Deceptive combined effects of short allele dominance and stuttering: an example with *Ixodes scapularis*, the main vector of Lyme disease in the U.S.A.

Thierry De Meeûs[1,*], Cynthia T. Chan[2,3], John M. Ludwig[2,4], Jean I. Tsao[5], Jaymin Patel[2,6], Jigar Bhagatwala[2,7], and Lorenza Beati[2].

1. Intertryp, IRD, Cirad, Univ Montpellier, Montpellier, France.
2. The U.S. National Tick Collection, Institute for Coastal Plain Science, Georgia Southern University, Statesboro, GA, USA.
3. College of Agricultural & Environmental Sciences, University of Georgia, Athens, GA, USA.
4. Department of Microbiology, University of Georgia, Athens, GA, USA
5. Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA
6. Division of Hospital Medicine, College of Medicine, University of Florida, Gainesville, FL, USA.
7. Medical College of Georgia, Augusta University, Augusta, GA, USA.

* Corresponding author: thierry.demeeus@ird.fr

Keywords: Microsatellite loci, short allele dominance, stuttering, heterozygote deficit, linkage disequilibrium, curing microsatellite data.

Running title: Short allele dominance and stuttering

Abstract

Null alleles, short allele dominance (SAD), and stuttering increase the perceived relative inbreeding of individuals and subpopulations as measured by Wright's F_{IS} and F_{ST}. Such amplifying problems are usually caused by inaccurate primer design (if developed from a different species or a distant population), poor DNA quality, low DNA concentration, or a combination of some or all these sources of inaccuracy. When combined, these issues can increase the correlation between polymorphism at concerned loci and, consequently, of linkage disequilibrium (LD) between those. In this note, we studied an original microsatellite data set generated by analyzing nine loci in *Ixodes scapularis* ticks from the eastern U.S.A. To detect null alleles and SAD we used correlation methods and variation measures. To detect stuttering, we evaluated heterozygote deficit between alleles displaying a single repeat difference. We demonstrated that an important proportion of loci affected by amplification problems (one with null alleles, two with SAD and three with stuttering) lead to highly significant heterozygote deficits (F_{IS}=0.12, *p*-value<0.0001). This occurred together with a prohibitive proportion (31%) of pairs of loci in significant LD and a significant variation in the measure of population subdivision across loci (Wright's F_{ST}). This suggested a strong Wahlund effect and/or homogenizing selection at several loci. By finding small peaks corresponding to previously disregarded larger alleles in some homozygous profiles for loci with SAD and by pooling alleles close in size for loci with stuttering, we generated an amended dataset. Except for one locus with null alleles, the analyses of the corrected dataset revealed a significant excess of heterozygotes (F_{IS}=-0.058) as expected in dioecious and strongly subdivided populations, with a more reasonable proportion (17%) of pairs of loci characterized by significant LD. Strong subdivision was also confirmed by the standardized F_{ST}' corrected for null alleles (F_{ST}'=0.28).

Introduction

Null alleles, short allele dominance (SAD) and stuttering are frequent consequences of poor PCR amplifications, in particular for microsatellite markers. Amplification problems usually arise when primers are designed by using DNA of a different species or a distant population, when DNA is degraded or at too low of a concentration (Chapuis and Estoup, 2007), or any combination of these listed causes.

Null alleles occur when a mutation on the flanking sequence of the targeted locus affects the hybridization of the corresponding primer, resulting in amplification failure. Heterozygous individuals with one null allele falsely appear to be homozygous, while homozygous individuals for null alleles are considered to be missing data.

SAD, also called long allele dropout (Van Oosterhout et al., 2004), known from minisatellite markers, was also discovered to occur in microsatellite loci (Wattier et al., 1998). In heterozygous samples, longer alleles are less successfully amplified than shorter alleles through competition for Taq polymerase. This can lead to misinterpreting heterozygous individuals as homozygous for the shortest allele (De Meeûs et al., 2007).

Stuttering is the result of inaccurate PCR amplification through Taq slippage of a specific DNA strand. This generates several PCR products that differ from each other by one repeat and can cause difficulties when discriminating between fake and true homozygotes, such as heterozygous individuals for dinucleotide microsatellite allele sequences with a single repeat difference.

Allelic dropout is akin to SAD, but occurs randomly to any allele irrespective of its size.

The consequence of these issues is a homozygous excess when compared to the expected Castle-Weinberg proportions (Castle, 1903; Weinberg, 1908) measured by Wright's F_{IS} (Wright, 1965). These problems, like all others associated with amplification, are locus specific (Van Oosterhout et al., 2004; De Meeûs et al., 2007; De Meeûs, 2018) and thus lead to locus specific variation (namely, an increase) of F_{IS}. A less well known, though well documented (Chapuis and Estoup, 2007; Séré et al., 2017; Manangwa et al., 2019) effect of such amplification problems consists of an increase of Wright's F_{ST} (Wright, 1965) that is commonly used to measure the degree of genetic differentiation between subpopulations.

While an analytical cure exists for null alleles (Chapuis and Estoup, 2007; Séré et al., 2017), such remediation is unavailable for SAD and stuttering. To the best of our knowledge, the impact of amplification problems on linkage disequilibrium (LD) between locus pairs has yet to be investigated. When combined, the effect of the occurrence of null

alleles, SAD, and stuttering may artificially generate a positive correlation between allele occurrences at affected loci and then increase the perceived LD between them.

In this note, we utilize an original data set generated through the analysis of nine microsatellite loci in *Ixodes scapularis*, sampled across the eastern U.S. to show that the combined effect of SAD, stuttering, and null alleles can induce an increase in the number of locus pairs in significant LD. We then propose and test an efficient way to amend such data.

Material and Methods

Sampling and DNA extraction

Larvae, nymphs, and adults of *I. scapularis* were sampled haphazardly from different sites across the eastern U.S. on different occasions, extending from November 2001 to May 2014, by means of dragging and flagging the vegetation (Figure 1 and Table 1) (Rulison et al., 2013).

Immatures, particularly the larval offspring belonging to a single female, tend to cluster on vegetation and can thus constitute different lineages within the same subsample (Kempf et al., 2011). Consequently, to avoid possible Wahlund effects, where a heterozygote deficit results from the admixture of individuals from genetically distant subpopulations (e.g. see (De Meeûs, 2018)), we removed immature stages from the present study. The remaining 387 adult ticks were subdivided into cohorts, with each cohort comprising samples collected across two consecutive years in the fall, the following winter and spring across the tick distribution range. This subdivision was based on observations showing that Northeastern adults active in Fall can undergo winter quiescence and resume activity in spring (Yuval and Spielman, 1990).

Many publications have emphasized the importance of mitochondrial clades in different populations of *I. scapularis* across the U.S. (Norris et al., 1996; Qiu et al., 2002; Sakamoto et al., 2014). Thus, to account for the mitochondrial clade representation and to (again) avoid possible Wahlund effects, all ticks were assigned clades by phylogenetic analysis of their 12S rDNA gene sequences. We identified 6 main clades in our dataset, the previously identified American clade was subdivided in two lineages (AMI and AMII), and the so-called southern clade was subdivided in 4 lineages (SOI, SOII, SOIII and SOIV) (Table 1).

In conclusion, the combination of Site-Clade-Cohort data defined 45 subsamples within the 387 individual adult ticks. Some subsamples included a small number of

individuals (1-4). While such subsamples were expected to exert a negligible weight on our analyses, they were not eliminated.

Procedures for all DNA extractions followed published modified protocols (Beati and Keirans, 2001) with a DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA).

Selection and characterization of microsatellite markers

Thirteen of the first batches of genome sequences of *I. scapularis* that were accessioned by VectorBase (www.vectorbase.org; Giraldo-Calderón *et al.* 2015) in GenBank (AC205653.1, AC205652.1, AC205650.1, AC205647.1, AC205646.1, AC205643.1, AC205642.1, AC205641.1, AC205638.1, AC205635.1, AC205634.1, AC205632.1, AC205630.1) were used to manually detect motifs with at least 6 repeats of AG, AT, CA, TA, TG, CT, GC, ACG, GTT, TTA, CAC, GAT, and AAAC. Primer pairs were selected in the flanking regions by using Oligo v.5 (Molecular Biology Insights, Colorado Springs, CO). DNA, extracted and pooled from six ticks from Connecticut, was used to test whether the selected primer sets successfully amplified fragments of the expected size. PCRs were performed using the 5-Prime Master PCR kit (5-Prime, Gaithersburg, MD) and a single touch-down amplification protocol consisting of 5 min. denaturation at 93°C; 5 cycles: 20 sec. denaturation at 93°C, 20 sec. annealing at 55°C-1.5°C/cycle, 30 sec. elongation at 72°C; 30 cycles: 20 sec. denaturation at 93°C, 25 sec. annealing at 47°C, 30 sec. elongation at 70°C; final extension at 70°C for 5 min. Amplicons were run on 4% E-gels (Life Technologies Co., Carlsbad, CA). The risk of having selected primers within repeated portions of the genome had to be considered due to the fact that large repeated genomic fragments are known to occur abundantly in *I. scapularis* (Gulia-Nuss et al., 2016). In order to confirm that the primers were amplifying the targeted loci, the amplicon of one randomly chosen tick was cloned with a TOPO-TA PCR cloning kit (Life Technologies Co, Carlsbad, CA) for each locus. Five cloned colonies were picked randomly for each tick and the insert amplified and sequenced (DNA Analysis Facility on Science Hill, Yale University). Finally, as microsatellite primers are known to amplify sometimes more than one closely related species, the same set of primers was tested on DNA samples of *Ixodes ricinus*, *Ixodes pacificus*, and *Ixodes persulcatus* (LB, personal collection), all taxa belonging to the *I. ricinus* complex of ticks (Keirans et al., 1999).

The primer pairs that yielded amplicons of the expected size were then used to individually amplify a subset of 67 DNA samples from ticks collected by flagging or dragging in Alabama (10), Georgia (15), Connecticut (16), Massachusetts (14), New York (2), Pennsylvania (2), and South Carolina (8) (Table 1). For these amplifications, forward

primers were labeled with fluorescent dyes (Applied Biosystems, Thermo Fisher Scientific, CA) as listed in Table 2. The amplicons were sent to the DNA Analysis Facility on Science Hill (Yale University, New Haven, CT) for genotyping. The allele peaks were scored using GeneMarker (SoftGenetics, State College, PA). All data were recorded in an Excel spreadsheet for further ease of conversion.

Figure 1: Sampling sites for *Ixodes scapularis* from the eastern U.S.A. (State codes as in Table 1).

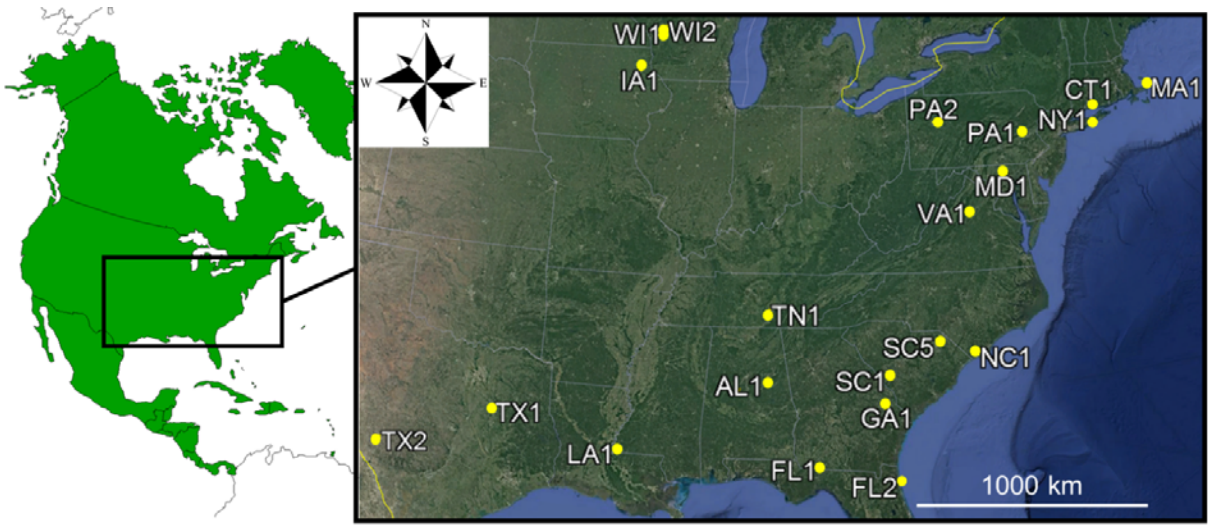


Table 1: State, site, GPS coordinates (decimal degrees), 12S clade membership, date of sampling, corresponding cohort membership and size (*N*) of *Ixodes scapularis* adult subsamples from the eastern U.S.A.

State	Site	latitude	longitude	Clade	Date	Cohort	N
Alabama	AL1	33.24	-86.13	AMI	2011 Jan	C7	21
	AL2	32.95	-87.14	AMI	2012 Dec	C9	3
	AL1	33.24	-86.13	AMII	2011 Jan	C7	18
	AL2	32.95	-87.14	AMII	2012 Dec	C9	2
	AL1	33.24	-86.13	SOI	2011 Jan	C7	6
	AL1	33.24	-86.13	SOII	2011 Jan	C7	6
Connecticut	CT1	41.35	-72.76	AMI	2001 Nov	C1	3
	CT1	41.35	-72.76	AMI	2003 Jun	C2	21
Florida	FL2	30.06	-81.37	AMI	2011 Jan	C7	2
	FL2	30.06	-81.37	AMII	2011 Jan	C7	3
	FL2	30.06	-81.37	SOI	2011 Jan	C7	1
	FL2	30.06	-81.37	SOII	2011 Jan	C7	4
	FL1	30.65	-84.21	SOII	2012 Dec	C9	4
Georgia	GA1	32.45	-81.78	AMI	2009 Dec	C5	11
	GA1	32.45	-81.78	SOI	2009 Dec	C5	5
	GA1	32.45	-81.78	SOII	2009 Dec	C5	17
	GA1	32.45	-81.78	SOIII	2009 Dec	C5	1
Iowa	IA1	42.67	-91.59	AMI	2007 May	C3	3
Louisiana	LA1	30.94	-91.46	AMI	2012 Dec	C9	3
	LA1	30.94	-91.46	AMII	2012 Dec	C9	2
	LA1	30.94	-91.46	SOIII	2012 Dec	C9	1
	LA1	30.94	-91.46	SOIV	2012 Dec	C9	2
Massachusetts	MA1	41.71	-69.92	AMI	2010 May	C5	6
Maryland	MD1	39.29	-76.88	AMI	2010 Oct	C7	11
North-Carolina	NC1	33.91	-78.39	AMI	2009 Jan	C4	22
	NC1	33.91	-78.39	SOI	2009 Jan	C4	4
	NC1	33.91	-78.39	SOIII	2009 Jan	C4	2
New-York	NY1	40.76	-72.83	AMI	2009 Oct	C5	10
Pennsylvania	PA2	41.06	-79.48	AMI	2014 May	C10	33
	PA1	40.67	-75.96	AMI	2010 Oct	C7	5

	SC1	33.33	-81.66	AMI	2011 Apr	C7	12
	SC5	34.29	-79.87	AMI	2012 Dec	C9	3
	SC1	33.33	-81.66	AMII	2011 Apr	C7	9
South-Carolina	SC1	33.33	-81.66	SOI	2011 Apr	C7	7
	SC1	33.33	-81.66	SOII	2011 Apr	C7	13
	SC5	34.29	-79.87	SOII	2012 Dec	C9	2
	SC1	33.33	-81.66	SOIII	2011 Apr	C7	1
Tennessee	TN1	35.37	-86.07	AMI	2010 Dec	C7	23
	TX2	30.24	-100.72	AMI	2012 Dec	C9	11
Texas	TX1	31.80	-96.23	SOI	2012 Dec	C9	6
	TX2	30.24	-100.72	SOIV	2012 Dec	C9	14
Virginia	VA1	38.29	-78.29	AMI	2010 Oct	C7	7
	VA2	37.32	-80.73	AMI	2012 Dec	C9	4
Wisconsin	WI1	43.95	-90.70	AMI	2011 Oct	C8	24
	WI2	44.04	-90.65	AMI	2012 Oct	C9	19

Table 1: List of primer sets used or developed for this study. PAC = *I. pacificus*, PER = *I. persulcatus*, RIC = *I. ricinus.* Approximate size range and number of alleles correspond to 67 individuals used for the initial tests.

Locus name	Repeat array	Primer name	Primer sequences (5'→3')	Reference or GenBank accession number	Dye	Approx. Size range (bp)	No of alleles	Cross-species amplification
IS1	(AG)10	Amy1-IsAG25a Amy2-IsAG25b	AAATGTCCGAACAGCCTTAT GCCCTTGAGTCTACCCACTA	Fagerberg *et al* (2001)	6 FAM	93-193	17	PAC/PER/RIC
IS3	$(GTT)_5$	bac1d_a bac1d_b	GCAGATCTCTTGGGCTAG AAGCTAAGGCGTTCGTTG	AC205653	VIC	76-100	7	none
IS4	(AT)21	bac1m_a bac1m_b	TGTCGGTTTGATGCCAA GGCTCCATTCACCAGTC	AC205653	VIC	88-126	17	PAC/PER/RIC
IS5	(CA)9	bac3dh_a bac3dh_b	TGCCTGTGACGAAACCA TCTCCCAAGAGATCTAGGTA	AC205650	NED	62-140	17	none
IS6	(TA)10	bac1j_a bac1j_b	TCTCCCAAGAGATCTAGGTA ATCTGTTCAGTGGGCACA	AC205653	VIC	100-186	13	PER
IS7	(TA)11	bac1k_a bac1k_b	GGGACTGGACACACGA CTAGGTGGCGCAAGTC	AC205653	VIC	48-170	26	none
IS8	(CA)14	bac3s_a bac3s_b	CGTTTCAAAGTCGGAGA GATGTGAGGGCGTGGT	AC205650	PET	96-194	11	PER
IS9	(AAAC)5	bac4cef_a bac4cef_b	CGCCTTTTGTCCCAACC GACTAACAGCATTGGAGCA	AC205647	6 FAM	85-125	12	PER
IS10	(TTA)9	bac5cf_a bac5cf_b	TCCCCCAACAAGATTGATG GAGACGACGTAGATTCTTG	AC205646	6 FAM	77-137	15	none
IS11	$(TTA)_6$	bac5g_a bac5g_b	GCTTTAGCGGGCTGGT TACGTGAATACGTCCTTGG	AC205646	PET	81-165	12	PER
IS12	(TA)43	bac6a_a bac6a_b	GCAAGCTTCGCTATTCTC CAGTAATTTCGCATCGGTT	AC205643	6 FAM	111-229	26	none
IS13	(TA)22	bac6c_a bac6c_b	TAGGTACAAGAAAACGTGCT CAAGGTAATTGTTCTCGTCA	AC205643	NED	37-91	17	none
IS14	(TA)5	bac6d_a	CCTTGCCTTACATGGTT	AC205643	HEX	57-105	13	PAC/PER/RIC

		bac6d_b	CGTACCAAACCAAAGCAAG					
IS15	$(AT)_8$	bac6e_a	TATTGTAACCGACGCTAGG					
		bac6e_b	GACAATCTCTACGCAAATCC	AC205643	NED	79-125	18	none
IS16	$(CA)_8$	bac6f_a	CCCCCAAACACGCACA					
		bac6f_b	TTGCTTCATGCAGGGAAC	AC205643	VIC	80-106	12	RIC
IS17	$(CA)_6$	bac7e_a	CCAGCATTTAACCCTCAAG					
		bac7e_b	TAGTGGGGTATGGCACTG	AC205642	HEX	139-197	12	PER/RIC
IS18	$(TG)_6$	bac8a_a	GTAGGTACCCTAAGAAGGAT		6			
		bac8a_b	TTGAGGAAGCAGAATGTAGG	AC205641	FAM	75-195	16	PER/RIC
IS19	$(CT)_7$	bac9a_a	AGAACCAGTTCAGCATTCC					
		bac9a_b	GAACATTTTCACGTGTTGC	AC205638	PET	94-166	6	PER
IS20	(GC)9	bac11a_a	CGCTCCCTTCGAAGTTC					
		bac11a_b	GAGAAGACAGTTTCCATCG	This study	HEX	76-106	13	PAC/PER/RIC
IS21	(ACG)6	bac11c_a	CGAATCGCGCACACTAG					
		bac11c_b	GCTGTGTTGCTGGTCAC	This study	NED	109-251	14	PAC/PER
IR27	(AC)9		ATACCCGTAGAACGAGAG	Delaye *et al*	6			
			GTTTTTCAAGATTTCCGCC	(2008)	FAM			RIC

Genotyping

Nine microsatellite loci (IR27, IS1, IS3, IS11, IS15, IS16, IS17, IS18, and IS19) were used for genotyping at the continental scale (Table 2). Of these, IR27 (Delaye et al., 1998) and IS1 (Fagerberg et al., 2001), were drawn from previously published studies. The loci were amplified and genotyped using the procedures described above, although PCR conditions had to be slightly optimized for markers IS11 and IS15 (touchdown annealing temperature decreased from 58°C to 50°C) and IR27 (touchdown annealing temperature decreased from 56°C to 53°C) (Table 2).

Population genetics analyses

The raw data set was coded and converted into all required formats using Create (Coombs et al., 2008).

To test for LD, we used the *G*-based test first described by Goudet et al. (Goudet et al., 1996), adapted for contingency tables of locus pairs with 10000 reshuffling of genotypes (or 15000 when needed). The *G* statistics obtained for each subsample were then summed over all subsamples to get a single statistic and hence, a single test across subsamples. This procedure was shown to be the most powerful (De Meeûs et al., 2009) and was implemented within Fstat 2.9.4 (Goudet, 2003) an updated version of the original 1995 Fstat software (Goudet, 1995). There are as many tests as locus pairs and these tests are correlated (one locus is used as many times as there is any other locus). To take into account this repetition of correlated tests, we used Benjamini and Yekutieli (BY) false discovery rate procedure (Benjamini and Yekutieli, 2001) with R version 3.5.1 (R-Core-Team, 2018). We also undertook a Fisher exact test with Rcmdr version 2.3-1 (Fox, 2005; Fox, 2007) when examining the eight tests in which each locus was involved in, in order to see if some pairs of loci were found in significant LD more often than by chance.

For a hierarchy with three levels (individuals, subsamples, and total sample), three *F*-statistics can be defined (Wright, 1965). F_{IS} measures inbreeding of individuals relative to inbreeding of subsamples or relative deviation of genotypic proportions from local random mating proportions. F_{ST} measures inbreeding of subsamples relative to total inbreeding or relative inbreeding due to the subdivision of the total population into several isolated subpopulations. F_{IT} measures inbreeding of individuals relative to total inbreeding. Under the null hypothesis (panmixia and no subdivision), all these statistics are expected to be null. Otherwise, F_{IS} and F_{IT} can vary from -1 (one heterozygote class) to +1 (all individuals homozygous) and F_{ST} from 0 (all subsamples share similar allele frequencies)

to +1 (all subsamples fixed for one or the other allele). These statistics were estimated with Weir and Cockerham's unbiased estimators (Weir and Cockerham, 1984) with Fstat.

In dioecious species (like ticks), heterozygote excess occurs over all loci (e.g. (De Meeûs et al. 2007)) and is proportional to subpopulation size (N_e=-1/(2×F_{IS})-F_{IS}/(1+F_{IS})) (Balloux, 2004). Therefore, the finding of homozygous excesses really represents a strong deviation from random mating expectations. Technical problems, like null alleles, stuttering, SAD or allele dropouts unevenly affects some loci, producing a positive F_{IS} with an important variation across loci with significant outliers (De Meeûs 2018). Significant departure from 0 of these *F*-statistics was tested with 10000 randomizations of alleles between individuals within subsample (deviation from local random mating test) or of individuals between subsamples within the total sample (population subdivision test). For F_{IS}, the statistic used was *f* (Weir and Cockerham's F_{IS} estimator). To test for subdivision, we used the *G*-based test (Goudet et al. 1996) over all loci, which is the most powerful procedure when combining tests across loci (De Meeûs et al. 2009).

To compute 95% confidence intervals (95%CI) of *F*-statistics, we used the standard error of F_{IS} (StrdErrFIS) and F_{ST} (StrdErrFST) computed by jackknifing over populations, or 5000 bootstraps over loci as described elsewhere (De Meeûs et al. 2007). For jackknives, the number of usable subsamples was restricted to subsamples with at least 5 ticks (23 subsamples) (e.g. (De Meeûs, 2012) p 73).

In case of significant homozygote excess and LD we tried to discriminate demographic from technical causes with the determination key proposed by De Meeûs (De Meeûs 2018). Null alleles better explain the data if the StrdErrFIS becomes at least twice as high as StrdErrFST; F_{IS} and F_{ST} are positively correlated; and a positive correlation links F_{IS} and the number of missing data (putative null homozygotes). The significance of correlations was tested with a unilateral (ρ>0) Spearman's rank correlation test with R. The presence of null alleles was also verified with MicroChecker v 2.2.3 (Van Oosterhout et al. 2004) and null allele frequencies estimated with Bookfield's second method (Brookfield 1996). The adjustment between observed and expected numbers of missing data was tested with a unilateral exact binomial test in R with the alternative hypothesis being "there is not enough missing data as expected if heterozygote deficits were entirely explained by null alleles under panmixia". MicroChecker also detects stuttering and SAD. Stuttering is detected when MicroChecker reveals an observed proportion of heterozygous individuals for alleles with one repeat difference significantly smaller than the expected value. The occurrence of SAD was also checked with an unilateral (ρ<0) Spearman's rank correlation test between allele size and F_{IT} as proposed by Manangwa et al. (Manangwa et al., 2019).

This test is more powerful than with F_{IS} as was proposed earlier (Wattier et al., 1998; De Meeûs et al., 2004). If previous tests are not significant and if StrdErrFIS > StrdErrFST, then a Wahlund effect better explains the data (De Meeûs, 2018), this is especially true in instances of significant LD (Manangwa et al., 2019). In these cases, a positive correlation between the number of times a locus is found in significant LD (NLD) and its total genetic diversity as measured by Nei's H_T (Nei and Chesser, 1983) (Spearman's correlation above 0.1) suggests an admixture of individuals from several subpopulations of the same species but with an important degree of genetic differentiation between admixed subpopulations (i.e. number of immigrants N_em=2). If the correlation is negative and the proportion of significant LDs is above 40%, an admixture of different strongly divergent entities (e.g. species) better explains the data (Manangwa et al., 2019). We thus undertook a bilateral Spearman's test.

When diagnosed, there is no analytical remedy for SAD or stuttering as there is for null alleles (Chapuis and Estoup, 2007; Séré et al., 2017). Nevertheless, SAD can be cured by going back to the chromatograms of homozygous individuals and trying to find a micro-peak (see the Results and discussion section), with a larger size, ignored in the first reading. If enough profiles can be corrected this might salvage the incriminated locus. Stuttering can be addressed by pooling alleles close in size. However, this procedure requires that none of the pooled allele groups is constituted of rare alleles only. Indeed, pooling rare alleles, usually found in heterozygosity with a more frequent allele, will tend to artificially generate heterozygous excesses. In order to avoid this consequence, each pooled group should contain at least one frequent allele (e.g. with $p \geq 0.05$).

In dioecious small populations, a heterozygote excess is expected. However, loci with null alleles may display heterozygote deficits. In such situations a bilateral test (F_{IS} is not different from 0) is needed and obtained as $p_{bilateral}=p_{mini}+1-p_{maxi}$, where p_{mini} is the minimum unilateral p-value (for heterozygote deficit or excess) and p_{max} is the maximum one.

Finally, a more accurate estimate of F_{ST} can be made for datasets with null alleles after recoding missing genotypes as homozygous for allele 999 (null allele) with the ENA method as implemented in FreeNA (Chapuis and Estoup, 2007). This estimate can then be corrected for polymorphism with the formula $F_{ST}'=F_{ST}/(1-H_S)$ (Hedrick, 1999).

Results and discussion

Primer selection and characterization of loci

The inspection of the GenBank genomic sequences revealed the presence of 67 short tandem repeated motifs. The program Oligo v.5 did not find suitable primers for 17 of them. Of the remaining 50 primer pairs, 22 amplified the pooled DNA sample and the sizes of the amplicons were approximately as expected. The cloned amplified inserts all contained the expected microsatellite repeats and flanking regions. The 22 primer sets consistently amplified DNA from the 67 *I. scapularis* ticks and some of them also amplified DNA of the other *Ixodes* species (Table 2).

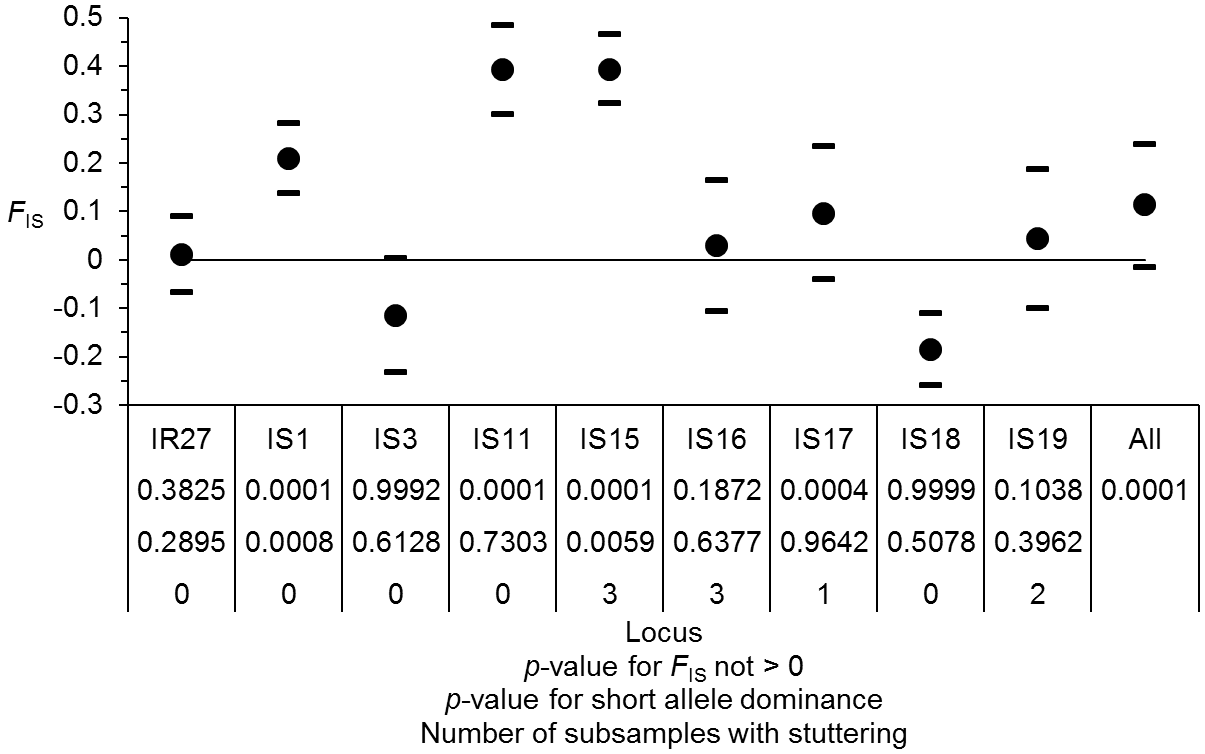
Raw data analyses

All data are available in the supplementary File S1.

In order to maximize precision, LD tests were implemented over 15000 permutations. There was a very important proportion of locus pairs in significant LD (~31%), with a negative correlation between NLD and H_T (ρ=-0.10, *p*-value=0.7974). None of these tests remained significant after BY correction, although one pair displayed a very low *p*-value (0.06). No single locus was found more often in significant LD than the others (*p*-value=0.1626).

There was a highly significant heterozygote deficit (F_{IS}=0.115, *p*-value<0.0001), with a considerable variation across loci (Figure 2).

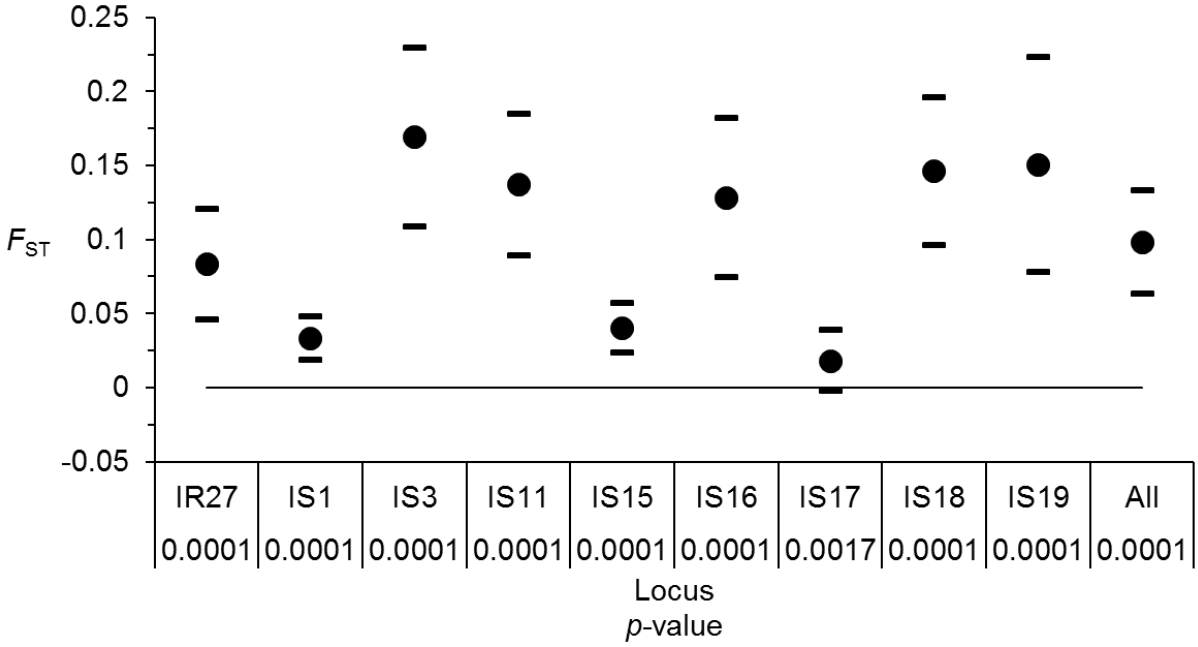
Figure 2: F_{IS} values for each locus and averaged across those (All) of *Ixodes scapularis* from the eastern U.S.A. with 95% jackknife confidence intervals over subsamples (for each locus) and bootstraps over loci (All). Results of tests for panmixia, short allele dominance and number of subsamples with stuttering for each locus are also indicated.



StdrdErrFIS (0.07) was almost four times higher than StrdErrFST (0.019); the correlation between F_{IS} and F_{ST} was negative (ρ=-0.55, *p*-value=0.9397) and the correlation between F_{IS} and the number of blanks was positive but not significant (ρ=0.18, *p*-value=0.3218). These results suggested that locus-specific effects were involved. Nevertheless, null alleles only weakly explained the observed pattern. The substantial proportion of significant LDs suggested the existence of a strong Wahlund effect. Small subsamples due to the partitioning of the data into cohorts and 12S clades was expected to considerably lower the power of these tests, which may also explain why none of them remained significant after BY correction, which is a rather stringent procedure.

Subdivision was substantial (F_{ST}=0.098, 95% CI [0.063,0.133], *p*-value<0.0001), but highly variable across loci (Figure 3). Some loci (IS1, IS15 and IS17) displayed very low values (Figure 3). Interestingly, loci IS15 and IS17 were found four and two times in significant LD, respectively. Homogenizing selection (overdominance, balanced selection) might have produced the pattern observed at loci IS15 and IS17.

Figure 3: F_{ST} values for each locus and averaged across those (All) of *Ixodes scapularis* from the eastern U.S.A. with 95% jackknife confidence intervals over subsamples (for each locus) and bootstrap over loci (All). Results of tests for significant subdivision (*p*-value) are also indicated.

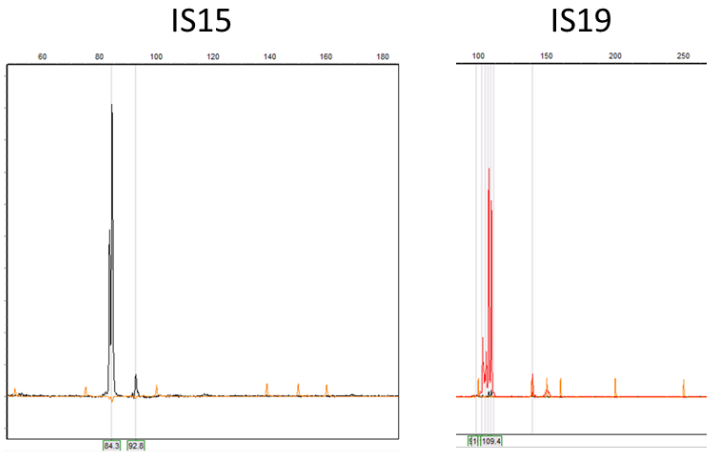


Two loci displayed highly significant SAD (Figure 2) (loci IS1 and IS15) and stuttering was diagnosed for four loci (IS15, IS16, IS17 and IS19).

Cured data set

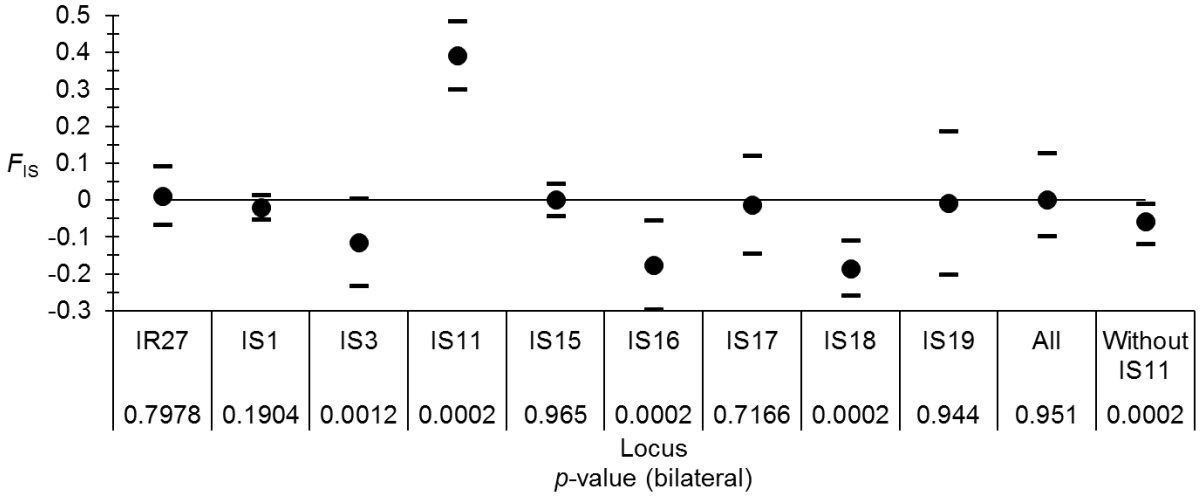
To prevent withdrawing five (out of nine) loci, due to stuttering, SAD and/or possible selection, we went back to the data. We first scanned the chromatograms for previously ignored micro-peaks in homozygous individuals at loci displaying SAD (IS1 and IS15) (Figure 4). SAD in IS15 might well have explained why we also found stuttering at this locus (see below). We then tried to pool alleles close in size as described above for loci IS16, IS17 and IS19. For IS16, alleles 88, 90 and 92 were recoded as 94; and allele 96 as 98. For locus IS17, alleles 170 and 172 were recoded as 174; alleles 178, 180, 182 and 184 were recoded as 186; alleles 188, 190 and 192 were recoded as 194; and alleles 196 and 198 as 200. Finally, for locus IS19, allele 91 was recoded as 93; alleles 97, 99, 101 and 103 as105; and alleles 107 and 109 as 111. The obtained amended dataset was called "Cured dataset" (Supplementary file S1).

Figure 4: Examples of an initially dismissed micro-peak that produced SAD at locus IS15 and of stuttering at locus IS19



With the cured dataset, the proportion of locus pairs in significant LD dropped to 17% and the smallest *p*-value after BY correction was 0.1052. In particular, for loci IS15 and IS17, significant LD occurred only once for the former and never for the latter. The correlation between NLD and H_T was much higher (ρ=0.27, *p*-value=0.4742). A heterozygote deficit was no longer observed (F_{IS}=0, bilateral *p*-value=0.951) (Figure 5). StrdErrFIS=0.063 was three times higher than StrdErrFST=0.021. There was no correlation between F_{IS} and F_{ST} (ρ=-0.28, *p*-value=0.78). The correlation between number of missing genotypes and F_{IS} was positive though marginally not significant (ρ=0.517, *p*-value=0.0809). MicroChecker diagnosed null alleles in 13 subsamples for locus IS11, which correspondingly displayed the highest F_{IS}=0.391, and in one subsample for IS19 (F_{IS}=-0.08). Null alleles probably explained these results, even if many missing data probably did not correspond to null homozygotes. If we remove IS11 from the data, global F_{IS} became significantly negative as expected for a small dioecious species (F_{IS}=-0.058, *p*-value=0.0002). As for subdivision, F_{ST} remained almost unaffected, even with the ENA method (F_{ST}=0.0964 in 95%CI=[0.0606..0.1338], *p*-value<0.0001). With H_S=0.66, the standardized value was F_{ST}'=0.28.

Figure 5: F_{IS} values for each locus and averaged across those (All) of *Ixodes scapularis* cured data from the eastern U.S.A. with 95% jackknife confidence intervals over subsamples (for each locus) and bootstraps over loci (All). Results of tests for panmixia for each locus are also indicated.



Conclusion

Combinations of amplification errors manifesting as null alleles, SAD or stuttering lead not only to heterozygote deficits but also to an overall increase in LD. For this, the different cures proposed are: in chromatograms, hunting for small and hard-to-detect peaks in homozygous individuals to correct for SAD and pooling alleles with close sizes to correct for stuttering. The obvious influence these problems can have would have led to the unnecessary withdrawal of the flawed loci. The proposed amendments, together with null allele's management with the ENA algorithm (Chapuis and Estoup, 2007), resulted in analyses which revealed heterozygote excesses that would be expected in small dioecious subpopulations, a more reasonable proportion of significant LD tests, and consequently more accurate estimates of the degree of population subdivision.

In this case, the relatively important global LD across loci is probably due to small effective sizes of the *I. scapularis* subpopulations, as confirmed by the important measure of subdivision between subsamples (F_{ST}≈0.3).

The correlation between mitochondrial clade allocation and genetic structure of *I. scapularis* is not within the scope of this work and will be treated in detail in a further study that will also include immature stages.

Acknowledgements

This work was funded by NSF Grant # EF0914390 to L.B. and EEID EF-0914476 to J.T. We thank members of the Lyme Gradient Consortium and many individuals who provided ticks. We are also grateful to Heather Walker, Jenny Dickson, Keely Duff, Laquita Burton, Alysha Benn and Nina Griffin, and the undergraduate students who provided field and laboratory assistance.

References

Balloux, F., 2004. Heterozygote excess in small populations and the heterozygote-excess effective population size. Evolution 58, 1891-1900.

Beati, L., Keirans, J.E., 2001. Analysis of the systematic relationships among ticks of the genera *Rhipicephalus* and *Boophilus* (Acari : Ixodidae) based on mitochondrial 12S ribosomal DNA gene sequences and morphological characters. J. Parasitol. 87, 32-48.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188.

Castle, W.E., 1903. The laws of heredity of Galton and Mendel, and some laws governing race improvement by selection. Proc Am Acad Arts Sci 39, 223-242.

Chapuis, M.P., Estoup, A., 2007. Microsatellite null alleles and estimation of population differentiation. Mol. Biol. Evol. 24, 621-631.

Coombs, J.A., Letcher, B.H., Nislow, K.H., 2008. CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs. Mol. Ecol. Res. 8, 578–580.

De Meeûs, T., 2012. Initiation à la génétique des populations naturelles: Applications aux parasites et à leurs vecteurs. IRD Editions, Marseille.

De Meeûs, T., 2018. Revisiting F_{IS}, F_{ST}, Wahlund effects and null alleles. J. Hered. 109, 446-456.

De Meeûs, T., Guégan, J.F., Teriokhin, A.T., 2009. MultiTest V.1.2, a program to binomially combine independent tests and performance comparison with other related methods on proportional data. BMC Bioinformatics 10, 443.

De Meeûs, T., Humair, P.F., Grunau, C., Delaye, C., Renaud, F., 2004. Non-Mendelian transmission of alleles at microsatellite loci: an example in *Ixodes ricinus*, the vector of Lyme disease. Int. J. Parasitol. 34, 943-950.

De Meeûs, T., McCoy, K.D., Prugnolle, F., Chevillon, C., Durand, P., Hurtrez-Boussès, S., Renaud, F., 2007. Population genetics and molecular epidemiology or how to "débusquer la bête". Infect. Genet. Evol. 7, 308-332.

Delaye, C., Aeschlimann, A., Renaud, F., Rosenthal, B., De Meeûs, T., 1998. Isolation and characterization of microsatellite markers in the *Ixodes ricinus* complex (Acari : Ixodidae). Mol. Ecol. 7, 360-361.

Fagerberg, A.J., Fulton, R.E., Black, W.C., 2001. Microsatellite loci are not abundant in all arthropod genomes: analyses in the hard tick, *Ixodes scapularis* and the yellow fever mosquito, *Aedes aegypti*. Insect Mol. Biol. 10, 225-236.

Fox, J., 2005. The R commander: a basic statistics graphical user interface to R. J. Stat. Software 14, 1–42.

Fox, J., 2007. Extending the R commander by "plug in" packages. R News 7, 46–52.

Goudet, J., 1995. FSTAT (Version 1.2): A computer program to calculate F-statistics. J. Hered. 86, 485-486.

Goudet, J., 2003. Fstat (ver. 2.9.4), a program to estimate and test population genetics parameters. Available at http://www.t-de-meeus.fr/Programs/Fstat294.zip, Updated from Goudet (1995).

Goudet, J., Raymond, M., De Meeûs, T., Rousset, F., 1996. Testing differentiation in diploid populations. Genetics 144, 1933-1940.

Gulia-Nuss, M., Nuss, A.B., Meyer, J.M., Sonenshine, D.E., Roe, R.M., Waterhouse, R.M., Sattelle, D.B., de la Fuente, J., Ribeiro, J.M., Megy, K., Thimmapuram, J., Miller, J.R., Walenz, B.P., Koren, S., Hostetler, J.B., Thiagarajan, M., Joardar, V.S., Hannick, L.I., Bidwell, S., Hammond, M.P., Young, S., Zeng, Q.D., Abrudan, J.L., Almeida, F.C., Ayllon, N., Bhide, K., Bissinger, B.W., Bonzon-Kulichenko, E., Buckingham, S.D., Caffrey, D.R., Caimano, M.J., Croset, V., Driscoll, T., Gilbert, D., Gillespie, J.J., Giraldo-Calderon, G.I., Grabowski, J.M., Jiang, D., Khalil, S.M.S., Kim, D., Kocan, K.M., Koci, J., Kuhn, R.J., Kurtti, T.J., Lees, K., Lang, E.G., Kennedy, R.C., Kwon, H., Perera, R., Qi, Y.M., Radolf, J.D., Sakamoto, J.M., Sanchez-Gracia, A., Severo, M.S., Silverman, N., Simo, L., Tojo, M., Tornador, C., Van Zee, J.P., Vazquez, J., Vieira, F.G., Villar, M., Wespiser, A.R., Yang, Y.L., Zhu, J.W., Arensburger, P., Pietrantonio, P.V., Barker, S.C., Shao, R.F., Zdobnov, E.M., Hauser, F., Grimmelikhuijzen, C.J.P., Park, Y., Rozas, J., Benton, R., Pedra, J.H.F., Nelson, D.R., Unger, M.F., Tubio, J.M.C., Tu, Z.J., Robertson, H.M., Shumway, M., Sutton, G., Wortman, J.R., Lawson, D., Wikel, S.K., Nene, V.M., Fraser, C.M.,

Collins, F.H., Birren, B., Nelson, K.E., Caler, E., Hill, C.A., 2016. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat Commun 7.

Hedrick, P.W., 1999. Perspective: Highly variable loci and their interpretation in evolution and conservation. Evolution 53, 313-318.

Keirans, J.E., Needham, G.R., Oliver, J.H., 1999. The *Ixodes* (*Ixodes*) *ricinus* complex worldwide. Diagnosis of the species in the complex, hosts and distribution, in: Needham, G.R., R., M., Horn, D.J., Welbourn, W.C. (Eds.), Acarology IX: Symposia. Ohio Biological Survey, Columbus, Ohio, pp. 341-347.

Kempf, F., De Meeûs, T., Vaumourin, E., Noel, V., Taragel'ová, V., Plantard, O., Heylen, D.J.A., Eraud, C., Chevillon, C., McCoy, K.D., 2011. Host races in *Ixodes ricinus*, the European vector of Lyme borreliosis. Infect. Genet. Evol. 11, 2043-2048.

Manangwa, O., De Meeûs, T., Grébaut, P., Ségard, A., Byamungu, M., Ravel, S., 2019. Detecting Wahlund effects together with amplification problems: cryptic species, null alleles and short allele dominance in Glossina pallidipes populations from Tanzania. Mol. Ecol. Res. 19, 757–772.

Nei, M., Chesser, R.K., 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47, 253-259.

Norris, D.E., Klompen, J.S.H., Keiransand, J.E., Black, I.W.C., 1996. Population genetics of *Ixodes scapularis* (Acari: Ixodidae) based on mitochondrial 16S and 12S genes. J. Med. Entomol. 33, 78–89.

Qiu, W.G., Dykhuizen, D.E., Acosta, M.S., Luft, B.J., 2002. Geographic uniformity of the Lyme disease spirochete (Borrelia burgdorferi) and its shared history with tick vector (*Ixodes scapularis*) in the Northeastern United States. Genetics 160, 833-849.

R-Core-Team, 2018. R: A Language and Environment for Statistical Computing, Version 3.5.0 (2018-04-23) ed. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Rulison, E.L., Kuczaj, I., Pang, G., Hickling, G.J., Tsao, J.I., Ginsberg, H.S., 2013. Flagging versus dragging as sampling methods for nymphal *Ixodes scapularis* (Acari: Ixodidae). J. Vector Ecol. 38, 163-167.

Sakamoto, J.M., Goddard, J., Rasgon, J.L., 2014. Population and demographic structure of *Ixodes scapularis* Say in the eastern United States. PLoS One 9, e101389.

Séré, M., Thévenon, S., Belem, A.M.G., De Meeûs, T., 2017. Comparison of different genetic distances to test isolation by distance between populations. Heredity 119, 55-63.

Van Oosterhout, C., Hutchinson, W.F., Wills, D.P.M., Shipley, P., 2004. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4, 535-538.

Wattier, R., Engel, C.R., Saumitou-Laprade, P., Valero, M., 1998. Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in *Gracilaria gracilis* (Rhodophyta). Mol. Ecol. 7, 1569-1573.

Weinberg, W., 1908. Über den Nachweis der Verebung beim Menschen. Jahresheft des Vereins fur Vaterlundische Naturkunde in Wurttemberg 64, 368-382.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. Evolution 38, 1358-1370.

Wright, S., 1965. The interpretation of population structure by F-statistics with special regard to system of mating. Evolution 19, 395-420.

Yuval, B., Spielman, A., 1990. Duration and Regulation of the Developmental Cycle of *Ixodes dammini* (Acari: Ixodidae). J. Med. Entomol. 27, 196-201.