NRC Research Press

ARTICLE

# Closed-Tube Barcoding[1]

Nicky M. Sirianni, Huijun Yuan, John E. Rice, Ronit S. Kaufman, John Deng, Chandler Fulton, and Lawrence J. Wangh

**Abstract:** Here, we present a new approach for increasing the rate and lowering the cost of identifying, cataloging, and monitoring global biodiversity. These advances, which we call Closed-Tube Barcoding, are one application of a suite of proven PCR-based technologies invented in our laboratory. Closed-Tube Barcoding builds on and aims to enhance the profoundly important efforts of the International Barcode of Life initiative. Closed-Tube Barcoding promises to be particularly useful when large numbers of small or rare specimens need to be screened and characterized at an affordable price. This approach is also well suited for automation and for use in portable devices.

*Key words:* Pan-phylum probes, LATE-PCR, Lights-On/Lights-Off probes, Closed-Tube Barcoding, fluorescent signatures.

**Résumé :** Les auteurs présentent une nouvelle approche pour augmenter le taux de succès et réduire le coût pour l'identification, le catalogage et le monitoring de la biodiversité globale. Ces avancées, que les auteurs ont nommé « Closed-Tube Barcoding » (codage à barres en tube fermé), représentent une application d'une suite de technologies PCR éprouvées inventées au sein de leur laboratoire. Le codage à barres en tube fermé s'appuie sur, et vise à étendre, les efforts extrêmement importants découlant de l'initiative du International Barcode of Life. Le codage à barres en tube fermé promet d'être particulièrement utile quand de grands nombres de spécimens rares ou de petite taille doivent être examinés et caractérisés à un prix raisonnable. Cette approche est également bien adaptée pour l'automatisation et l'emploi avec des appareils portatifs. [Traduit par la Rédaction]

*Mots-clés :* sondes pan-spécifiques, LATE-PCR, Sondes noir et blanc (« Lights-On/Lights-Off probes »), codage à barres en tube fermé (« Closed-Tube Barcoding »), signatures en fluorescence.



## Introduction

This paper presents new ideas and new possibilities for combining a large body of information with a proven technology for amplification and analysis of single-stranded DNA. The information comes from the extensive and ever-growing database of barcoding sequences that are being collected by the International Barcode of Life project. The technology known as LATE-PCR is a method for non-symmetrically amplifying and characterizing single-stranded DNA targets in closed-tube reactions. LATE-PCR can be combined with a suite of synergistic methods and chemistries for the characterization of single-stranded DNA targets in closed-tube reactions. These technologies have been used extensively to construct diagnostics assays for infectious diseases (Carver-Brown et al. 2012; Pierce et al. 2013) and cancer gene targets (Tetrault et al. 2014). We have demonstrated the feasibility and reliability of these technologies for analysis of barcoding sequences in a modest set of proof-of-principle experiments (Rice et al. 2014), and we have now decided to name this method "Closed-Tube Barcoding". The present paper describes in both experimental terms and theoretical terms how Closed-Tube Barcoding assays might be used to conveniently identify very large numbers of species at various taxonomic levels. Closed-Tube Barcoding is not a replacement for barcoding by sequencing because the species-specific fluorescent signature generated by Closed-Tube Barcoding cannot be read at the nucleotide level. Rather, Closed-Tube Barcoding offers a way to increase the rate and lower the cost of identifying, cataloguing, and monitoring global biodiversity by eliminating the need to sequence every specimen. Closed-Tube Barcoding will be particularly cost effective for rapid analysis of specimens whose barcode

Received 29 January 2016. Accepted 15 July 2016.

Corresponding Editor: Dirk Steinke.

**N.M. Sirianni, H. Yuan, J.E. Rice, R.S. Kaufman,* J. Deng, C. Fulton, and L.J. Wangh.** Brandeis University, Department of Biology, Waltham, MA 02453-9110, USA.

**Corresponding author:** Lawrence J. Wangh (email: Wangh@brandeis.edu).

*Present address: Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA.

[1]This paper is part of a special issue entitled Barcodes to Biomes.

sequences are already in a database. Sequences that have been observed before will have recognizable fluorescent signatures and will therefore not have to be sequenced again. When a novel signature is found, it will indicate the presence of an unknown sequence that can then be analyzed at the nucleotide level via our convenient and cost-effective Dilute-'N'-Go protocol (Rice et al. 2007; Jia et al. 2010). As described in this paper, we envision that Closed-Tube Barcoding will also reduce costs by permitting the use of smaller samples with standardized sets of reagents and, in the future, will also permit analysis of samples in the field.

## Materials and methods

This paper describes both actual Closed-Tube Barcoding assays and thought-based assays determined by in silico calculations of the melting temperatures ($T_m$s) of probe/target hybrids generated by using published cytochrome *c* oxidase 1 (CO1) sequences as targets and sets of probes that are short DNA sequences complementary to contiguous segments of these target sequences. In the case of the thought-based assays, no amplification reactions were carried out to verify the temperature-dependent patterns of the calculated probe/target hybrids. Our confidence in the utility of these assays comes from our extensive experience using LATE-PCR to design actual Closed-Tube Barcoding assays, as well as many assays in the field of human in vitro diagnostics (Carver-Brown et al. 2012; Pierce et al. 2013; Tetrault et al. 2014).

### LATE-PCR

Linear-After-The-Exponential (LATE) PCR is an advanced form of non-symmetric PCR (Sanchez et al. 2004; Pierce et al. 2005). It is used to generate single-stranded amplicons from one or more target genes of interest. Over the course of 30–70 thermal cycles, the single-stranded amplicons accumulate in 10–20 fold excess over the double-stranded amplicons produced during the initial rounds of amplification. The single-strands are analyzed at end-point in the same closed-tube by simply lowering the temperature, thereby allowing sets of fluorescently labeled probes to coat each single-stranded target in a temperature-dependent manner. The probes of choice for construction of closed-tube assays are Lights-On/Lights-Off probes (Rice et al. 2012).

### Lights-On/Lights-Off probes

Each Lights-On probe is a low-temperature dual-labeled Molecular Beacon (Rice et al. 2012), comprised of a target binding sequence of ~20–35 nucleotides (nt) flanked by 1–2 nt on the two ends that can hybridize to each other to stabilize the interaction of the fluorophore on one end and the quencher on the other when the probe is not bound to its target sequence. Thus the fluorescent signal emitted by a Lights-On probe in solution is lower than that emitted when the probe is bound to its target sequence because the fluorophore and the quencher are rigidly separated by hybridization. Under these circumstances, the quencher on the Lights-Off probe completely turns off the fluorescence of the bound Lights-On probe.

The analytical power of Lights-On/Lights-Off probes rests in the fact that each probe within a set of probes is mismatch tolerant, meaning that it can hybridize to its fully complementary target sequence, as well as to increasingly mismatched variants of that sequence, when the detection temperature is lowered. Probe/target hybrids that differ by 1 °C or more can readily be distinguished. In the case of drug-sensitive versus drug-resistant bacteria, the sequence of the drug-sensitive strains is considered the wild-type sequence and sets of Lights-On/Lights-Off probes that are perfectly complementary to this sequence can be designed and used to look for sequence variations in other strains (see Fig. 1). In contrast, in the case of Closed-Tube Barcoding for species identification there is no wild-type sequence. Under this circumstance, all of the probes in a set can be designed as consensus probes generated via statistical analysis of equivalent regions in large numbers of CO1 target sequences. Although none of the probes in a set of consensus probes is likely to be perfectly matched to any particular target, sequence variations among different targets nevertheless generate distinct patterns of probe/target hybridization.

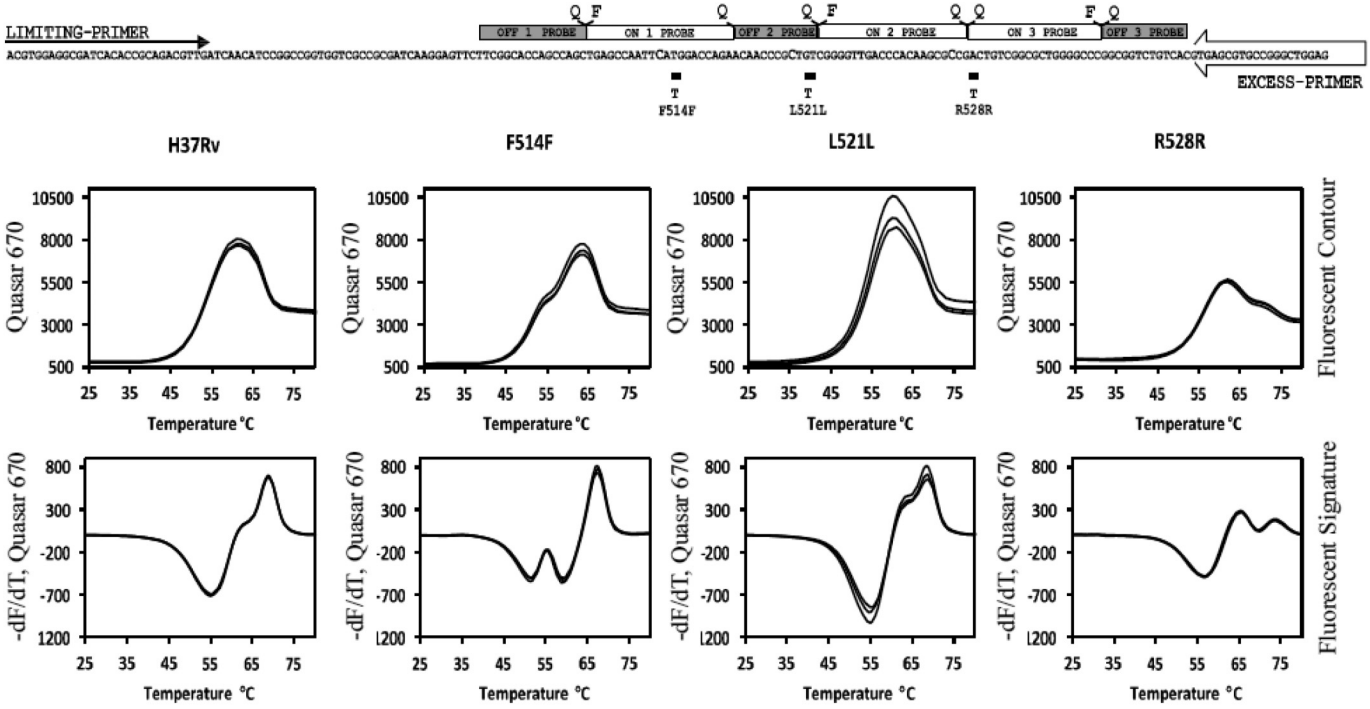
Each probe is mismatch tolerant and can be used to distinguish >10 possible variants over a detection temperature of at least 20 °C (Pierce and Wangh 2014). Since each probe in a set has its own pattern of increasing or decreasing $T_m$s to its own regional variants, sets of 2, 3, …*n* Lights-On/Lights-Off probes can potentially distinguish $>10^n$ related target sequences. For instance, three probes can theoretically distinguish >1000 sequence variants and 15 probes (the number needed to coat an entire CO1 amplicon of ~650 nt long) can theoretically distinguish $>10^{15}$ sequence variants.

Clearly, it is not possible to empirically test this prediction, but the large magnitude of this number suggests that it is plausible that a single set of probes could be designed to uniquely identify each species of animals (or plants) on earth that has a unique DNA barcode sequence. As discussed below, we have taken the first step to designing such a set of probes in silico, and we have used these probes to calculate the individual $T_m$ values for each of 15 adjacent probes aligned on a wide range of CO1 barcode sequences across many phyla. We have also used the same in silico approach to design smaller sets of probes that target portions of full-length CO1 barcodes that are sufficiently informative to distinguish all of the species in a smaller group of organisms.

### Sequences and methods used for the in silico calculations

The species names and acquisition numbers for all of the in silco assays were all downloaded from Barcode of

**Fig. 1.** The *rpoB* gene target of *Mycobacterium tuberculosis* was amplified with a Limiting Primer and an Excess Primer to generate a single-stranded DNA amplicon. Four different strains were used, each with a different drug-sensitive genotype due to the presence or absence of a single neutral point mutation. Each product strand was analyzed at end-point by lowering the temperature to coat the target strand with three pairs of Lights-On/Lights-Off probes. The Lights-On probes were all labeled with Quasar 670. The probes were then melted off to generate four distinct fluorescent contours that were then converted mathematically into four fluorescent signatures.




Life Data System (BOLD) v.3 (Ratnasingham and Hebert 2007) and are provided in the supplementary data[2]. All of the probe/target $T_m$s in both constructed assays and the in silico assay were calculated using Visual OMP software version 7.8.42.0 (DNA Software, Inc. Ann Arbor, Mich.) with the following parameters turned on: assay temperature = 65 °C, monovalent = 0.05M, $Mg^{++}$ = 0.003M, duplex polymer salt correction = OFF.

In the case of the in silico assay of species of *Naegleria*, we used sequences generated in our laboratory during the course of a study aimed at comparing sequence-relatedness of these species using two target sequences, CO1 and the previously used ITS1 region of the ribosomal plasmid (DeJonckheere 2004). The purpose of this ongoing study was to determine the extent to which these two trees were in agreement or disagreement regarding the relatedness of different strains of *Naegleria* collected all over the world. A major conclusion from that study is that the trees of sequence-relatedness using the two sequences are similar. A manuscript describing our analysis and providing all of the sequence data is now in preparation (R.S. Kaufman, N.M. Sirianni, J.E. Rice, B.S. Robinson, L.J. Wangh, and C. Fulton). For the purpose of the present paper, the *Naegleria* tree of sequence-relatedness as shown below in Fig. 7 was constructed using MEGA6 (build #6140226) (Tamura et al. 2013).

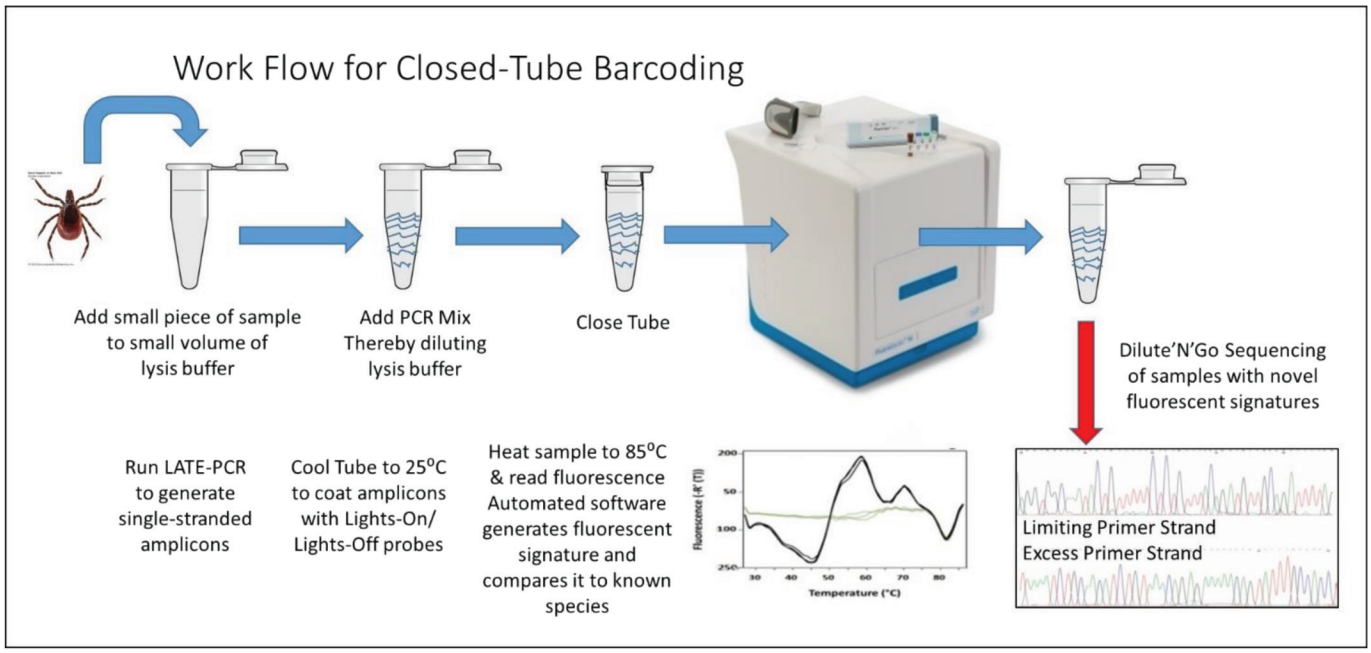
**Primer design**

In the case of the *rpoB* gene of *Mycobacterium tuberculosis*, the Limiting Primer, the Excess Primer, and the target sequence are as described in Rice et al. 2012. In the case of the CO1 gene of *Ixodes scapularis* there were three Limiting Primers (5′-ACGACATAGTTATACCTGTAATAATTGGGGGTTTTGG-3′, 5′-CACGACATAGTTATACCAATTATAATCGGAGGATTTGG-3′, 5′-CGACATAGTTATACCTGTTATGATTGGTGGGTTTGG-3′) and two Excess Primers (5′-CTGTAATTAAAACTGATCATAC AAATAAAGGTATTCG-3′, 5′-GCTATGACTGATCATACAAA TAAAGGTATTCG-3′) that were designed by comparison of the CO1 sequences from the 20 species of ticks listed in the supplementary data[2]. In addition, there were four consensus probes having the sequences shown in Fig. 3. The LATE-PCR primers referred to in the in silico assay for the CO1 sequences of *Naegleria*, *Wallardia*, and *Tetramitis* were Limiting Primer 5′-AGTAATGCCAATCTTATTTGGAGG ATTTGGTAAC-3′ and Excess Primer 5′-TACTGGGTCATAG AAAGAAGTATTAAAATTAC-3′.

As used in this paper, the term Folmer primers means any pair of primers that target the relatively conserved

[2]Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/gen-2016-0026.

**Fig. 2.** The work flow for Closed-Tube Barcoding. Sample preparation can either be carried out in a separate tube or directly in the same tube in which PCR amplification is carried out. In either case, the enzymes or chaotropic salts used for sample preparation are simply diluted away by combining a small sample volume and a larger volume of the PCR mix. The tube is then closed, LATE-PCR amplification takes place at annealing and extension temperatures above that of the probe melting temperature ($T_m$). Product detection takes place at end-point by simply dropping the temperature down to coat it with probes and then melting them off to create a fluorescent contour that is then converted mathematically into a fluorescent signature. Not shown here are the downstream steps of data entry into BOLD, including the fluorescent signature for each species.




sequences that are typically used to amplify the full-length CO1 sequence used for conventional barcoding-by-sequencing experiment (Folmer et al. 1994), except that these primers have been converted into a Limiting Primer and an Excess Primer for LATE-PCR in agreement with the formula $T_{m_{[0]}}{}^{L} - T_{m_{[0]}}{}^{X} \geq 0$, as explained in Sanchez et al. 2004. The pan-phylum in silico assay described here will likely require use of several pairs of Folmer primers.

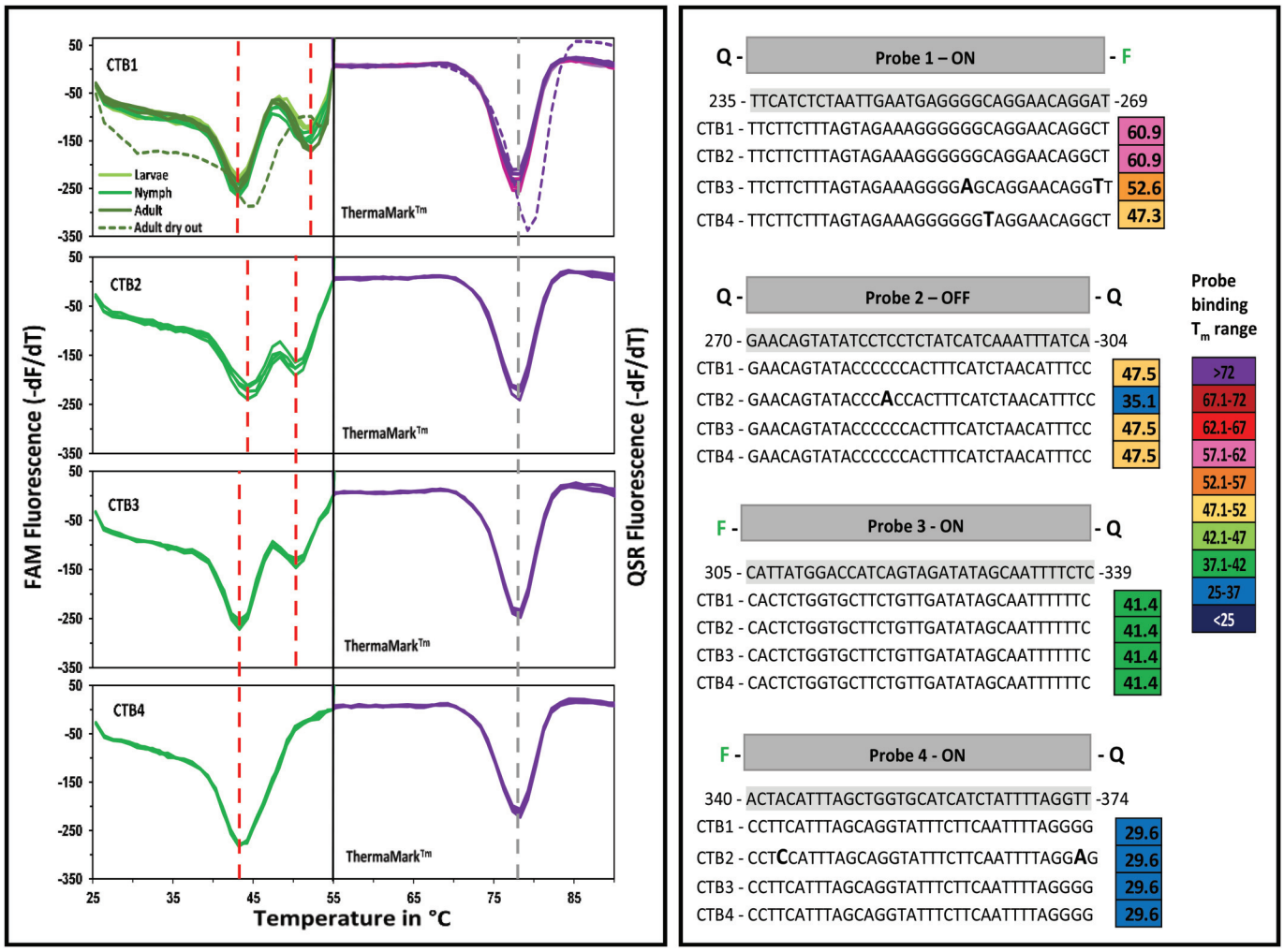
## Results

Figure 2 is a workflow diagram that shows how a specimen collected in the field is processed in the laboratory and is then amplified and analyzed in a closed-tube reaction in about 3.5 h. These assays can be run in several real-time PCR devices. We prefer the Stratagene 3005Mx Pro (Agilent, Santa Clara, Calif.) and the soon-to-be-released, Fluorocycler96® from Hain Lifescience. Both of these machines carry out 10–25 μL reactions in 250 μL conical PCR tubes or 96-well plates.

Once the specimen reaches the laboratory and its metadata are recorded in the database, the entire specimen or a fragment of it (even as small as a single cell) is lysed in appropriate lysis buffer with or without mechanical disruption as needed. We have shown, for instance, that samples containing 1 – 10 000 eukaryotic cells can readily be disrupted using either protease K (Pierce et al. 2002) or chaotropic salts (Hartshorn et al. 2005) in buffers that are compatible with subsequent PCR amplification if they are diluted sufficiently. No clean-up is required because mitochondrial genomes are present in high copy number per cell and because separation of DNA from its bound proteins is sufficient to permit amplification. The advantage of this approach is that preparation is quick and sensitive down to single molecules (Osborne et al. 2009). The ease of this approach is further enhanced by use of LATE-PCR and Lights-On/Lights-Off probes because probe/target hybridization is measured at the end of the reaction, rather than in real-time. This means that partial inhibition of amplification due to the presence of PCR inhibitors can usually be overcome by use of more cycles of single-strand DNA synthesis. Highly reproducible fluorescent signatures are obtained as long as about 10 cycles of singled-stranded amplification have taken place during LATE-PCR (Rice et al. 2012).

Once the specimen and the LATE-PCR mix are combined in the PCR tube, the lid is closed and amplification is carried out by thermal cycling for 30–70 cycles depending on the starting concentration of the target. Our LATE-PCR reaction mixtures typically contain ThermaStop™, ThermaGo™, and ThermaMark™, which are three types of modified oligonucleotides developed in our laboratory to suppress mis-priming before, during, and after amplification, as well as to rule out evaporation in every reaction (see Fig. 3).

**Fig. 3.** DNA was prepared from six deer ticks, *Ixodes scapularis,* collected in New England and was amplified using a set of three Limiting Primers and two Excess Primers designed to amplify many species of North American ticks (see supplementary data[2]). Five replicate reactions for each sample were analyzed at end-point using four Lights-On/Lights-Off probes that were designed as consensus probes. All Light-On probes in the set were labelled with FAM. Each reaction also contained ThermaMark™, a non-amplifiable double-Quasar labelled oligonucleotide that has a defined melting temperature ($T_m$) of 80 °C and therefore serves as an internal control for reaction fidelity. The results show that these reactions generated four different Closed-Tube Barcodes, and samples of each of these reactions had four different sequences. In silico analysis of the $T_m$ of each probe/target hybrids established that each Closed-Tube Barcode had its own pattern of pseudo-colors, even though consensus probe 4 was too mis-matched to its target sequence to hybridize at a high enough temperature for melt curve analysis.






At the end of LATE-PCR amplification, the temperature of the reaction is lowered to about 25 °C to allow all of the Lights-On/Lights-Off probes to hybridize to their respective adjacent regions of the target sequence. The bound probes are then melted off by heating the reaction slowly. Total fluorescence in the system is comprised of background signals as well as increases and decreases over background, all of which are condition and temperature dependent. The fluorescent contour of a sample is its total fluorescence as a function of temperature and is characteristic of that sample. We transform this contour into a fluorescent signature by calculating the rate at which fluorescence changes as a function of temperature. Fluorescent signatures are easier to "read" than fluorescent contours, as their high and low values are more dramatic and are independent of signal strengths (see Fig. 1; Rice et al. 2012).

The primary goal of this paper is to illustrate some of the ways in which Closed-Tube Barcoding could be used to enhance the goals of iBOL. Since the method of Closed-Tube Barcoding is largely unknown to investigators familiar with barcoding-by-sequencing, we begin by illustrating how LATE-PCR and Lights-On/Lights-Off probes can be combined to build highly informative close-tube assays. Figure 1 shows a 150 nt-long sequence from the *rpoB* gene of *M. tuberculosis* and the Limiting Primer and the Excess Primer used to amplify this sequence. The central portion of this target is known as the RRDR sequence because almost 100 mutations in this region are known to cause resistance to rifampicin, a first line antibiotic.

By convention, the wild-type sequence of this gene comes from H37Rv, a drug sensitive laboratory strain. Figure 1 also shows the positions of three single nucleotide polymorphisms that are rare neutral mutations at the protein level (F514F, L521L, and R528R) and are therefore also drug sensitive. In the realm of in vitro diagnostics, it is important not to confuse these variants with the many mutations that cause drug resistance. To accomplish this objective, we coat the target strand with three pairs of Lights-On/Lights-Off probes (for details see Rice et al. 2012). All of these Lights-On probes are labelled with Quasar 670. Figure 1 shows that the three variants fall under three different probes. The replicate reactions for these sequences generated four easily distinguished, highly reproducible fluorescent contours and their corresponding fluorescent signatures. These results are consistent with our extensive experience that >95% of single nucleotide polymorphisms in several *M. tuberculosis* gene targets generate distinct fluorescent signatures compared to the wild-type sequences.

Figure 3 demonstrates that informative sets of Lights-On/Lights-Off probes can be designed in the absence of a "wild-type" sequence and that Closed-Tube Barcoding can be designed as consensus probes derived by statistical analysis of many CO1 sequences. In this preliminary analysis, we used deer tick, *I. scapularis*, collected in the woods of New England. A 140 nt-long portion of *I. scapularis* CO1 gene sequence was amplified using a set of primers that has the capacity to amplify the same target in any one of 20 different species of North American ticks (see Materials and methods and the supplementary data[2]). The amplicon was probed using four consensus probes whose sequences were chosen on the basis of the nucleotide frequency at each position in the 20 target sequences.

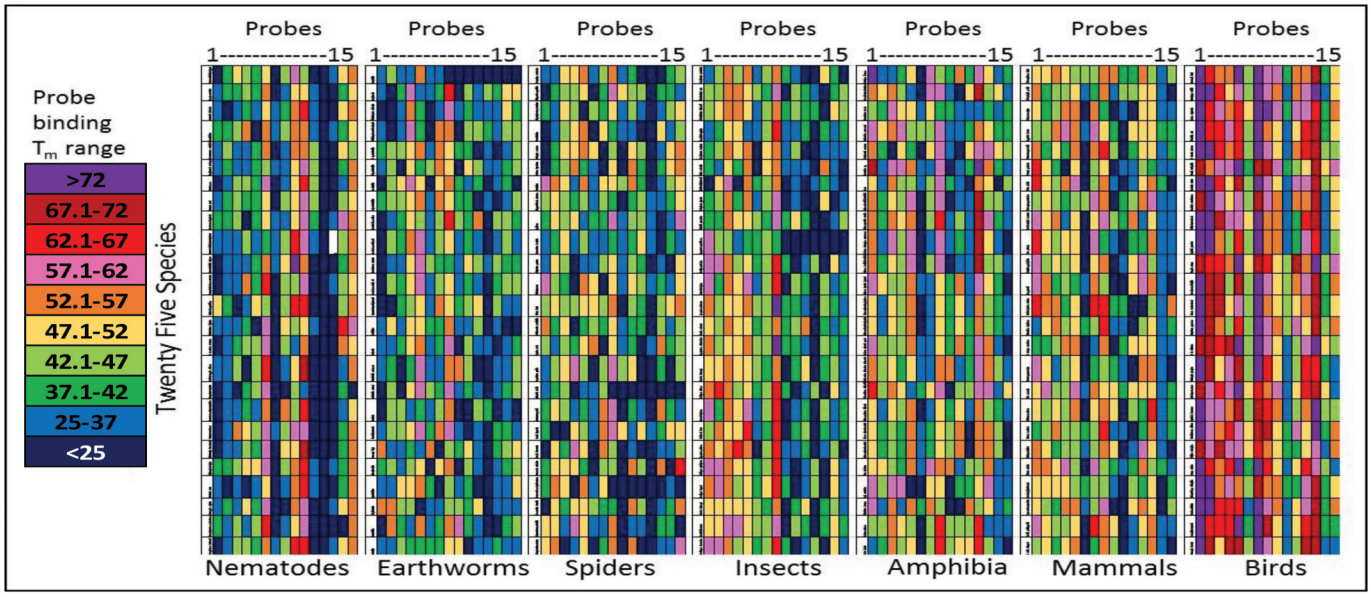
The results in Fig. 3 illustrate the power of the Closed-Tube Barcoding approach. Six ticks were examined using five replicate reactions of each. Four different fluorescent signatures were observed, and all of the replicates were highly reproducible except for one replicate from the adult tick in CTB-1. The fluorescent signature in this specimen is shifted to the right. This artifact is due to evaporation of a sample during PCR thermal cycling. The $T_m$s of all probe/target hybrids in such samples shift to the right because of the increase in the salt concentration. This fact is borne out by examination of the ThermaMark™ "valley" in this same sample. ThermaMark™ is labeled using Quasar 670, while the three Lights-On probes used to analyze the CO1 target sequence are all labelled with FAM. ThermaMark™ fidelity in all of the other samples serves to validate the differences in the four signatures. In the case of CTB-1, our specimen collection included a larva, a nymph, and an adult. All three specimens had identical fluorescent signatures.

Samples from each of the four fluorescent signatures in Fig. 3 were further analyzed by Dilute-'N'-Go protocol sequencing (Rice et al. 2007). The results show that the sequence for each fluorescent signature is different and that even a single point mutation is sufficient to dramatically alter the fluorescent signature (compare probe 1 target sequences for CTB-1 and CTB-4). The fact that this small set of data for only six specimens of *I. scapularis* revealed several different fluorescent signatures is fully consistent with reports in the literature that there are 55 haplotypes of *I. scapularis* based on variations within this 140-nt stretch to CO1 and its flanking regions (Mechai et al. 2013; Sakamoto et al. 2014). Indeed, each of the sequences we obtained is perfectly matched to one or more haplotype in the data base (BOLDSYSTEMS, http://www.boldsystems.org/index.php/Public_SearchTerms?query=%22Ixodes%20scapularis%22[tax]).

Figure 3 also serves to explain the approach used in the other experiments in this paper for in silico analysis of probe/target hybrids. The $T_m$ of each probe/target hybrid was calculated using Visual OMP software (see Material and methods). A pseudo-color scale comprised of 10 steps was then created and each calculated $T_m$ was assigned a corresponding pseudo-color. The seven pseudo-colors in the middle of the scale span steps of 5 °C each, while the steps at the top and the bottom of the scale are wider because $T_m$s in these ranges are inherently uninformative, see below. Probe 4 in Fig. 3 is an example of an uninformative probe, as its $T_m$ is too low to bind to any of the *I. scapularis* sequences. Thus, probe 4 cannot currently distinguish the two mutations in CTB-2 that are different from the target sequences of CTB-1, CTB-3, and CTB-4. As discussed below, we call this phenomenon probe dropout. This limitation of probe 4 is due to the fact that it is a consensus probe rather than an *I. scapularis* specific probe. Probe 4 will be redesigned and additional primers and probes covering more of the CO1 sequence will be included in the reaction for future studies that focus on detecting all of the haplotypes of *I. scapularis*.

All of the Lights-On probes in Figs. 1 and 3 are labeled using a single-colored fluorophore. In our experience, this approach works well when there are up to four Lights-On probes in a set. The advantage of using just one color for all Lights-On probes in a set is that the other colors in a four to five color PCR machine can be used to analyze additional target sequences. The disadvantage of using just one color is that it is not as informative as using more colors. For instance, a full-length CO1 target sequence of about 650 nt can be analyzed by coating it with 15 probes, see Fig. 4. If the probe set is subdivided into three subsets, each labeled with a different color and each covering a portion of the long target sequence, the $T_m$ range is used three times; this makes it possible to observe sequence differences and similarities of contiguous regions of the long target sequence in closely related Closed-Tube Barcodes. Examples of using both one color and three colors in the same set of 10 probes are

**Fig. 4.** Pseudo-color map of the approximate melting temperature ($T_m$) calculated for each probe in an initial set of 15 pan-phylum probes hybridized to 25 CO1 sequences for each of seven classes of animals.




shown in our proof-of-principle closed-tube analysis of five species of nematodes (Rice et al. 2014).

### Big picture thought-based Closed-Tube Barcoding assays

#### *Can a single set of pan-phylum Closed-Tube Barcoding probes be designed for identification of all animal species?*

A single set of probes for identification of any animal species on earth would be very useful. Such a pan-phylum set of probes would greatly reduce the cost of analysis, as the mass-produced reagents could be mixed and dried in a reaction vessel for analysis of any specimen. Such an approach would be ideal for Closed-Tube Barcoding analysis because all that would be required to start the reaction is the addition of the mitochondrial DNA template prepared as described above.

A unique strategy for designing a set of pan-phylum Closed-Tube Barcoding probes is summarized in Table 1. Fifteen probes were chosen from adjacent regions of 15 different CO1 sequences drawn from different species in multiple taxa. This strategy insures that this set of probes will not be perfectly complementary to any particular species in the animal kingdom, but at least some probes in the set would probably bind to every species, regardless of its taxonomic identity. The $T_m$ of each probe/target hybrid was assigned a pseudo-color as described in pseudo-color map Fig. 3 to obtain the colored barcodes shown in Fig. 4. The results are consistent with our basic expectation that at least one of the 15 mismatch-tolerant probes in the set would hybridize at some temperature above 25 °C to all of the species tested (see the supplementary data[2] for genus and species names, as well as their CO1 sequence identifiers in BOLD, for each group). In fact, all 25 species in each of seven different types of animals bound several probes at various temperatures above 25 °C. Thus, all animals within the same group could be distinguished from each other. The data in Fig. 4 also demonstrate that $T_m$s of many probes varied widely both within one group of animals and among different groups of animals. For instance, probe 12 failed to bind to almost all of the nematode species, but it bound at mid-to-high $T_m$s to all of the birds. This observation is in accord with the fact that the probes used here are mismatch tolerant, plus the fact that each region of a CO1 sequences is mutable.

On average, the distribution of colors in Fig. 4 shifts from the dark blue and blues for nematodes to the oranges and reds for the birds (left to right in Fig. 4). This shows that on average these 15 probes are more perfectly matched to bird CO1 sequences than they are to nematode CO1 sequences. This conclusion is borne out by calculating the average $T_m$ value per probes across these species (Table 2). The results in Table 2 are highly instructive because they point the way forward to the design of an ideal set of pan-phylum probes. Specifically, we estimate that the average $T_m$ of all probes in the ideal set of pan-phylum probes would have an average of about 45–55 °C. Fortunately, designing such an ideal set of probes should be relatively easy since any nucleotide in any probe can be altered to increase or decrease the $T_m$ of a particular probe, and probes can also be built using non-natural nucleotides. Further design work on an optimal set of pan-phylum probes is underway in our laboratory.

### Smaller picture thought-based Closed-Tube Barcoding assays

#### *Closed-Tube Barcoding of mosquitoes and malaria* Plasmodium *– an example of medical interest*

When a probe fails to bind to a target at a temperature above 25 °C we call it a probe dropout. Multiple probe dropouts in a target, such as the nematode targets in

**Table 1.** Full length CO1 sequences from different, randomly chosen species (see supplementary data[2]) aligned and divided into contiguous segments 40 nucleotides long.

| Strategy for initial design of 15 universal CO1 probes for pan-phylum analysis | | | | |
|---|---|---|---|---|
| Nucleotide position | Phylum | Class | Genus and species | Probe sequence |
| 20–60 | Chordata | Amphibia | *Atelopus limosus* | GGGGCATGGGCTGGCATAGTGGGCACAGCTCTAAGTCTATT |
| 61–100 | Annelida | Clitellata | *Eunicida* | TATTCGAGTAGAATTAGGGCAACCTGGGTCTTTTCTCGGA |
| 101–140 | Nematoda | Enoplea | *Enoplida* | TCGGATCAAATTTATAATGTTCTTGTTACGGCGCATGCCT |
| 141–180 | Nematoda | Adenophorea | *Xiphinema breviocolle* | GGTTGTGGCGCTGGAACTGGATGAACTATTTATCCTCCGCTT |
| 181–220 | Chordata | Reptillia | *Homopus signatus* | ATACCAATCATAATTGGTGGATTTGGAAACTGACTCACACC |
| 221–260 | Chordata | Reptillia | *Strongylidae* | ACCAGATATAAGTTTTCCTCGTTTAAATAATTTAAGGTTT |
| 261–300 | Chordata | Reptillia | *Orlitia borneensis* | AGACATGGCATTCCCACGCATAAATAATATGAGCTTTTGA |
| 301–340 | Arthropoda | Insecta | *Argia moesta* | GCTTAGTAGAAAGAGGGGCAGGGACTGGGTGAACTGTTTACC |
| 341–380 | Nematoda | Secementea | *Caenorhabditis elegans* | TTGTAGATATAGGTTGTGGGACTAGGTGAACAGTCTACCCA |
| 381–420 | Chordata | Mammalia | *Schistonchus caprifici* | TGAATACTTCTTTTTTTGATCCTAGGATAGGGGGTAATCCT |
| 421–460 | Chordata | Aves | *Tadarida teniotis* | CTCAATCCTTGGTGCCATTAATTTTATTACTACTATTATTA |
| 461–500 | Chordata | Amphibia | *Centrolene sp.* | AACATAAAACCCCCATCAATAACTCAATATCAAACCCCAC |
| 501–540 | Chordata | Mammalia | *Molossus molossus* | ATACCAAACACCACTATTTGTATGATCTGTATTAATTACA |
| 541–580 | Nematoda | Secementea | *Spauligodon* | GCTTCTTTGCCGGTGTTAGCTGGTGCTCTGACTATACTTA |
| 581–620 | Arthropoda | Arachnida | *Hermania reticulata* | TTAACCGATCGTAATTTTAACACTTCTTTTTTTGATCCTA |

**Note:** One segment from each sequence was used to design a complementary probe.

**Table 2.** Average melting temperature ($T_m$) value of pan-phylum probes across tested species.

| | Average $T_m$ (°C) | SD |
|---|---|---|
| Nematodes | 36 | 3.2 |
| Earthworms | 37 | 3.6 |
| Spiders | 37 | 4.4 |
| Insects | 41 | 4.9 |
| Amphibia | 43 | 2.1 |
| Mammals | 42 | 2.6 |
| Birds | 57 | 1.3 |

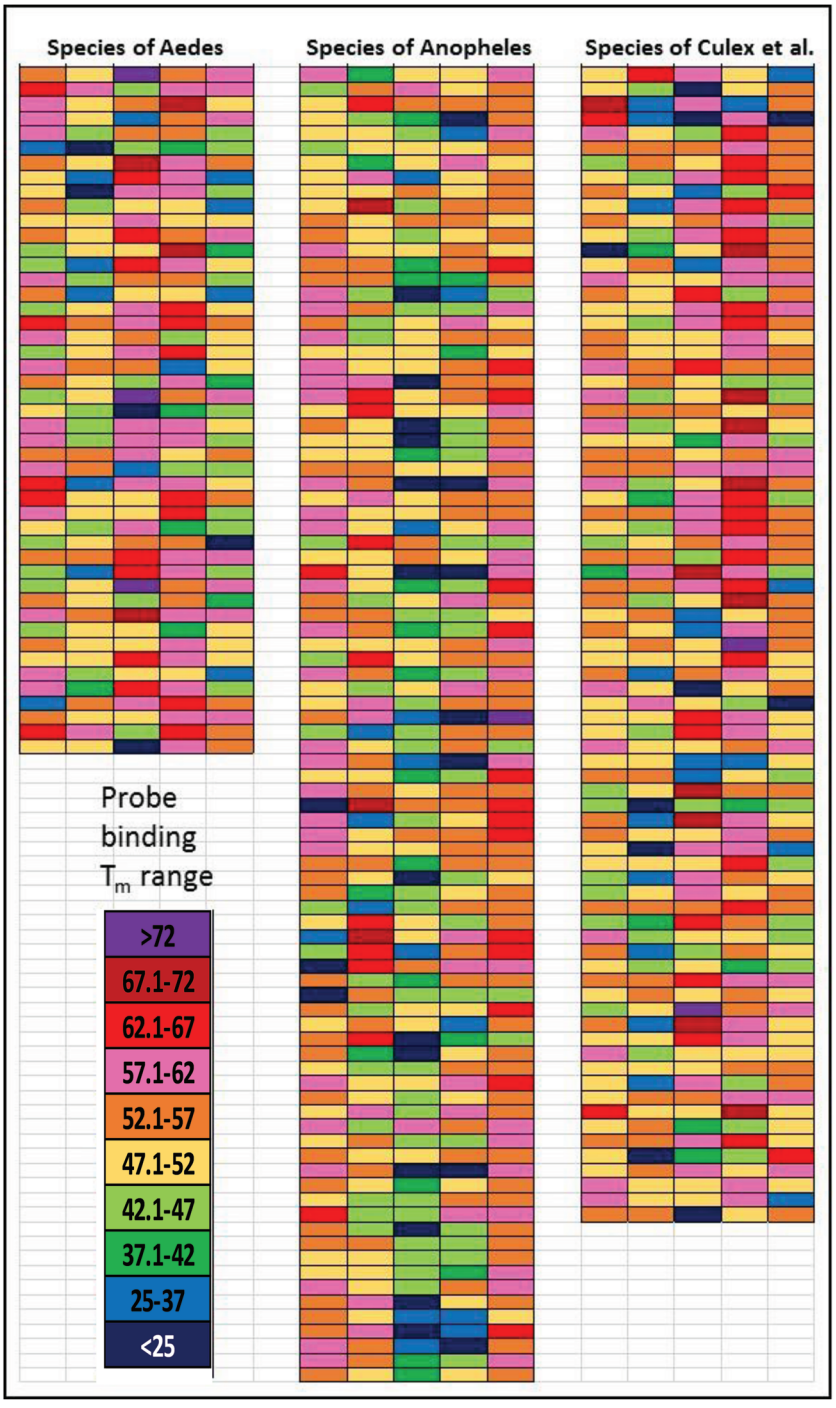
**Note:** Each animal type was comprised of 25 representative species (see Fig. 4; supplementary data[2]). The $T_m$ value for each of the 15 universal probes was calculated for each probe/target hybrid in each of the 25 species, and these 15 values were summed to generate a total $T_m$ value for each of the 25 species. These total $T_m$ values were then divided by 15 to give an average probe $T_m$ value for each of the representative species. The average probe $T_m$ value and standard deviation of that value among the 25 species in each set was then calculated.

Fig. 4, detract from the usefulness of that probe set, as there is a risk that the sequence differences between closely related species might go undetected using such a set of probes. In contrast, rare cases of probe dropout do not detract from the usefulness of a probe set, as they simply contribute to the unique fluorescent signature for that species. Figure 5 illustrates this point by focusing on a portion of the CO1 sequences in 216 species of mosquitoes from 14 genera. A unique fluorescent signature for each of these species is generated using just five consensus probes. The average $T_m$ of each probe in this set for each genus is within the range for an ideal set of probes, Table 3.

The mosquito Closed-Tube Barcoding assay shown in Fig. 5 could be used for rapid and accurate, one-by-one identification of mosquitoes. Analysis would not require expertise in mosquito morphology and, as demonstrated in the case of ticks, the species-specific virtual barcode would be the same for eggs, larvae, and adult mosquitoes in the same haplotype. The mosquito assay could be refined in two ways to make it even more informative. First, it could be combined with a malaria *Plasmodium* Closed-Tube Barcoding assay. All that is required is to add a second pair of primers and a second set of five probes to amplify and identify each of the CO1 sequences of the species of malaria *Plasmodium*, Fig. 6. The malaria probes and the mosquito probes would be labelled in different fluorescent colors so the reactions could be completed in the same well. In silico analysis reveals that the five probes for mosquito CO1 sequences do not cross-hybridize to the *Plasmodium* CO1 sequences, and vice versa. This is a case of extreme, yet a very useful example of, probe dropout. A single tube assay can thus be built for identification of mosquito species, as well as detection and identification of malaria species that may be present in each mosquito.

A second refinement that could be applied to the above assays would be to make them mosquito-species specific. This can be achieved by replacing the primers for the CO1 Folmer sequences with one or more pairs of LATE-PCR primers to species-specific sequences within the full-length CO1 sequences. Under these circumstances, DNA could be prepared from all of the insects collected in a trap. Stringent amplification conditions that make use of ThermaStop™, ThermaGo™, and carefully designed primers would make it possible to amplify

**Fig. 5.** Pseudo-color map of the approximate melting temperature ($T_m$) calculated for each probe in a set of five consensus probes hybridized to the CO1 sequences from each of 216 species of mosquitoes across 14 genera.



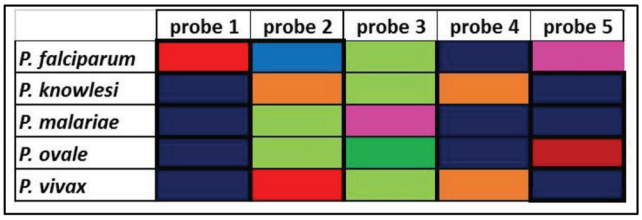

a particular target of interest despite the background of other genomes. For instance, we routinely detect five copies of the *rpoB* gene target of *M. tuberculosis* (a small bacterial genome) in reactions that also contain 50 000 copies of human genomic DNA (a large eukaryotic genome). If the genome of interest is present it will be amplified and its single-stranded amplicons will be recognized by the mosquito probes. Addition of the malaria primers and probes will likewise "pick-out" and amplify any *Plasmodium* CO1 sequence that may be present in the sample, and the fluorescent signature of this amplicon will identify its species. Selective amplification and characterization of particular species of mosquito would be useful for many mosquito control programs that are tasked with monitoring for invasive species. Clearly, the proposed logic could also be applied to improve monitoring of many other unwanted invasive species that could be present in a background of otherwise benign insects.

**Table 3.** Average melting temperature ($T_m$) value of mosquito probes across tested species among different genera.

| | Average $T_m$ (°C) |
|---|---|
| *Aedes* | 46.2±4.9 |
| *Anopheles* | 45.8±5.0 |
| *Culex* et al. | 47.1±4.3 |

**Note:** See Fig. 5 and the supplementary data[2] for representative species in each group. See Table 2 for method of analysis.

**Fig. 6.** Pseudo-color map of the approximate melting temperature ($T_m$) calculated for each probe in a set five probes hybridized to a portion of the CO1 sequences from each of five species of malaria *Plasmodium*.




**Closed-Tube Barcoding of *Naegleria* could allow the identification of all species within a diverse protist genus**

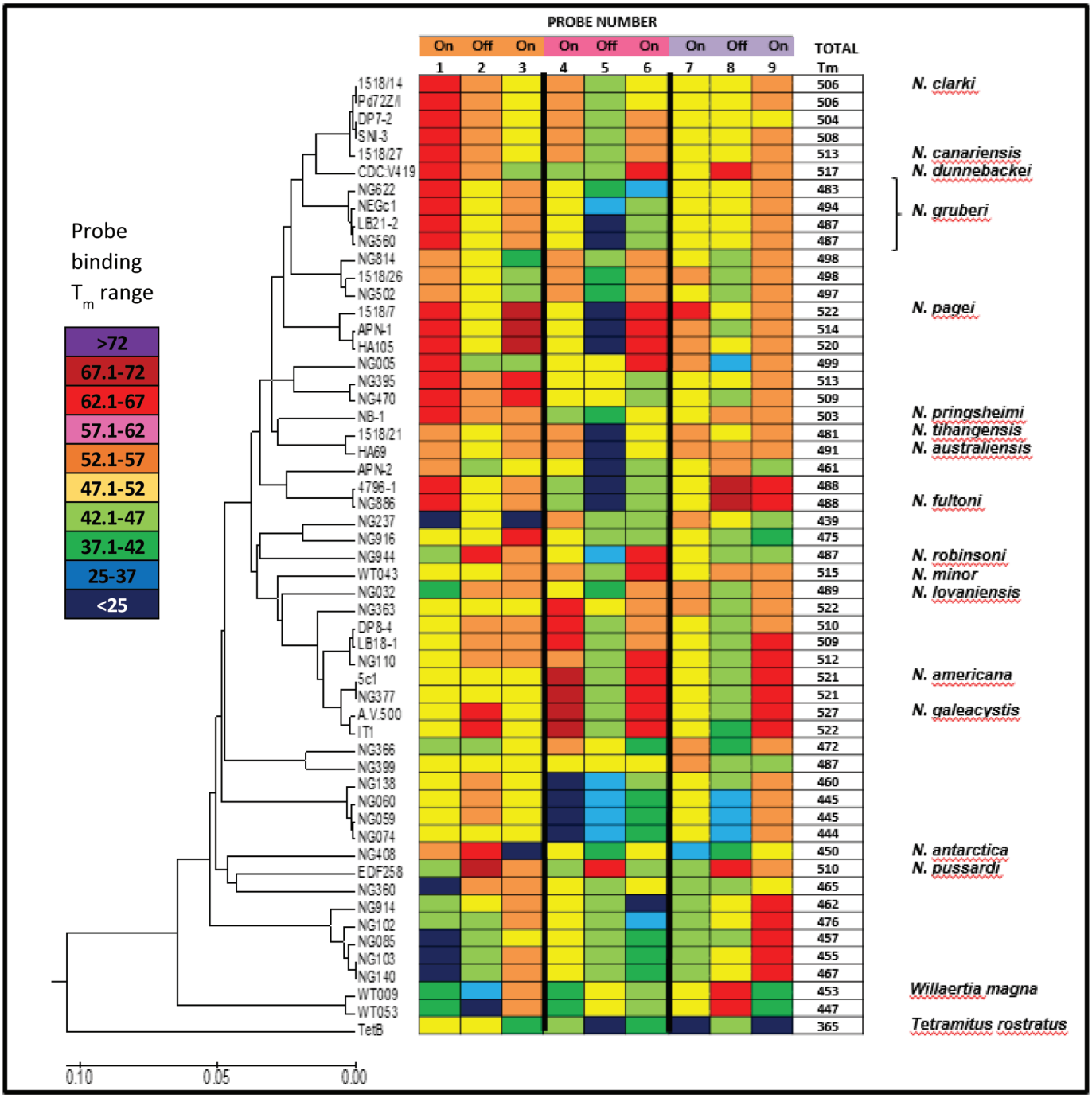
The fact that mosquito primers and probes do not cross-hybridize with the CO1 sequences of malarial *Plasmodium*, and vice versa, is very useful, but it is not surprising in light of the enormous evolutionary distance between mosquitoes (multicellular Metazoans) and malaria *Plasmodium* (parasitic protists). Indeed, even though the distant connections between the major branches of Eucarya remains unclear, it appears that the major supergroups of protists became separated from each other billions of years ago (Fritz-Laylin et al. 2010). One branch includes the Heterolobosea, and specifically the amoeboflagellate genus *Naegleria*, which are unicellular protists that can differentiate their morphology from amoebae, their reproductive form, into temporary swimming flagellates (Fulton 1970, 1993; De Jonckeere 2002). Members of the genus *Naegleria* are abundant in freshwater and permanently wet soil all over the earth (Fulton 1970, 1993). Originally, all *Naegleria* were grouped in the species *N. gruberi*, but then an opportunistic human pathogenic species was found and named *N. fowleri* (Fulton 1970). Studies quickly showed that the genus is extremely diverse (Fulton 1993), a conclusion confirmed by the marked differences in gene sequences and gene order between the genomes of *N. gruberi* (Fritz-Laylin et al. 2010) and *N. fowleri* (Zysset-Burri et al. 2014). Studies by several methods, but mostly using sequences of the ITS1 region of the ribosomal plasmid present in all species, have indicated that there are about 40 species in the genus *Naegleria* (De Jonckheere 2004, 2008). Many of the species tend to look very similar, but also differ in the temperature conditions favoring their growth and in their enzymes (Robinson et al. 1992).

To further characterize members of this genus we decided to examine the same strains and species using the CO1 sequence and the ITS1 sequence (R.S. Kaufman, N.M. Sirianni, J.E. Rice, B.S. Robinson, L.J. Wangh, and C. Fulton, in preparation). Here, we use a sample of that data to demonstrate in silico that Closed-Tube Barcoding would provide an ideal technology for identifying species of *Naegleria*, as well as for distinguishing *Naegleria* from other protist genera. Figure 7 shows a tree of sequence relatedness constructed from the CO1 sequences of 52 stains of *Naegleria* that we sequenced using one pair of primers that originated with *N. gruberi* and using the same PCR conditions for all *Naegleria* samples (see Materials and methods). The species of the tested strains were initially determined comparing their ITS1 sequences with published results (De Jonckheere 2004). The tree also shows one strain of each of two related genera of amoeboflagellates, *Willaertia* and *Tetramitus*. *Naegleria* is quite distant from most studied organisms, but *Willaertia* and *Tetramitus* are known to be relatives (Fulton 1993; De Jonckheere 2002). The CO1 sequence of *Willaertia magna* was amplified using the *Naegleria* primers and same PCR conditions for primer annealing and extension. In contrast, amplification of *Tetramitus rostratus* CO1 required a reduction in PCR stringency. Both of these observations are consistent with the relative proximities of these two genera to *Naegleria*.

Using the *Naegleria* CO1 sequences (GenBank submission ID: 1928341), we designed nine consensus probes that would hybridize with similar $T_m$s to all species regardless of their distribution within the *Naegleria* tree. Figure 7 displays the $T_m$ pseudo-color map for the nine consensus probes to the CO1 sequences for all 52 strains of *Naegleria* plus *Willaertia* and *Tetramitis*. Only those strains that have identical sequences also have identical pseudo-color maps, whereas strains that have nearly identical sequences have distinguishable probe pseudo-color maps. The average $T_m$ for the nine probes hybridized to each of the different species of *Naegleria* listed in Table 4 is optimal for such a probe set. The same probes have a similar, but lower, average $T_m$ for *Willaertia magna* and a suboptimal average $T_m$ for *Tetramitus rostratus*. These results are also consistent with their apparent evolutionary distance from the genus *Naegleria*.

Taken together, these results suggest that a Closed-Tube Barcoding assay containing one pair of primers and nine probes can be used to rapidly screen large numbers of amoeboflagellates that meet the morphological definition of *Naegleria*. Such an assay would be constructed using three subsets of Lights-On probes in three fluorescent colors. Each subset would be comprised of two Lights-On probes of the same color flanking a double-quencher Lights-Off probe (Fig. 7). The proposed assay

**Fig. 7.** Pseudo-color map of the approximate melting temperature ($T_m$) calculated for each probe in a set of nine consensus probes hybridized to the CO1 sequences for each of 52 strains of *Naegleria*, plus two strains of *Willaertia* and one strain of *Tetramitis*.




would greatly enhance current capacities to catalogue very large numbers of species of *Naegleria* from samples of diverse environments all over the earth, without the need for DNA sequencing. We hope to test this approach experimentally using the *Naegleria* data set. Similar Closed-Tube Barcoding assays could be designed for rapid screening of new species in other major protist genera, simply by designing the optimal pan-genus primers and probes.

## Discussion

### The future of Closed-Tube Barcoding

This paper describes strategies, methods, and criteria for construction of Closed-Tube Barcoding assays using LATE-PCR and Lights-On/Lights-Off probes. Most significantly this paper illustrates that sets of probes can be designed to generate unique fluorescent signatures for many species, whose amplification is determined using sequence discriminating primers and conditions. Clear

**Table 4.** Average melting temperature ($T_m$) value of *Naegleria* specific probes across tested strains/species and nearest protists.

| | Average $T_m$ (°C) |
|---|---|
| *Naegleria* spp. | 55±2.7 |
| *Wallaertia magna* | 50 |
| *Tetramitus rostratus* | 41 |

**Note:** See Fig. 7 for representative strains and species of *Naegleria*. See Table 2 for method of analysis.



benefits are accrued from using standardized sets of primers and probes for analysis of many species. These include a substantial reduction in reagent costs, as well as a significant improvement in reagent quality due to economies of scale. Further reductions in costs and time can be achieved by improved instrumentation. Conventional real-time PCR instruments use 96-well plates with reaction volumes of 10–25 μL per well and take 3–4 h to run. We have already shown, however, that the simple LATE-PCR assays can be run in 10 nL volumes in under an hour and that all of the required reagents (including primers, probes, and Taq polymerase) can be dried down in advance; all that is needed to start the reaction is addition of a DNA template in water (Jia et al. 2010). More recently, Jia and her colleagues have used electrode-based digital microfluidics to build chips that use mismatch-tolerant Molecular Beacons for high resolution melt curve analysis of single nucleotide polymorphisms in 1.2 μL droplets, in an astoundingly short period of 7 s (Chen et al. 2016). This is 300× faster than melt curve analysis of the same targets in 25 μL reactions in standard real-time machines.

We foresee a future for Closed-Tube Barcoding that combines automated preparation of extremely small samples with chaotropic salts followed by dilution and then rapid amplification and analysis in very small volumes, using portable instruments that include algorithms for rapid comparison of the fluorescent signature of the specimen with those of already known specimens. Closed-Tube Barcoding will significantly reduce the cost, and the time of many barcoding projects that currently depend on sequencing.

In the hands of a trained naturalist, a Closed-Tube Barcoding platform will make it possible to screen many specimens on location, such as insects in the tropics. This in turn will make it possible to choose which specimens should be collected and which are not needed for repositories back home. Moreover, specimens having novel fluorescent signatures in the field can be flagged and stored for follow-up analysis using Dilute-'N'-Go sequencing as soon as they reach the laboratory. In addition, rapid and inexpensive analysis of many small samples, even single cell samples, will open up the field of barcoding to analysis of unicellular eukaryotes and even microbes.

Most immediately, we propose that Closed-Tube Barcoding should be used to pre-screen many samples now processed in standard barcoding pipelines, to distinguish novel specimens from those that have been previously observed. Both classes of samples are interesting for an in-depth understanding of biodiversity that depends on knowing both "what is out there" and "how often it is out there". Finally, Closed-Tube Barcoding in a hand held device can be used for rapid, inexpensive, on-sight monitoring of natural products all over the world and at all points of transition from collection, through processing, to marketing.

## Acknowledgements

Funded by a teaching grant and undergraduate fellowships from Brandeis University, as well as a grant for basic research Hain Lifescience. We thank Nicola Casali for the *rpoB* data she generated using our assay. We thank Eric Olson for providing us with the ticks used in this study.

## References

Carver-Brown, R.K., Reis, A.H., Jr., Rice, L., and Wangh, L. 2012. A LATE-PCR single tube multiplex assay for the detection of septicemia using Lights-On/Lights-Off probes. J. Pathogens, **2012**: Article ID 424808. doi:10.1155/2012/424808.

Chen, T., Jia, Y., Dong, C., Gao, J., Mak, P., and Martins, R.P. 2016. Sub-7-second genotyping of single-nucleotide polymorphism by high-resolution melting curve analysis on a thermal digital microfluidic device. Lab Chip, **16**: 743–752. doi:10.1039/C5LC01533B. PMID:26781669.

De Jonckheere, J.F. 2002. A century of research on the amoeboflagellate genus *Naegleria*. Acta Protozool. **41**: 309–342.

De Jonckheere, J.F. 2004. Molecular definition and the ubiquity of species in the genus *Naegleria*. Protist, **155**: 89–103. doi:10.1078/1434461000167. PMID:15144061.

De Jonckheere, J.F. 2008. *Naegleria*. Version 21 September 2008 (under construction). http://tolweb.org/Naegleria/124653/2008.09.21 in The Tree of Life Web Project, http://towweb.org/ [accessed 10 January 2016].

Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. 1994. DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotechnol. **3**(5): 294–299. PMID:7881515.

Fritz-Laylin, L.K., Prochnik, S.E., Ginger, M.L., Dacks, J.B., Carpenter, M.L., Field, M.C., et al. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell, **140**: 631–642. doi:10.1016/j.cell.2010.01.032. PMID:20211133.

Fulton, C. 1970. Amebo-flagellates as research partners: the laboratory biology of *Naegleria* and *Tetramitus*. Methods Cell Physiol. **4**: 341–476. [*Edited by* D.M. Prescott.]

Fulton, C. 1993. *Naegleria*: a research partner for cell and developmental biology. J. Euk. Microbiol. **40**: 520–532. doi:10.1111/j.1550-7408.1993.tb04945.x.

Hartshorn, C., Anshelevich, A., and Wangh, L.J. 2005. Rapid, single-tube method for quantitative preparation and analysis of RNA and DNA in samples as small as one cell. BMC Biotechnol. **5**: 2. doi:10.1186/1472-6750-5-2. PMID:15649321.

Jia, Y., Osborne, A., Rice, J.E., and Wangh, L.J. 2010. Dilute-'N'-Go dideoxy sequencing of all DNA strands generated in multi-

plex LATE-PCR assays. Nucleic Acids Res. **38**: e119. doi:10.1093/nar/gkq111. PMID:20189962.

Mechai, S., Feil, E.J., Gariepy, T.D., Gregory, T.R., Lindsay, L.R., Millien, V., and Ogden, N.H. 2013. Investigation of the population structure of the tick vector of lyme disease *Ixodes scapularis* (Acari: Ixodidae) in Canada using mitochondrial cytochrome *c* oxidase subunit I gene sequences. J. Med. Entomol. **50**(3): 560–570. doi:10.1603/ME12178. PMID:23802450.

Osborne, A., Reis, A.H., Jr., Bach, L., and Wangh, L.J. 2009. Single-molecule LATE-PCR analysis of human mitochondrial genomic sequence variations. PLoS ONE, **4**(5): e5636. doi:10.1371/journal.pone.0005636. PMID:19461959.

Pierce, K.E., and Wangh, L.J. 2014. Rapid detection and identification of hepatitis C virus (HCV) sequences using mismatch-tolerant hybridization probes: a general method for analysis of sequence variation. BioTechniques, **55**(3): 125–132. doi:10.2144/000114076.

Pierce, K.E., Rice, J.E., Sanchez, J.A., and Wangh, L.J. 2002. QuantiLyse™: reliable DNA amplification from single cells. BioTechniques, **32**: 1106–1111. PMID:12019784.

Pierce, K.E., Sanchez, J.A., Rice, J.E., and Wangh, L.J. 2005. Linear-After-The-Exponential (LATE)-PCR: Primer design criteria for high yields of specific single-stranded DNA and improved real-time detection, Proc. Natl. Acad. Sci. U.S.A. **102**: 8609–8614. doi:10.1073/pnas.0501946102.

Pierce, K.E., Peter, H., Bachmann, T.T., Volpe, C., Mistry, R., Rice, J.E., and Wangh, L.J. 2013. Rapid detection of TEM-type extended-spectrum β-lactamase (ESBL) mutations using lights-on/lights-off probes with single-stranded DNA amplification. J. Mol. Diagn. **15**: 291–298. doi:10.1016/j.jmoldx.2013.02.002. PMID:23518216.

Ratnasingham, S., and Hebert, P.D.N. 2007. BOLD: The Barcode of Life Data System (www.barcodinglife.org). Mol. Ecol. Notes, **7**(3): 355–364. doi:10.1111/j.1471-8286.2007.01678.x.

Rice, J.E., Sanchez, J.A., Pierce, K.E., Reis, A.H., Jr., Osborne, A., and Wangh, L.J. 2007. Monoplex/multiplex linear-after-the-exponential-PCR assays combined with PrimeSafe and Dilute-'N'-Go sequencing. Nat. Protocols, **2**(10): 2429–2438. PMID:17947984.

Rice, J.E, Reis, A.H., Jr., Rice, L.M., Carver-Brown, R.K., and Wangh, L.J. 2012. Fluorescent signatures for variable DNA sequences. Nucleic Acids Res. **40**(21): e164. doi:10.1093/nar/gks731. PMID:22879378.

Rice, L.M., Reis, A.H., Jr., and Wangh, L.J. 2014. Virtual barcoding using LATE-PCR and Lights-On/Lights-Off probes: identification of nematode species in a closed-tube reaction. Mitochondrial DNA, **27**(2): 1358–1363. doi:10.3109/19401736.2014.947581. PMID:25109627.

Robinson, B.S., Christy, P., Hayes, S.J., and Dobson, P.J. 1992. Discontinuous genetic variation among mesophilic *Naegleria* isolates: further evidence that *N. gruberi* is not a single species. J. Protozool. **39**: 702–712. PMID:1453360.

Sakamoto, J.M., Goddard, J., and Rasgon, J.L. 2014. Population and demographic structure of *Ixodes scapularis* Say in the eastern United States. PLoS ONE, **9**(7): e101389. doi:10.1371/journal.pone.0101389. PMID:25025532.

Sanchez, J.A., Pierce, K.E., Rice, J.E., and Wangh, L.J. 2004. Linear-After-The-Exponential (LATE)-PCR: an advanced method of asymmetric PCR and its uses in quantitative real-time analysis. Proc. Natl. Acad. Sci. U.S.A. **101**(7): 1933–1938. doi:10.1073/pnas.0305476101. PMID:14769930.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol. Biol. Evol. **30**: 2725–2729. doi:10.1093/molbev/mst197. PMID:24132122.

Tetrault, S.M., Rice, J.E., Wangh, L.J., and Aquiles Sanchez, J. 2014. Single-tube mutation scanning of the epidermal growth factor receptor gene using multiplex LATE-PCR and Lights-On/Lights-Off probes. J. Mol. Biomark Diagn. **5**: 175. doi:10.4172/2155-9929.1000175.

Zysset-Burri, D.C., Muller, N., Beuret, C., Heller, M., Schurch, N., Gottstein, B., and Wittwer, M. 2014. Genome-wide identification of pathogenicity factors of the free-living amoeba *Naegleria fowleri*. BMC Genomics, **15**(1): 496. doi:10.1186/1471-2164-15-496. PMID:24950717.