Zoonotic *Babesia microti* in the northeastern U.S.: evidence for the expansion of a specific parasite lineage

Short title: Expansion of one lineage of *Babesia microti*

Heidi K. Goethert[1,2], Philip Molloy[2], Victor Berardi[2], Karen Weeks[2], and Sam R. Telford III[1,2*]

[1] Department of Infectious Diseases and Global Health, Cummings School of Veterinary Medicine, Tufts University, North Grafton, Massachusetts, United States

[2] Imugen, Inc., Norwood, Massachusetts, United States

*Corresponding author

Sam.telford @tufts.edu (SRT)

Abstract

The recent range expansion of human babesiosis in the northeastern United States, once found only in restricted coastal sites, is not well understood. This study sought to utilize a large number of samples to examine the population structure of the parasites on a fine scale to provide insights into the mode of emergence across the region. 228 *B. microti* samples collected in endemic northeastern U.S. sites were genotyped using published VNTR markers. The genetic diversity and population structure were analysed on a geographic scale using Phyloviz and TESS. Three distinct populations were detected in northeastern US, each dominated by a single ancestral type. In contrast to the limited range of the Nantucket and Cape Cod populations, the mainland population dominated from New Jersey eastward to Boston. Ancestral populations of *B. microti* were sufficiently isolated to differentiate into distinct populations. Despite this, a single population was detected across a large geographic area of the northeast that historically had at least 3 distinct foci of transmission, central New Jersey, Long Island and southeastern Connecticut. We conclude that a single *B. microti* genotype has expanded across the northeastern U.S. The biological attributes associated with this parasite genotype that have contributed to such a selective sweep remain to be identified..

Author summary

Babesiosis is a disease caused by a protozoan parasite, *Babesia microti*, related to malaria. The disease is acquired by the bite of the deer tick, the same tick that transmits Lyme disease. Although Lyme disease rapidly emerged over a wide range within the last 40 years, babesiosis remained rare with an extremely focal distribution. Within the last

47 decade, the number of reports of babesiosis cases has increased from an expanded area of
48 risk, particularly across the mainland of southern New England. We determined whether
49 the expanded risk may be due to local intensification of transmission as opposed to
50 introduction of the parasite. Historical fragmentation of the landscape suggests that sites
51 of *B. microti* transmission should have been isolated and thus evidence of multiple
52 genetically distinct populations should be found. By a genetic fingerprinting method, we
53 found that samples from the new mainland sites were all genetically similar. We
54 conclude that one parasite genetic lineage has recently expanded its distribution and
55 now dominates, suggesting that it has some phenotypic attribute that may confer a
56 selective advantage over others.
57

58 ## Introduction

59 Human babesiosis due to *Babesia microti* was first recognized on Nantucket
60 Island nearly 50 years ago [1], and a few years later the first cases of Lyme arthritis were
61 described from Old Lyme, Connecticut [2]. Both infections were found to be transmitted
62 by the deer tick (*Ixodes dammini*; American clade of *I. scapularis*), which had started to
63 be locally recognized as a human-biting pest [3]. In the 1970s and 80s, cases of either
64 were restricted to coastal New England sites, as well as foci in Wisconsin and Minnesota
65 [4–6]. Over the next 20 years, the number of Lyme disease cases significantly increased
66 and zoonotic risk spread rapidly across the northeastern United States. Lyme disease is
67 now endemic all the way north into Canada, west to Ohio, and south as far as Virginia.
68 Babesiosis, in constrast, lagged Lyme disease across these sites in time and in force of
69 transmission [7,8] and most cases were reported from coastal sites in the northeastern

U.S. However, in the last two decades, risk for babesiosis has intensified across the northeastern U.S. [9,10].

The 20 year lag between the range expansion of Lyme disease and that of babesiosis is not fully understood but in part relates to the difficulty with which *B. microti* may be transported. The two key facts that pose a paradox for range expansion are (1) only rodents and insectivores are known to be competent reservoirs of *B. microti* (may pass infection to uninfected ticks; [11]; and (2) *B. microti* is not inherited by ticks [11]. Larval ticks transported long distances by migratory birds, a critical mode of introduction for the agent of Lyme disease (for which certain passerines are competent reservoirs; [12]), do not develop into infected nymphs after they engorge on a bird because birds are not likely to be reservoir competent for *B. microti*. A *B. microti*-infected nymph (which acquired infection as a larva feeding on a mouse) transported by a bird could develop into an infected adult tick, but because that stage feeds only on medium to large sized mammals, especially deer, would not pass infection to a reservoir competent animal during the adult bloodmeal; deer are not competent reservoirs and carnivores are not likely to be competent. Hence, *B. burgdorferi* is said to travel on the backs of birds but *B. microti* on mice. Mice or other small mammals are unlikely to travel large distances. These considerations argue that the range expansion for *B. microti* babesiosis is not due to introductions of infected ticks by migratory birds.

The existence of silent natural foci of transmission is suggested by early rodent serosurveys for *B. microti* in Connecticut [13] and the detection of zoonotic clade parasites from sites in Maine where human babesiosis had not been recorded [14]. However, ecological surveillance has not been conducted across the northeastern U.S.

with sufficient detail across the likely range to provide much data of utility in understanding the tempo and mode of babesiosis risk. Longitudinal analyses of cases reported to state departments of public health are useful because case reports are based on a standard surveillance case definition and data are comparable between states. In Rhode Island, risk diminished from south to north [15]. In New York, babesiosis case reports gradually expanded from Long Island up the Hudson River valley. Similarly, in Connecticut, case reports expanded through the years from the southeastern coast first extending westward along the coast and then moving inland. [7,16–18]. The expansion of risk has been limited and incremental, with no long-distance introduction events such as those documented for Lyme disease, exemplified by its introduction into Canada. [19] A recent model for the emergence of babesiosis in New England suggests a "stepping-stone" model: a strong predictor of a town reporting babesiosis cases was the presence of a neighboring town reporting cases and that Lyme disease risk was a prerequisite [8]. Two stepping stone scenarios might have been operating concurrently in the last 20 years. (1) The force of *B. microti* transmission increased, slow and wave-like, across the northeastern landscape with the coastal earliest known zooonotic sites seeding adjacent more northerly sites. (2) Multiple cryptic enzootic sites (natural foci) with little zoonotic risk existed across the region, with local intensification of the force of *B. microti* transmission as tick densities increased to a threshold (estimated to be more than 20 nymphal deer ticks collected per hour; [20], and subsequent spread to adjacent areas.

The population structure of *B. microti* may provide evidence for tempo and mode of the expansion of babesiosis risk across the northeast. At the very basic level, new demes will be related genetically to their parent populations. In expanding populations,

116 genetic diversity may be low be due to bottlenecks and founder effects at the expanding
117 front [21,22]. In fact, observed patterns of diversity will vary depending on the process
118 of population expansion, viz., whether the population is being "pushed" or
119 "pulled"[21,23]. A "pulled" expansion occurs when pioneers are seeding new
120 populations ahead of the source population, such as would occur if individual infected
121 ticks are being introduced into a new site. This causes the genetic diversity to be lower at
122 the edge than the main body of the population due to successive founder effects. By
123 contrast, a "pushed" expansion occurs when a population expands at the edges of the
124 source location due to population growth. This expansion is usually slower and allows for
125 diversity in the source population to keep pace with geographical spread. A skewed
126 population diversity can occur near the expanding front of the population due to “allele
127 surfing”, that is high rates of reproduction can increase mutation and allow an allele to
128 surf the wave of population growth and become prevalent when it might not have become
129 fixed in a stationary population [21,24–26].
130 We have previously described variable number tandem repeat (VNTR) markers
131 for analyzing the population structure of *B. microti* and detected 3 distinct populations in
132 ticks and rodents across New England [27]. Whole genome sequencing of ecological
133 and clinical samples determined that these *B. microti* populations were strongly
134 differentiated, suggesting that they were geographically isolated [28]. However, neither
135 study analyzed sufficient samples to provide detail on the mode of expansion of the range
136 of B.microti in the northeastern U.S. Accordingly, we leveraged >200 diagnostic blood
137 samples from patients suspected of having acute babesiosis presenting to several clinical
138 practices across the northeastern U.S. and analyzed them with the VNTR assay. In

139 particular, we sought to determine the population structure of these parasites, and whether
140 range expansion was best represented by a “pulled” expansion model by introductions
141 into small founder populations, or a “pushed” model consistent with stepping stone
142 expansion.
143

144 ## Materials and methods

145 ### *B. microti* blood samples

146 De-identified discarded blood samples were collected from specimens that had
147 been sent to Imugen, Inc. for diagnosis of *B. microti* infection during the transmission
148 season of 2015. The town of the submitting doctor's office or hospital was associated
149 with each sample but no other data was available. Samples with a Ct>34 on the
150 diagnostic real time PCR performed at Imugen were excluded from the analysis because
151 they would not have had enough parasite DNA to yield reliable VNTR typing results.
152

153 ### Ethics Statement

154 This study was considered not to comprise human subjects research by the Tufts
155 University institutional review board.
156

157 ### Genotyping

158 DNA was extracted using a commercial spin column method (Qiagen Inc.). *B.*
159 *microti* was typed as described [27], with the exception that the hypervariable locus,
160 BMV4, was excluded. Samples were excluded from the final analysis if more than 1
161 locus failed to amplify. To avoid erroneously scoring stutter peaks, multiple peaks were

scored only if the size of the minor peak was almost equal to that of the major peak. *B. microti* merozoites infecting humans are haploid [29]; so all analyses were done under the assumption of haploidy. Samples that had multiple peaks in more than one locus were excluded, as it was impossible to determine the individual haplotypes needed for assigning a haplotype to a population using Phyloviz (see below). Samples that had multiple peaks in only a single locus were retained in the analysis and treated as two separate haplotypes.

Data analysis

VNTR haplotypes were analyzed with two programs (Phyloviz [30] and TESS [31]) that utilize different algorithms for assigning them to a population. Phyloviz determines mutually exclusive related groups by using the eburst algorithm on haplotype data to identify founder haplotypes and then predicts the descent from the founder to the other haplotypes without any predefined assumptions of populations or geographic location. TESS uses a Bayesian clustering algorithm to determine population structure from geographically defined haplotypes without assuming predefined populations.

TESS requires that a unique geographic location be associated with each sample. Because samples were de-identified and only the location of the contributing clinical practice was known, we created random locations for each sample within a standard deviation of 0.05 degrees longitude and 0.025 degrees latitude from the town associated with each sample using the tool provided by TESS. To ensure that nearest neighbor connections could not occur over the ocean, 23 dummy points, i.e. points at which sampling cannot occur, were added in the Atlantic Ocean along the shoreline. In

addition, the spatial network was altered to remove any remaining nearest neighbor connections that spanned the ocean. Geographic distances between each sample point were calculated using TESS. The program was then run for 10 permutations for K populations, from 2 to 8, with allowance for admixture. The mean deviance information criterion (DIC) was calculated across runs for each K population in order to choose the best fit among alternate models. The output from the 10 individual runs of the chosen K was downloaded into CLUMPP [32] which compiled them together. The resulting ancestry coefficients were displayed as a bar graph. An ancestry coefficient of 0.80 or greater for a single population was determined to be a member of that population. Any sample with a coefficient less than 0.80 for any single population was determined to have significant admixture from more than 1 source population. The ancestry coefficients were spatially interpolated onto a map of New England using R [33]. Fst estimates were calculated with Genepop on the web [34,35], PhiPT estimates were calculated using GenAlEx [36], and the Shannon Index of Diversity was calculated using PAST [37] on samples grouped by region. The Outline map of the northeastern United States was downloaded from dmaps.com (http://dmaps.com/carte.php?num_car=3895&lang=en)

Results

B. microti was typed from 234 specimens from 24 towns throughout New England during 2015 (Fig 1 and Table 1). Of these samples, 42 had multiple alleles in one locus and 6 had multiple alleles for more than 1 locus. The latter were excluded from the analysis because we were unable to accurately determine the haplotype necessary for analysis by Phyloviz. From the 228 samples used in the study, 113 unique haplotypes

were obtained. The samples were grouped by geographic region (Table 1) and the Shannon Index (H) was calculated for each region (Fig 2). The diversity for most regions ranged from 1.8-2.5 and was not significantly different from each other. However, the diversity from the New Jersey (NJ) samples was significantly lower ($H=0.9$, $p=0.02$) and the diversity from southeastern Massachusetts (SeMA) samples was significantly higher ($H=3.3$, $p<0.001$) than the rest. Population differentiation estimates, PhiPT, suggest isolation between some regions and almost none between others (Table 2). Samples from Nantucket (N) and Cape Cod (CC) have significant amounts of population differentiation between each other and each of the other geographic groups. (Table 2) In contrast, there is no evidence of any population differentiation between samples from NJ, Long Island (LI), Connecticut (CT) and Rhode Island (RI). Samples from SeMA and western Massachusetts (WMA) show moderate amounts of population differentiation between each other and those from NJ, LI, CT and RI (Table 2).

Table 1. Sites from which samples were collected and the number of haplotype identified from each site.

Region	No. Samples	No. per region	No. haplotypes
Boston (Bos)		19	11
Acton, MA	6		
Beverly, MA	1		
Boston, MA	3		

Norwood, MA	6		
Norwell, MA	1		
Southeastern MA (SEMA)		40	36
Fall River, MA	13		
Plymouth, MA	8		
New Bedford, MA	11		
Wareham, MA	4		
Dartmouth, MA	4		
Western MA (WMA)		7	7
Great Barrington, MA	1		
Pittsfield, MA	6		
Cape Cod (CC)		23	18
Falmouth, MA	6		
Hyannis, MA	17		
Nantucket (N)		16	13
Nantucket, MA	16		
Rhode Island (RI)		22	15
Providence, RI	4		

Wakefield. RI	18		
Connecticut (CT)		27	16
Putnam, CT	8		
Norwich, CT	19		
Long Island (LI)		46	19
Greenport, NY	2		
Hicksville, NY	15		
Riverhead, NY	12		
Southhampton, NY	17		
New Jersey (NJ)		15	5
Flemmington, NJ	15		
Total:	228		113

225

226

227 **Table 2: PhiPT estimates for *B. microti* from human patients by region.** [a]

228

Region	Boston	Cape Cod	Long Island	Nantucket	Connecticut	New Jersey	Rhode Island	SeMA[b]
Cape Cod	**0.51**							
Long Island	0.05	**0.64**						

Nantucket	**0.49**	**0.62**	**0.59**					
Connecticut	0.07	**0.62**	0.02	**0.54**				
New Jersey	0.07	**0.66**	0.01	**0.67**	0.04			
Rhode Island	0.03	**0.58**	0.04	**0.47**	0.01	0.05		
SeMA[b]	0.04	**0.33**	0.15	**0.33**	0.13	0.14	0.10	
WMA[c]	0.003	**0.51**	0.12	**0.48**	0.09	0.17	0.06	0.04

[a] Significant population structure PhiPT>0.25 are shown in bold. Moderate population structure PhiPT = 0.1-0.25 is underlined.

[b] Southeastern Massachusetts

[c] Western Massachusetts

The eBurst algorithm of Phyloviz grouped the samples into 3 main clusters consisting of samples primarily from Nantucket (N population), samples primarily from Cape Cod (CC population) and those from all other sites except for SEMA (Mainland population) (Fig 3). Samples from SEMA were divided among all 3 populations. About 6% of the samples remained unresolved and were not connected to any of the 3 major groups; the majority of these (>75%) were from SEMA and RI.

By plotting the mean DIC for K populations from 2-8, we determined that 3 populations, K=3, best fit the data from TESS (Fig 4). Ancestry coefficients from 10 runs for K=3 were estimated for each sample, and the CLUMPP algorithm was used to combine the data from all the runs (Fig 5). These coefficients indicate the probability of membership into each of the 3 populations and corresponded well with the results from

Phyloviz (Fig 3). Many samples that remained unresolved with Phyloviz showed significant amount of admixture, which would explain the inability of that algorithm to decisively place them into any single cluster (Table 3 and Fig 3 inside pink circle). However, the agreement between the two methods was not unanimous. There were a few samples that Phyloviz was unable to assign to a cluster that TESS had >85% certainty of inclusion into one of the populations (see unconnected bubbles inside larger circles Fig 3), as well as samples that Phyloviz connected to major populations that TESS could not determine to >85% probability (see bubbles with grey connections stretched to fit into pink circle Fig 3 and Table 3).

Table 3. Ancestry coefficients from TESS of samples that showed significant admixture.

Haplotype	Region	M	CC	N
197	N	0.59	0.01	0.39
272	WMA	0.44	0.18	0.38
286	WMA	0.69	0.19	0.12
315	SEMA	0.57	0.27	0.16
314	SEMA	0.51	0.37	0.13
327	SEMA	0.79	0.03	0.18
312	SEMA	0.67	0.17	0.16
307	SEMA	0.73	0.05	0.22
308	SEMA	0.62	0.15	0.23

232	RI	0.56	0.14	0.30
233	RI	0.43	0.17	0.40
289	Bos	0.56	0.14	0.30
290	Bos	0.43	0.17	0.40
331	SEMA	0.20	0.68	0.11
330	SEMA	0.41	0.49	0.10
310	SEMA	0.22	0.71	0.08
287	SEMA	0.22	0.71	0.06
235	RI	0.22	0.08	0.70
283	N	0.22	0.03	0.75
285	SEMA	0.20	0.09	0.70
234	SEMA	0.21	0.05	0.74

The geographically placed ancestry coefficients produced by TESS were spatially interpolated onto a map of New England (Fig 6). Haplotypes from the Nantucket population are primarily found on Nantucket. There has been some introduction into southeastern MA. The CC population also has limited scope: these haplotypes are found primarily on CC with some extending along the eastern coast of MA south of Boston. Contrary to the limited range of the N and CC populations, the mainland population dominates all of NJ, LI, CT, RI and MA, other than Cape Cod and Nantucket. It should be noted that this study did not include any data from Martha's Vineyard; so it may be that the predicted populations included in this figure are erroneous.

Each of the 3 populations has a dominant haplotype that is also the putative ancestral type (as determined by Phyloviz), type 4 for mainland, type 49 for Nantucket, and type 88 for Cape Cod (Table 4). Type 49 is present in 48% of Nantucket samples; Type 88 is found in 37% of Cape Cod samples, and type 4 ranges from 33% to 75% in the regions included in the mainland population (Figure 7). SEMA is the only region with a mixture of the dominant types; type 4 was detected in 22% of samples and type 88 detected in 7%. All other haplotypes in this study are detected only once or twice from any given region, with the exception of type 91 from LI which was found 4 times (8% of the observed haplotypes). Type 91 differs from the dominant type 4 by only the BMV1 locus (335bp instead of 340bp) of type 4.

Table 4. The microsatellite amplicon sizes of the 3 major haplotypes in base pairs.

Haplotype	Pop	BMV1	BMV2	BMV5	BMV8	BMV10	BMV13	BMV23	BMV20
4	M	340	405	317	241	305	396	243	695
49	N	340	405	317	241	305	520	248	713
88	CC	346	398	389	271	305	351	243	713

Discussion

Our analysis provides data to help reconstruct the tempo and mode of the processes that have led to the current epidemic population structure of *B. microti* in northeastern US. There are at least 3 distinct populations of *B. microti* in New England,

as we suggested previously [27,28] in analyses of ecological as well as clinical samples, with PhiPT ranging from 0.32-0.67 between them (Table 2). Each of the three populations has a single dominant haplotype that is found in at least 30% of the samples from each site and is the presumed ancestral strain; type 4 for mainland, type 49 for Nantucket and type 88 for CC. Southeastern MA is currently experiencing a natural experiment as the 3 populations, CC, N and M, are zoonotic in this area. The CC population is moving northward and westward along the eastern coast of MA, the N type is invading from the southern coast, and the mainland type is invading from the west. The genetic signature from all 3 populations can be clearly detected in clinical samples from this area, and significant admixture is occurring (Fig 6). For this reason, the diversity of *B. microti* from SEMA is significantly greater than that from all other regions in our study. Although we do not know when each of the *B. microti* populations were first introduced into SEMA, nor which one arrived first, type 4 is found more often in this area and the majority of samples harbor loci that originate from type 4. This dominance is clearly represented in the map of the ancestry coefficients from TESS, and suggests that type 4 parasites have some attribute that allows for greater amplification than do the other *B. microti* populations. It may be that type 4 parasites are more transmissible.

If the expansion of *B. microti* in New England was caused by individual founders "pulling" the population, we would have expected the diversity estimates from ancestral sites (Nantucket; Cape Cod; Long Island; [11], where cases have been diagnosed since the 1970s, to be greater than those from incipient sites with more recent emergence of cases. However, this was not the case; the diversity estimates of *B. microti* from the regions we sampled across the northeastern United States were not significantly different.

311 In fact, the diversity of *B. microti* from ancestral sites, such as Nantucket and Long
312 Island, were no greater than those from more newly established sites. Furthermore, the
313 diversity from coastal CT was not significantly different than that from northern CT
314 where babesiosis cases were first detected 15 years later. The maintenance of diversity
315 across New England supports the theory that expansion was the result of a "pushing"
316 population expansion, consistent with the stepping-stone hypothesis inferred by Walter
317 and colleagues [8]. Notably different, however, were samples from NJ; their diversity
318 was significantly less than those from every other site in our study; more than 70% of the
319 parasite samples comprised the dominant type 4. The lack of genetic diversity is
320 consistent with the New Jersey foci representing newly established populations that have
321 experienced significant founder effects. However, *B. microti*-infected ticks were
322 documented from northern New Jersey in the early 1990s [38] and human cases shortly
323 thereafter [39]. New Jersey became endemic for babesiosis at the same time as northern
324 CT and northern RI, but the diversity of *B. microti* from those states are similar to those
325 from the rest of the study populations. The biological basis for the limited diversity
326 found in New Jersey *B. microti* samples remains to be described.
327 Some patient samples may have been mistakenly assigned to location because we
328 used convenience samples that were de-identified other than for site of the contributing
329 clinical practice. We assumed that a case became exposed near the healthcare provider
330 who provided the sample to Imugen for analysis. Residents of any of our sites are likely
331 to travel within the northeast, and may vacation or visit in sites where risk is similar to
332 where they live. We are confident, for example, that two samples from our Nantucket
333 cohort acquired infection elsewhere. Each of these samples contained parasite haplotypes

that grouped with the mainland population. We have analyzed sufficient numbers of ecological samples from Nantucket Island and have never detected the other lineages [27]. Despite this clear example of mistaken assignment, the outcome of our analysis did not appear to be effected; TESS correctly concluded that Nantucket Island is dominated solely by the Nantucket population and the other sites by their respective parasite populations. Accordingly, we believe that our analysis is robust enough to be unaffected by other unknown errors in geographic assignment of samples and that our conclusions about the population structure of *B. microti* in the northeastern U.S. are reasonable.

It is also possible that focusing our analysis solely on parasites derived from presumably symptomatic patients (those presenting to a healthcare provider who in turn requested analysis of a sample for confirmation of a diagnosis) does not capture variation of all those that may be present in the enzootic cycle of the mainland parasites. There is as yet no published evidence that the diversity of *B. microti* infectious for humans differs from that in local mice or ticks, i.e., that only a subset of naturally occurring strains are zoonotic. However, such an argument would need to apply across all sites and we note that there is much variation evident in parasites from patients presenting to healthcare providers on Nantucket, Cape Cod, or Southeastern Massachusetts.

Significant differentiation (PhiPT >0.36) between each of the 3 populations implies that they have been isolated from each other and remain so. We have previously speculated that the microbial guild transmitted by *I. dammini* had been maintained in relict or refugial foci during glaciation [11]. Then too, postcolonial deforestation likely provided a fragmented landscape that only allowed for perpetuation of ticks and their hosts in small less-disturbed natural foci. The lack of differentiation among parasites

357 from the mainland sites, from central NJ westward to RI, appears to be inconsistent with
358 a scenario of multiple relict foci across the mainland northeastern landscape, with
359 coalescence of the isolated demes occurring as a result of amplification and expansion of
360 the foci as successional habitat increased over the last 100 years. In the 1990s, babesiosis
361 was documented from 3 distinct sites within the area where the mainland population
362 parasites have been detected, viz., Long Island, southeastern CT and central NJ. Each of
363 these foci was isolated from the others; few cases were identified in areas between them.
364 Ecological sampling, where it was done, supports the inference that *B. microti* was indeed
365 absent or very rare [7,13,15,16,18,40]. We expected to detect a distinct genetic signature
366 of multiple small isolated foci within parasites from the mainland lineages but there is
367 little differentiation among LI , CT, RI and NJ, and our analyses group these sites
368 together into a single population. In fact, the mainland haplotype, type 4, dominates
369 from NJ eastward through NY, CT and RI and northward towards Boston, creating an
370 epidemic population structure.

371 It may be that these sites were not isolated for sufficient time for genetic drift to
372 operate, thereby explaining the lack of differentiation among mainland parasites. It is
373 also possible that the epidemic population structure occurred purely by chance, i.e.
374 genetic drift has occurred as *B. microti* has expanded leading to an overabundance of a
375 single haplotype. Some alleles may reach a high frequency because of repeated founder
376 events [22], a process called genetic surfing [26]. We assume that our VNTR loci are
377 neutral or are not linked with loci under selection and thus the observed lack of variation
378 is not due to selective constraints. The alternative hypothesis for the lack of diversity
379 among mainland *B. microti* is that there were no refugial or relictual sites within

fragments of forest, and that the parasite populations have not actually been isolated from each other, allowing sufficient gene flow within the various sites comprising the mainland. However, the population structure of *I. dammini* suggests otherwise. A seminal study of the population structure of this vector tick and *B. burgdorferi* infecting them [41] sampled 12 sites in the northeast from Massachusetts to Virginia; 5 of these overlap with our area of study. Mitochondrial 16SrDNA haplotypes demonstrated that the New York-CT region may have contained refugial tick populations that served as a source for expansion of the range of *I. dammini*. Although tick populations that were sampled were structured, this was not observed for *B. burgdorferi*, although the borrelial genes that were analyzed were likely to have been influenced by balancing selection [41]. Additional studies are required to identify the relative contributions of selective and demographic processes that serve as the basis for biogeographic variation in northeastern populations of *B. microti*.

We believe the most likely scenario is that type 4 parasites have selectively swept across the mainland landscape, replacing and erasing historic genetic signatures of other lineages. Such a hypothesis is not without precedent with the microbial guild maintained by *I. persulcatus*-like ticks. The population structure of *B. afzelii* (an Eurasian agent of Lyme disease that appears restricted to rodent hosts) in Sweden is essentially clonal, which may be the result of the epidemic spread of a single genotype [42]. Across Europe, however, *B. afzelii* has significant population structure [43], similar to what we have found in this study. There are likely public health implications of a specific *B. microti* lineage that appears to be rapidly expanding its range.

403

404

405 ## Acknowledgements

406 Many clinicians and clinical practices submit diagnostic samples to
407 Imugen Inc for testing. Samples for this study were de-identified and thus we do not
408 know the identities of their submitters, but we thank them for their contribution to this
409 study.

410

411 ## References

412 1. Western KA, Benson GD, Gleason NN, Healy GR, Schultz MG. Babesiosis in a
413 Massachusetts Resident. N Engl J Med. 1970;283(16):854–6.

414 2. Steere AC, Malawista SE, Snydman DR, Shope RE, Andiman WA, Ross MR, et al.
415 An epidemic of oligoarticular arthritis in children and adults in three Connecticut
416 communities. Arthritis Rheum. 1977;20(1):7–17.

417 3. Spielman A, Wilson ML, Levine JF, Piesman J. Ecology of *Ixodes dammini*-borne
418 human babesiosis and Lyme disease. Annu Rev Entomol. 1985;30(1):439–460.

419 4. Steere AC, Malawista S. Cases of Lyme Disease in the United States: Locations
420 Correlated with Distribution of *Ixodes dammini*. Ann Intern Med. 1979 Nov
421 1;91(5):730–3.

422 5. Piesman J, Mather TN, Donahue J, Levine J, Campbell JD, Karakashian SJ, et al.
423 Comparative prevalence of *Babesia microti* and *Borrelia burgdorferi* in four

populations of *Ixodes dammini* in eastern Massachusetts. Acta Trop. 1986 Sep;43(3):263–70.

6. Lastavica CC, Wilson ML, Berardi VP, Spielman A, Deblinger RD. Rapid Emergence of a Focal Epidemic of Lyme Disease in Coastal Massachusetts. N Engl J Med. 1989 Jan 19;320(3):133–7.

7. Joseph JT, Roy SS, Shams N, Visintainer P, Nadelman RB, Hosur S, et al. Babesiosis in Lower Hudson Valley, New York, USA. Emerg Infect Dis. 2011 May;17(5):843–7.

8. Walter KS, Pepin KM, Webb CT, Gaff HD, Krause PJ, Pitzer VE, et al. Invasion of two tick-borne diseases across New England: harnessing human surveillance data to capture underlying ecological invasion processes. Proc R Soc B Biol Sci. 2016 Jun 15;283(1832):20160834.

9. Herwaldt BL, Montgomery S, Woodhall D, Bosserman E. Babesiosis Surveillance — 18 States, 2011. Morb Mortal Wkly Rep. 2012;61(27):505–9.

10. Brinkerhoff RJ, Gilliam WF, Gaines D. Lyme Disease, Virginia, USA, 2000–2011. Emerg Infect Dis. 2014 Oct;20(10):1661–8.

11. Telford III SR, Spielman A. Enzootic transmission of *Babesia microti.* In: Tick Borne Pathogens at the Host-Vector Interface. St. Paul, MN: University of Minnesota College of Agriculture; 1992. p. 259–64.

12. Anderson JF, Johnson RC, Magnarelli LA, Hyde FW. Involvement of birds in the epidemiology of the Lyme disease agent *Borrelia burgdorferi*. Infect Immun. 1986 Feb;51(2):394–6.

13. Anderson JF, Magnarelli LA, Kurz J. Intraerythrocytic Parasites in Rodent Populations of Connecticut: Babesia and Grahamella Species. J Parasitol. 1979 Aug 1;65(4):599–604.

14. Goethert HK, Telford III S. What is *Babesia microti*? Parasitology. 2003 Oct;127(04):301–309.

15. Rodgers SE, Mather TN. Human *Babesia microti* Incidence and *Ixodes scapularis* Distribution, Rhode Island, 1998–2004. Emerg Infect Dis. 2007 Apr;13(4):633–5.

16. Stafford KC, Williams SC, Magnarelli LA, Bharadwaj A, Ertel S-H, Nelson RS. Expansion of Zoonotic Babesiosis and Reported Human Cases, Connecticut, 2001-2010. J Med Entomol. 2014 Jan 1;51(1):245–52.

17. Xue L, Scoglio C, McVey DS, Boone R, Cohnstaedt LW. Two Introductions of Lyme Disease into Connecticut: A Geospatial Analysis of Human Cases from 1984 to 2012. Vector-Borne Zoonotic Dis. 2015 Sep;15(9):523–8.

18. Kogut SJ, Thill CD, Prusinski MA, Lee J-H, Backenson PB, Coleman JL, et al. *Babesia microti*, Upstate New York. Emerg Infect Dis. 2005 Mar;11(3):476–8.

19. Scott JD, Anderson JF, Durden LA. Widespread Dispersal of *Borrelia burgdorferi*–Infected Ticks Collected from Songbirds Across Canada. J Parasitol. 2011 Aug 24;98(1):49–59.

20. Mather TN, Nicholson MC, Hu R, Miller NJ. Entomological correlates of *Babesia microti* prevalence in an area where *Ixodes scapularis* (*Acari: Ixodidae*) is endemic. J Med Entomol. 1996;33(5):866–870.

21. Goodsman DW, Cooke B, Coltman DW, Lewis MA. The genetic signature of rapid range expansions: How dispersal, growth and invasion speed impact heterozygosity and allele surfing. Theor Popul Biol. 2014 Dec;98:1–10.

22. Edmonds CA, Lillie AS, Cavalli-Sforza LL. Mutations arising in the wave front of an expanding population. Proc Natl Acad Sci U S A. 2004;101(4):975–979.

23. Roques L, Garnier J, Hamel F, Klein EK. Allee effect promotes diversity in traveling waves of colonization. Proc Natl Acad Sci. 2012 Jun 5;109(23):8828–33.

24. Ogden NH, Mechai S, Margos G. Changing geographic ranges of ticks and tick-borne pathogens: drivers, mechanisms and consequences for pathogen diversity. Front Cell Infect Microbiol. 2013 ;3. Available from: http://www.frontiersin.org/Journal/10.3389/fcimb.2013.00046/full

25. Klopfstein S. The Fate of Mutations Surfing on the Wave of a Range Expansion. Mol Biol Evol. 2005 Dec 20;23(3):482–90.

480 26. Excoffier L, Foll M, Petit RJ. Genetic Consequences of Range Expansions. Annu
481 Rev Ecol Evol Syst. 2009 Dec;40(1):481–501.

482 27. Goethert HK, Telford SR. Not “out of Nantucket”: *Babesia microti* in southern New
483 England comprises at least two major populations. Parasit Vectors. 2014 Dec
484 10;7(1):546.

485 28. Lemieux JE, Tran AD, Freimark L, Schaffner SF, Goethert H, Andersen KG, et al.
486 A global map of genetic diversity in *Babesia microti* reveals strong population
487 structure and identifies variants associated with clinical relapse. Nat Microbiol.
488 2016 Jun 13;1(7):16079.

489 29. Rudzinska MA, Spielman A, Lewengrub S, Trager W, Piesman J. Sexuality in
490 piroplasms as revealed by electron microscopy in *Babesia microti*. Proc Natl Acad
491 Sci. 1983 May 1;80(10):2966–70.

492 30. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA.
493 PHYLOViZ: phylogenetic inference and data visualization for sequence based
494 typing methods. BMC Bioinformatics. 2012 May 8;13(1):87.

495 31. Caye K, Deist TM, Martins H, Michel O, François O. TESS3: fast inference of
496 spatial population structure and genome scans for selection. Mol Ecol Resour. 2016
497 Mar;16(2):540–8.

498 32. Jakobsson M, Rosenberg NA. CLUMPP: a cluster matching and permutation
499 program for dealing with label switching and multimodality in analysis of
500 population structure. Bioinformatics. 2007 Jul 15;23(14):1801–6.

501 33. François O. Running Structure-like Population Genetic Analyses with R. 2016 ;
502 Available from: http://www-timc.imag.fr/Olivier.Francois/tutoRstructure.pdf

503 34. Raymond M, Rousset F. GENEPOP (version 1.2): population genetics software for
504 exact tests and ecumenicism. J Hered. 1995;86(3):248–249.

505 35. Rousset F. genepop'007: a complete re-implementation of the genepop software for
506 Windows and Linux. Mol Ecol Resour. 2008 Jan;8(1):103–6.

507 36. Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in Excel. Population genetic
508 software for teaching and research--an update. Bioinformatics. 2012 Oct
509 1;28(19):2537–9.

510 37. Hammer Ø, Harper DAT, Ryan PD. PAST: Paleontological statistics software
511 package for education and data analysis. Palaeontol Electron. 2001;4(1):9.

512 38. Varde S, Beckley J, Schwartz I. Prevalence of tick-borne pathogens in *Ixodes*
513 *scapularis* in a rural New Jersey County. Emerg Infect Dis. 1998;4(1):97.

514 39. Herwaldt BL, McGovern PC, Gerwel MP, Easton RM, MacGregor RR. Endemic
515 babesiosis in another eastern state: New Jersey. Emerg Infect Dis. 2003;9(2):184.

516 40. Prusinski MA, Kokas JE, Hukey KT, Kogut SJ, Lee J, Backenson PB. Prevalence of
517 *Borrelia burgdorferi* (*Spirochaetales: Spirochaetaceae*), *Anaplasma*
518 *phagocytophilum* (*Rickettsiales: Anaplasmataceae*), and *Babesia microti*
519 (*Piroplasmida: Babesiidae*) in *Ixodes scapularis* (*Acari: Ixodidae*) Collected From

Recreational Lands in the Hudson Valley Region, New York State. J Med Entomol. 2014 Jan 1;51(1):226–36.

41. Qiu W-G, Dykhuizen DE, Acosta MS, Luft BJ. Geographic Uniformity of the Lyme Disease Spirochete (*Borrelia burgdorferi*) and Its Shared History With Tick Vector (*Ixodes scapularis*) in the Northeastern United States. Genetics. 2002 Mar 1;160(3):833–49.

42. Hellgren O, Andersson M, RåBerg L. The genetic structure of *Borrelia afzelii* varies with geographic but not ecological sampling scale: Genetic structure of *Borrelia afzelii*. J Evol Biol. 2011 Jan;24(1):159–67.

43. Vollmer SA, Feil EJ, Chu C-Y, Raper SL, Cao W-C, Kurtenbach K, et al. Spatial spread and demographic expansion of Lyme borreliosis spirochaetes in Eurasia. Infect Genet Evol. 2013 Mar;14:147–55.

Figure Legends

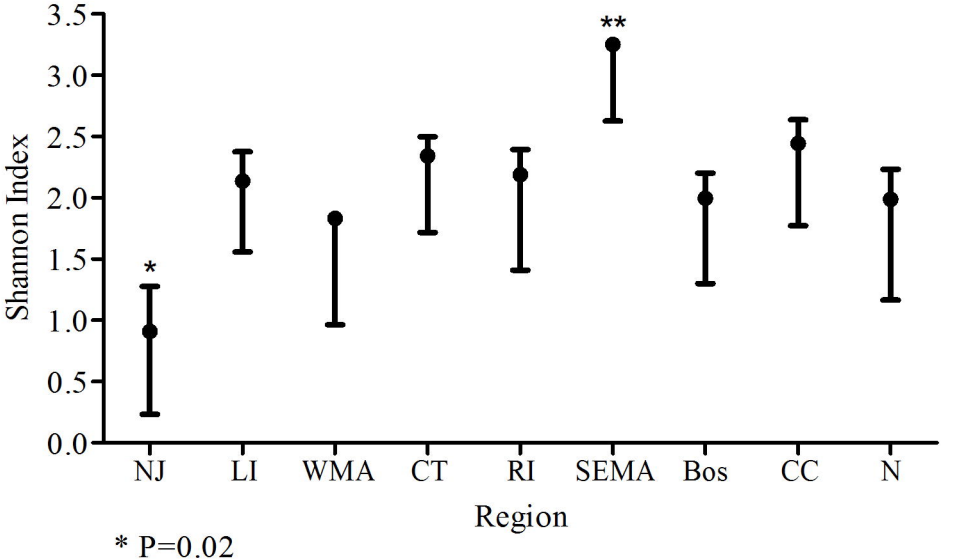
Figure 1. Map of the Northeastern United States labeled with the sites from which samples were collected.

Figure 2. Shannon's Index of diversity with standard error for *B. microti* haplotypes found in each region. New Jersey (NJ), Long Island (LI), Western Massachusetts (WMA), Connecticut (CT), Rhode Island (RI), southeastern Massachusetts (SEMA), Boston (Bos), Cape Cod (CC) and Nantucket (N).

542 **Figure 3. Cluster analysis of *B. microti* samples using Phyloviz.** Each small bubble
543 represents a unique haplotype. Bubbles are colored to correspond with the region from
544 which the sample originated. The size is not directly correlated with the number of
545 samples. Haplotypes that differ by a single locus are connected with a gray line. The
546 large circles correspond with the population groupings calculated by TESS; blue is the
547 Nantucket population, red is the Cape Cod population, green is mainland population and
548 pink are the haplotypes that showed significant admixture and could not be placed solely
549 in any of the 3 populations. Bubbles that are unconnected to the major groups are placed
550 in the larger circles according to the ancestry coefficients from TESS.
551
552 **Figure 4. Graph of the mean DIC**. Mean DIC was calculated from 10 individual
553 TESS runs for population size 2-8. Three populations, K=3, best fit the data.
554
555 **Figure 5. Ancestry coefficients from TESS for K=3 populations**. Geographic
556 distances between each sample point were calculated using TESS. Green corresponds to
557 the mainland population, red is Cape Cod and blue is Nantucket. Lines beneath the bar
558 chart indicate the source of the sample. Black= Nantucket, light blue= RI, dark blue=
559 CT, purple= WMA, pink= Bos, red=CC, yellow= SeMA, dark green = NJ and light
560 green= LI
561
562 **Figure 6. Geographic interpolation of the ancestry coefficients showing the**
563 **distribution of each population of *B. microti*.** Cluster 1(green) = mainland population,
564 cluster 2 (red) = Cape Cod population, and cluster 3 (Blue) = Nantucket population.

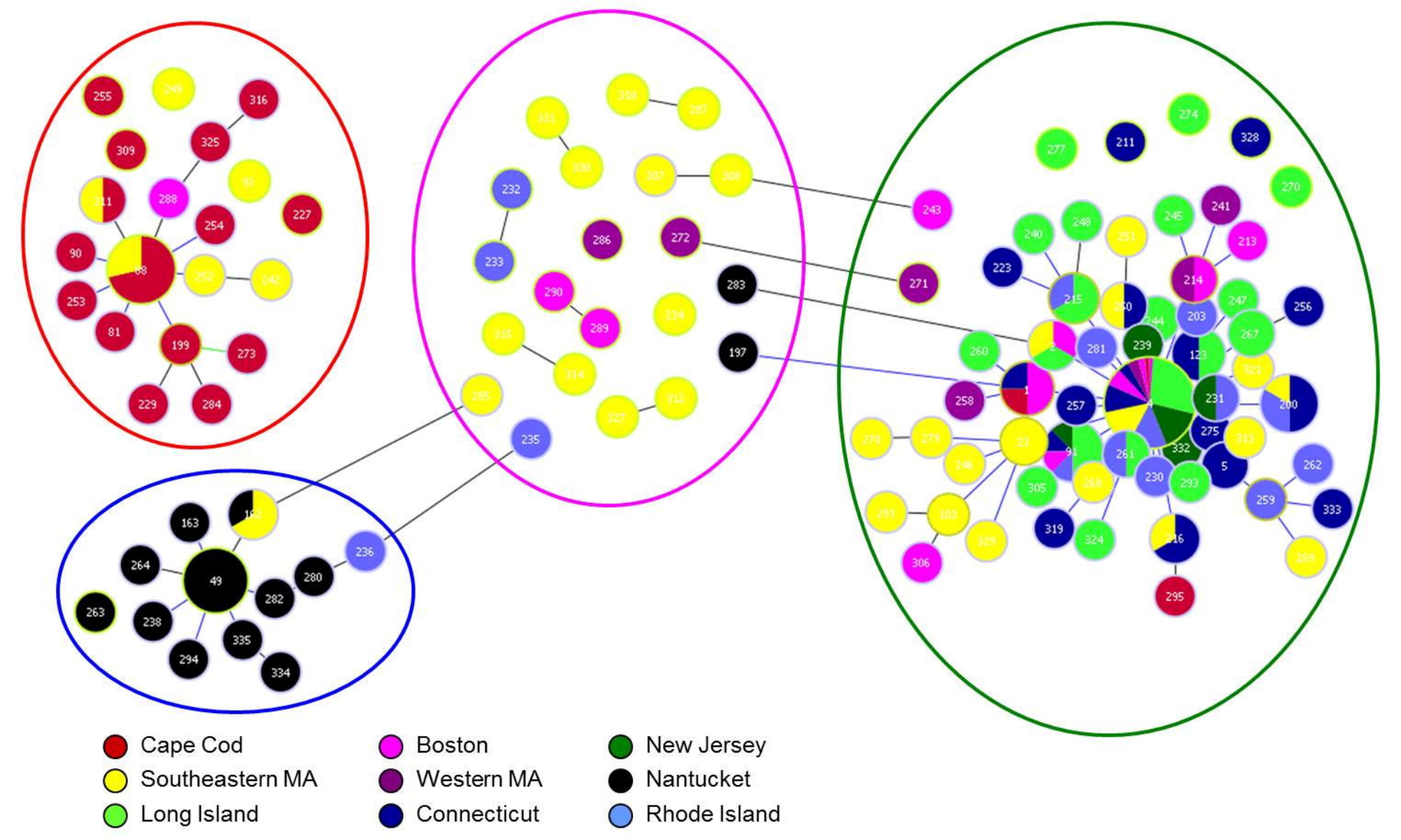
565 Areas with samples that have a high admixture coefficient, ie a high probability of
566 membership to that population, are shaded darker. Lighter shades indicate areas where
567 there the ancestry coefficients are lower, indicating areas where mixing is occurring. This
568 study did not include data from Martha's Vineyard; so the predicted populations on that
569 island may be erroneous.
570
571 **Figure 7. The percent of the total samples for each region for each of the main**
572 **haplotypes: type 4, type 88 and type 49.**
573

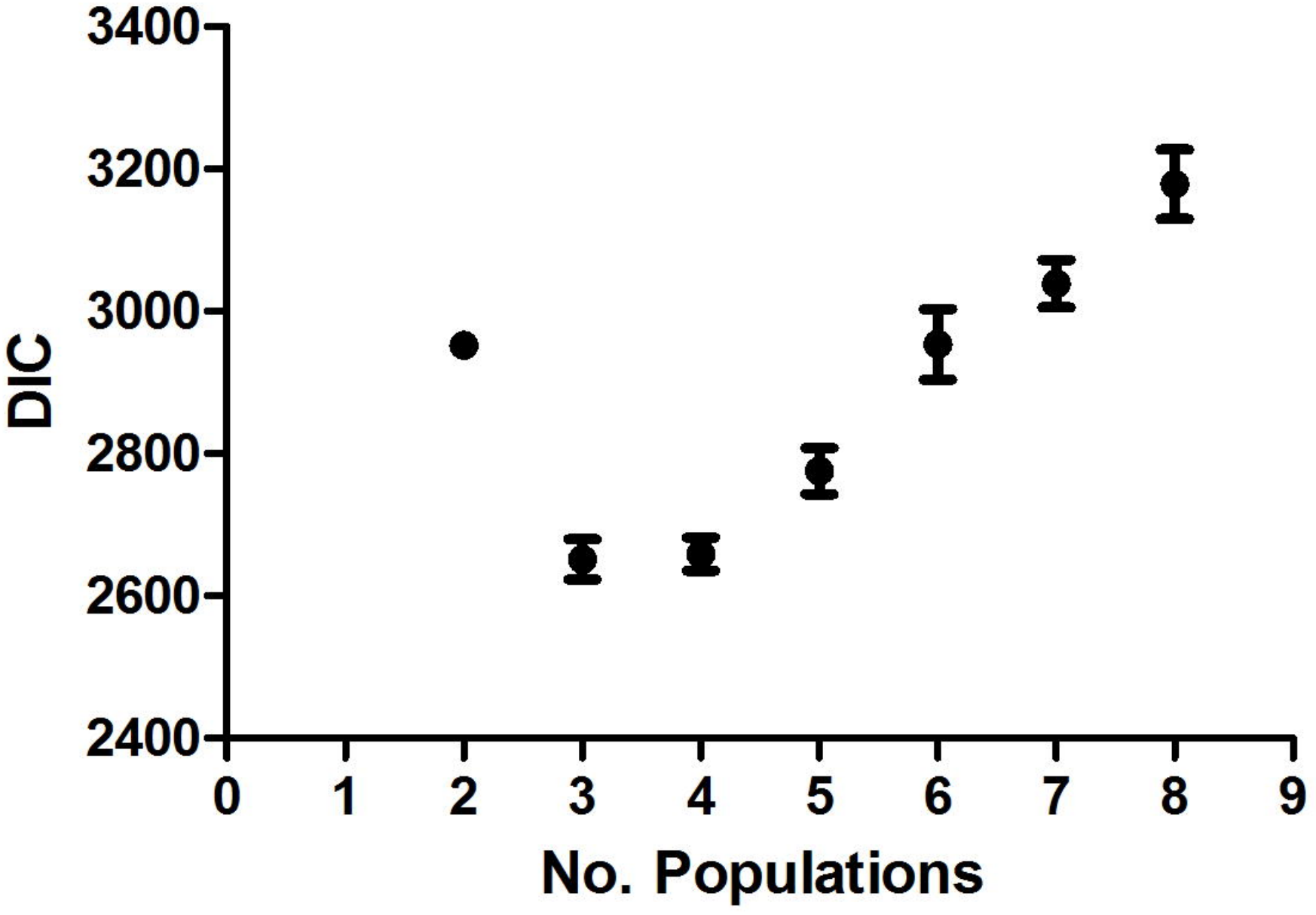

Pittsfield
Great Barrington
Beverly
Acton
Boston
Norwell
Norwood
Putnam
Providence
Fall River
Plymouth
Wareham
Hyannis
New Bedford
Falmouth
Dartmouth
Wakefield
Norwich
Nantucket
Greenport
Riverhead
Hicksville
South Hampton
Flemmington

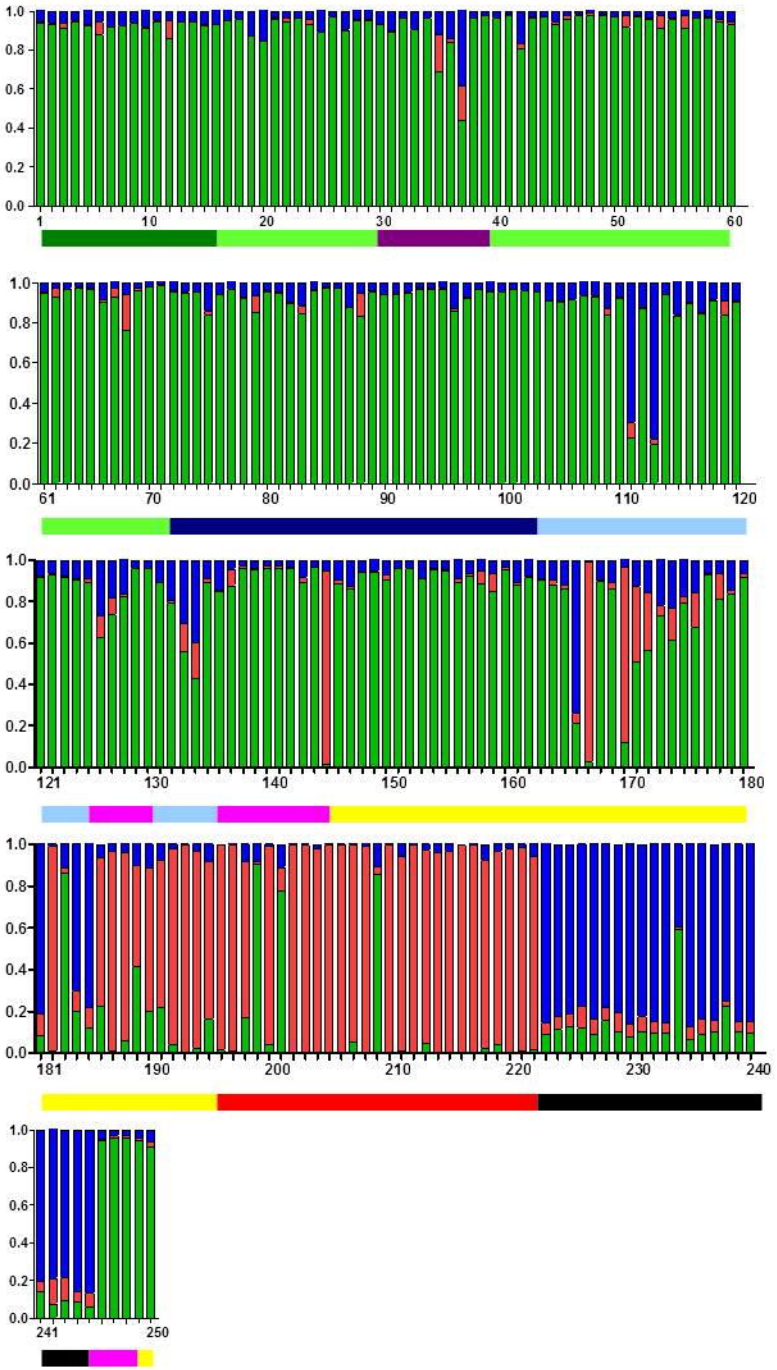

3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Shannon Index
*
**
NJ
LI
WMA
CT
RI
SEMA
Bos
CC
N
Region

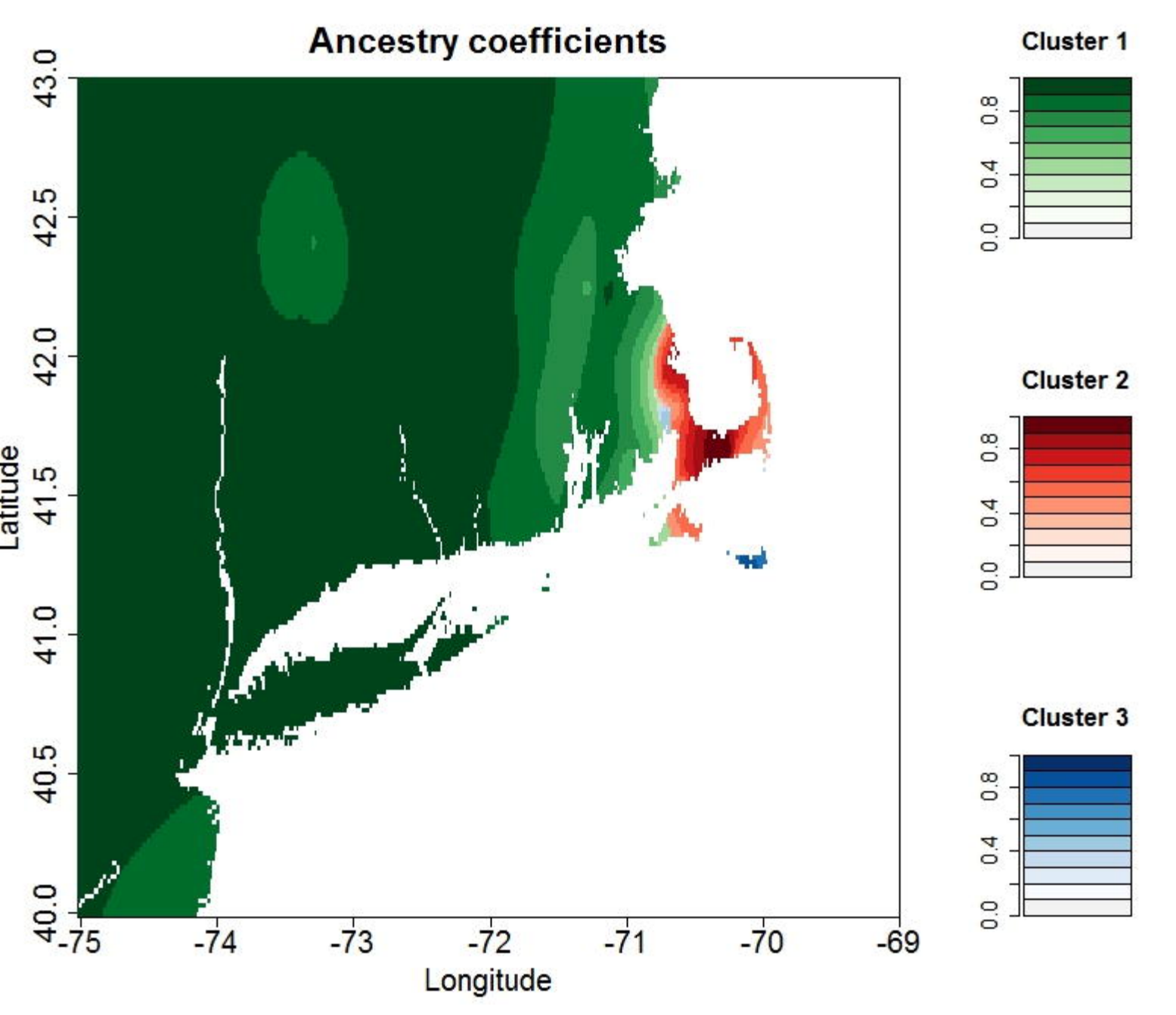
* P=0.02

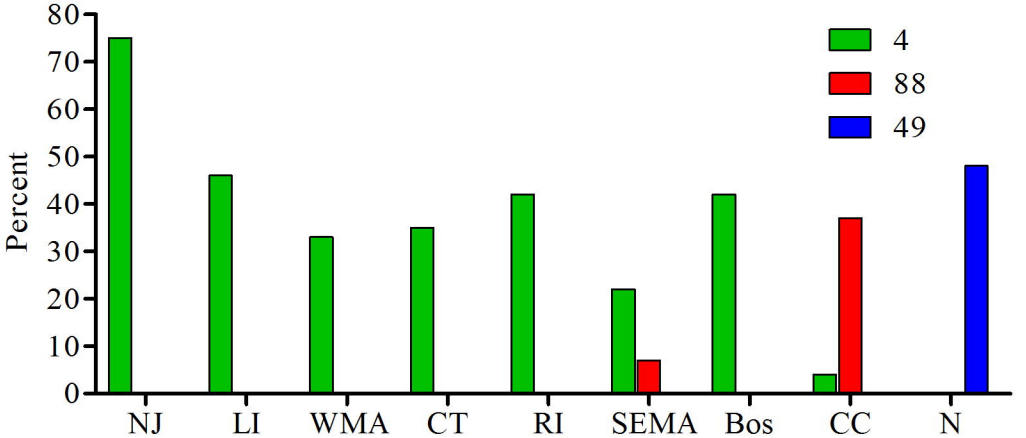
**P<0.001


Cape Cod
Southeastern MA
Long Island
Boston
Western MA
Connecticut
New Jersey
Nantucket
Rhode Island


3400
3200
3000
2800
2600
2400
DIC
0
1
2
3
4
5
6
7
8
9
No. Populations


1.0
0.8
0.6
0.4
0.2
0.0
1
10
20
30
40
50
60
61
70
80
90
100
110
120
121
130
140
150
160
170
180
181
190
200
210
220
230
240
241
250


Ancestry coefficients
Cluster 1
Cluster 2
Cluster 3
0.0
0.4
0.8
Latitude
43.0
42.5
42.0
41.5
41.0
40.5
40.0
Longitude
-75
-74
-73
-72
-71
-70
-69


80
70
60
50
40
30
20
10
0
Percent
4
88
49
NJ
LI
WMA
CT
RI
SEMA
Bos
CC
N