MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 135–143 doi: 10.1111/1755-0998.12034

Direct sequencing of haplotypes from diploid individuals through a modified emulsion PCR-based single-molecule sequencing approach

BRIAN PATRICK HANSEN METZGER,* GREGORY WILLIAM GELEMBIUK and CAROL EUNMI LEE
Center of Rapid Evolution, University of Wisconsin, 430 Lincoln Drive, Birge Hall, Madison, WI 53706, USA

Abstract

While standard DNA-sequencing approaches readily yield genotypic sequence data, haplotype information is often of greater utility for population genetic analyses. However, obtaining individual haplotype sequences can be costly and time-consuming and sometimes requires statistical reconstruction approaches that are subject to bias and error. Advancements have recently been made in determining individual chromosomal sequences in large-scale genomic studies, yet few options exist for obtaining this information from large numbers of highly polymorphic individuals in a cost-effective manner. As a solution, we developed a simple PCR-based method for obtaining sequence information from individual DNA strands using standard laboratory equipment. The method employs a water-in-oil emulsion to separate the PCR mixture into thousands of individual microreactors. PCR within these small vesicles results in amplification from only a single starting DNA template molecule and thus a single haplotype. We improved upon previous approaches by including SYBR Green I and a melted agarose solution in the PCR, allowing easy identification and separation of individually amplified DNA molecules. We demonstrate the use of this method on a highly polymorphic estuarine population of the copepod *Eurytemora affinis* for which current molecular and computational methods for haplotype determination have been inadequate.

Keywords: copepod, *Eurytemora affinis*, haplotype, phasing, single-molecule sequencing

Received 7 August 2012; revision received 8 October 2012; accepted 11 October 2012

Introduction

The common approach of PCR followed by Sanger sequencing is widely used for obtaining DNA sequence data from biological samples (Sanger & Coulson 1975; Mullis & Faloona 1987). However, conventional PCRs are generally designed for locus specificity, such that amplification of multiple homologous chromosomes in nonhaploid organisms results in a mixture of haplotype sequences. Thus, when multiple heterozygous sites are sequenced in the same reaction, the actual sequence of a single chromosome cannot be unambiguously obtained without additional investigation. The sequences of individual chromosomes represent distinct evolutionary histories and contain additional information regarding population parameters relative to genotypic data alone, such as recombination, mutation and migration rates, ancestry, demographic parameters, and signatures of natural selection (e.g. Akey *et al.* 2001; Sabeti *et al.* 2002; The International HapMap Consortium 2005; Bagos 2011; Tewhey *et al.* 2011).

The typical molecular approach for obtaining haplotype sequences has been the cloning of DNA fragments into a vector followed by Sanger sequencing. As each vector contains only a single inserted piece of DNA, the resulting sequence should accurately reflect only a single DNA fragment, and thus a haplotype, of the individual sequenced. However, the highly polymorphic nature of many nonmodel systems means that individuals have a high probability of containing novel haplotypes, thus requiring large sample sizes to capture all haplotypes and obtain reliable frequency estimates within populations. More problematically, an initial PCR step must often be performed to obtain a sufficient amount of DNA prior to cloning. This PCR step can introduce biases in amplification, potentially skewing the expected one-to-one ratio of the haplotypes in diploid DNA. This problem results in an (unknown) large number of clones that require sequencing to obtain both haplotypes (Suzuki & Giovannoni 1996; Becker *et al.* 2000). Additionally, allelic dropout due to random nonamplification of specific

Correspondence: Brian Metzger, Fax: 734-763-0544;
E-mail: bpmetz@umich.edu
*Present address: University of Michigan, 1300 Natural Science Building, 830 North University, Ann Arbor, MI 48109, USA

alleles can lead to incorrect assignment of homozygosity and result in potentially missing haplotypes. This phenomenon is particularly problematic for environmental or museum samples where DNA quantity and quality are low, but also poses problems for high-quality DNA samples (Soulsbury *et al.* 2007). PCR can also introduce errors into haplotype sequences due to PCR recombination (Meyerhans *et al.* 1990; Wang & Wang 1997). Due to large regions of homology, incompletely extended fragments of one haplotype from early PCR cycles can reanneal during later PCR cycles to other haplotypes and act as large primers, resulting in single DNA molecules with two distinct biological origins. Cloning and sequencing of such single molecules would result in incorrect inference of haplotype identity for individuals, potentially yielding haplotypes that do not exist in the population (Wang & Wang 1997; Flot *et al.* 2006). Thus, cloning procedures are not ideal for determining haplotypes because they are both labour intensive and rely on a bulk PCR step that can lead to artifactual PCR recombination.

As an alternative and complementary approach, statistical methods have been developed for haplotype determination from genotypic data, through a process known as 'phasing' (Browning & Browning 2011). These methods use information from multiple samples to determine the haplotype phase for individuals. In particular cases, these methods are known to have high accuracy (Adkins 2004; Avery *et al.* 2005). However, the process of phasing fundamentally represents a missing data problem, and can create complex statistical artefacts that are difficult to detect and can heavily influence downstream analyses (Levenstien *et al.* 2006; Lin & Huang 2007; Browning & Browning 2009). Additionally, these methods can be statistically inconsistent, potentially converging on a set of haplotypes and haplotype frequencies that are incorrect, even as more data are added (Andrés *et al.* 2007; Uddin *et al.* 2008). Finally, many of these methods require that various assumptions regarding the sample and its underlying population history be met, such as constant population size or Hardy–Weinberg equilibrium (Stephens *et al.* 2001; Marchini *et al.* 2006). These assumptions are not always satisfied for samples and populations under complex demographic and selective regimes and are especially problematic when the regions to be phased were chosen for an *a priori* reason, such as a potential causal locus for a trait under adaptive evolution. These methods can thus introduce biases in the inferred haplotypes and their frequencies, especially when model assumptions are not met, but even when they are (Lin & Huang 2007; Higasa *et al.* 2009).

Recently, next-generation sequencing has been used to determine genome-wide haplotype identities within a single individual (Kitzman *et al.* 2011). For projects requiring phase information for large regions of the genome, next-generation sequencing is a cost-effective approach. However, for smaller regions, such as single loci, next-generation sequencing technologies present a number of limitations. While the cost per base pair of sequence is low, the total cost is still prohibitively expensive for large numbers of samples due to the cost of individual libraries and the bioinformatics analyses necessary (Garvin *et al.* 2010). Additionally, current access to the latest technology is limited and samples can often wait several months in the queue prior to sequencing. Overall, next-generation sequencing technologies offer a number of advantages, but are not the best approach available for all situations (see http://www.molecularecologist.com/next-gen-fieldguide for up to date comparisons of the benefits and limits of sequencing technologies).

Because of the limitations stated above, much research has focused on amplifying a single DNA molecule, and hence haplotype, using PCR. However, low concentrations of DNA in a bulk reaction, such as a single initial template strand, result in primarily nonspecific PCR amplification leading to an abundance of primer-dimers in lieu of a specific product (Ruano *et al.* 1989). The most widely used cost-effective strategies to prevent this problem employ a water-in-oil emulsion to create millions of microreactors (vesicles) at the nanolitre scale (Tawfik & Griffiths 1998; Nakano *et al.* 2003; Musyanovych *et al.* 2005; Hori *et al.* 2007). The number of template molecules and microreactor volumes are adjusted such that individual microreactors are unlikely to contain multiple DNA molecules, allowing amplification of individual molecules at concentrations similar to those of bulk PCR. However, to obtain DNA sequence from these vesicles, the emulsion must be broken and the DNA from all individually amplified template molecules must be pooled. Typically, this pooled DNA is cloned and individual clones sequenced to obtain haplotypes. While this approach of single-molecule PCR followed by cloning resolves the aforementioned problems of PCR amplification bias, allelic dropout, and PCR recombination, it cannot be performed easily on large numbers of samples.

Thus, our goal was to modify previous emulsion PCR approaches to develop a practical method that could be applied to many samples with ease. We augmented prior methods to include SYBR Green I and melted agarose in the PCR cocktail to facilitate rapid identification and isolation of individual vesicles containing successful amplification. Because SYBR Green I is highly specific to double-stranded DNA, vesicles containing successful amplification products are readily detectable from vesicles without amplification. The presence of melted agarose does not interfere with amplification, but upon

completion of PCR, the agarose solidifies and allows physical separation of microreactors. Combined, SYBR Green I and agarose allow for easy detection and manipulation of individual microreactors containing DNA amplified from single initial template molecules. Individual microreactors with successful amplification contain sufficient quantities of DNA for conventional PCR and Sanger sequencing.

We used the diploid copepod *Eurytemora affinis* to test our novel method. *E. affinis* is an estuarine and salt marsh copepod (crustacean) that has recently (70 years) invaded freshwater lakes and reservoirs throughout the Northern Hemisphere (Lee 1999). Attempts to examine shifts in haplotype frequencies between ancestral saline and derived freshwater populations, to detect signatures of natural selection during invasions, have proved difficult. This difficultly arises in part because *E. affinis* populations often contain high haplotype diversity, including numerous indel length polymorphisms. This high haplotype diversity makes *E. affinis* populations representative of many highly polymorphic nonmodel systems and ideal for testing our novel method of haplotype determination.

Materials and methods

DNA extractions

DNA was extracted from *E. affinis* individuals following the methods of Lee & Frost (2002). Extractions were performed overnight at 37°C using 50 μL of lysis buffer and resulted in DNA concentrations of $\sim$0.4 ng/μL. Because large genomes contain only a few copies of each locus per ng of DNA, while small genomes contain many more copies of each locus for the same amount of DNA, an estimate of genome size is needed to calculate the total number of copies of each haplotype present. Assuming the average weight of a single DNA base pair is 650 Da, the number of copies of each locus can be calculated from the genome size and the total amount of DNA (in ng) used for the PCR (http://www.uri.edu/research/gsc/resources/cndna.html).

The genome size of *E. affinis* is approximately 640 MB (2C) (Rasch *et al*. 2004) and thus contains approximately 1450 copies of each haplotype per ng of DNA or 580 copies of each haplotype per μL of extracted DNA ($\sim$2900 copies per ng and $\sim$1160 copies per μL of extracted DNA for each loci).

Emulsion PCR approach

The emulsion PCR consisted of two components, an aqueous phase and an oil phase (see below for exact details on composition). The aqueous phase contained all PCR reagents, as well as the agarose and SYBR Green I necessary for detection and isolation of vesicles containing successful amplification. The oil phase consisted of a mixture of three oils. The oil and aqueous phases were combined by vortexing, forming millions of small aqueous vesicles separated by the oil phase and each directly capable of PCR amplification.

Aqueous phase of the emulsion component for PCR

The aqueous phase of the emulsion contained two parts: a modified PCR cocktail and a melted agarose solution. For the modified PCR cocktail, all PCR reagents were doubled in concentration relative to a conventional PCR protocol (Table 1). This conventional PCR protocol was optimized to robustly generate a single, bright band. Due to the high temperatures needed to maintain the agarose in a melted state, a hot start DNA polymerase was used (Platinum Taq High Fidelity; Invitrogen^{TM}). The highly efficient double-stranded binding dye SYBR Green I was added at a final concentration of 40$\times$ to the modified PCR cocktail (Invitrogen^{TM}; SYBR Green I is obtained as a 10 000$\times$ solution, the exact molar concentration is not given).

To determine the total amount of DNA to use, the number of molecules of each haplotype per individual microreactor was assumed to follow a Poisson distribution (Nakano *et al*. 2000, 2003). The mean of this distribution was determined by the number of copies of each haplotype present in the reaction (as determined by the amount of DNA added and genome size) and the total number of vesicles created (Fig. S1). Because vesicles are approximately spherical, a mean vesicle diameter of 90 μm results in vesicles containing approximately $\sim$0.4 nL of reagents, and a total of 66 000 individual microreactors for a 25 μL reaction. Based on our estimate of 1160 copies of each loci per μL of DNA, we used a Poisson distribution with mean of 0.0176 copies per vesicle. Thus, vesicles should contain at least a single template strand of the locus of interest greater than 1.74% of the time, but should have a low probability of having more than a single template ($\sim$0.015%). Finally, vesicles of this size should result in approximately 1150 fluorescent vesicles per reaction, with each fluorescent vesicle having less than a 1% chance of amplification from multiple initial template molecules.

Our modified PCR amplification method was tested on two loci, *Cuticle Protein 4761* and *V-type ATPase*. Primers 4761-A (AAAGCAGCTGGATACTGAGC) and 4761-B (CAACAACTTTGGACACAAGG) were designed to amplify *Cuticle Protein 4761* (Table 1A), whereas primers V-TYPE-B-I360 (GTAAACAAGTGCCGACCGATTC) and V-TYPE-A26 (CCCGTATCACCTACAAGACC) were designed to amplify *V-type ATPase*, (Table 1B). The same primers were used for sequencing the PCR products (Table 1).

Table 1 PCR cocktails used for conventional and emulsion PCR to amplify two loci from the copepod *Eurytemora affinis*

Reagent	Concentration	(I) Conventional PCR cocktail (μL)	(II) Emulsion PCR cocktail (μL)
(A) Cuticle protein 4761			
ddH_2O		17.65	11.3
High fidelity buffer*	10×	2.5	5
$MgSO_4$	50 mM	1.25	2.5
dNTPs	10 mM each	0.5	1
Primer 4761-A	10 μM	0.5	1
Primer 4761-B	10 μM	0.5	1
Platinum *Taq* High Fidelity†	5 U/μL	0.1	0.2
DNA	0.4 μg/μL	2	2
SYBR Green 1	1000×	0	1
Total volume		25	25
(B) V-Type ATPase			
ddH_2O		18.5	12
Picomaxx buffer‡	10×	2.5	5
dNTPs	10 mM each	0.5	1
Primer A-26	10 μM	1	2
Primer B-I360	10 μM	1	2
Picomaxx High Fidelity	2.5 U/μL	0.5	1
DNA	0.4 μg/μL	1	1
SYBR Green I	1000×	0	1
Total volume		25	25

*Contains 600 mM Tris-SO_4 (pH 8.9), 180 mM ammonium sulphate.
†Contains 20 mM Tris-HCl (pH 8.0), 0.1 mM EDTA, 1 mM DTT, stabilizers and 50% (v/v) glycerol.
‡Contains 25 mM $MgCl_2$.

This modified PCR cocktail (shown in Table 1A-II and Table 1B-II) was mixed with a melted agarose solution. More specifically, a 1-3% solution of SeaKem® Gold Agarose (Lonza Group Ltd.) in PCR-grade water was autoclaved to completely melt the agarose, and then maintained in a 60–70°C hot water bath prior to addition to the PCR cocktail. This solution was used within 30 min of being removed from the autoclave to minimize solidification. The accurate pipetting of melted agarose solutions was difficult due to high viscosity and rapid cooling and solidification in small volumes. Thus, the appropriate pipetter settings were determined to dispense the desired volume (generally an additional 5-15% relative to water). The addition of agarose to the PCR cocktail in a 1:1 ratio created an aqueous phase which contained all necessary PCR reagents at standard concentrations and a final SYBR Green I concentration of 20×. After addition of the melted agarose solution to the PCR cocktail, the mixture was vortexed for 1–2 s at high speed (maximum setting on a vortex-genie, model K-550-G S8223) and maintained between 60°C and 70°C.

Oil phase of the emulsion component for PCR

The oil phase of the emulsion consisted of the following in a 7:2:1 ratio: DC 5225C Formulation Aid (Dow Corning Corporation), DC 749 Fluid (Dow Corning Corporation) and Silicone Oil AR 20 (Fluka, Buchs, Switzerland). The mixture was vortexed for ~5 s at high speed to ensure complete mixing and was warmed to the same temperature as the aqueous phase (60–70°C). As with the melted agarose, accurate pipetting of any of the three oils, or their mixture, was difficult due to their viscosity and was handled in the same manner as described previously.

Creation of emulsion and PCR conditions

The emulsion was generated by adding the oil phase to the aqueous phase in a 2:1 ratio, and vortexing the mixture for 5–7 s at moderate speeds (setting 5 on a vortex-genie). For emulsion PCR amplification of *Cuticle Protein 4761,* the PCR profile was 1 min at 94°C, followed by 15 cycles of 94°C (15 s), 65°C (30 s), 70°C (60 s) with a 1°C decrease in annealing temperature at each cycle, followed by 34 cycles of 94°C (15 s), 50°C (30 s), 70°C (60 s) and with a final hold at 4°C. For emulsion PCR amplification of *V-Type ATPase,* PCR cycling conditions were 5 min at 95°C, followed by five cycles of 94°C (45 s), 60°C (45 s), 72°C (60 s) with a 1°C decrease in annealing temperature at each cycle, followed by 40 cycles of 94°C (45 s), 55°C (45 s), 72°C (60 s), followed by 5 min at 72°C and a final hold at 4°C.

Visualization of vesicles containing DNA

Following PCR amplification, fluorescent vesicles were visualized to allow for detection and collection. Initial visualization of vesicles was performed using a Zeiss Axioplan EL-Einsatz microscope with a mercury arc lamp and Zeiss filter set number 17. Images were taken using Fujichrome 400 slide film under both bright field and fluorescence using a Nikon NFX-35 camera mounted to the microscope. Film exposure times of 1–10 min were necessary to obtain fluorescent images. This setup allowed for easy detection of fluorescent vesicles and was used during initial testing of emulsion PCR. However, the nearness of the objective lens to the stage made manipulation and collection of fluorescent beads difficult. Thus, a standard stereoscope capable of between 1× and 5× magnification was used with a custom built light source for vesicle manipulation. The light source con-

sisted of a Luxeon V Emitter Blue Lambertian LED with a Fraen Narrow Beam Lens (with holder) and a Xitanium 120VAC, 12 W, 700mA LED Driver (Luxeon http://www.luxeonstar.com). The LED and lens were attached to an appropriate heat sink and mounted with a small fan blowing perpendicular to the heat sink fins. The Luxeon® LED emits light around 470 nm (blue light), near the peak excitation of SYBR Green I of 488 nm. To remove background wavelengths from the LED, flexible Roscolux filters (−12 Straw) were inserted into the ocular lenses of the microscope (Rosco Laboratories, Inc., http://www.rosco.com). Details of designing the custom light source were derived from the following website: http://130.15.90.245/gfp_stereoscope.htm.

Reamplification of DNA from vesicles

Vesicles were checked after the initial PCR amplification step to ensure they maintained a spherical shape and were relatively uniform in size. Five microlitres of each emulsion reaction was placed onto a standard microscope slide, forming a thin oil layer with agarose vesicles spread out and only one layer deep; i.e. vesicles were not stacked on top of one another. Vesicles showing successful amplification (i.e. showing fluorescence) under the custom light source were individually collected using standard gel loading pipet tips and placed individually into 5 μL of PCR-grade water. Isolation of several vesicles from a single emulsion took approximately 1–2 min, and was performed on up to 10 samples at once. The entire 5 μL of water containing a single fluorescent vesicle was used as a DNA template for a subsequent PCR to obtain sufficient DNA for sequencing. Because the initial PCR amplified DNA from a single molecule, the second PCR was performed as a conventional PCR and was not performed in an emulsion (Table 1A-I and Table 1B-I). The initial denaturation step of the subsequent PCR was sufficient to melt the agarose and release the DNA for amplification.

Testing the modified single-molecule sequencing approach

Our single-molecule sequencing approach was tested in several ways. PCR amplification within individual microreactors in the presence of melted agarose and SYBR Green I was first confirmed, and DNA sequence information was obtained using our modified single-molecule sequencing approach. An *E. affinis* individual that possessed a single heterozygous site for a region of the *V-Type ATPase* gene and thus had unambiguously distinct haplotypes was used as the DNA source. PCR amplification was performed using both our single-molecule approach and conventional PCR, and PCR products from both approaches were sequenced. For this, and all other experiments, PCR products were sequenced in both directions as a control against sequencing errors.

The reproducibility of haplotype calls from multiple vesicles and the absence of chimeras using our approach were confirmed using a single individual that was heterozygous at six sites within a portion of the *Cuticle Protein 4761* gene. PCR product from a conventional PCR and 12 independent fluorescent vesicles from our approach were sequenced.

Lastly, to demonstrate the practical application of this approach, our modified single-molecule sequencing approach was applied to a pooled DNA sample of four individuals from two populations (two individuals each from Saint-Jean Port Joli, Quebec, Canada and Racine Harbor in Lake Michigan, USA) (Lee 2000; Winkler *et al.* 2008) for a small portion of the *Cuticle Protein 4761* gene (130 bp). The two populations segregated at several different sites within this segment of DNA and three of the four individuals chosen were heterozygotes. The single homozygous individual acted as a control against PCR chimeras because this haplotype was known.

Results and discussion

To develop a cost-effective method for single-molecule sequencing, we modified previous emulsion PCR approaches by including SYBR Green I and a melted agarose solution in the PCR. We designed an oil-in-water-emulsion so that the vast majority of vesicles contained zero copies of the locus of interest (98.3%), whereas a small percentage contained only a single copy (1.7%), and thus haplotype, of the locus of interest. The presence of SYBR Green I allowed for easy identification of the vesicles containing successful amplification and the agarose allowed us to isolate individual vesicles with successful amplification without destroying the emulsion. Because the emulsion was not destroyed, PCR products derived from amplification of different initial DNA templates were kept separate throughout the entire procedure. Thus, haplotype sequences could be determined directly from sequencing the PCR product without an intermediate cloning step. We tested our approach on regions of the *Cuticle 4761* and *V-Type ATPase* genes for several populations of the diploid copepod *E. affinis* and were able to repeatedly obtain individual haplotype sequences, showing no evidence of heterozygosity at known heterozygous sites.

Vesicle formation and stability are crucial for successful amplification

Among the most crucial factors influencing success of our approach was the ability to create optimal vesicle sizes for amplification of single DNA molecules. There

were trade-offs in efficacy between small and large vesicle size. Emulsions with large average vesicle size resulted in fewer vesicles, almost all of which were fluorescent and likely to contain multiple initial template molecules (Fig. S1a). Because vesicles that contained multiple initial template molecules probably contained multiple haplotypes, we chose to use vesicles with a mean diameter less than 100 μM. Alternatively, we detected lower SYBR Green 1 fluorescence in smaller than in larger vesicles (Fig. S1b). In many cases, fluorescence was almost nonexistent in vesicles that were less than 10 μM in diameter, probably indicative of low levels of amplification due to limitation of PCR reagents in such tiny volumes (Nakano *et al.* 2003). We thus chose to use vesicles with a mean diameter of 90 μM and empirically determined the vortex speed and time required to create these appropriately sized vesicles, with faster speeds and longer vortexing times on average creating smaller vesicles. We observed that replicate emulsions (formed under identical conditions) resulted in varying average vesicle size, likely due to variability in vortexing. One alternative to vortexing is the use of a commercial bead-beater for more consistent vesicles sizes. However, given that many research groups do not have access to bead-beaters, we simply created 3-4 independent emulsions per sample to obtain one with the appropriate vesicle size.

Conditions under which the water-in-oil emulsions were formed also influenced the stability of emulsion during PCR amplification. We observed a positive correlation between length of vortexing time used to generate the emulsion and stability of the emulsion (5–7 s was optimal), and a negative correlation between vesicle size and emulsion stability (with less stable emulsions having larger vesicles). Emulsions created rapidly (<2 s) or using low vortex speeds (settings less than 2 on a vortex-genie) were generally unstable. Using a higher ratio of the oil phase to aqueous phase resulted in no increase in emulsion stability, whereas lower ratios, down to (1.3:1) of the oil phase to aqueous phase, were generally stable. Ratios lower than this (1.3:1) occasionally resulted in oil-in-water, instead of water-in-oil emulsions, that were unstable during PCR cycling (Murakawa *et al.* 2008). Different ratios of oils (DC 5225C Formulation Aid, DC 749 Fluid, and Silicone Oil AR 20) were tested, but resulted in either lower stability of the emulsion or greater opacity, making visualization of individually fluorescent vesicles difficult in subsequent steps.

Overall, the procedure could be performed quickly and inexpensively. Each emulsion PCR resulted in several hundred to a thousand fluorescent vesicles. We typically reamplified two or three vesicles from a single sample for sequencing to guard against PCR errors early in the reaction resulting in incorrect sequences. The amount of oil and SYBR Green I required individually for each reaction was small, and the total cost was essentially only double or triple that of conventional PCR for a single sample due to the two PCR steps and several sequences needed per sample. The entire procedure was routinely completed in a single day, including the emulsion PCR, reamplification PCR, and PCR cleanup for sequencing. Overall, optimization time for a new locus was equivalent to that of traditional PCR once initial setup of the light source was completed.

Isolating individual haplotypes using emulsion PCR

Following amplification, visualization of vesicles from our emulsion procedure clearly distinguished fluorescent from nonfluorescent vesicles (Fig. 1). The amount of fluorescence was proportional to vesicle size. We observed roughly the expected number of fluorescent vesicles based on amount of template DNA and average vesicle size. The addition of agarose to the PCR made it easy to collect fluorescent vesicles without requiring disruption of the emulsion. Reamplification and sequencing of an individual fluorescent vesicle from an individual copepod containing a single heterozygous site for the *V-Type ATPase* gene gave no indication of heterozygosity at the expected site, revealing that only an individual haplotype was captured by each microreactor (Fig. 2c,d). In contrast, sequencing of the same sample amplified using conventional PCR clearly revealed the heterozygous position (Fig. 2a,b). Thus, PCR amplification within microreactors in the presence of melted agarose and SYBR Green I is

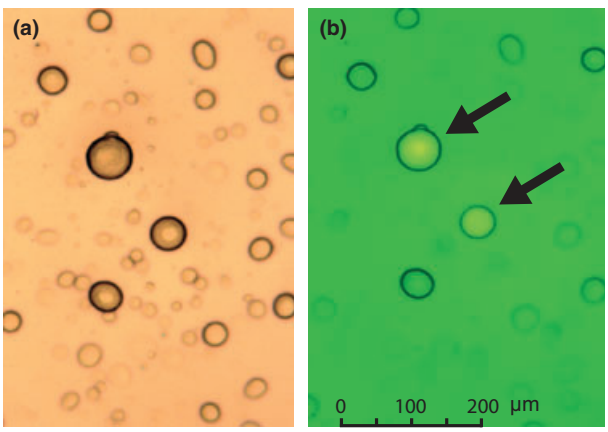

Fig. 1 Individual fluorescent vesicles generated from our emulsion PCR approach, shown under (a) brightfield light and (b) fluorescent light. Arrows indicate fluorescent vesicles. Vesicles were approximately 75 μM in diameter. Note: fluorescent vesicles were manipulated to be near one another for the purposes of this photograph and are typically spaced farther apart. Arrows point to fluorescent vesicles.

possible and this method can yield sequences amplified from single individual DNA templates.

We next tested our ability to obtain haplotype sequences consistently and in the absence of PCR chimeras. Using our modified single-molecule sequencing approach on an individual copepod containing multiple heterozygous sites, we recovered only two distinct haplotype sequences from 12 independent fluorescent vesicles indicating that our method could reliably and consistently uncover the sequences of individual haplotypes without the presence of PCR chimeras (Fig. S2). In contrast, and as expected, we recovered all heterozygous sites using a conventional PCR approach.

Previous work has revealed an extensive amount of haplotype diversity both within and between populations of *E. affinis* (Winkler *et al*. 2008). This large amount of diversity has made haplotype determination difficult. We demonstrated an application of our method by identifying haplotypes for a portion of the *Cuticle 4761* gene from two populations of *E. affinis* (Racine Harbor, Lake Michigan and St. Jean Port Joli, Quebec). For a larger region of this gene, conventional PCR amplification and sequencing of numerous individuals from the two populations did not reveal any shared genotypes (data not shown). However, because almost all individuals were heterozygous, haplotypes could not be directly determined for most individuals. Additionally, the high levels of heterozygosity prevented statistical phasing methods from arriving at a fully resolved set of haplotypes when applied to individual populations. Lastly, combining information across multiple populations for phasing resulted in inconsistent results, with both the actual haplotypes and their frequencies dependent on which populations were chosen to be phased together. We thus used our modified single-molecule sequencing approach to resolve the haplotypes of a small region of this gene which harboured numerous nearby segregating sites within and between populations. We recovered exactly three haplotypes, which were consistent with each individual's genotype (Individual = Haplotype 1:Haplotype 2; J1 = I:II, J2 = I:II, R1 = III:III, R2 = II:III) (Fig. 3). It should be noted that given only these four individuals, the correct haplotypes could be derived using statistical phasing algorithms; however, this was not true when using larger numbers of individuals.

Potential pitfalls and problems

We encountered several common problems when implementing our single-molecule sequencing approach. Most notably, about 10% of sequences obtained by single-molecule sequencing still contained heterozygous sites (as compared to our estimate of 1% based on a Poisson distribution), indicating that a larger than expected proportion of fluorescent vesicles contained more than a single strand of DNA (see Fig. S2, k and m for representative examples). We expect that this elevated percentage was due to a combination of factors, including differences in the total amount of DNA per individual, differences in DNA extraction efficiency between individuals, our inability to accurately quantify the exact number of DNA strands, and variation in vesicle size. However, this problem was readily detectable and sequencing of several vesicles was sufficient to obtain individual haplotype information.

We additionally found that if vesicles appeared spherical and uniform in size and shape, yet were not fluorescent after PCR, using a new SYBR Green I stock was sufficient for restoring fluorescence in all cases, presumably because SYBR Green I is both light and heat-sensitive. If vesicles appeared abnormally shaped or otherwise nonuniform in appearance after PCR, the agarose solution was discarded and remade. Even at temperatures between 60°C and 70°C, the agarose became more viscous over time due to water loss, making formation of an appropriate emulsion difficult. Additionally, changes in either the orientation of the tube, speed, or numerous stops and starts when creating the emulsion resulted in

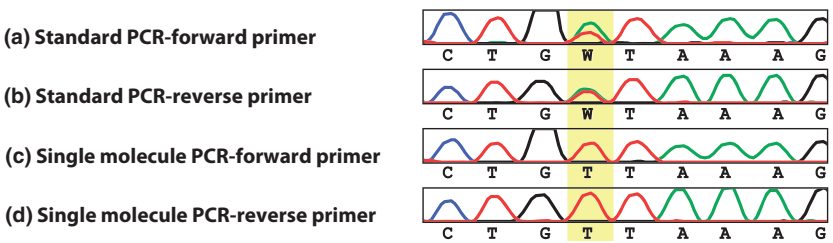

Fig. 2 DNA sequence chromatograms amplified from heterozygous DNA using conventional PCR and our modified single-molecule sequencing approach. Top two chromatograms show conventional PCR products (control) containing the heterozygous site (highlighted in yellow) amplified using the (a) forward primer and (b) reverse primer. Bottom two chromatograms show PCR products from a single fluorescent vesicle, generated using our single-molecule PCR approach. These products show no indication of heterozygosity using both the (c) forward primer and (d) reverse primer.

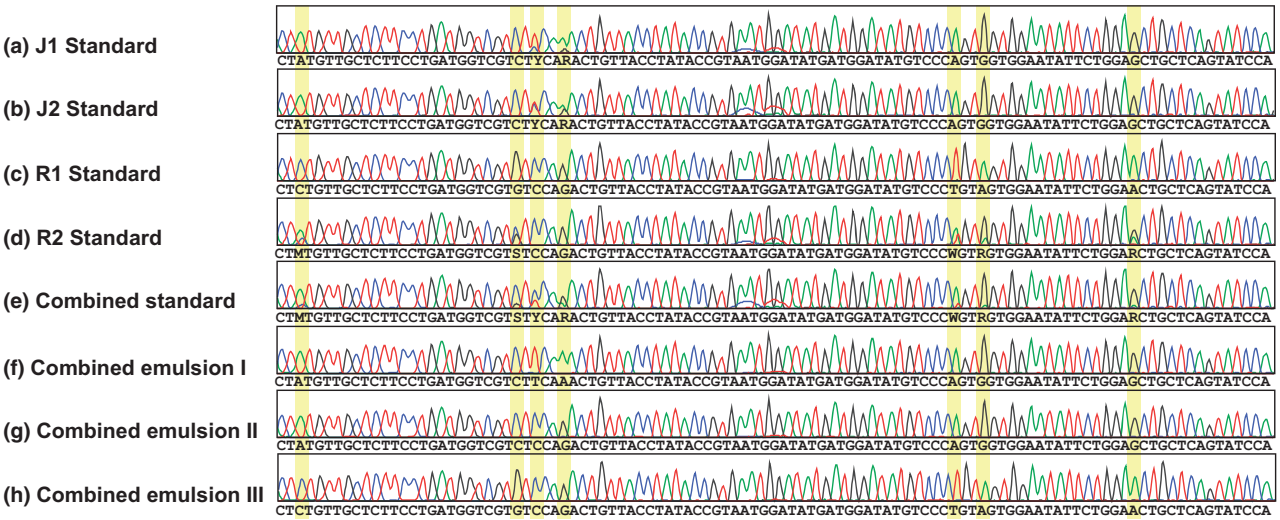

Fig. 3 Sequences of heterozygous diploid individuals generated using conventional PCR (a–e) and our modified single-molecule sequencing approach applied to a sample combining DNA from all individuals (f–h). We sequenced four individual *E. affinis* copepods using conventional PCR (a–d), as well as a combined sample containing all individuals (e), and identified numerous segregating sites (highlighted in yellow). When we subjected the combined sample of all individuals to our modified single-molecule sequencing approach, all segregating sites were resolved into three distinct haplotypes. Representative chromatograms from these three haplotypes are shown (f–h). Genotypes: Individual = Haplotype 1:Haplotype 2; J1 = I:II, J2 = I:II, R1 = III:III, R2 = II:III.

uneven vesicle formation, which reduced emulsion stability and increased the variance in vesicle size.

We also occasionally noticed that SYBR Green I could leak out of the aqueous phase and into the oil phase, resulting in strong background fluorescence with increased SYBR Green I concentrations. Using the custom light source, this background often appeared as an orange haze that readily obscured faint/small fluorescent vesicles. We found no consistent explanation for this phenomenon, and simply repeating the procedure was sufficient to eliminate this problem.

Overall, our modified single-molecule sequencing approach is simple, cost-effective for numerous samples, and can be implemented using basic molecular biology equipment. This method avoids common problems with cloning, such as PCR bias and recombination, and is not subject to artefacts that affect statistical phasing methods. The read lengths for the most common next-generation sequencing platforms are still shorter than those of traditional Sanger sequencing and while cost-effective for genomic studies are not ideal for investigating a single locus. We modified and improved on previous emulsion-based DNA amplification methods by using a melted agarose solution and SYBR Green I, allowing for rapid identification and isolation of individual vesicles with successful PCR amplification. We demonstrated the efficacy of this technique by sequencing DNA from individual vesicles to obtain individual haplotype sequences from an estuarine copepod, for which existing haplotype assessment methods had been difficult to apply.

Acknowledgements

This research was funded by NSF DEB-0745828 to Carol E. Lee, and University of Wisconsin, Hilldale Undergraduate Research Fellowship, College of Agricultural and Life Sciences Research Fellowship, and Microbiology Major Research Awards to Brian Metzger. Dr. Linda Graham provided useful advice on fluorescence microscopy.

References

Adkins RM (2004) Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. *BMC Genetics*, **5**, 22.

Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics*, **9**, 291–300.

Andrés AM, Clark AG, Shimmin L, Boerwinkle E, Sing CF, Hixson JE (2007) Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genetic Epidemiology*, **31**, 659–671.

Avery CL, Martin LJ, Williams JT, North KE (2005) Accuracy of haplotype estimation in a region of low linkage disequilibrium. *BMC Genetics*, **6**(Suppl. 1), S80.

Bagos PG (2011) Meta-analysis of haplotype-association studies: comparison of methods and empirical evaluation of the literature. *BMC Genetics*, **12**, 8.

Becker S, Boger P, Oehlmann R, Ernst A (2000) PCR bias in ecological analysis: a case study for quantitative *Taq* nuclease assays in analyses of microbial communities. *Applied and Environmental Microbiology*, **66**, 4945–4953.

Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, **84**, 210–223.

Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, **12**, 703–714.

Flot J-F, Tillier A, Samadi S, Tillier S (2006) Phase determination from direct sequencing of length-variable DNA regions. *Molecular Ecology Notes*, **6**, 627–630.

Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, **10**, 915–934.

Higasa K, Kukita Y, Kato K, Wake N, Tahira T, Hayashi K (2009) Evaluation of haplotype inference using definitive haplotype data obtained from complete hydatidiform moles, and its significance for the analyses of positively selected regions. *PLoS Genetics*, **5**, e1000468.

Hori M, Fukano H, Suzuki Y (2007) Uniform amplification of multiple DNAs by emulsion PCR. *Biochemical and Biophysical Research Communications*, **352**, 323–328.

Kitzman JO, Mackenzie AP, Adey A *et al*. (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotechnology*, **29**, 59–63.

Lee CE (1999) Rapid and repeated invasions of fresh water by the copepod *Eurytemora affinis*. *Evolution*, **53**, 1423–1434.

Lee CE (2000) Global phylogeography of a cryptic copepod species complex and reproductive isolation between genetically proximate "populations." *Evolution*, **54**, 2014–2027.

Lee CE, Frost BW (2002) Morphological stasis in the *Eurytemora affinis* species complex (Copepoda: Temoridae). *Hydrobiologia*, **480**, 111–128.

Levenstien MA, Ott J, Gordon D (2006) Are molecular haplotypes worth the time and expense? A cost-effective method for applying molecular haplotypes. *PLoS Genetics*, **2**, e127.

Lin D, Huang B (2007) The use of inferred haplotypes in downstream analyses. *The American Journal of Human Genetics*, **80**, 577–579.

Marchini J, Cutler D, Patterson N *et al*. (2006) A comparison of phasing algorithms for trios and unrelated individuals. *The American Journal of Human Genetics*, **78**, 437–450.

Meyerhans A, Vartanian J-P, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Research*, **18**, 1687–1691.

Mullis KB, Faloona FA (1987) Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Methods in Enzymology*, **155**, 335–350.

Murakawa K, Takiguchi S, Kambara H(2008) Method and apparatus for sample preparation. *US Patent App 2008/0241841 A1*.

Musyanovych A, Mailänder V, Landfester K (2005) Miniemulsion droplets as single molecule nanoreactors for polymerase chain reaction. *Biomacromolecules*, **6**, 1824–1828.

Nakano H, Kobayashi K, Ohuchi S, Sekiguchi S, Yamane T (2000) Single-step single-molecule PCR of DNA with a homo-priming sequence using a single primer and hot-startable DNA polymerase. *Journal of Bioscience and Bioengineering*, **90**, 456–458.

Nakano M, Komatsu J, Matsuura S, Takashima K, Katsura S, Mizuno A (2003) Single-molecule PCR using water-in-oil emulsion. *Journal of Biotechnology*, **102**, 117–124.

Rasch EM, Lee CE, Wyngaard GA (2004) DNA-Feulgen cytophotometric determination of genome size for the freshwater-invading copepod *Eurytemora affinis*. *Genome*, **47**, 559–564.

Ruano G, Fenton W, Kidd KK (1989) Biphasic amplification of very dilute DNA samples via "booster" PCR. *Nucleic Acids Research*, **17**, 5407.

Sabeti PC, Reich DE, Higgins JM *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, **94**, 441–448.

Soulsbury CD, Iossa G, Edwards KJ, Baker PJ, Harris S (2007) Allelic dropout from a high-quality DNA source. *Conservation Genetics*, **8**, 733–738.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, **68**, 978–989.

Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, **62**, 625–630.

Tawfik DS, Griffiths AD (1998) Man-made cell-like compartments for molecular evolution. *Nature Biotechnology*, **16**, 652–656.

Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ (2011) The importance of phase information for human genomics. *Nature Reviews Genetics*, **12**, 215–223.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Uddin M, Sturge M, Griffin C, Benteau S, Rahman P (2008) Variability of haplotype phase and its effect on genetic analysis. *Canadian Conference on Electrical and Computer Engineering*, 000595–000600. IEEE.

Wang GCY, Wang Y (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Applied and Environmental Microbiology*, **63**, 4645–4650.

Winkler G, Dodson JJ, Lee CE (2008) Heterogeneity within the native range: population genetic analyses of sympatric invasive and noninvasive clades of the freshwater invading copepod *Eurytemora affinis*. *Molecular Ecology*, **17**, 415–430.

All authors contributed to the writing of the paper and preparation of tables and figures.

Data accessibility

Original chromatograms: dryad doi:10.5061/dryad.rh7f4

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Theoretical probability that a single vesicle contains either zero or one copy of a specific haplotype as a function of the total number of vesicles and the total number of copies of that haplotype.

Figure S2 Haplotype sequences of a diploid individual generated using conventional PCR (a) and our modified single-molecule sequencing approach (b–m).