MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2009) **9**, 1353–1358 doi: 10.1111/j.1755-0998.2009.02692.x

PERMANENT GENETIC RESOURCES ARTICLE

MalAvi: a public database of malaria parasites and related haemosporidians in avian hosts based on mitochondrial cytochrome *b* lineages

STAFFAN BENSCH,* OLOF HELLGREN† and JAVIER PÉREZ-TRIS‡
*Department of Ecology, Lund University, Ecology Building, S-22362 Lund, Sweden, †Department of Zoology, The Edward Grey Institute, South Parks Road, Oxford OX1 3PS, UK, ‡Departamento de Zoología y Antropología Física, Facultad de Biología, Universidad Complutense de Madrid, 28040 Madrid, Spain

Abstract

Research in avian blood parasites has seen a remarkable increase since the introduction of polymerase chain reaction-based methods for parasite identification. New data are revealing complex multihost–multiparasite systems which are difficult to understand without good knowledge of the host range and geographical distribution of the parasite lineages. However, such information is currently difficult to obtain from the literature, or from general repositories such as GenBank, mainly because (i) different research groups use different parasite lineage names, (ii) GenBank entries frequently refer only to the first host and locality at which each parasite was sampled, and (iii) different researchers use different gene fragments to identify parasite lineages. We propose a unified database of avian blood parasites of the genera *Plasmodium, Haemoproteus* and *Leucocytozoon* identified by a partial region of their cytochrome *b* sequences. The database uses a standardized nomenclature to remove synonymy, and concentrates all available information about each parasite in a public reference site, thereby facilitating access to all researchers. Initial data include a list of host species and localities, as well as genetic markers that can be used for phylogenetical analyses. The database is free to download and will be regularly updated by the authors. Prior to publication of new lineages, we encourage researchers to assign names to match the existing database. We anticipate that the value of the database as a source for determining host range and geographical distribution of the parasites will grow with its size and substantially enhance the understanding of this remarkably diverse group of parasites.

Keywords: *Haemoproteus*, host range, host–parasite interactions, *Leucocytozoon*, *Plasmodium*

Received 22 December 2008; revision accepted 25 February 2009

Avian malaria parasites (*Plasmodium*) and related haemosporidians (*Haemoproteus* and *Leucocytozoon*) have received considerable attention in ecological and evolutionary studies during the last 10 years, as a result of the ease with which these parasites can be studied using polymerase chain reaction (PCR) techniques on samples of DNA extracted from bird blood. The first PCR-based protocol for avian haemosporidians (Feldman *et al.* 1995) was designed to amplify *Plasmodium* parasites from birds in Hawaii using primers for 18S rRNA, but this protocol has not been used broadly because it only works for a small group of *Plasmodium* parasites. The first general protocol for both avian *Plasmodium* and *Haemoproteus* parasites was published by Bensch *et al.* (2000), where a portion of the cytochrome *b* gene of the mitochondria was the target molecule. This protocol together with slightly modified protocols (Hellgren *et al.* 2004; Waldenström *et al.* 2004) is still among the most widely used protocols. Following the publication of the initial cytochrome *b* protocol, three key studies (Perkins & Schall 2002; Ricklefs & Fallon 2002; Waldenström *et al.* 2002) took advantage of the newly available molecular information to gain new ecological and evolutionary insight on this host–parasite system, thus kick-starting the field of molecular studies of avian haemosporidians. Since

Correspondence: Staffan Bensch, Fax: +46 46 2224716; E-mail: staffan.bensch@zooekol.lu.se

then, more than 60 publications have used molecular analyses of these parasites' cytochrome *b* gene in studies of taxonomy, systematics, ecology, biogeography and evolution, which has greatly expanded the field (Fig. 1).

The great number of host–parasite associations identified with the new methodology has made it possible to reveal various ecological and evolutionary patterns (Fallon *et al.* 2005; Beadell *et al.* 2006; Pérez-Tris *et al.* 2007; Martinsen *et al.* 2008). Because of the massive information generated by different research groups worldwide, we need unified criteria of identification and labelling of parasite lineages. This would guarantee that the rapidly growing information produced by different researchers could be easily compared.

To date, avian haemosporidian parasites include approximately 40 morphologically distinct species of the genus *Plasmodium*, 130 species of the genus *Haemoproteus* and 35 species of the genus *Leucocytozoon* (Valkiūnas 2005). All these parasites multiply as haploid clones in avian hosts and undergo the sexual process in Dipteran vectors. Since 2000, approximately 800 unique cytochrome *b* lineages of avian *Plasmodium*, *Haemoproteus* and *Leucocytozoon* have been deposited to GenBank. Although part of this variation is expected to reflect intra-specific polymorphisms, there are several lines of evidence suggesting that the vast majority of these mtDNA lineages correspond to good species, if one applies the biological or the phylogenetic species concepts. Hence, the species richness is massively higher than that inferred from parasite morphology (Valkiūnas 2005). The strongest evidence for the lineages representing species comes from parallel studies of mtDNA (cyt *b*) and a comparatively fast-evolving nuclear gene (DHFR-TS), showing congruent gene-trees and an apparent absence of recombination for many very closely related lineages (Bensch *et al.* 2004; Hellgren *et al.* 2007; Pérez-Tris *et al.* 2007). Moreover, mtDNA lineages of parasites differing by as little as one nucleotide in the cytochrome *b* gene frequently exhibit distinctly different areas of transmission and range of host species (Waldenström *et al.* 2002; Beadell *et al.* 2004; Reullier *et al.* 2006). That most of this diversity is representing species should not be taken as evidence that all lineages are good species. To prove that two lineages are representing good species requires in each case detailed analyses of multiple genetic markers.

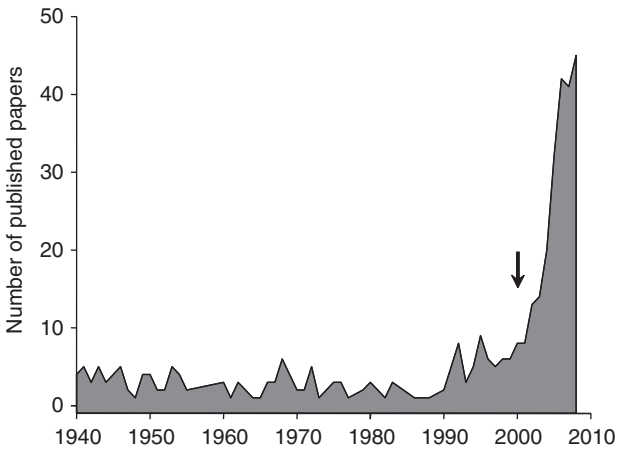

Fig. 1 Number of papers on avian haemosporidians published per year since 1940, according to a search in ISI-Web of Science (Thomson, December 2008) with the general keyword 'avian malaria'. The arrow indicates the year when the first general protocol for amplification of cytochrome *b* sequences was published (Bensch *et al.* 2000).

Molecular methods have also been able to identify a remarkable diversity of parasites within single host species, for example, the blackcap *Sylvia atricapilla* harbours a lineage rich and host-specific clade of parasites (Pérez-Tris *et al.* 2007) belonging to the same morphospecies (*Haemoproteus parabelopolskyi*; Valkiūnas *et al.* 2007). These closely related lineages frequently co-infect the same host individual, and sometimes as many as five lineages (phylogenetically cryptic species) can occur together in the same bird, apparently without resulting in recombination (Pérez-Tris & Bensch 2005).

Although the new molecular methods have greatly boosted research on avian haemosporidians, simple but important questions like 'in which host species and geographic regions has parasite x been found?' cannot easily be answered today due to three main problems. First, there is no common naming of lineages. For example, identical sequences to the *Haemoproteus* lineage 22 (Ricklefs & Fallon 2002) can also be found published as AP21 (Beadell *et al.* 2004), COLL2 (Krizanauskiene *et al.* 2006), SWTH.H2 (Svensson *et al.* 2007) and WAH24 (Beadell *et al.* 2009). Second, GenBank entries frequently lack information about the geographical origin of the sample, or just report the first locality where the parasite was observed, and although most GenBank entries contain information about the host species, additional host species for that lineage in the same study or other studies are in many cases given only in the published paper, thus making it difficult to find and compare such information. Third, different research groups use protocols that cover partial, more or less overlapping regions of the cytochrome *b* gene, which complicates the assessment of lineages. This lack of synchronization is strongly hindering broad analyses of patterns of parasite distribution.

To mitigate these problems, we have compiled a simple database in Microsoft Access (MalAvi.mbd) that can be downloaded at http://mbio-serv4.mbioekol.lu.se/avianmalaria/index.html. The database includes avian blood parasites of the genera *Plasmodium*, *Haemoproteus* and *Leucocytozoon* identified by a partial region of their cytochrome *b* sequences. We have decided to follow the

traditional nomenclature (see for example, Valkiūnas 2005), although recent studies (Martinsen *et al.* 2008) suggest that two distinct clades of *Haemoproteus* should be split into two genera (*Haemoproteus* and *Parahaemoproteus*). The database uses a standardized nomenclature to remove synonymy, and concentrates all available information about each parasite in a public reference site, thereby facilitating access to all researchers. The database will be initially launched with not only information about the list of host species and geographical localities each parasite lineage has been found in, but we also suggest directions to further develop this resource as a permanent repository. The database has been compiled from the information contained in 68 papers and many unpublished GenBank entries. It presently contains 864 unique lineages of parasites, obtained from 579 species of hosts from all continents except Antarctica, roughly corresponding to the screening of 35 000 host specimens. The site also has an associated file in fasta format available for downloading (MalAvi.fas), which contains an alignment of the sequences of all unique lineages included in the database. This file can easily be used for searching matching lineages using local BLAST within the program BIOEDIT (Hall 1999).

To define unique lineages, we chose the 479 base pairs between the primers HAEMF and HAEMR2 for *Plasmodium* and *Haemoproteus* parasites (Bensch *et al.* 2000), with the corresponding 480 bp for *Leucocytozoon* parasites between primers HAEMFL and HAEMR2L (Hellgren *et al.* 2004). The primers frequently used in studies covering the region between HAEMF and HAEMR2 can be found in Table 1 and Fig. 2. Although the mtDNA genome of the parasites is small (about 6000 bp; Omori *et al.* 2007), a short sequence like our target region will not distinguish all mtDNA haplotypes. It has been shown that the region between HAEMF and HAEMR2 captures most haplotypes that would have been found had the full cytochrome *b* gene been analysed (Hellgren *et al.* 2007) and should therefore be sufficient for the present purpose. However, it is important to emphasize that some of the lineages identified as being shared across a range of host species based on this partial fragment may actually be two or more genotypes when sequenced over a larger portion of the mtDNA genome.

The Microsoft Access database is organized in 'tables' that can be linked to each other by common identifiers. Summary information can then be extracted by 'queries' from the tables. The different tables in the avian haemosporidian database are summarized in Table 2. We present a more precise description of each of the eight different 'tables':

1 The main table *Master* contains all unique lineages. Each unique lineage has been given an acronym that most often corresponds to an abbreviation of the host species name (in Latin or in English) in which the lineage was first identified, followed by a number (e.g. GRW01, great reed warbler lineage 1). The naming system is not fully consistent as we have tried to keep names of published lineages when possible. A simple alternative to this naming system would be a strict accession number system, but most people seem to prefer an acronym that communicates some information (like SGS1, first lineage found in Sudan Golden Sparrow) rather than a flat number (e.g. 0784). Nevertheless, the basis of naming is not essential as long as the lineages are uniquely assigned. Sequences that are identical in the region between the primers HAEMF and HAEMR2 may appear to be different if sequenced for a larger portion of the cytochrome *b* gene. In the few cases when we have encountered sequences that correspond to different haplotypes outside our target region, these sequences have been lumped to a single name as identified by the sequence between HAEMF and HAEMR2. For each unique lineage, the 'Master' table contains fields for the GenBank accession number with a corresponding link to the GenBank entry, parasite genus, host species of the first isolate and a bibliographical reference.
2 The *Sequences* table contains all DNA sequences.
3 The *Hosts and sites* table contains all records of the unique lineages with information about where (geographical site) and in which host species they have been found, with references to the original publications.
4 The table *MorphoSpecies* contains the information of scientific names for lineages whose morphological identity has been verified by microscopic analyses, also with a reference to the original paper describing the association. Recently, most researchers on avian haemosporidians employ molecular analyses of DNA extracted from blood samples. Nonetheless, there is often a need to relate the identified lineages to described morphospecies to access the rich knowledge already existing based on morphological analyses (Valkiūnas 2005). A main problem is that the majority of published lineages have not been identified morphologically, and even worse, GenBank contains many entries with erroneous parasite species names (Valkiūnas *et al.* 2008), which if used would give wrong directions to the traditional literature.
5 *References* contains all the bibliographical references cited in the database.
6 *AlternativeLineageNames* provides a key to translate between previously published lineage names and those defined for the database.
7 *HostTaxonomy* provides the host species and its associated taxonomic information (genus, family, order).

Table 1 Primers for amplifying avian haemosporidian cytochrome *b* sequences that cover the region between the primers HAEMF and HAEMR2

Protocol/parasite genera	Primers	Sequence (5′–3′)	Reference
Nested PCR, 479 bp for *Haemoproteus* and *Plasmodium*			
Primer pair 1	HAEMNF	CATATATTAAGAGAATTATGGAG	2
	HAEMNR2	AGAGGTGTAGCATATCTATCTAC	2
Primer pair 2	HAEMF	ATGGTGCTTTCGATATATGCATG	1
	HAEMR2	GCATTATCTGGATGTGATAATGGT	1
Nested PCR, 479 bp for *Haemoproteus*, *Plasmodium* and *Leucocytozoon*			
Primer pair 1	HAEMNFI*	CATATATTAAGAGAAITATGGAG	3
	HAEMNR3	ATAGAAAGATAAGAAATACCATTC	3
Primer pair 2 (H&P)	HAEMF	ATGGTGCTTTCGATATATGCATG	1
	HAEMR2	GCATTATCTGGATGTGATAATGGT	1
Primer pair 2 (L)	HAEMFL	ATGGTGTTTTAGATACTTACATT	3
	HAEMR2L*	CATTATCTGGATGAGATAATGGIGC	3
Single PCR, 533 bp for *Haemoproteus* and *Plasmodium*			
	3760F	GAGTGGATGGTGTTTTAGAT	4
	4292Rw2	TGGAACAATATGTARAGGAGT	4
Nested PCR, 1138 bp for *Haemoproteus*, *Plasmodium* and *Leucocytozoon*			
Primer pair 1	DW2	TAATGCCTAGACGTATTCCTGATTATCCAG	5
	DW4	TGTTTGCTTGGGAGCTGTAATCATAATGTG	5
Primer pair 2	DW1	TCAACAATGACTTTATTTGG	5
	DW6	GGGAGCTGTAATCATAATGTG	5

1 (Bensch *et al*. 2000), 2 (Waldenström *et al*. 2004), 3 (Hellgren *et al*. 2004), 4 (Beadell *et al*. 2004), 5 (Perkins & Schall 2002). *The letter I in the primer sequences of HAEMNFI and HAEMR2L corresponds to the modified base Inosine.

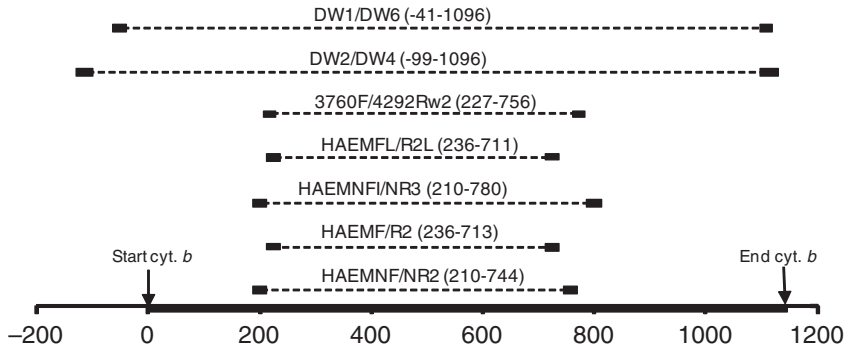

Fig. 2 Locations of amplified fragments of cytochrome *b*. for seven different primer pairs. Boxes represent primer positions. Numbers in parentheses represent the start and end of the amplified fragment in relationship to the start of the cytochrome *b* gene.

Table 2 Main organization of the MalAvi Access database for avian malaria parasites and related haemosporidians

Table name	Primary key	General description of a record
Master	LineageName	A unique lineage
Sequences		A unique lineage
Hosts and sites		The host species and the geographical site* of sampling of a lineage
HostTaxonomy	HostSpecies	A host species and its associated taxonomic information (genus, family, order)
MorphoSpecies		The scientific name of a morphologically identified parasite lineage and a reference to the paper describing it
AlternativeLineageNames		A previously used name for a unique lineage and a reference to the paper where it is used
Geography	Country	A geographical site* in the table 'Hosts and Sites' and its classification to one of 9 global regions
References	Reference	A paper referenced to in any of the above Tables

*For practical reasons, and depending on the information in the source article, the localities have been assigned to countries, states (USA) or islands.

8 *Geography* lists the different geographical localities where lineages have been found and groups them into nine main regions. The tables 'HostTaxonomy' and 'Geography' provide the necessary data used in either the preset query 'Grand Lineage Summary' or other queries depending on the question of interest. Summary tables in Access can further be exported easily into Microsoft Excel and via any text editor, adapted to the infile format of programs for phylogenetical analyses of the selected lineages. Robust phylogenetical analyses require longer sequences than the 479 bp between HAEMF and HAEMR2. Future phylogenetical studies of these parasites should therefore strive to include longer sequences and from multiple genes.

Needless to say, this database may contain errors and we encourage all users to check the original papers before drawing major conclusions from the extracted results, and when errors are encountered, report these to SB. To maintain this database useful, it needs to be updated and new lineage names need to remain unique. Ideally, this should be developed into an automated online system, but such a project would take resources not presently available. In the short term, we offer the simple solution of manually assigning new lineage names and adding new data sets. Preferably, new data should be emailed to SB as an aligned fasta file for the lineages to be named, and as an Excel file structured as the Access table 'Hosts and sites' (template files can be downloaded from the website hosting the database). In the long term, we hope that the research community studying these parasites will find the resources to develop and manage a user-friendly online database.

Future developments of the database could preferably also include information about prevalence, i.e., proportion of infected individuals among screened hosts for each species and geographical site. Many of the published studies contain this information, but entering the data would be a huge commitment beyond our present capacity. Another aspect of importance would be information about vector species for the parasite lineages; however, this data is virtually lacking for most but a few haemosporidian lineages. The initiative of the MalAvi database was inspired by the Barcode of Life Data System (BOLD) for species of the animal kingdom (Ratnasingham & Hebert 2007). In addition to being much more sophisticated, the identification in BOLD is based on another mtDNA gene, the cytochrome *c* oxidase I (COI). A natural extension of MalAvi would, therefore, be to obtain sequences for the COI gene from the identified lineages with a goal to incorporate MalAvi in the global initiative of BOLD.

The value of the database as a source for determining host range and geographical distribution of the parasites will grow with its size. We therefore hope that the presence of a common database will synchronize activities across research groups to the benefit of the understanding of the ecology, evolution and taxonomy in this remarkably diverse group of parasites.

Acknowledgements

We thank the numerous colleagues who contributed with detailed information about their studies and unpublished sequences. In particular, we thank Gediminas Valkiūnas, Asta Križanauskienė and Vaidas Palinauskas for useful discussions during the development of the database. We thank Åke Lindström and Keith Larson for help with Access. The work was supported by the Swedish Research Council (VR), the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas) and the Spanish Ministry of Science and Technology (CGL2007-62937/BOS).

References

Beadell JS, Gering E, Austin J *et al.* (2004) Prevalence and differential host-specificity of two avian blood parasite genera in the Australo-Papuan region. *Molecular Ecology*, **13**, 3829–3844.

Beadell JS, Ishtiaq F, Covas R *et al.* (2006) Global phylogeographic limits of Hawaii's avian malaria. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **273**, 2935–2944.

Beadell JS, Covas R, Gebhard C *et al.* (2009) Host associations and evolutionary relationships of avian blood parasites from West Africa. *International Journal for Parasitology*, **39**, 257–266.

Bensch S, Stjernman M, Hasselquist D *et al.* (2000) Host specificity in avian blood parasites: a study of *Plasmodium* and *Haemoproteus* mitochondrial DNA amplified from birds. *Proceedings of the Royal Society of London B, Biological Sciences*, **267**, 1583–1589.

Bensch S, Pérez-Tris J, Waldenström J, Hellgren O (2004) Linkage between nuclear and mitochondrial DNA sequences in avian malaria parasites – multiple cases of cryptic speciation? *Evolution*, **58**, 1617–1621.

Fallon SM, Bermingham E, Ricklefs RE (2005) Host specialization and geographic localization of avian malaria parasites: a regional analysis in the lesser antilles. *The American Naturalist*, **165**, 466–480.

Feldman RA, Freed LA, Cann RL (1995) A PCR test for avian malaria in Hawaiian birds. *Molecular Ecology*, **4**, 663–673.

Hall TA (1999) BIOEDIT: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Hellgren O, Waldenström J, Bensch S (2004) A new PCR assay for simultaneous studies of *Leucocytozoon*, *Plasmodium*, and *Haemoproteus* from avian blood. *Journal of Parasitology*, **90**, 797–802.

Hellgren O, Krizanauskiene A, Valkiūnas G, Bensch S (2007) Diversity and phylogeny of mitochondrial cytochrome *b* lineages from six morphospecies of avian *Haemoproteus* (Haemosporida, Haemoproteidae). *Journal of Parasitology*, **93**, 889–896.

Krizanauskiene A, Hellgren O, Kosarev V *et al.* (2006) Variation in host specificity between species of avian Haemosporidian

parasites: evidence from parasite morphology and cytochrome b gene sequences. *Journal of Parasitology*, **92**, 1319–1324.

Martinsen ES, Perkins SL, Schall JJ (2008) A three-genome phylogeny of malaria parasites (*Plasmodium* and closely related genera): evolution of life-history traits and host switches. *Molecular Phylogenetics and Evolution*, **47**, 261–273.

Omori S, Sato Y, Isobe T, Yukawa M, Murata K (2007) Complete nucleotide sequences of the mitochondrial genomes of two avian malaria protozoa, *Plasmodium gallinaceum* and *Plasmodium juxtanucleare*. *Parasitology Research*, **100**, 661–664.

Pérez-Tris J, Bensch S (2005) Diagnosing genetically diverse avian malaria infections using mixed-sequence analysis and TA-cloning. *Parasitology*, **131**, 15–23.

Pérez-Tris J, Hellgren O, Krizanauskiene A *et al.* (2007) Within-host speciation of malaria parasites. *PLoS ONE*, **2**, e235. DOI: 10.1371/journal.pone.0000235.

Perkins SL, Schall JJ (2002) A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. *Journal of Parasitology*, **88**, 972–978.

Ratnasingham S, Hebert PDN (2007) BOLD: the Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Reullier J, Pérez-Tris J, Bensch S, Secondi J (2006) Diversity, distribution and exchange of blood parasites meeting at an avian moving contact zone. *Molecular Ecology*, **15**, 753–763.

Ricklefs RE, Fallon SM (2002) Diversification and host switching in avian malaria parasites. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **269**, 885–892.

Svensson LME, Ruegg KC, Sekercioglu CH, Sehgal RNM (2007) Widespread and structured distributions of blood parasite haplotypes across a migratory divide of the Swainson's thrush (*Catharus ustulatus*). *Journal of Parasitology*, **93**, 1488–1495.

Valkiūnas G (2005) *Avian Malaria Parasites and Other Haemosporidia*. CRC Press, Boca Raton.

Valkiūnas G, Krizanauskiene A, Iezhova TA, Hellgren O, Bensch S (2007) Molecular phylogenetic analysis of circumnuclear hemoproteids (Haemosporida, Haemoprotidae) of Sylviid birds, with description of *Haemoproteus parabelopolskyi*, nov. *Journal of Parasitology*, **93**, 680–687.

Valkiūnas G, Atkinson CT, Bensch S, Sehgal RNM, Ricklefs RE (2008) Parasite misidentifications in GenBank: how to minimise their number? *Trends in Parasitology*, **24**, 247–248.

Waldenström J, Bensch S, Kiboi S, Hasselquist D, Ottosson U (2002) Cross-species infection of blood parasites between resident and migratory songbirds in Africa. *Molecular Ecology*, **11**, 1545–1554.

Waldenström J, Bensch S, Hasselquist D, Östman Ö (2004) A new nested PCR method very efficient in detecting *Plasmodium* and *Haemoproteus* infections from avian blood. *Journal of Parasitology*, **90**, 191–194.