African Zoology 2016: 1–12
Printed in South Africa — All rights reserved
Copyright © Zoological Society of Southern Africa

AFRICAN ZOOLOGY
ISSN 1562-7020        EISSN 2224-073X
http://dx.doi.org/10.1080/15627020.2016.1151377

African Zoology is co-published by NISC (Pty) Ltd and Taylor & Francis

This is the final version of the article that is published ahead of the print and online issue

A review of over a decade of DNA barcoding in South Africa: a faunal perspective

Jessica M da Silva1,2,3* and Sandi Willows-Munro1

1 School of Life Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa.  
2 School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa  
3 Current address: Kirstenbosch Research Centre, South African National Biodiversity Institute, Cape Town, South Africa  
* Corresponding author, email: jessica.m.dasilva@gmail.com

For over a decade, molecular short standardised DNA fragments, termed DNA barcodes, have been developed for species discrimination around the world. As of 2010, the vast majority of barcoding research was biased toward particular taxonomic groups and geographic regions largely because researchers in developed countries were the ones with the resources and capacity to carry out such work. To rectify this, the International Barcode of Life Project was launched with the intent to extend the geographic and taxonomic coverage of the barcode reference library. South Africa committed to this mission in an attempt to catalogue all of its known biodiversity and, possibly, help identify new species. To date, approximately 48 000 South African faunal barcodes are housed in the Barcode of Life Data System (BOLD), which represent only 2.3% of all known South African animal species. Although insects are the best represented in absolute terms, with over 37 000 samples recorded, they are still grossly lacking with just over 1% representation. Much like the global trend, there is a general taxonomic bias, with fish, birds and mammals showing the greatest representation. Moreover, geographic bias is also present, with the Free State province particularly under-represented on BOLD, likely owing to limited human capacity. Although few studies have been published with respect to barcoding, the majority reveal that the cytochrome c oxidase 1 (CO1) gene, used in isolation or in conjunction with other molecular markers, can greatly benefit South African biodiversity research. Several limitations of DNA barcoding are discussed and recommendations specific to South Africa provided.

Keywords: animals, biodiversity, CO1 gene, conservation, taxonomy

Online supplementary information: Supporting information for this paper is available as online supplementary material at http://dx.doi.org/10.1080/15627020.2016.1151377

Introduction

South Africa’s landscape is the third most biologically diverse in the world (WCMC 1992; Mittermeier et al. 1999). In fact, it is because of this vast diversity that it has three of the 35 globally recognised biodiversity hotspots – the Cape Floristic Region, the Succulent Karoo and the Maputoland–Pondoland–Albany hotspot (Mittermeier et al. 2004; Zachos and Habel 2011). These hotspots were classified based on the high percentage of threatened, endemic flora; however, South Africa is also home to a vast array of fauna, with 6% of the world’s mammal species, 8% of bird species and 5% of reptile species, many of which are also endemic (Driver et al. 2012). Currently, approximately 65 500 animal species are known from South Africa, the vast majority being insects (~44 000 species) (Scholtz 1999; Hamer 2013). Although these numbers seem extensive, they are thought to represent less than half of the actual faunal richness in South Africa, with as many as 80 000 animal species (~45 000 insects) left to be discovered or described (Hamer 2013). It has taken about 250 years to reach the current level of taxonomic knowledge of South African fauna (Scholtz 1999), and although new animal species continue to be described at an increasing rate (Hamer 2013), the prospect of documenting the remaining fauna in the foreseeable future is unrealistic. This is not only because of the diversity of animals in South Africa, but also because of the growing shortage of human taxonomic capacity and financial resources. This is further exacerbated by the loss of species through habitat transformation and other human-mediated influences, resulting in many species that will never be studied and protected. This is not a local problem, but a global challenge known as the Taxonomic Impediment (SCBD 2010). Consequently, there has been a strong, global impetus for a rapid, standardised, replicable technique that will improve species data acquisition and quality. DNA barcoding was identified as such a technique (Hebert et al. 2003). For animals, the DNA barcoding gene region is the 658 base-pair segment in the gene encoding the
mitochondrial cytochrome c oxidase 1 (CO1) (Hebert et al. 2003). As of October 2015, the Barcode of Life Data System (BOLD; http://boldsystems.org) – a resource tool to assist in all stages of barcoding research, from specimen collection to tightly validated barcode libraries (Hajibabaei et al. 2005; Ratnasingham and Hebert 2007) – contained barcode records for over 4.5 million specimens and has recognised close to 430 000 animal barcode index numbers (BINs; i.e. operational taxonomic units [OTUs] recognised through sequence variation in the COI DNA barcode region; Ratnasingham and Hebert 2013). Of these BINs, more than 160 000 represent formally described species from around the world – approximately 10% of the world’s known animal diversity (Costello et al. 2013). The remaining 270 000 BINs likely represent species not yet described, and thus denote specimens requiring taxonomic attention (Schindel and Miller 2005).

Considerable literature exists documenting the effectiveness of DNA barcoding in cataloguing the Earth’s diversity (e.g. refer to the Barcoding Bibliography; http://ibol.org/barcoding-bibliography/); however, as of 2010, the vast majority of barcoding research was limited to particular taxonomic groups and geographic regions, largely because researchers in developed countries were the ones with the resources and capacity to carry out such work (Vernooy et al. 2010). To rectify this, the International Barcode of Life Project (iBOL; http://www.ibolproject.org) was launched in October 2010. Its main mission is to extend the geographic and taxonomic coverage of the barcode reference library, by first focusing on known species that are endangered by human activity, that are of particular socio-economic importance or that are used in environmental assessments (International Barcode of Life 2014).

South Africa has committed to assisting in iBOL’s mission and, in 2011, a formal South African node was established on iBOL. Accordingly, the purpose of this review is (1) to provide an assessment of South Africa’s current contribution towards the global barcoding initiative and (2) to determine to what extent DNA barcoding can realistically contribute to a practical understanding of South Africa’s biodiversity and to successful conservation efforts. By examining the taxonomic information and spatial distribution of barcode records currently available on BOLD, this review highlights knowledge gaps and sets priorities for future barcoding.

Materials and methods

All publically available South African faunal records were mined from BOLD in October 2014 and all accompanying information (specifically taxonomic and locality data) was downloaded and used in subsequent analyses. For comparative purposes, the faunal groups listed here mirror those highlighted in South Africa’s National Strategy for Zoological Taxonomy (Hamer 2013). These include amphibians, annelids, arachnids, birds, cnidarians, crustaceans, echinoderms, fish, insects, mammals, molluscs, platyhelminthes, reptiles and sponges. All other groups were classified under ‘other’.

Next, the barcoding contribution of each of South Africa’s nine provinces was assessed. Although South Africa as a nation has committed to iBOL’s mission of extending the geographic and taxonomic coverage of the global barcode reference library, each of South Africa’s nine provinces have their own biodiversity mandates and ordinances, which can influence the resources and effort made available for reaching this mission.

Lastly, the data exchange between BOLD and GenBank was examined by calculating the number of BOLD records that have been mined from GenBank. GenBank is run by the National Center for Biotechnology Information (NCBI) and is the largest open-access annotated repository of publicly available genetic sequence data. In 2005, a partnership between BOLD and GenBank was initiated to facilitate the availability, maintenance and bulk transfer of barcoding data (CBOL 2005).

Results and discussion

DNA barcoding in South Africa

As of October 2014, BOLD housed approximately 48 000 records of South African animals, of which over 38 000 (~80%) were the result of the iBOL initiative. These are considerable numbers given that there are a total of 65 571 known animal species in South Africa (Hamer 2013). However, not all of the records in BOLD represent individual species. In fact, only 2.3% of South Africa’s known animal species are represented in BOLD (Table 1).

Fish are the best represented taxonomic group, with approximately 36% of all known South African fish species catalogued in the database. This is impressive considering they are a diverse group with over 2 000 known species. The comparative success is likely due to both global and local initiatives, such as FISH-BOL (Ward 2012; http://www.fishbol.org/index.php). Following fish are birds and mammals, with 5.4% and 4.9% representation, respectively. Although insects have the largest number of samples uploaded onto BOLD, with over 37 000 samples recorded, this amounts to just over 1% of all known South African insect species. All other taxonomic groups have less than 2% representation in the database.

All of these values could be underestimations as they do not take into account the different BINs present within the database. These BINs represent taxa not assigned to a particular species. If all of these BINs denote species currently unknown to science, then the percentage of represented South African animals would increase to 16.3% (Table 1). Amphibians would show the greatest gain, with an additional 54% representation. Insect numbers would also be significantly bolstered by approximately 17%. Unfortunately, until each BIN has been taxonomically assessed by taxonomic experts and assigned to known species, these numbers cannot be verified.

Unfortunately, taxonomic expertise is not always readily available. Globally, a bias exists with a large percentage of taxonomists studying well-documented taxa. The situation is no different in South Africa, with 28 individuals working on vertebrates (mammals, birds, reptiles, fish and amphibians), which have less than 4 000 species in South Africa, and only 23 working on insects with more than 44 000 species (Hamer 2013). There is a strong impetus to increase the range of taxonomic expertise across animal
Table 1: Comparison of the number of known animal species in South Africa and that made publically available on BOLD. SA = South Africa

<table>
<thead>
<tr>
<th>Taxonomic group</th>
<th>Number of known species in SA*</th>
<th>Number of known SA species in BOLD</th>
<th>Percentage of known SA species in BOLD</th>
<th>Number of BINS in BOLD</th>
<th>Percentage of BINS to known SA species</th>
<th>Total number of samples in BOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insects</td>
<td>44 088</td>
<td>513</td>
<td>1.16</td>
<td>8 238</td>
<td>18.69</td>
<td>37 165</td>
</tr>
<tr>
<td>Arachnids</td>
<td>6 630</td>
<td>7</td>
<td>0.11</td>
<td>555</td>
<td>8.37</td>
<td>1 721</td>
</tr>
<tr>
<td>Molluscs</td>
<td>3 154</td>
<td>22</td>
<td>0.70</td>
<td>214</td>
<td>6.79</td>
<td>734</td>
</tr>
<tr>
<td>Crustaceans</td>
<td>2 559</td>
<td>15</td>
<td>0.59</td>
<td>86</td>
<td>3.36</td>
<td>259</td>
</tr>
<tr>
<td>Fish</td>
<td>2 395</td>
<td>857</td>
<td>35.78</td>
<td>1 142</td>
<td>47.68</td>
<td>7 073</td>
</tr>
<tr>
<td>Annelids</td>
<td>1 110</td>
<td>3</td>
<td>0.27</td>
<td>99</td>
<td>8.92</td>
<td>357</td>
</tr>
<tr>
<td>Cnidarians</td>
<td>926</td>
<td>1</td>
<td>0.11</td>
<td>3</td>
<td>0.32</td>
<td>15</td>
</tr>
<tr>
<td>Birds</td>
<td>854</td>
<td>46</td>
<td>5.39</td>
<td>46</td>
<td>5.39</td>
<td>130</td>
</tr>
<tr>
<td>Echinoderms</td>
<td>410</td>
<td>2</td>
<td>0.49</td>
<td>6</td>
<td>1.46</td>
<td>197</td>
</tr>
<tr>
<td>Reptiles</td>
<td>381</td>
<td>2</td>
<td>0.52</td>
<td>2</td>
<td>0.52</td>
<td>2</td>
</tr>
<tr>
<td>Sponges</td>
<td>346</td>
<td>1</td>
<td>0.29</td>
<td>6</td>
<td>1.73</td>
<td>8</td>
</tr>
<tr>
<td>Mammals</td>
<td>307</td>
<td>15</td>
<td>4.89</td>
<td>53</td>
<td>17.26</td>
<td>179</td>
</tr>
<tr>
<td>Platyhelminthes</td>
<td>246</td>
<td>1</td>
<td>0.41</td>
<td>6</td>
<td>2.44</td>
<td>10</td>
</tr>
<tr>
<td>Amphibians</td>
<td>123</td>
<td>2</td>
<td>1.63</td>
<td>70</td>
<td>56.91</td>
<td>205</td>
</tr>
<tr>
<td>Other</td>
<td>1 218</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>64 747</td>
<td>1 487</td>
<td>2.30</td>
<td>10 526</td>
<td>16.26</td>
<td>47 995</td>
</tr>
</tbody>
</table>

* Hamer (2013)

As part of the greater eThekwini Municipality–University of KwaZulu-Natal Joint Research Partnership (Cockburn 2014; Cockburn et al. in press), the Urban DNA barcode project was initiated in 2011 to help establish an all-taxa biodiversity inventory for eThekwini. The project currently focuses on invertebrate taxa, including keystone species such as bees, flies and spiders, all of which have been collected from several open spaces within the city of Durban. Undergraduate, Honours and Masters students have been involved in gathering and analysing the data, which has contributed over 8 000 specimens to BOLD thus far. These specimens will act as the reference library for future monitoring. Because of the ongoing nature of many of the associated subprojects, these data have not yet been made publicly available on BOLD. However, once available, they will contribute significantly to South Africa’s contribution to BOLD and iBOL. For example, 1 032 Hemiptera specimens representing 312 BINs have been added to BOLD, 92% of which are new to the database (Willows-Munro 2013). These barcode records need to be verified to ensure the accuracy and reliability of the database.

Consequently, researchers may choose other more reliable and accurate markers. Even without such issues, some researchers have opted for using other molecular markers simply because they have traditionally been used in previous studies. In this way, comparisons can be made with sequences already available on GenBank, thereby allowing researchers to broaden the distribution and scope of their study. This is common practice for reptiles, with researchers typically using the ribosomal RNA marker 16S and mitochondrial markers, ND2 and ND4. Similarly, mammalian molecular studies are typically based on the cytochrome b gene.

A final reason for the low numbers on BOLD may be due to the contentious nature of DNA barcoding. Some researchers have simply not ‘bought in’ to the concept, as
is evident by the extensive literary debates on the promise, utility and shortcomings of DNA barcoding (e.g. Moritz and Cicero 2004; DeSalle et. al. 2005; Ebach and Holdrege 2005; DeSalle 2006, 2007; Rubinoff et al. 2006; Taylor and Harris 2012; Collins and Cruickshank 2013). The main criticism involves the notion that DNA barcoding can assist with species discovery. While they have helped reveal cryptic taxa (e.g. Hebert et al. 2004; Witt et al. 2006; Porco et al. 2012), barcodes should rather be considered as an initial screening step for species discovery (Rubinoff 2006; Goldstein and DeSalle 2011). However, given the simplicity and rapidity involved in generating DNA barcodes, there is an inherent risk of them becoming the sole source of information for policy makers, which can bring about misguided conservation actions (Rubinoff 2006).

**Provincial representation**

Overall, insects constitute the majority of South African animal samples made publically available on BOLD. The greatest contribution of insects has been from Gauteng, with over 25 000 samples submitted, which is approximately 99% of all Gauteng samples. Although to a lesser extent, insects also make up the majority of samples from all other provinces, with proportions ranging from 44% to 88% (Figure 1, Supplementary Table S1).

After Gauteng, KwaZulu-Natal (KZN) province has contributed the greatest number of animal samples (11 863). While over half of these are insects, this province has also contributed a significant number of fish (3 713; c. 31% of the provincial total) and arachnid specimens (1 174; c. 10% of the provincial total). Following KZN are the Western and Eastern Cape provinces, which have made significant Collembola and fish contributions. The samples from these three provinces are also the most diverse, including the full spectrum of South African animal clades available on BOLD (Figure 1, Supplementary Table S1). The Free State has contributed the fewest specimens onto BOLD, with a total of 117 samples.

The disparity in contributions between provinces is likely associated with the amount of interest and expertise

---

**Figure 1:** Proportion of animal barcodes made publicly available on BOLD by province. Size of pie charts is approximately proportional to the number of barcode contributions made by each province. WC = Western Cape, NC = Northern Cape, NW = North West, GT = Gauteng, LP = Limpopo, MP = Mpumalanga, FS = Free State, KN = KwaZulu-Natal, EC = Eastern Cape. Category ‘Other’ is not included as the data were unavailable (refer to Table 1)
available within them, and less likely to do with different levels of diversity, although that may affect contributions to a degree. For example, although Gauteng is the smallest of South Africa’s provinces and not as species rich as the coastal provinces (KwaZulu-Natal, Western Cape and Eastern Cape), it is the most populous of the nine provinces largely owing to it being an industrial and economic hub (Stats SA 2014). Consequently, Gauteng has a significant amount of human research capacity, with 34 individuals involved in some form of taxonomic research from seven different institutions (M Hamer, SANBI, pers. comm.), making it the largest concentration of taxonomic expertise in the country. The coastal provinces – KwaZulu-Natal, Western Cape and Eastern Cape – which are also highly populated, all fall within global biodiversity hotspots due to their rich biodiversity. This diversity attracts considerable attention, and can explain why so many researchers and research institutions are based within these provinces (Western Cape: 26 individuals from seven institutions; Eastern Cape: 17 individuals from five institutions; KwaZulu-Natal: 15 individuals from five institutions; M Hamer, pers. comm.). In addition to the researchers within these top four contributing provinces are several students assisting with sample collecting and processing, which is likely to bolster the barcoding numbers associated with each.

In contrast, the Free State, which is the second least populated South African province after the Northern Cape (Stats SA 2014), has only eight individuals involved in taxonomic research based at two institutions (M Hamer, pers. comm.). This could explain why it is the province with the fewest contributions to BOLD. It is unlikely that the Free State contains four of South Africa’s seven vegetative biomes (Grassland, Nama Karoo, Forest and Savanna) and a variety of faunal diversity (National Department of Rural Development and Land Reform 2013). None of this diversity, however, is known to be endemic to the province. This lack of endemic species could help explain the limited contributions from this province, as species may have been collected and submitted from other localities. Consequently, this assessment suggests that the Free State may simply be the least sampled for faunal barcoding of all provinces.

**Contributing institutions: South Africa vs the world**
As of October 2014, 122 institutions (universities, museums and research centres) and private research collections contributed records of South African animals onto BOLD (Table 2), of which South African institutions provided approximately 36%. The main South African contributors were the University of KwaZulu-Natal (>6 300 records) and the South African Institute for Aquatic Biodiversity (~4 000). Overall, the primary contributor as identified by BOLD was the Biodiversity Institute of Ontario (BIO) in Canada, responsible for more than 24 000 records (c. 52%). The BIO is the birthplace of DNA barcoding and home of the Canadian Centre for DNA Barcoding (CCDB: the primary analytical facility for iBOL) and BOLD (BIO 2015). Stating that the BIO is the main contributor is misleading considering that all of the more than 24 000 samples were actually collected by South African researchers who took photographs and extracted the DNA. The DNA was then sent to the CCDB for sequencing, which they provided free of charge. This free sequencing was made possible by a grant from Canada’s International Development Research Centre, which started in 2010. The intention behind this funding was to make African biodiversity a much more significant part of the iBOL research program and to provide a major incentive for South African institutions to meet the requirements of the Regional Node of iBOL (iBOL 2010). Given that the grant has ended, it is expected that the contributions from South African institutions will increase – a likely prospect considering the extensive barcodes being generated from ongoing projects throughout the country (see Box 1).

**GenBank and BOLD**
In total, 1 438 South African animal records on BOLD were mined from GenBank since their partnership was established in 2005 up until October 2014. GenBank and BOLD share a tightly integrated data exchange; however, the exchange has its limitations. BOLD allows for the automatic submission of data to GenBank, to which GenBank responds by providing the user, as well as BOLD, accessions for the submitted records. As new information is obtained, updates can be made to these records increasing the functionality of the partnership. This data exchange may lead some to assume that the two databases are mirror-images of each other, at least with respect to CO1 sequence data; however, this is not the case. BOLD only mines CO1-5P records from GenBank that contain a country feature, which is what BOLD uses to assign records to a country (J Robertson, BOLD systems, pers. comm.).

Table 2: Contributors of South African faunal records on BOLD. Ni = Number of institutions or individuals, Nr = number of records on BOLD

<table>
<thead>
<tr>
<th>Contributor</th>
<th>Museums</th>
<th>Research centres</th>
<th>Universities</th>
<th>Private research collections</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ni</td>
<td>Nr</td>
<td>Ni</td>
<td>Nr</td>
</tr>
<tr>
<td>South Africa</td>
<td>8</td>
<td>976</td>
<td>6</td>
<td>5 080</td>
</tr>
<tr>
<td>Kenya</td>
<td></td>
<td></td>
<td>7</td>
<td>10 580</td>
</tr>
<tr>
<td>Australia</td>
<td>2</td>
<td>5</td>
<td></td>
<td>1 225</td>
</tr>
<tr>
<td>South America</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>5</td>
<td>109</td>
<td>2</td>
<td>34</td>
</tr>
<tr>
<td>Middle East</td>
<td>1</td>
<td>9</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>UK</td>
<td>2</td>
<td>38</td>
<td>1</td>
<td>82</td>
</tr>
<tr>
<td>Rest of Europe</td>
<td>11</td>
<td>762</td>
<td>7</td>
<td>2 066</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>701</td>
</tr>
</tbody>
</table>
If this information is missing, then BOLD will not recognise them. For example, when performing a search on GenBank using the query ‘South Africa’ AND reptilia[organism] AND (COI[gene name] OR COX1[gene name] OR COI[gene name] OR COX1[gene name]), 268 records were found. Most of these records do not have a country feature in the flat file, likely explaining why a search on BOLD for South African reptiles only produces two records. Considering this, users should mine both databases for relevant sequence data and, when submitting records to GenBank (if doing so directly), ensure that all relevant spatial information accompanies the data.

Is DNA barcoding useful in South African biodiversity conservation?

The most commonly discussed benefits of barcoding are to the field of taxonomy by assisting taxonomists with single-species identifications when traditional morphological traits are not sufficient or when decisions on the similarities and differences of specimens are viewed as subjective. For example, around the world, barcoding has helped uncover phenotypic plasticity, identify cryptic, sexually dimorphic and multiple life-stage species, as well as species that are damaged or only partially available (e.g. stomach contents and faecal remains) (e.g. Hebert et al. 2004; Weigand et al. 2011; Zeale et al. 2011; Meiklejohn et al. 2013; Pramual and Wongpakam 2014; Meheust et al. 2015; Przybyłowicz and Tarcz 2015). It can also help clarify taxonomic inconsistencies or seemingly subjective species descriptions resulting from specimens described and classified by different taxonomists who employ different criteria or species concepts – a situation commonly found with birds (Newton 2003). Moreover, barcoding can expedite the taxonomic process by identifying groups of species that require more detailed studies, as well as propose hypotheses about new species (Goldstein and DeSalle 2011). Describing new species based on DNA sequence divergence alone has been criticised as being too simplistic (e.g. DeSalle 2006). Instead, barcoding should be used in conjunction with other molecular markers and, where possible, traditional morphology-based taxonomy (Prendini 2005), as well as other lines of evidence in order to seek corroboration for lineage diversification to more accurately describe species.

Few South African studies have been published with respect to barcoding, yet the majority of the available examples reveal that the CO1 gene, used in isolation or in conjunction with other molecular markers, has helped resolve taxonomic issues, such as the case of the Yellowtail blenny, *Cirripespectes heemstraorum* (Williams 2010), the intertidal parasitoid wasp, *Echthrodessis lamaroli* (van Noort et al. 2014), and the bloom-forming jellyfish, *Crambionella stuhlmanni* (Neethling et al. 2011). The CO1 gene has also helped identify cryptic speciation within scale insects (*Sethusa* et al. 2014) and the *Eisenia* earthworms (Otomo et al. 2013), as well as phenotypic plasticity within the coral, *Stylophora pistillata* (Keshavmurthy et al. 2013). To date, only one study reports the CO1 gene failing to distinguish between two taxonomically identified species, *Manta alfredi* and *M. birostris* (Kashiwagi et al. 2012). Considering these rays occur in sympatry, the authors attribute this to post-divergence gene flow in the recent past.

In addition to taxonomy, DNA barcoding has also assisted with biosecurity through the identification of invasive and pest species that can have serious ecological and economic implications (Pieterse et al. 2010; Jones et al. 2013; Marsberg et al. 2015). Barcoding has also been advantageous in monitoring the trade of wildlife products in South Africa by identifying unknown zoological material, such as in the case of mislabelled fish and meat in commercial markets (Cawthorn et al. 2012; D’Amato et al. 2013; Cawthorn et al. 2015) and confiscated meat (e.g. carcasses or dried or frozen meat) (Dalton and Kotze 2011). Considering that South Africa has become a favourite destination for poachers and wildlife traffickers, in 2013 the South African National Biodiversity Institute (SANBI) together with the National Zoological Gardens of South Africa and the South African Institute for Aquatic Biodiversity (SAIAB) became involved in the Barcode of Wildlife Project with the hopes of deterring poaching and wildlife trafficking (SANBI 2015). Currently, DNA material from 130 of the 200 target threatened plants and animals have been sourced for barcoding.

By assisting in the protection of threatened species from invasive species and illegal trade, DNA barcoding is also directly involved in species conservation. Another way in which barcoding can benefit conservation is through conservation planning; however, to date, there are no reported studies from South Africa. This approach would involve applying phylogenetic diversity, which is a measure of the taxonomic divergence between species (Faith 1992, 1994, 1996), to boost predictions about biodiversity patterns. Phylogenetic diversity would become the key criterion on which to design, prioritise and extend conservation areas, instead of relying on species counts (e.g. Faith and Baker 2006; Kress et al. 2014; Pollock et al. 2015; Shapcott et al. 2015; Vargas et al. 2015). Indeed, phylogenetic diversity defined by DNA barcodes is likely to be the most important measure for comparing diversity and establishing protected areas across landscapes in regions known for their especially unique diversity, such as South Africa (Kress et al. 2014; Shapcott et al. 2015). Such analyses can be further enhanced by using metabarcoding, metagenomics (reconstruction of whole genomes) or environmental DNA (the sequencing of free DNA from soil, water and/or air) (see Yu et al. 2012; Bohamm et al. 2014; Cristescu 2014; Escalante et al. 2014).

Other areas in which barcoding can be beneficial are public health (e.g. Sweeney et al. 2011; Paramasivan et al. 2013), paleoecology (Valentini et al. 2009) and dietary studies (e.g. Valentini et al. 2009; Jo et al. 2014; Santos et al. 2015). However, such research has not yet been attempted, or rather published, in South Africa. Although numerous studies have been published in each of these fields using other genetic techniques and markers, none have used the CO1 gene. Whether this gene has proven ineffective in these fields in South Africa, thereby explaining its lack of use, has yet to be reported.

Limitations and recommendations

As much as barcoding has its advantages, it also has its limitations. These mainly involve using a single-locus to identify and ‘discover’ species, employing a ‘universal
barcode gap', the cost and soon-to-be redundancy of Sanger sequencing, and the disconnect between the data generators and end-users.

**The single-locus approach**

The CO1 gene is highly conserved across species and has a high mutation rate compared with other DNA sequences, making it ideal for barcoding. These qualities allow for the identification and discrimination of species, including closely related species. However, using a single locus to answer questions on species boundaries and discovery can be problematic. Different genes and lineages evolve at different rates, which can make a genetic marker indispensable for one level of analysis but uninformative in another, or informative for one group of organisms and not others (Rubinoff 2006). For example, numerous studies have documented the discordance between mitochondrial and nuclear genealogies, resulting in very different interpretations of species status, rarity and conservation importance (e.g. Monsen and Blouin 2003; Ballard and Whitlock 2004; Rubinoff and Sperling 2004; Spinks et al. 2012). This is because of their different modes of inheritance, with the mitochondrial genome being maternally inherited (Shaw 2002). Issues such as hybridisation and incomplete lineage sorting are considered the most likely explanations for this discordance, often resulting in mtDNA genes being shared between taxa. Consequently, DNA barcoding would fail to recognise such taxa as different species, whereas nuclear genes may, resulting in different phylogenies (e.g. DeSalle and Giddings 1986; Funk and Omland 2003; Spinks et al. 2012). This can have serious conservation implications as different species status, rarity and conservation importance of species can be problematic. Different genes and lineages evolve at different rates, which can make a genetic marker indispensable for one level of analysis but uninformative in another, or informative for one group of organisms and not others (Rubinoff 2006). For example, numerous studies have documented the discordance between mitochondrial and nuclear genealogies, resulting in very different interpretations of species status, rarity and conservation importance (e.g. Monsen and Blouin 2003; Ballard and Whitlock 2004; Rubinoff and Sperling 2004; Spinks et al. 2012). This is because of their different modes of inheritance, with the mitochondrial genome being maternally inherited (Shaw 2002). Issues such as hybridisation and incomplete lineage sorting are considered the most likely explanations for this discordance, often resulting in mtDNA genes being shared between taxa. Consequently, DNA barcoding would fail to recognise such taxa as different species, whereas nuclear genes may, resulting in different phylogenies (e.g. DeSalle and Giddings 1986; Funk and Omland 2003; Spinks et al. 2012). This can have serious conservation implications based on taxonomic divergence (Faith 1994).

To rectify this, we suggest using an integrative taxonomic approach, by combining multiple genetic markers (mitochondrial and nuclear, if possible), in conjunction with morphological, ecological and geographical data (Dayrat 2005; Schlick-Steiner et al. 2010). This combination of data would provide a more comprehensive and accurate understanding of the evolution of the taxa under investigation, which is essential when assessing and managing biodiversity.

**A 'universal barcode gap'**

Many attribute the success or failure of DNA barcoding to the presence or absence, respectively, of a barcode gap. This gap is defined as interspecific variation exceeding intraspecific variation. Several attempts have been made to establish a standard limit of sequence divergence, and hence define a ‘universal barcode gap’, with which to delimit species. Most barcoding researchers have adopted a 2% divergence threshold as this cut-off value (e.g. Hebert et al. 2003, 2008; Hubert et al. 2008; Ward et al. 2009; Pereira et al. 2013; Ratnasingham and Hebert 2013; Telfer et al. 2015). This value is based on the distribution of the mean inter- and intraspecific Kimura-2-parameter genetic distance values from the thousands of species that have been barcoded (Pereira et al. 2013), and is the method employed by BOLD (Ratnasingham and Hebert 2007). However, this threshold cannot be generalised for all organisms as coalescent depths vary among species, and intraspecific distances for one species can exceed interspecific distances for other species, as has been found in several invertebrate species (Meier et al. 2006; Collins and Cruickshank 2013; Nunes et al. 2014). As such, the failure to detect a barcoding gap is rather more of a failure of defining and employing a universal cut-off value for this gap.

Instead of using the 2% divergence as an absolute threshold, Pereira et al. (2013) suggested that it should rather be considered a starting point to investigate divergence among specimens, and considered alongside other characteristics, such as a group’s evolutionary history, before defining species limits. Other researchers have also sought alternative methods for investigating the presence of a barcode gap, such as using a scatterplot to illustrate the distance between the furthest conspecific and nearest non-conspecific, where a 1:1 slope represents the point at which there is no barcoding gap (see Robinson et al. 2009), or using uncorrected p-distances and calculating the difference between the smallest interspecific and the largest intraspecific distance (Meier et al. 2008; Collins et al. 2012; Srivathsan and Meier 2012; Collins and Cruickshank 2013), or employing the Automatic Barcode Gap Discovery (ABGD) method, which automatically detects the first significant gap used to partition sequence data (Puillandre et al. 2012).

In addition to the technique used, sampling effort and geographic scale have been shown to affect the detection of a barcode gap in some organisms, resulting in people questioning the objectiveness and, thus, utility of the barcode gap (e.g. Meyer and Paulay 2005; Wiemers and Fiedler 2007; Bergsten et al. 2012). DeSalle (2007) has even questioned the use of genetic distances altogether and recommends using diagnostic genetic characters (DeSalle et al. 2005). In an attempt to minimise the chances of species or sequences being overlooked using an isolated DNA barcoding approach, it is recommended that barcoding be used in an integrative taxonomic context, as discussed above. Corroboration of lineage separation from multiple lines of evidence provides a more definitive and objective means of identifying and delimiting species.

**Cost and redundancy of Sanger sequencing**

For almost three decades, Sanger sequencing was the only approach used to generate DNA sequences, and as such became the conventional way of obtaining a DNA barcode for a species (Sanger et al. 1977; Shokralla et al. 2014). It can generate a single sequence of up to 1 000 bases from a DNA concentration greater than 20 ng µl⁻¹ to avoid inherent biases and errors (Sanger Institute 2010). Currently, the cost of generating a DNA barcode is US$5 per sample, which equates to approximately R65. According to a recent study, this equates to between 1.7 and 3.4 times the cost of traditional taxonomic identification (Stein et al. 2014). However, the authors did not consider the time required to identify specimens without available taxonomic keys (E Stein pers. comm.). Nevertheless, with ever-evolving technology, processing costs will continue to reduce (e.g. Creer et al. 2010; Bohmann et al. 2014).

Since the completion of the first human genome project (International Human Genome Sequencing Consortium
2001), demand for cheaper and faster sequencing methods have increased, which has driven the development of next-generation sequencing (NGS). NGS is capable of parallel sequencing at a grand scale, thereby facilitating the integrative taxonomic approach mentioned above through high-throughput sequencing. Millions of DNA fragments from a single sample are sequenced in unison, allowing for an entire genome to be sequenced in less than a day (see Valentini et al. 2009). This technology has already been implemented successfully in various studies (e.g. Creer et al. 2010; van Bers et al. 2010; Ruiz-González et al. 2013; Giampaoli et al. 2014; Diaz-Real et al. 2015).

Moreover, the cost of NGS can be less than US$1 per specimen (c. R13), which is cheaper than or on par with morphology-based taxonomic identifications (Stein et al. 2014). Consequently, many genomics facilities have phased out their Sanger sequencers and switched to NGS (Shokralla et al. 2014). This does not mean that NGS will replace DNA barcoding, but rather NGS can greatly benefit barcoding by being able to capture all representative sequences present in a complex mixture of species (such as soil, water, diet and faecal samples), which can then be mapped to a reference DNA barcode database (Taylor and Harris 2012; Kress et al. 2015). This metabarcoding approach will save considerable time and expense, and lead to tremendous growth in available sequence data and, ideally, our understanding of global biodiversity (Cristescu 2014). However, such an approach also has its challenges, including how to manage the very large amounts of data produced (Bik et al. 2012; Cristescu 2014) and identifying the plethora of sequences produced when many taxa still have incomplete reference libraries (Meyer and Paulay 2005; Wiemers and Fielder 2007). With that said, NGS has been used to sequence century-old type specimens, thereby providing invaluable genetic information to a reference library and the only certain connection between the application of a Linnean name and a physical specimen (Prosser et al. 2015).

**Disconnect between data generators and end-users**

The usefulness of DNA barcoding to South African biodiversity research has been highlighted above; however, there is often a disconnect between the individuals generating the data (i.e. DNA sequences) and those that are responsible for making informed policy and management decisions. This is not a novel concept. Serious effort has been placed on bridging the research-implementation gap within the past decade (e.g. Knight et al. 2008; Cook et al. 2013).

BOLD is a repository of DNA barcodes that is freely available to the public; yet, many of the intended downstream users of biodiversity data (for example, policy-makers and managers in municipalities, custom officials monitoring wildlife trade, among others) are not molecular ecologists and, hence, do not know how to analyse and interpret sequence data. Unlike land-cover data, for example, sequence data cannot simply be inputted into GIS and conservation planning software. As such, much of the data produced is not being used by the end user.

An obvious response would be to ensure that those skilled in molecular methods and analysis work closely with downstream users, as is the case with the eThekwini Municipality–University of KwaZulu-Natal Joint Research Partnership (Cockburn 2014; Cockburn et al. in prep; see Box 1). However, this is not often available or feasible. In such instances, a possible solution would be to provide training courses to end users to empower them to utilise the plethora of information already available or establish an online resource that informs end users of basic genetic concepts in non-technical language, provides a toolbox so that practitioners can identify how DNA barcoding can help them address familiar management issues and questions, and provides a forum to enable open and ongoing discussion about common issues and questions, as well as sharing tips and data (Hoban et al. 2013). Such programs do exist, albeit not in South Africa (e.g. the Urban Barcode Research Programme in New York: http://www.urbanbarcodeproject.org/ubresearch.html; ConGRESS: http://www.congressgenetics.eu/)

**The future of DNA barcoding in South Africa**

In the past decade, Africa has contributed the fewest number of publications describing new species compared with the rest of the world; yet, within this region South Africa has been the major contributor (Tancoigne et al. 2011; Grieneisen et al. 2014). Despite the overall low publication record, there has been a general increase in the number of species described in South Africa; however, the contributions from local taxonomists have declined (Hamer 2013). In fact, within the last three decades, 66% of new South African species have been described by foreign taxonomists (Hamer 2013). This may be partly because there has been a general move away from descriptive and revisionary taxonomy within the country towards more phylogenetic analyses. The reason behind this likely stems from the lack and loss of taxonomic positions, the low salaries and limited potential for promotions, and the general loss of a research culture in museums (von Staden et al. 2013). Moreover, the general speed at which molecular expertise can be developed compared with taxonomic expertise, the fact that molecular studies can be published in higher-impact journals and receive high citation ratings compared with taxonomic identifications and revisions (key factors affecting promotion and funding opportunities) and the general perception that molecular research is of higher quality than morphological investigation may further exacerbate the issue (Hamer 2013; von Staden et al. 2013).

To help improve the rate of species discovery within South Africa, and in turn Africa as a whole, strategic and collaborative data collection, storage and dissemination needs to be put into place within South Africa that supports all end users, not just individuals well-versed in molecular data. South Africa’s commitment to iBOL is one way of doing this; in many ways, BOLD, the barcoding database, is an important portal for compiling specimen data as it requires that photographs, GPS coordinates, and other collection information (such as date, collectors and voucher locality) accompany all sequence data. Moreover, as already mentioned, DNA barcoding has identified a significant number of unique sequences, to
which BOLD has assigned unique BINs. Each BIN may represent a novel species; however, the associated specimens need further validation through additional molecular markers and taxonomic assessments. For many groups of taxa, the need for taxonomic verification poses a significant problem because of the limited taxonomic capacity in South African institutions, as well as the general lack of curatorial staff in South African museums (Hamer 2013). This lack of curatorial staff is critical as voucher specimens of all newly described species need to be housed in a museum so that they can be the reference sample and basis for future research. These specimens need to be maintained and catalogued, which requires care, attention and expertise.

To help rectify this, links or collaborations with global partners should be established to facilitate the training of more taxonomists and attract more researchers to southern Africa. In this way, a network of specialists will be created to share knowledge, which should greatly enhance South Africa’s publication productivity as well as increase the number of species described.

Through this specialist network, coordinated sampling efforts could also be put in place to fill in the geographic and faunal gaps mentioned in this review. Of course, all faunal groups require greater representation in BOLD, as no group is completely represented. For the larger-bodied, more well-known groups, such as birds, fish and mammals, it should be feasible to barcode all individuals within the near future. However, the remaining groups have less than 1% representation in BOLD; and, as such, all are in great need of added representation. However, if one takes into account the unnamed BINs, there are three groups that are particularly underrepresented. These are arachnids, cnidarians and reptiles. As such, we recommend that increased efforts be put into these groups, first focusing on species of conservation concern, such as those involved in illegal trade or protected under CITES. For reptiles, this does not have to be a daunting task as many specimens are already housed in DNA banks and museum collections throughout the country; one simply has to generate their barcoding sequence. With that being said, increased sampling is always encouraged. Overall, sampling efforts should also be targeted toward the underrepresented regions of the country, such as the Free State and North West provinces.

In addition to amassing and verifying ‘new species’ data, greater attention should also be put into unifying and disseminating knowledge of known species. Species-level information for South African animals on the internet is limited and the information that does exist is fragmented across hundreds of different websites (Hamer 2013). Although to a far lesser extent, the same is true for barcoding data, which is available on BOLD and/or GenBank. A central repository of all South African species information would greatly help all end users have access to all current species information. A resource analogous to ConGRESS for South Africa may be one way of doing this (see Hoban et al. 2013; http://www.congressgenetics.eu/).

Acknowledgements — We would like to thank Michelle Hamer for her invaluable input into this manuscript, as well as the editors and three anonymous reviewers for their constructive comments, which helped us to improve the manuscript. This work was supported by the South African Research Chairs Initiative of the Department of Science and Technology and the National Research Foundation of South Africa.

References


Meheust E, Alfonsi E, Le Menec P, Hassani S, Jung J-L. 2015. DNA barcoding for the identification of soft remains of prey in the stomach contents of grey seals (Halichoerus grypus) and...


