoding, and for which are already reference libraries available.

5.4 Quantification of results

an important issue for (e)DNA metabarcoding is the quantification of results. The techniques used for analysis might affect the relative abundances of the DNA and therefore bias the results. PCR and NGS might influence the relative DNA abundances because some DNA might amplify better during PCR and

emulsion PCR (which is one of the steps prior to sequencing in NGS, see chapter 4). With PCR, differences in temperature are used to separate DNA strands and to ligate primers to the DNA. Different target groups might have different optimal temperature conditions, and therefore influence the relative DNA abundances. Secondly, relative species abundances might not be directly related to relative DNA abundance because size can differ between and within species and therefore cause differences in DNA quantities (de Groot 2012, pers. com.).

5.5 other limitations related to PCR and DNA sequencing

The (e)DNA metabarcoding approach involves the use of an array of different techniques: (e)DNA needs to be collected, the barcode regions need to be amplified (usually with PCR) and sequenced. When (e)DNA is collected and extracted, it needs to be amplified prior to sequencing. The use of a PCR amplification step is one of the major limitations of NGS. Errors could occur during the PCR reaction, and these errors could lead to misidentification of specimens. DNA polymerase could for instance produce artificial mutations, and chimeric DNA fragments could be produced which combine parts of different DNA templates (Carlsen et al. 2012). Another remaining challenge of PCR dependent sequencing lays in the fact that because of the limitations of the PCR method (see also §5.4), groups of organisms need to be analyzed separately (Taberlet et al. 2012c).

5.6 Bioinformatics tools for analyzing data

Due to the ongoing development of NGS, it is possible to produce large sequencing outputs. Because of this, data processing now became the biggest challenge in working with NGS data (e.g. Kahn 2011). Most of the tools for processing these data are primarily designed for classic DNA barcoding and it has never been tested if they also work for (e)DNA metabarcoding, so it is unsure whether these tools work or not (Coissac et al. 2012). Furthermore, data processing often requires very specialized software, and this software is often without *GUI*. There are initiatives such as Galaxy (www.galaxyproject.org) where complete analyses can be performed, shared and reproduced. This way, Galaxy serves as a tool for bioinformatics analysis when bioinformaticians are not available for researchers to work with. Some labs have their own bioinformatics specialist and other labs can benefit from cooperations with other research groups, but it is still a challenge because bioinformatics knowledge might not be available to each lab (Li & Homer 2010). Possible solutions to facilitate bioinformatics tools for more researchers could be more freeware, more computer networks and more cooperation among research groups. Li & Homer (2010) suggest cloud computing, where data could be stored and analysed on a shared cloud. However, in order for this to work properly, cooperation of entire communities is needed.

Other remaining bioinformatics challenges concern finding suitable tools for species alignment. similar sequences need to be aligned in operational clusters, and those clusters should be matched to species in the reference library (Casiraghi et al. 2010). The use of different bioinformatics tools gives different results and in order to develop a reliable eDNA metabarcoding method, research is needed to improve these species assignment methods.

5.7 Additional issues for eDNA

Paragraphs 5.1 – 5.6 addressed the limitations concerning metabarcoding, both with eDNA and with normal DNA. Using eDNA involves some extra limitations, specifically for eDNA. Although the experiment is done for aquatic environments, Dejean et al. (2011) gives a good overview of the limitations of eDNA for barcoding analysis. Most of the limitations concern the detectability of eDNA in a sample (Dejean et al. 2011). For eDNA to be detected and to be analyzed without biases, it is important that the eDNA is released in the environment and that this eDNA is not too much degraded (Dejean et al. 2011). It is unsure whether every organism present in the environment excretes eDNA. Furthermore, if it excretes eDNA, it is unsure if the quantities are sufficient for proper DNA sequencing analysis. If there are species missing it is difficult to find out how many species are missing and in which quantities they are missing (de Groot 2012 pers. com.).

Another issue is the degradation of eDNA. Environmental conditions and the original length of the eDNA affect eDNA degradation. When eDNA is too much degraded, this biases the results (Dejean et al.2011). A different challenge related to this is caused by the nature of eDNA: the eDNA found in the samples might not be of species which are still present but of species which are already extinct in the area. This also biases the results.

All the above mentioned issues related to eDNA might affect the final results quantitatively; it is unsure how the limitations affect the relation between relative species abundances and relative eDNA abundances.

The limitations mentioned in the paragraphs above are up until now preventing the (e)DNA (meta)barcoding from working perfectly. To illustrate the present state of research, in the next paragraph an example of present day research is listed.

5.8 Monitoring mite diversity in European soils

Alterra (Wageningen UR, the Netherlands) collaborates with several partners in Europe in a project which focuses on among others the development of high-throughput species identification tools for eDNA extracted from soil samples, to map soil diversity. The final goal of the project is to investigate the links between the soil diversity and the functions an ecosystem could fulfill (i.e. ecosystem services). The different collaborators of the project each focus on a different group of species, and Alterra focuses on mites (*Acari*).

First, soil samples from grasslands around Europe are collected and the mites are extracted from the soil (I in figure 4). The soil samples are used to construct a reference library (II) and to test the eDNA barcoding method (III). To create a reference library; per species, multiple specimens from multiple locations are selected. Each specimen is identified morphologically, and its DNA is extracted for DNA barcoding. As DNA barcoding-region, a fragment of the mitochondrial CO1-gene of 209 bp long is chosen. The barcode region is sequenced to create a reference library.

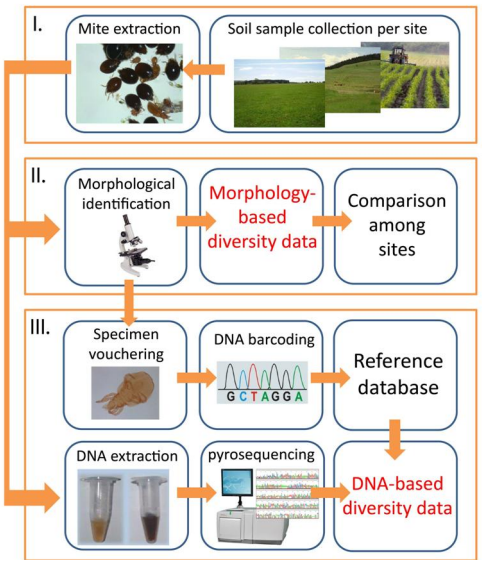

Figure 4: workflow of the (e)DNA (meta)barcoding research conducted at Alterra, Wageningen. De Groot et al. 2012

In step III, the samples collected at each location are split in two. One half is used to extract the DNA from the mites. The barcode region is amplified and with 454-pyrosequencing the DNA is sequenced. The sequenced barcodes are compared to the DNA barcodes stored in the reference library, to investigate which species are present in the samples and how many individuals of each species are present. To test if the created method works, the second half of the samples is morphologically identified and compared with the results from the DNA barcoding identification method.

So far, the researchers cannot perform eDNA metabarcoding analysis; they only succeed in "classis" DNA barcoding. eDNA and metabarcoding are (hopefully) future steps, but there are several limitations that have to be overcome. For metabarcoding, sampling issues such as mentioned In paragraph 5.1 have to be overcome. Researchers have difficulties with extracting all the target groups (including the large species) from a soil sample. The Alterra research group collaborates with other research groups in Europe, and every research group deals with the above mentioned limitations in its own way and it is hard to put all the results together for protocols for metabarcoding. For eDNA, there are some extra challenges; it is not known if all species that can be found in the soil excrete eDNA. A lot of testing is needed before eDNA metabarcoding could be performed.

Chapter 6: Conclusion/summary

The classic DNA barcoding approach, first described by Hebert et al. in 2003, was not able to fulfill all the promises made at the time. Some challenges remained; for instance the ability of DNA barcoding to identify specimens accurately up to species level. Now, almost 10 years later, the research field of DNA barcoding has diversified itself and new DNA barcoding methods have been developed. These methods were not only applicable to taxonomic research, but scientists realized that new forms of DNA barcoding (eDNA barcoding, metabarcoding and eDNA metbabarcoding) could be very well used in ecological research. For this thesis, the main promises of (e)DNA (meta)barcoding and the remaining limitations have been investigated in order to find a realistic perspective on the (e)DNA (meta)barcoding approach.

eDNA could be used for the identification of species when only part of the species is present in the environment. Another promise of eDNA barcoding is that with eDNA barcoding it should be possible to identify species which cannot be morphologically identified. eDNA usually involves shorter, degraded DNA samples and this is highly compatible to a new, advanced method for DNA sequencing: Next-Generation Sequencing. With NGS, multiple DNA samples could be analyzed simultaneously, fast and accurately. The development of NGS facilitated the emergence of (e)DNA metabarcoding, which could potentially identify all the organisms present in an ecosystem. For these new forms of DNA barcoding, new techniques had to be developed. Ordinary barcodes were too long and minibarcodes were needed; the use of NGS involves shorter read lengths than Sanger sequencing. Since the emergence of NGS techniques around 2005, DNA sequencing becomes faster and less expensive and the amount of reads per run is increasing. All this data needs to be analyzed and bioinformatics tools could help with this.

Although (e)DNA (meta)barcoding seems very promising for the future, there are still many limitations and gaps that need to be filled. Some limitations concern sampling issues: how large should the sample be in order to represent the biodiversity in the area correctly? Furthermore, the (e)DNA needs to be extracted from the soil and this could influence the results. Other limitations concern the use of different barcodes and reference databases. There is an ongoing search for new primers and barcodes, and new reference libraries have to be constructed. Techniques used for (e)DNA (meta)barcoding could also be limiting for achieving good results. PCR and NGS could influence the relative quantities of (e)DNA and the ratio DNA:species. Bioinformatics tools have to be improved to process the large amounts of data generated by NGS and for species assignment. Lastly, there are some extra limitations when (e)DNA is used. Not all species excrete eDNA in their environment, and when the eDNA is too much degraded it will bias the results. Furthermore, it is hard to be sure whether the eDNA found in the samples belongs to species which are still present in the area or if it belongs to species which are already extinct.

Due to the limitations mentioned above, it is not yet possible to conduct a reliable eDNA metabarcoding analysis. However, research is going full steam ahead and the current limitations do not seem insuperable. Combined with the fact that every year, NGS is becoming faster and cheaper, the eDNA metabarcoding approach could become available to ecologists around the world.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Molecular Biology of the Cell, 4th edition. Garland Science New York, 1463 pp

Ansorge WJ (2009) Next-generation DNA sequencing techniques. New Biotechnology vol. 25 no. 4 pp. 195-203

Armstrong KF & Ball SL (2005) DNA barcodes for Biosecurity: Invasive Species Identification. Philosophical Transactions vol. 360 no. 1462 pp. 1813-1823

Barcode of Life project, www.barcodeoflife.org, last visited on January 11 2013

Bienert F, De Danieli S, Miquel C, Coissac E, Poillot, Brun JJ, Taberlet P (2012) Tracking eartworm communities from soil DNA. Molecular Ecology vol. 21 pp. 2017-2030

Callaway E (2012) A bloody boon for conservation. Nature vol 484 pp. 424-425

Cameron S, Rubinoff D, Will K (2006) Who will actually use DNA barcoding and what will it cost? Systematic Biology vol. 55 no. 5 pp. 844-847

Carlsen T, Aas AB, Lindner D, Vrålstad T, Schumacher T, Kanserud H (2012) Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? Fungal Ecology vol. 5 pp. 747-749

Casiraghi M, Labra M, Ferri E, Galimberti a, De Mattia F (2010) DNA barcoding: a six-question tour to improve users' awareness about the method. Briefings in Bioinformatics vol. 11 no. 4 pp. 440-453

CBOL Plant Working Group, http://www.barcoding.si.edu/plant_working_group.html last visited January 7 2013

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Molecular Ecology vol. 21 pp. 1834-1847

de Groot GA, During HJ, Maas JW, Schneider H, Vogel JC, et al. (2011) Use of rbcL and trnL-F as a Two-Locus DNA Barcode for Identification of NWEuropean Ferns: An Ecological Perspective. PLoS ONE 6(1): e16371. doi:10.1371/journal.pone.0016371

De Groot GA, Laros I, Dimmers W, Beentjes K, Doorendweerd C, Faber J (2012) Monitoring Mite diversity in European Soils. EcoFINDERS project www.EcoFINDERS.eu

De Groot GA pers. com.: during the writing process of this thesis, I visited A de Groot several times, among others to talk about his research on (e)DNA (meta)barcoding and the limitations he encounters.

Dejean T, Valentini A, Duparc A, Pellier-Cuit S, Pompanon F, Taberlet P, Miaud C (2011) Persistance of environmental DNA in freshwater ecosystems. PLoS ONE 6(8): e23389. doi:10.1371/journal.pone.0023398

Ebach MC & Holdrege C (2005) DNA barcoding is no subsitute for taxonomy. Nature vol. 434 pp. 697

Elias M, Hill RI, Willmott KR, Dasmahapatra KK, Brower AVZ, Malles J, Jiggins CD (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proceedings of the Royal Society vol. 274 pp. 2881-2889

Epp LS, Boessenkool S, Bellemain EP, Haile J, Esposito A, Riaz T, Erséus C, Gusarov VI, Edwards ME, Johnsen A, Senøien HK, Hassel K, Kauserud H, Yoccoz NG, Bråthen KA, Willerslev E, Taberlet P, Coissac E, Brochmann C (2012) New environmental metabarcodes for analyzing soil DNA: potential for studying past and present ecosystems. Molecular Ecology vol. 21 pp. 1821-1833

Galaxy, www.galaxyproject.org last visited on January 13, 2013

Gregory TR (2005) DNA barcoding does not compete with taxonomy. Nature vol. 434 pp. 1067

Hajibabaei M, deWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, Macke PM, Hebert PDN (2005) Critical factors for assembling a high volume of DNA barcodes. Philosophical Transactions of the Royal Society vol. 360 pp. 1959-1967

Hajibabaei M, Smith MA, Janzen DH, Rodriguez JJ, Whitfield JB, Hebert PDN (2006) A minimalist barcode can identify a specimen whose DNA is degraded. Molecular Ecology Notes vol. 6 pp. 959-964

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2008) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. TRENDS in genetics vol. 23 no. 4 pp. 167-172

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications Using River Benthos. PLoS ONE 6(4): e17497. Doi:10.1371/journal.pone.0017497

Hajibabaei M (2012) The golden age of metasystematics. Trends in Genetics vol. 28 no. 11 pp. 535-537

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology vol. 10 no. 3 article R32

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological Identifications through DNA Barcodes. Proceedings: Biological Sciences vol. 270 no. 1512 pp. 313-321

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of Birds through DNA Barcodes. PLoS Biol 2 (10): e312

Hebert PDN & Gregory TR (2005) The Promise of DNA Barcoding for Taxonomy. Systematic Biology vol. 54 no. 5 pp. 852-859

Human Genome Project, http://ornl.gov/sci/techresources/Human_Genome/home.shtml. Last visited on January 10, 2013

Janzen DH (2004) Now Is The Time. Philosophical Transactions: Biological Sciences vol. 359 no. 1444 pp. 731-732

Kahn SD (2011) On the future of genomic data. Science vol. 331 pp. 728-729

Li H and Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Briefings in bioinformatics vol. 2 no. 5 pp. 473-483

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends in Genetics vol. 24 no. 3 pp. 133-141

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature vol. 437 pp. 376-380

Marshall E (2005) Will DNA Bar Codes Breathe Life Into Classification? Science vol. 307 pp. 1037

Metzker ML (2005) Emerging technologies in DNA sequencing. Genome Research vol. 15 pp. 1767-1776

Meusnier I, Singer GAC, Landry J-F, Hickey DA, Hebert PDN, Hajibabaei M (2008) A universal DNA mini-barcode for biodiversity analysis. BMC Genomics vol. 9 no. 214

Meyer CP & Paulay G (2005) DNA Barcoding: Error Rates Based on Comprehensive Sampling. PLoS Biol 3(12): e422

Moritz C & Cicero C (2004) DNA barcoding: promise and pitfalls. PLoS Biol 2(10): e354

Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Molecular Ecology vol. 21 pp. 1864-1877

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA vol. 74 no. 12 pp. 5463-5467

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Philosophical Transactions of the Royal Society vol. 360 pp. 1805-1811

Shokralla S, Zhou X, Janzen DH, Hallwachs W, Landry J-F, Jacobus LM, Hajibabaei M (2011) Pyrosequencing for Mini-Barcoding of Fresh and Old Museum Specimens. PLoS ONE 6(7): e21252. Doi:10.1371/journal.pone.0021252

Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Molecular Ecology vol. 21 pp. 1794-1805

Stoeckle M (2003) Taxonomy, DNA, and the Bar Code of Life. BioScience vol. 53 no. 9 pp. 796-797

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012a) Environmental DNA. Molecular Ecology vol. 21 pp. 1789-1793

Taberlet P, Prud'homme SM, Campione E, Roy J, Miquel C, Shehzad W, Gielly L, Rioux D, Choler P, Clément JC, Melodelima C, Pompanon F, Coissac E (2012b) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Molecular Ecology vol. 21 pp. 1816-1820

Taberlet P, Coissac e, Pompanon F, Brochmann C, Willerslev E (2012c) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology vol. 21 pp. 2045-2050

Taylor HR & Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular Ecology Resources vol 12 pp. 377-388

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy

Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. Trends in Ecology and Evolution vol. 24 no. 2 pp. 110-117

Will KW & Rubinoff D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics vol. 20 pp. 47-55

Will KW, Mishler BD, Wheeler QD (2005) The Perils of DNA Barcoding and the Need for Integrative Taxonomy. Systematic Biology col. 54 no. 5 pp. 844-851

Yoccoz NG (2012) The future of environmental DNA in ecology. Molecular Ecology vol. 21 pp. 2031-2038

Appendix 1: Summary for laymen

This thesis focused on the realistic perspective on eDNA metabarcoding. eDNA metabarcoding evolved from classic DNA barcoding, and this method to identify species based on differences in their DNA, was first described by Hebert and his colleagues in 2003. All living organisms have DNA in their cells, and DNA contains all the inherited genetic information in the form of A, T, C and G's. The letters (or in other words: base pairs) can form 3-letter words, and each word codes for a different amino acid. Multiple amino acids can form enzymes and other proteins, and in this way the order of base pairs in the DNA determines how living organisms are build and how they function.

The order of the base pairs can be read out, this is called DNA sequencing. It was first performed by F. Sanger in 1977, and the method he used is nowadays known as Sanger sequencing. When certain genes (that is, parts of the DNA that code for a certain protein) are sequenced in different organisms and then compared to each other, evolutionary relationships can be mapped: closely related species have more similar DNA than distantly related species. If an unknown species is found and that same gene is sequenced, it can be compared to the other DNA of known species. If it resembles closely enough to a species who's name is already known, the unknown sample can be identified.

The method described above, which uses DNA comparisons to identify species, is an example of DNA barcoding. The genes used for species comparison and identification are called "barcodes". Ideally, the barcode only needs to be "scanned" (that is, sequenced and compared to a reference database which contains barcodes of known species) to identify the species. So, when an unknown sample is collected, its DNA needs to be extracted, the barcode (so the gene which is used for comparison with the reference database) needs to be amplified, and this amplified DNA needs to be sequenced . This sequenced DNA is then compared to a reference database which contains sequenced barcodes from as much species as possible.

The presentation of DNA barcoding as a concept in 2003 has caused great uproar in the scientific world. Some researchers were so excited about the concept that they wrote scientific-like articles which described visions in which all the biodiversity in the world could be identified in no-time, on the spot, with some sort of "tricoder" or portable barcoding device (just like in Star Trek). However, not everyone was as excited and positive. After a few years of research, the DNA barcoding method still faced some major limitations. One of the most important limitations was its inability to determine samples up to species level. At the same time, researchers discovered new (sequencing)methods and applications of DNA barcoding, and the method gradually evolved.

Over the years, DNA sequencing became cheaper and faster. A new method for DNA sequencing was introduced; Next Generation Sequencing (NGS). Apart from the fact that it was cheaper and faster, it became possible to sequence very short strands of DNA and multiple different DNA strands could be sequenced in one run. Combined with the fact that for application in research fields such as ecology, it is not crucial to determine samples up to species level, the discovery of NGS opened doors to a new range of applications. One of these applications was investigated more thoroughly in this thesis: environmental DNA metabarcoding. Environmental DNA (eDNA) is DNA which is shed in the environment by organisms, such as DNA from nails or hair. Metabarcoding means that not just one DNA sample is concerned, but multiple. With this eDNA metabarcoding method, theoretically, it should be

possible to identify all the species present in an environmental sample (such as soil or water) and to determine in which quantities these identified species occurred.

The eDNA metabarcoding method is relatively new, and there are still some limitations which have to be overcome before the method works properly. The techniques used for the sampling of DNA need to be further developed. Furthermore, there is a search for good, universal barcodes and there is a need for good reference databases. When all the data is collected, it needs to be translated to useful information, with so called bioinformatics tools. Lastly, more research is needed on eDNA, because not all species excrete DNA in their environment, and it is hard to be sure whether the DNA found in the samples belongs to species which are still present in the area or if it belongs to species which are already extinct.

Due to the limitations mentioned above, it is not yet possible to conduct a reliable eDNA metabarcoding analysis. However, research is going full steam ahead and the current limitations do not seem insuperable. Combined with the fact that every year, NGS is becoming faster and cheaper, the eDNA metabarcoding approach could become available to ecologists around the world.

Appendix 2: Timeline

Stoeckle 2003 – Taxonomy, DNA, and the bar code of life

Hebert et al. 2003 – Biological identifications through DNA barcodes

Tautz et al. 2003 – A plea for DNA taxonomy

Janzen 2004 - Now is the time

Moritz & Cicero 2004 – DNA barcoding: promise and pitfalls

Margulies et al. 2005 - Genome sequencing in microfabricated high-density picolitre reactors

Gregory 2005 – DNA barcoding does not compete with taxonomy

Savolainen et al. 2005 - Towards writing the encyclopaedia of life: an introduction to DNA barcoding

Marshall 2005 - Will DNA bar codes breathe life into classification?

Armstrong et al. 2005 – DNA barcodes for Biosecurity: invasive species identification

Hebert & Gregory 2005 – The promise of DNA barcoding for taxonomy

2003 2004 2005 2006

Hebert et al. 2004 – identification of birds through DNA barcodes

Will & Rubinoff 2004 – Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification

Ebach & Hioldrege 2005 – DNA barcoding is no substitute for taxonomy

Hajibabaei et al. 2005 – Critical factors for assembling a high volume of DNA barcodes

Will et al. 2005 – The perils of DNA barcoding and the need for integrative taxonomy

Meyer & Paulay 2005 – DNA barcoding: Error rates based on comprehensive sampling

Metzker 2005 – Emerging technologies in DNA sequencing

Ansorge et al. 2005 - Next-generation DNA sequencing techniques

Camaron et al. 2006 – Who will actually use DNA barcoding and what will it cost?

Hajibabaei et al. 2006 – A minimalist barcode can identify a speciment whose DNA is degraded

Elias et al. 2007 – Limited performanxe of DNA barcoding in a diverse community of tropical butterflies

Mardis 2008 - The impact of next-generation sequencing technology on genetics

Meusnier et al. 2008 – A universal DNA mini-barcode for biodiversity analysis

Hajibabaei et al. 2008 – DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics

Harismendy et al. 2009 – Evaluation of next generation sequencing platforms for population targeted sequencing studies

Valentini et al. 2009 – DNA barcoding for ecologists

Casiraghi et al. 2010 – DNA barcoding: a six-question tour to improve users'awareness about the method

Li H and Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Briefings in bioinformatics vol. 2 no. 5 pp. 473-483

Shokralla et al. 2011 – Pyrosequencing for mini-barcoding of fresh and old museum specimens

De Groot et al. 2011 - Use of rbcL and trnL-F as a Two-Locus DNA Barcode for Identification of NWEuropean Ferns: An Ecological Perspective

2009 2010 2011 2012

Hajibabaei et al. 2011 – Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos

Kahn SD (2011) On the future of genomic data. Science vol. 331 pp. 728-729

Taylor & Harris 2012 – An emergent science on the brink of irrelevance: a review of the past 9 years of DNA barcoding

Shokralla et al. 2012 – next-generation sequencing technologies for environmental DNA research

Taberlet et al. 2012a – Environmental DNA

Taberlet et al. 2012c – Towards next-generation biodiversity assessment using DNA metabarcoding

Hajibabaei 2012 – the golden age of metasystematics

Callaway 2012 – A bloody boon for conservation

Yoccoz NG (2012) The future of environmental DNA in ecology. Molecular Ecology vol. 21 pp. 2031-2038

Bienert et al. (2012) Tracking earthworm communities from soil DNA. Molecular Ecology vol. 21 pp. 2017-2030

Dejean et al. (2011) Persistence of environmental DNA in freshwater ecosystems. PLoS ONE 6(8

2012

2013

Epp et al. (2012) New environmental metabarcodes for analyzing soil DNA: potential for studying past and present ecosystems. Molecular Ecology vol. 21 pp. 1821-1833

De Groot et al. (2012) Monitoring Mite diversity in European Soils. EcoFINDERS project www.EcoFINDERS.eu

Carlsen et al. (2012) Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? Fungal Ecology vol. 5 pp. 747-749

Coissac et al. (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Molecular Ecology vol. 21 pp. 1834-1847

Puillandre et al. (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Molecular Ecology vol. 21 pp. 1864-1877

Taberlet et al. (2012b) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Molecular Ecology vol. 21 pp. 1816-1820