Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Qunna Li Date

Associations between Schistosomiasis Intermediate Host Snail's Recent Migration Rates and Geographical Distances in Sichuan, China

By

Qunna Li

MSPH

Emory University

Rollin School of Public Health

Department of Biostatistics and Bioinformatics

_____ [Thesis Advisor's signature]

Howard H. Chang, PhD

_____ [Reader's signature]

Yijuan Hu, PhD

Associations between Schistosomiasis Intermediate Host Snail's Recent Migration Rates and Geographical Distances in Sichuan, China

By

Qunna Li

B.M., Peking University, 2000

M.MS. Peking University, 2003

MSPH Emory University

Rollins School of Public Health

2014

Advisor: Howard H. Chang, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics and Bioinformatics

2014

Abstract

Associations between Schistosomiasis Intermediate Host Snail's Recent Migration Rates and Geographical Distances in Sichuan, China

By Qunna Li

Schistosomiasis is a parasitic disease caused by several species of trematode. More than 200 million people are infected worldwide. China had made great efforts to control schistosomiasis and 60% of endemic counties had achieved interruption of Schistosoma transmission**.** However, the disease has reemerged in previously controlled regions in Sichuan, China. *S. japonicum* is the causal agent of schistosomiasis in this region. And *O. hupensis* snail is the sole intermediate host for *S. japonicum*. Understanding how environmental factors influencing *O. hupensis* snail migration may offer insights on strategy for in controlling schistosomiasis transmission. In this study, a Bayesian multilocus genotyping method was first used to estimate *O. hupensis* snail recent migration rates between populations. Mixed models were then used to assess the association between geographic distances and snail recent migration rates. Four geographic distances were modeled, namely, Euclidean distance, incline distance, stream-only distance and land-use distance. 833 *O. hupensis* snails from 29 villages of Sichuan province, China were sampled. The estimated median migration rate between two villages was 0.0072(IQR: 0.0034). Also 12 (2.24%) out of 536 pairwise migration rates were greater than 10%. Overall, the association between snail recent migration rates and each of the four geographical distances were very similar. All models indicate that as the geographic distances increase, snail recent migration rates decreased. There was considerable village specific heterogeneity in migration rates, which indicates that village characteristics might play an important role on snail recent migration. Further study will need to address the effect of village specific characteristics on snail recent migration. Furthermore, it will be valuable to investigate the combined effects of geographic distances, hydrological connections between villages, land-use, as well as social network on *O. hupensis* snail's recent migration rates.

Associations between Schistosomiasis Intermediate Host Snail's Recent Migration Rates and Geographical Distances in Sichuan, China

By

Qunna Li

B.M., Peking University, 2000

M.MS. Peking University, 2003

MSPH Emory University

Rollins School of Public Health

2014

Advisor: Howard H. Chang, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics

2014

Acknowledgments

I would like to express my sincere appreciation to my thesis advisor, Dr. Howard H. Chang for his guidance, inspiration and continuous support throughout the course of this work. The extensive knowledge, cheerful heart and creative thinking have inspired me along the way. I would also like to thank my thesis reader, Dr. Yijuan Hu for her valuable time and insightful advice. I would also like to thank Dr. Justin Remais and Jessica H. Belle for their kindly sharing the data and explaining the project. I would like to give my special thanks to Dr. Edmund K. Waller for his continuous support, encouragement and opportunities for my career development. I am very grateful to my colleagues, Hilary Rosenthal, Wayne Harris, Chunzi Huang, Lauren Owens and all the members from Dr. Waller's lab, for their help when I was in classroom and encouragement while I was busy between classes and work. I would also like to thank all the faculties, staff members and fellow students from the Department of Biostatistics and bioinformatics for their help in my study and research. I thank Dr. Quyyumi and his group from Cardiology for their understanding and support. Last but not least, I would like to extend my deepest appreciation to my family for their unconditional love and support.

Table of Contents

List of tables

List of figures

Associations between Schistosomiasis Intermediate Host Snail's Recent Migration Rates and Geographical Distances in Sichuan, China

Introduction

Schistosomiasis is a parasitic disease caused by several species of trematode belonging to the genus Schistosoma. More than 200 million people are infected worldwide and it is considered one of the neglected tropical diseases[1]. China had made great efforts to control schistosomiasis where 5 of 12 endemic provinces and 60% of endemic counties in China had achieved interruption of Schistosoma transmission. Currently, China is aiming to eliminate the disease. If its program is successful, China's program may serve as a model for schistosomiasis control elsewhere. However, schistosomiasis has reemerged in previously controlled regions. In Sichuan province, schistosomiasis was identified in 8 of 46 counties that had met transmission control [2]. Little is known about the epidemiology of reemerging schistosomiasis, including how infections are distributed across human populations, as well as the distribution of intermediate host snails and other mammalian reservoirs. Snail surveys such as assessment of snail densities and detection of *Schistosoma japonicum* infected snails are one of the strategies of schistosomiasis surveillance in controlled area [3]. Therefore, studies that examine snail migration between different regions and the environmental determinant of snail migration may play an important role in understanding schistosomiasis epidemiology and control.

In order to better understand snail's role in the transmission of schistosomiasis, we first briefly summarize schistosoma's life cycle. There are three main species that infect humans including *Schistosoma haematobium, S. japonicum,* and *S. mansoni. S. japonicum* is the causal

agent of the schistosomiasis in East and Southeast Asia. Mainly three genera of snails serve as intermediate hosts of human Schistosoma parasites including *Biomphalaria, Bulinus* and *Oncomelania hupensis*. *Oncomelania hupensis* is the sole intermediate host for *S. japonicum*. Oncomelania live in water as well as out of water. Humid areas such as sluggish streams, swamp, poorly tilled rice fields, secondary and tertiary canals of irrigation systems and roadside ditches are most suitable for snail, but they can also survive periods of drought. Schistosoma infection occurs when skin comes in contact with contaminated freshwater in which the intermediate snails that carry the parasite are present. Freshwater becomes first contaminated by *Schistosoma* eggs when infected individuals' wastes excrete into the water. Then the eggs hatch to miracidium, which infect the appropriate species of snails. The miracidium then infect and reproduce many times asexually inside the snails until thousands of new form (cercariae) break out of the snail into the water. The cercariae can live for up 48 hours outside of the snail. Within that time they must penetrate the skin of a human being in order to continue their life cycle. People who come in contact with contaminated freshwater, typically when wading, swimming, bathing, or washing become infected. Over several weeks, the parasites migrate through host tissue and develop into adult worms inside the blood vessels of the body. Once mature, the worms mate and females produce eggs. Some of these eggs travel to the bladder or intestine and are passed into the urine or stool. Only about half of the eggs leave the body in the feces or urine; the rest remain embedded in the body where they cause damage to organs [4]

The *O. hupensis* snails are amphibious and largely inhabit the margins of irrigation canals, rice fields, and small streams where they are subject to advective transport as well as active dispersal. The vegetation in these sites serves to maintain suitable microenvironment, including temperature, humidity and food resources. They are seldom found in large rivers and fast-

flowing streams. Juveniles are submerged in water during early stages of development, while adults are often found above water line on vegetation and on moist soil[5, 6]. Adult snails have coping mechanisms to resist dry conditions. The snail can migrate both upstream and downstream directions. Lab observations of *O. hupensis* have confirmed that snails actively move against an elevation gradient when threatened by simulated flooding. And *O. hupensis* has been observed to spread via irrigation systems that link one village to next. Akullian et al. carried out a mass mark release (MMR) study in Gongqiao village, Sichuan province, China to direct estimate the distribution of distances dispersed by *O. hupensis* from a source population per unit time. They then use this MMR data to simulate the passive downstream diffusion of cercariae, the waterborne, human-infective parasite stage of *O. hupensis*. The simulation results suggested that *O. hupensis* migration can substantially increase the concentration of cercariae reaching downstream locations, relative to no snail dispersal, and putting downstream sites at increased risk of exposure to cercariae from upstream sources[7]. The above study indicates that *O. hupensis* snail migration plays a role of spreading schistosomiasis.

There is other evidence showing that vector dispersal plays an important role of infectious diseases control. Killeen et al. studied the migration of mosquitoes and the effectiveness of insecticide-treated nets on malaria. The efficacy measurements of insecticide-treated nets are much lower when assessed in the field than in the experimental settings. That is because adult mosquitoes are capable of travelling several kilometers. If the intervention of insecticide-treated nets are only implemented in an isolated village, the effectiveness of intervention can be underestimated, which indicates that high coverage of nets is necessary for maximizing the effectiveness of such powerful malaria-control tools[8].

The main objective of this thesis is to (1) estimate recent snail migration rates between 29 villages in Sichuan, China using microsatellite genotypes, and (2) examine associations between snail migration rates and functional variables that characterize ecological distance between villages.

There are variety ways to estimate gene flow between populations. One way is direct estimates of migration rates based on mark-recapture[9]. This method estimates the actual "instantaneous" migration rate of individual and can be time consuming and impractical for large populations that exchange small numbers of migrants because the expected number of recaptures is too low. Another way of estimate gene flow is indirect estimates by using genetic data such as restriction fragment length polymorphisms (RFLPs), microsatellite markers, single-nucleotide polymorphisms (SNPs) and DNA sequencing data. The indirect methods include the fixation-index F_{ST}, examining rare alleles, and using gene frequency distributions. In a review by Slatkin and Barton[10], the authors argued that the gene frequency distributions methods tend to yield biased estimates when relatively small numbers of locations are sampled. F_{ST} and rare alleles yield comparable estimates, but F_{ST} is preferred since it uses all of the gene-frequency data and thereby making it less sensitive to particular loci. However, both F_{ST} and rare alleles methods require the populations to reach genetic equilibrium within population and do not utilize all the information that molecular data provided in order to infer the levels of gene flow more accurately. The development of coalescent theory[11] , which traces the ancestral genealogy of a sample can be used in developing indirect estimates of gene flow with less restrictive models. Some new methods have been developed based on coalescent theory to accommodate recent population expansion, nonsymmetrical migration, and other complexities that are typical of

biological process[12-14]. However, even those new methods assume that population size and migration rates did not change over time, which may not be reasonable for snail migration study, since snail population and migration rates may change because of snail controlling efforts, land use and other social activities. However, methods based on coalescent theory have focused on the long-term effects of immigration on allele frequency distribution and do not measure contemporary migration rates, which are more relevant for developing disease control strategies. Recently, non-equilibrium approaches have been proposed to estimate recent migration from transient disequilibrium observed at individual multilocus genotypes of migrants or individuals who have recent immigrant ancestry[15, 16]. Bayesian inference has been used and focus was put on identifying individual migrants and their source populations but not on migration rates between populations. All above indirect approaches assume genotypes are in Hardy-Weinberg equilibrium within populations.

In this study, a Bayesian multilocus genotyping method proposed by Wilson and Rannala [17] was used to estimate recent migration rates between populations. This method assumes linkage equilibrium but relaxes Hardy-Weinberg equilibrium. It allows arbitrary genotype frequency distribution within populations by incorporating population-specific inbreeding coefficients. Migration rates among populations can be asymmetric but are constant over short periods of time (within the last one to three generations). In addition, it also assumes that migration rates are small (<1/3). Thus it imposes a constraint that nonmigrant proportions must be in the interval of 2/3 to1. The method also assumes that genetic drift and migration during the last few generations do not change subpopulation allele frequency. Moreover missing genotype data are allowed. The software BayesAss was developed using Markov chain Monte Carlo (MCMC) to implement this method. Faubet et al. [18] validated this method by using multi-allelic markers and scenarios with

varying number of populations. They found that if the assumptions of negligible change in allele frequencies due to migration and/or genetic drift over a few generations are not violated, and if the genetic differentiation is not too low ($F_{ST} \geq 0.05$) then the method can give fairly accurate estimate of migration rates even when they are close to the threshold (about 0.1). However, when the above assumptions are violated, accurate estimates are obtained only if migration rates are very low ($m = 0.01$) and genetic differentiation is high ($F_{ST} \geq 0.10$). BayesAss was used by Anderson[19] to estimate generalist rodents' migration rates between 18 forest patches. They demonstrated that the migration rates estimated by BayesAss (~20%) were smaller than the ones estimated by GENECLASS (~50%). The reason could be that BayesAss restricts the migration rates to be less than 1/3, and GENECLASS assumes that population reach Hardy-Weinberg equilibrium which was probably violated in that study. Bertrand et al. also used BayesAss to estimate migration rates in an island bird and found that the migration rates were small which agree with their expectations[20]. Finally, Howes et al. found that gene flow estimated by BayesAss was smaller than the ones estimated by F_{ST} for black ratsnake and Blanding's turtle. The authors claimed that the different estimate of migration rates was attributed to the assumptions applied to the methods[21].

As discussed above, snail migration may play an important role in Schistosomiasis transmission between regions (nodes). Identifying measures of distance between two nodes that are associated with snail migration can be used to predict snail migration from an endemic region develop appropriate monitoring and control strategies. Several previous studies have focused on measuring the degree of connection between nodes. The Euclidean model is a useful tool for the measurement of distance and simple connections, but it may not sufficiently represent social

distances that mediate transmission processes[22] . Furthermore, Euclidean model alone cannot measure epidemiological distance when environmental pathways lie on heterogeneous landscapes. Therefore, other models have been proposed, such as overland distance model and watershed model. Overland distance model is a Euclidean distance model corrected for the distance travelled when moving over sloped topography. Watershed model considers the distance along flow paths to stream. Remais et al.[23] used those models to test if Schistosomiasis re-emergent in one village is close to another re-emergent village in terms of Euclidean distance and watershed distance in Sichuan, China. It was found that the watershed model performs better than Euclidean model. For schistosomiasis, there is evidence that land use can change the disease transmission. Diama dam in Senegal serves as an example where the development of irrigation channels following the construction of the dam resulted in increased transmission of *S. haematobium* and the introduction of *S. mansoni*. On the contrary, the destruction of the Dabara dam in Madagascar and its associated irrigation network resulted in reductions of *S. mansoni* transmission in the absence of any systematic chemotherapy treatment[24]. Schistosomiasis transmission can attribute to the larval forms of the parasite dispersal or vector migration. Since water is essential for *S. japonicum*'s life cycle and indispensable for its intermediate host *O. hupensis* snails, hydrological connections between villages are important for transmission of schistosomiasis and migration of snails.

Modeling of *O. hupensis* snail is rare. There are some studies focused on *O. hupensis* snail seasonal abundance fluctuations and temperature change or precipitation[25, 26]. In this study, we will use mixed model to test the following hypotheses: (1) are estimated *O. hupensis* snail migration rates between villages associated with Euclidean distance? (2) Are *O. hupensis* snail

migration rates better associated with a modified distance measure that accounts for topography, streams, or land use? Random intercepts were used to account for village-specific propensity to act as a source, as well as its propensity to act as a target. As indicated by Remais[5], some villages might serve as "sinks" in the environmental network where they lie at the bottom of a watershed of numerous connected upstream villages. And some villages might reside on the upstream of hydrological network.

The thesis is organized as following. Firstly, descriptions of data collection including study sites, multilocus information and environmental function were presented. Secondly, migration rates estimation using BayesAss were described. Thirdly, statistical modeling between migration rates and environmental geographic distance was presented. Fourthly, presented the results. Finally, further steps were discussed in conclusions.

Methods

Data collection

The study was conducted in 32 villages within 3 counties in the Chuanbei region of Sichuan Province, People's Republic of China. Most villages within each county are connected by rivers. The stream flows from north to south. The characteristics of this region and selection of villages were described in detail elsewhere[23]. The villages lie on the mountainous area where intense, irrigated agricultural cultivation is the dominant landscape. Use of human waste (termed nightsoil) for crop fertilization is pervasive in this region which facilitates schistosomiasis transmission. According to Chinese Ministry of Health guidelines, Schistosoma japonicum has re-emerged in these areas that had previously attained transmission control[2]. Villages were not

selected at random, but focus was first placed on the availability of data on the presence of schistosomiasis infection in snails, acute human cases or infected children under 12 years old since control status was attained.

Snail and its locus information: From April 2008 to April 2010, snails were sampled from 29 out of those 32 villages. 833 snails were sampled and microsatellite genotyped. Microsatellite genotyping was done for 11 loci of *Oncomelania hupensis* snail DNA. Those 11 loci were OH12, OH150, OH157, OH211, OH212, OH235, OH47, OH573, OH68, OH73 and OH08.

Functional environmental models were used to obtain measures of distance between each pair of villages, accounting for potential environmental determinants of snail migrations. Specifically, each measure was developed based on a cost-path using the isoclines method. The method utilizes geographical information system (GIS) software to calculate a set of least-cost paths from village i to village j. This method works by drawing a line that represents the halfway point for all possible paths from village i to village j within the area of calculation. This line is known as the isoclines or allocation boundary. Then, the lowest-cost path value known as 'effective' geographic distance (EGD) value was calculated for the path that passes through a single point along isoclines on its way from village i to village j taking into account for topography and additional resistance to movement which may be associated with various aspects of the local environment. This process is then repeated for all points along the isoclines. Thus create a distribution of EGD values between village i and village j. The larger the EGD values, the more resistance for snails to migrate. EGD distribution for each pair of 32 villages (n=496) were calculated.

Cost path processing were used to calculate EGD values for nine different functional environmental models, namely, Euclidean model, topography model, incline model, land use

model, distance from watershed model, wetness model, stream only model, stream velocity model, and streams and channels model. Each model accounts for different environmental variables that will encourage or limit snail migration. For example, the Euclidean model gives equal resistance to all cells and generates straight line distances between villages. Whereas the incline model is a Euclidean distance corrected for the distance travelled when moving over sloped topography. Incline model, land use model, stream only model, stream velocity model and streams and channels model is bi-directional. Euclidean model, topography model, distance from watershed model and wetness model has no direction. Median and 25th percentile of all EGD values along isoclines between each pair of villages was used to predict snail migration between each pair of villages.

Recent migration rates estimation

Snail migration rates were estimated from microsatellite genotype data by using Bayesian inference[17]and they were bidirectional between villages. We assume there was no migration between county AN and JI, and between AN and ZH because of large distance and boundary barrier. The recent migration approach relaxes the Hardy-Weinberg equilibrium assumption within population. This method assumes that some proportion of an individual's alleles originate via a single migrant ancestor that arrived at the current (or past) generation. And populations only include non-immigrants, first-generation migrants and second-generation migrants. If the individual itself is a migrant, then 100% of its genome is of migrant origin. It also assumes that the loci are unlinked because individuals are sampled randomly. Assume the populations are large enough that there is negligible genetic drift over two, or three generations. The method also assumes that the total migration rate out of a village is 1/3.

Let matrix **m**=$\{m_{lq}\}$ be the migration rates between villages, where m_{lq} is the fraction of snails in population q that are migrants from village l. Let **M**=$\{M_h\}$, where M_h is the source of migrant ancestry for snail h. Let **t**=$\{t_h\}$, where t_h is the generation at which a migrant ancestor of snail h arrived. If t_h=0 then the snail has no migrant ancestry (non-migrant), if t_h=1 then the snail itself is the migrant, and if t_h=2 then the snail is a descendant of a migrant. Let **F**=$\{F_l\}$, where F_l is the inbreeding coefficient for population l and $-1 \le F_l \le 1$. Let **p**=$\{p_{ijl}\}$ be the population frequencies of marker alleles, where p_{ijl} is the frequency of allele i at locus j in population l. Let **X**=$\{X_{hj}\}$ be the multilocus genotypes observed at J marker loci in a random sample of n diploid snail, where X_{hj} is the genotype of snail h at locus j. Let **S**=$\{S_h\}$ be the population source for each sampled snail, wher S_h is the population that snail h was sampled from. Finally, n_l is the number of individuals sampled from the l^{th} population. **M**, **t, p, m and F** are unobserved variables, and **X** and **S** are the observed data. In our analysis, the parameter of interest is the matrix **m** of migration rates between populations.

The likelihood function of the data is $\Pr(\mathbf{X}|\mathbf{S};\mathbf{M},\mathbf{t},\mathbf{F},\mathbf{p})=\prod_{h=1}^{n}\prod_{j=1}^{J} Pr(X_{hj}|S_h;M_h,t_h,F,p)$ (1)

where

$$\Pr(X_{hj}|S_h;M_h,t_h,\mathbf{F},\mathbf{p})=$$

$$\begin{cases} \Phi(X_{hj},\mathrm{g}) & \text{if } M_h = S_h = \mathrm{g} \text{ and } t_h = 0 \\ 0 & \text{if } M_h \neq S_h = \mathrm{g} \text{ and } t_h = 0 \\ \Phi(X_{hj},\mathrm{r}) & \text{if } M_h = \mathrm{r}, S_h = \mathrm{g}, \text{ and } t_h = 1 \\ \left(1 - {}^{1}/_{2}^{\,t_h-2}\right)\Phi(X_{hj},\mathrm{g}) + \left({}^{1}/_{2}^{\,t_h-2}\right)\varphi(X_{hj},\mathrm{r},\mathrm{g}) & \text{if } M_h = \mathrm{r}, S_h = \mathrm{g}, \text{ and } t_h > 1 \end{cases}$$

$$\Phi(X_{hj},\mathrm{r})=\begin{cases} (1-F_r)\mathrm{p}_{\mathrm{ijr}}^2 + \mathrm{F_r}\mathrm{p}_{\mathrm{ijr}} & \text{if } X_{hj}(1) = X_{hj}(2) = \mathrm{i} \\ 2(1-F_r)\mathrm{p}_{\mathrm{ijr}}\,\mathrm{p}_{\mathrm{kjr}} & \text{if } X_{hj}(1) = \mathrm{i} \text{ and } X_{hj}(2) = \mathrm{k} \text{ for } \mathrm{i} \neq \mathrm{k} \end{cases}$$

$$\varphi(X_{hj}, \mathrm{r}, \mathrm{g}) =$$

$$\begin{cases} \mathrm{p_{ijr}p_{ijg}} & if\ X_{hj}(1) = X_{hj}(2) = \mathrm{i} \\ \mathrm{p_{ijr}p_{kjg} + p_{kjr}p_{ijg}} & if\ X_{hj}(1) = \mathrm{i\ and}\ X_{hj}(2) = \mathrm{k\ or}\ X_{hj}(2) = \mathrm{i\ and}\ X_{hj}(1) = \mathrm{k\ for\ i} \neq \mathrm{k} \end{cases}$$

and $X_{hj}(1)$ denotes the allele present on the maternal chromosome and $X_{hj}(2)$ denotes the allele present on the paternal chromosome. If $M_h = S_h$, then the individual h does not have immigrant ancestry and $t_h = 0$.

Prior distributions of parameters:

$$\Pr(\mathrm{M,t|m}) = \prod_{l=1}^{I} n_l! \left(\prod_{t=1}^{2} \prod_{q \neq l}^{I} \left(\frac{(2^{t-1} m_{lq})^{n_{lqt}}}{n_{lqt!}}\right)\right) \times \prod_{l=1}^{I} \left(\frac{m_{ll}^{n} l l 0}{n_{ll0!}}\right) \quad (2)$$

where m_{ll}=1-$\sum_{t=1}^{2} \sum_{q \neq l} 2^{t-1} m_{lq}$

n_{lqt}=$\sum_{h=1}^{n} \Im(M_h, t_h, S_h)$

$\Im(M_h, t_h, S_h)$=$\begin{cases} 1 & if\ M_h = l, S_h = q, and\ t_h = t \\ 0 & otherwise \end{cases}$

Uniform uninformative priors were used for $f_p(\boldsymbol{p})\ and\ f_m(\boldsymbol{m})$ subject to the constraints:

$\sum_{i=1}^{k_{lj}} p_{ijl}$=1, for all j =1,2,…, J and l=1,2,…,I, where klj is the total number of alleles at locus j in population l and $\sum_{q=1}^{I} m_{ql}$=1, for all l=1,2,…,I. For $f_F(\boldsymbol{F})$, a uniform prior on the interval (-1,1) is used.

Posterior distributions of parameters using Bayes' theorem is,

$$\mathrm{f(\mathbf{m,M,t,F,p,|X,S}) = \frac{\Pr(\boldsymbol{X|S};\boldsymbol{M},\boldsymbol{t},\boldsymbol{F},\boldsymbol{p}) \times \Pr(\boldsymbol{M},\boldsymbol{t}|\boldsymbol{m}) f_p(\boldsymbol{p}) f_m(\boldsymbol{m}) f_F(\boldsymbol{F})}{\Pr(\boldsymbol{X|S})}} \quad (3)$$

$\Pr(\boldsymbol{X|S})$ from equation (3) involves high-dimensional sum and intergral, so MCMC method is carried out to estimate the joint posterior probability density of equation (3)

The joint posterior distributions of the parameters **M**, **t, p, m and F** are estimated numerically using Markov chain Monte Carlo (MCMC) method using the software BayesAss. The estimated posterior distributions are used to make inferences.

To estimate the posterior distributions of parameters, the MCMC was run for a total of 10^7 iterations, discarding the first 2×10^6 as burn-in for AN county and a total of 1.5×10^7 iterations, discarding the first 5×10^6 as burn-in for JI+ZH county to allow the chain to reach stationarity. Samples were collected every 100 iterations to obtain posterior mean and posterior standard deviation of the migration rate.

Statistical analysis

Effective geographic distance and snail migration rates

We assume that snail migration between two villages is determined by the distance and stream between two villages. Topography and land use may also be predictive of snail migration. The EDG values of Euclidean model, topography model, incline model and wetness model are highly correlated. Stream only model, stream velocity model and streams and channels model are highly correlated. We test the hypotheses that distance between two villages (Euclidean model) can predict snail migration, and if this relationship can be enhanced by considering topography (Incline model). Other hypotheses we examined include if stream only and land use model can better predict snail migration.

Simple linear regression model (model 1) and random intercept models accounting for village specific migration propensity were fitted to predict snail migration (model 2-4). Since pair-wise migration rates were modeled as outcomes, an exchangeable correlation structure was used in random intercept models. Because migration rate lies between 0 and 1, we modeled its logit transformation which has an unrestricted range. For each distance measure, the following four models were fitted

$$\text{logit}(y_i)=\beta_0+\beta_1 x_i+\varepsilon_i \qquad (1)$$

$$\text{logit}(y_i) = \beta_0+\beta_1 x_i +b_{01i}+\varepsilon_i \quad (2)$$

$$\text{logit}(y_i) = \beta_0+\beta_1 x_i +b_{02i}+\varepsilon_i \quad (3)$$

$$\text{logit}(y_i) = \beta_0+\beta_1 x_i +b_{01i}+ b_{02i} +\varepsilon_i \quad (4)$$

where y_i=migration rate, x_i=ln(median or 25th percentile of EGD value) for different EGD models, b_{01i}=random effect of migration to village i, b_{02i}=random effect of migration from village i. The random effects are assumed to be Gaussian: $b_{01i}\sim N(0, \tau_1^2)$, $b_{02i}\sim N(0, \tau_2^2)$ and $\varepsilon_i\sim N(0, \sigma^2)$. In Model 4, the two random effects b_{01i} and b_{02i} are assumed to be independent. τ^2 represents variance of random effects and σ^2 represents residual variance.

Model AIC, marginal R^2 (m_R^2), conditional R^2 (c_R^2), leave-one-out cross-validation R^2 (cv_R^2) and root-mean-square error (cv_RMSE) were used to compare models.

First, we defined as AIC=-2log likelihood+2k, where k is the total number of parameters in the model. The lower AIC, the better the model. For mixed effect model, marginal R2 represents the variance explained by fixed effect (EGD values), and conditional R2 is interpreted as variance explained by both fixed and random effects. Cross validation R2 is defined as the R2 from fitting a simple linear regression between the set of observed migration rate from ith observation (i=1, 2……n) and the migration rate predicted from model without ith observation ($\hat{y}$i). To obtain the predicted migration rate for ith observation, we first obtained model parameter estimates using all the data except the except ith observation. We then calculated logit(yi) and obtained predicted migration rate $\hat{y}$i =exp(logit(yi))/(1+exp(logit(yi))). cv_RMSE represents the difference between predicted migration rate from model without ith observation and migration rate from ith observation. cv_RMSE=$\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}i-y(i))^2}{n}}$. We also estimated baseline migration rate for each model, $\hat{p}1$=exp(β0c)/(1+exp(β0c)) when xi centered at median of the covariate (log(median)or log(25th percentile)).

Sensitivity analysis

To assess the sensitivity of estimating the migration rates using BayesAss, we considered different seeds, iterations and mixing parameters for MCMC. In this report, migration rates were estimated by two separate MCMC processing. One for AN county and the other for JI and ZH combined due to the large distance and boundary barrier between those two regions. In order to assess the sensitivity, we also run a MCMC processing combining all counties together as well as all counties separately.

We also considered logarithm of EGD median, logarithm of EGD 25th percentile and logarithm of EGD minimum are used as covariates to predict migration rates separately. Finally, the impacts of outlying migration rate were also assessed.

Migration rates were estimated by BayesAss 3.0. Data analysis was performed by using SAS 9.3 and RStudio.

Results

Descriptive analyses

Snail data were available in 29 out of 32 villages within 3 counties. 833 *Oncomelania hupensis* snails were sampled (table 4). Microsatellite genotypes of the 11 loci were summarized in supplement table 2. 536 pairs of snail migration path between villages were estimated, most snail migration rates were small. The maximum snail migration rate was 0.26 and minimum migration rate was 0.0047. The median migration rate was 0.0072 (IQR: 0.0034). Also 12 (2.24%) out of 536 migration rates were greater than 10% (table 1). With respect to the study region, we found that the direction of large migration rates were unidirectional and, mainly from

up-stream to down-stream. Non-migrant rates were also estimated from BayesAss with a maximum 0.92 and minimum 0.67. The median non-migrant rate was 0.77 (IQR: 0.18). Table 2 shows the 29 non-migrant rates.

The majority of the migration rates estimated from the MCMC processing combining all counties together were similar to those estimated from the method used in the paper (AN county separately, JI and ZH combined), but the migration rates outliers (≥0.1) were less common. This is probably due to the fact that Bayesass assumes the non-migrant rates to be greater than or equal to 2/3. With greater numbers of villages, the migration rates will get smaller as the number of villages increases because of the constraint that they need to add up to 1. Most of the migration rates estimated from the MCMC processing by running all three counties separately agree with the results from the method used in the report. Those sensitivity analyses indicate the Bayesian inference using BayesAss is robust. Considering that the boundary between JI and ZH is not clear and the distance of villages between those two counties is not necessary further than the villages within those counties, our final analyses were based on migration rates estimated from combining those two villages together.

Median and 25th percentile of EGD distribution between each pair of villages ($\binom{32}{2}$=496) was summarized for each functional environmental model. Since incline model, land use model, stream-only model, stream-velocity model and streams, and channels model are bi-directional, the total sample size of EGD distribution is 992 (supplement table 3). The distributions of median and 25th percentile of EGD are right-skewed for all nine environmental models. Consequently, a natural log transformation of median and 25th percentile was used in the analysis of relationship between EGD models and snail migration. From the scatter plots between logit of migration rates and logarithm of 25th percentile of cost function, we can see that as 25th

percentile of cost function increase, migration rates decrease (figure 1). The scatter plots between logit of migration rates and median of cost function are very similar to the logit of migration rates and 25th percentile of cost function, so only the logit of migration rates and 25th percentile of cost function are shown. High-pair wise correlations between the nine environmental models. Specifically, Euclidean model, topography model, incline model and wetness model are all highly correlated ($r>0.9$) for both median and 25th percentile. Stream only model, stream velocity model and streams and channels model are all highly correlated ($r>0.9$) for both median and 25th percentile (Supplement figure 1,2, supplement table 4,5). Scatter plots between logit of migration rates and EGD original scale, and between logit of migration rates and logarithm tranformation of EGD for different models all show that as EGD increase, the logit of migration rates decrease. This pattern was consistent between logarithm of EGD median, logarithm of EGD 25th percentile and logarithm of EGD minimum.

Modeling results

Logarithm of medians and 25th percentiles of EGD distribution from four environmental models- Euclidean model, incline model, stream only model and land use model were used as covariate to predict snail migration rates. Overall, the four environmental models gave similar results (table3). Results from Incline model were almost identical to the results from Euclidean model. All models indicate that EGD was associated with snail migration rates. When cost distance increases, snail migration rates decrease (all $\exp(\beta)<1$). For example, for the model with two random intercepts using the log 25th as the covariate, one IQR increase in log25th EGD is associated with an odds ratio of migration of 0.497 (95%CI: 0.453-0.545). There was considerable village specific heterogeneity in migration rates. When we include both terms of migration to specific village and migration from specific village, model gives best AIC and R^2.

Moreover, the random effect estimates (and the associated heterogeneity variance) for "from village" are consistently larger than the random effect estimates for "to village". This indicates that source villages explain more variation in the high migration rates than the presence of target villages. There is also evidence that the log of 25th percentile of EGD value predicts migration rates better than log of median of EGD value. The estimated migration rate for a typical village ($\hat{p}_1$) was less than 0.01 for every model when log of EGD at the median value. Although environmental EGD values were negatively associated with migration rates, there were migration rate outliers that can't be explained by EGD values alone. Supplement figure 3 shows migration rates can be predicted by EGD fairly well when migration rates are small, but for large value of migration rates, EGD predict poorly. We conducted a sensitivity analysis for the incline random effect "to and from" model using log25th as covariate by excluding outliers (migration rate $\geq$0.1).The resulted model fits better (significantly smaller AIC and CV_RMSE, larger cv_R^2) compare to the same model with outliers (table 3). However, the negative association between migration rates and EGD remains.

Discussion

Most snail migration rates were small. There were only 12 out of 536 pairwise migration rates that were greater than 10% and a few of which, that were greater than 20%. BayesAss assumes the sum of all migration rates to be less than 1/3, resulting in non-migrant rate to be between 2/3 and 1. Most migration rates between villages were small and the total migration rates from a specific village were less than 1/3. However, some non-migrant rates were very close to 2/3 indicating that we might underestimate migration rates for those villages because of the constraint. We hope future study can develop a statistical model which can relax the

constraint of total migration rate being less than 1/3. Nevertheless, we did a series of sensitivity analysis by estimating migration rates with three counties combined or separately- the results indicated that Bayesian inference using BayesAss was robust. Also Faubet et al.[18] validated BayesAss by using multi-allelic markers and scenarios with varying number of populations. They found that if the assumptions of negligible change in allele frequencies due to migration and/or genetic drift over a few generations are not violated, and if the genetic differentiation is not too low ($F_{ST} \geq 0.05$) then the method can give fairly accurate estimate of migration rates even when they are close to the threshold (about 0.1). The study collected snail data between 2008 and 2010. If we can obtain yearly multiloci data, we will be able to estimate allele frequencies for each year and assess if there were changes in allele frequencies during the study period given that snails can survive for about a year. It is also possible to estimate F_{ST} from multiloci data [27], which enables us to assess if the genetic differentiation is indeed large enough for accurate estimation of migration rates by BayesAss.

From random effect models between logit of snail migration rates and cost function, we observed very high heterogeneity in the village random effect both for source and target. The two random intercept models suggested that about 50% of the variation is attributed to the individual villages. This implies village characteristics may help explain between-village snail migration rates in addition to ecological distance. Future analysis may consider village characteristics, such as up-stream or down-stream of water flow, agricultural type and social connections in predicting migration rates. Furthermore, for the random intercept models, we assume random effects are independent. We might want to allow the random effects to be spatially correlated. Since villages are hydrologically connected, if stream plays an important role on snail migration, it is likely that villages from same hydrological connection are more correlated than villages from further apart.

Linear models and random effect models using the Euclidean distance, incline distance, stream only distance and land use distance all suggested that as the environmental distance increase, snail migration rates decrease. Guo et al.[28] used a geographic information and remote sensing based model to predict *O. hupensis* snail's habitats in the Poyang Lake Area in China. Poyang Lake is a known habitat for *O. hupensis* snails. Their study indicated that the snail density decreased as the distance from the centroid of the lake increased. Clearly, environmental distance is a good predictor for snail migration. However, when we inspect scatter plots (figure 1) and model fit (supplement figure 3), the relationship between logit of snail migration rates and environmental distance is not perfectly linear, and model fit improves as we remove migration rate outliers. We should consider non-linear effect of cost distance such as polynomial splines or piecewise regression in the future study.

For those villages with outlying snail migration rates, it would be of interest to investigate the infectious status of snails, individuals, and animal host as well as historical epidemiological data. If the destination villages attained schistosomiasis transmission control or transmission interruption in the past while source villages did not, and schistosomiasis re-emerged in the destination villages, then it indicates that snail migration plays a significant role on schistosomiasis transmission. This implies that isolated controlling efforts for Schistosomiasis may not be as effective and avmore systematic regional strategy should be considered. This has been observed in the case of controlling mosquito-transmitted malaria. Mosquitoes are the vectors for transmitting malaria and they are capable of travelling several kilometers. Insecticide-treated nets is an effective way of controlling malaria in the experimental settings, but field experiences showed that if the insecticide-treated nets only covered an isolated village, the treatment was not as effective as expected[8].

From EGD and migration rate model fit (table 3), we can see that overall Euclidean model, incline model, stream-only model and land-use model gave similar results. Results from incline model were almost identical to the results from Euclidean model. The EGD distance values from Euclidean model is highly correlated with incline model (slope close to 1, supplement figure 1 &2) indicating that the topography features used by the incline model may provide limited additional information beyond Euclidean distance for villages located in mountainous areas. Furthermore, Akullian et al.[7] suggested that although *O. hupensis* snails are capable of moving up-stream and down-stream, snails dispersed further, on average, down-stream than up-stream. And even slow-moving flows in the small irrigation channels were observed to facilitate snail dispersal in the down-stream direction. Thus accounting for flow direction of stream, river and irrigation system would improve the model. Although in this study, geographic distances from incline model, land-use model, stream-only model, stream-velocity model, and streams and channels model are bi-directional, the EGD distance values are not so different for one direction from the other. It will be useful to include a covariate for the direction in the model. Finally, we considered the effect of each geographical model one-at-a-time. Future study should consider multiple environmental determinants along with social network in the same model.

Conclusion

Overall, the *Oncomelania hupensis* snail migration rates were small. We observed very high heterogeneity in the village random effects both for source and target. Euclidean model, incline model, stream-only model and land-use model gave similar results where as EGD distance increased, snail migration rates decreased. Future study may consider multiple environmental determinants, village-specific characteristic and social network in the same model.

Tables

Table 1. Large migration rates computed from BayesAss

From	To	Estimate (SE)
AN_BG1	AN_BG2	0.2543(0.0297)
AN_JQ3	AN_JQ1	0.2616(0.0232)
JI_SB1	JI_GH3	0.1265(0.0280)
JI_DS4	JI_DS5	0.1242(0.0243)
ZH_XZ4	JI_LJ3	0.1432(0.0265)
JI_XQ5	JI_XQ3	0.1963(0.0187)
JI_XQ5	JI_XQ7	0.1069(0.0275)
ZH_HQ2	ZH_FX7	0.1204(0.0290)
ZH_HQ4	ZH_HQ1	0.1718(0.0242)
ZH_XD3	ZH-ZH5	0.1552(0.0223)
ZH_HQ2	ZH_WX3	0.1106(0.0237)
ZH_XD3	ZH_XL5	0.1118(0.0219)

Table 2. Non-migration rates computed from BayesAss

Village	Estimated non-migrant rate (SE)
AN-320210_JG10	0.7988(0.0257)
AN-330302_BG2	0.6739(0.0071)
AN-330401_JQ1	0.6762(0.0093)
AN-330403_JQ3	0.8602(0.0281)
AN-350603_HT3	0.9176(0.0240)
AN-330301_BG1	0.9211(0.0235)
JI-110103_GH_B	0.6764(0.0094)
JI-110107_GH_AB	0.8794(0.0225)
JI-110202_GQ_AB	0.8680(0.0228)
JI-110203_GQ_B	0.6760(0.009)
JI-110501_SB_AB	0.7989(0.0254)
JI-120704_DS_AB	0.8743(0.0218)
JI-120705_DS_A	0.6765(0.0095)
JI-120805_DC_AB	0.7946(0.0247)
JI-130408_GH_B	0.6765(0.0096)
JI-130903_LJ_AB	0.6743(0.0075)
JI-141003_XQ_AB	0.673(0.0061)
JI-141005_XQ_AB	0.8568(0.0231)
JI-141007_XQ_B	0.6763(0.0094)
ZH-210107_FX_B	0.6767(0.0098)
ZH-210301_HQ_AB	0.6742(0.0073)
ZH-210302_HQ_AB	0.7792(0.0241)
ZH-210304_HQ_AB	0.8373(0.0247)

ZH-220403_XD_AB	0.8798(0.0226)
ZH-220406_XD_AB	0.7739(0.0239)
ZH-220504_XZ_A	0.7945(0.0271)
ZH-220605-ZG_AB	0.7108(0.0165)
ZH-230703_WX_AB	0.6737(0.0069)
ZH-240805_XL_AB	0.7423(0.0204)

Table 3. Model fit for snail migration rates and environmental models

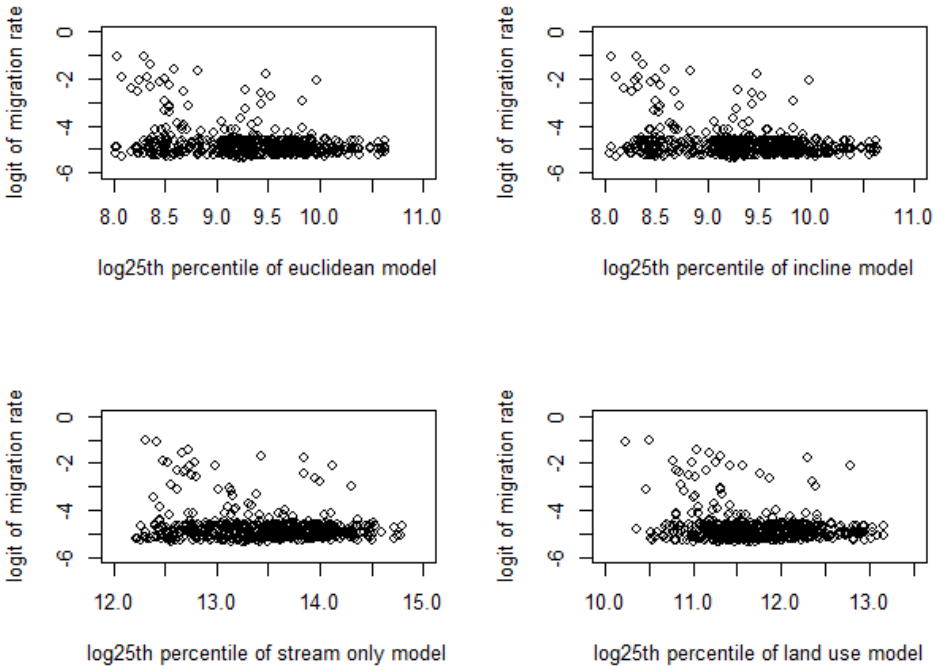
Models	Covariate	Statistical models	AIC	m_R^2	c_R^2	cv_R^2	CV-RMSE	IQR of covariate	IQR *exp($\hat{\beta}$)	95%CI of IQR*exp($\hat{\beta}$)	$\hat{p}_1$	95%CI of $\hat{p}_1$	τ_1^2 to	τ_2^2 from	σ^2
Incline	logmedian	Linear model	1027.4	0.062		0.041	0.025	0.630	0.424	(0.372,0.482)	0.0074	(0.0069, 0.0079)			
model		Random intercept to	975.2	0.058	0.310	0.038	0.025		0.424	(0.372,0.483)	0.0080	(0.0069, 0.0092)	0.116		0.317
		Random intercept from	968	0.058	0.430	0.118	0.024		0.412	(0.362,0.469)	0.0080	(0.0067, 0.0095)		0.196	0.305
		Random intercept to and from	909.9	0.089	0.530	0.103	0.024		0.372	(0.324,0.427)	0.0077	(0.0064, 0.0093)	0.063	0.172	0.249
	log25th	Linear model	1017.9	0.078		0.061	0.025	0.738	0.536	(0.489,0.587)	0.0073	(0.0068, 0.0078)			
		Random intercept to	967.1	0.071	0.310	0.049	0.025		0.539	(0.492,0.591)	0.0079	(0.0069,0.0091)	0.111		0.312
		Random intercept from	957.8	0.070	0.440	0.136	0.024		0.527	(0.481,0.577)	0.0079	(0.0067, 0.0094)		0.194	0.298
		Random intercept to and from	899.3	0.097	0.540	0.115	0.024		0.497	(0.453,0.545)	0.0077	(0.0064, 0.0093)	0.062	0.172	0.244
Stream	logmedian	Linear model	1051.8	0.018		0.013	0.025	0.654	0.528	(0.465,0.600)	0.0082	(0.0078, 0.0087)			
only model		Random intercept to	988.7	0.036	0.330	0.028	0.025		0.480	(0.421,0.549)	0.0088	(0.0076, 0.0103)	0.141		0.322
		Random intercept from	991.8	0.025	0.410	0.096	0.024		0.497	(0.435,0.568)	0.0090	(0.0075, 0.0107)		0.211	0.318
		Random intercept to and from	920	0.090	0.590	0.100	0.024		0.386	(0.331,0.449)	0.0086	(0.0071, 0.0105)	0.087	0.178	0.251
	log25th	Linear model	1044.1	0.032		0.028	0.025	0.873	0.709	(0.645,0.779)	0.0081	(0.0076, 0.0086)			
		Random intercept to	981.5	0.048	0.330	0.036	0.025		0.672	(0.612,0.739)	0.0087	(0.0075, 0.0100)	0.134		0.318
		Random intercept from	982.6	0.039	0.420	0.113	0.024		0.678	(0.616,0.747)	0.0088	(0.0074, 0.0105)		0.207	0.312
		Random intercept to and from	909.2	0.096	0.560	0.111	0.024		0.585	(0.528,0.648)	0.0084	(0.0070, 0.0102)	0.082	0.178	0.245
Euclidean	logmedian	Linear model	1027.5	0.062		0.041	0.025	0.626	0.421	(0.370,0.479)	0.0074	(0.0069, 0.0079)			
model		Random intercept to	975.2	0.058	0.310	0.038	0.025		0.421	(0.369,0.480)	0.0080	(0.0069, 0.0092)	0.116		0.317

Models	Covariate	Statistical models	AIC	m_R^2	c_R^2	cv_R^2	CV-RMSE	IQR of covariate	IQR *exp($\hat{\beta}$)	95%CI of IQR*exp($\hat{\beta}$)	$\hat{p}_1$	95%CI of $\hat{p}_1$	τ_1^2 to	τ_2^2 from	σ^2
		Random intercept from	967.8	0.058	0.428	0.119	0.024		0.409	(0.359,0.466)	0.0080	(0.0067, 0.0095)		0.197	0.305
		Random intercept to	909.7	0.089	0.532	0.104	0.024		0.369	(0.321,0.424)	0.0077	(0.0064, 0.0093)	0.063	0.173	0.249
		and from													
	log25th	Linear model	1018.0	0.078		0.061	0.025	0.726	0.527	(0.481,0.578)	0.0073	(0.0069, 0.0078)			
		Random intercept to	967.3	0.071	0.314	0.049	0.025		0.531	(0.485,0.582)	0.0079	(0.0069, 0.0091)	0.111		0.312
		Random intercept from	957.7	0.071	0.437	0.112	0.025		0.518	(0.473,0.568)	0.0079	(0.0067, 0.0094)		0.194	0.298
		Random intercept to	899.4	0.097	0.539	0.115	0.024		0.489	(0.446,0.537)	0.0077	(0.0064, 0.0093)	0.062	0.173	0.244
		and from													
Land use	logmedian	Linear model	1043.8	0.033		0.022	0.025	0.594	0.454	(0.402, 0.512)	0.0083	(0.0078,0.0087)			
model		Random intercept to	988.1	0.042	0.294	0.027	0.025		0.435	(0.381, 0.496)	0.0088	(0.0077, 0.0101)	0.116		0.325
		Random intercept from	990.2	0.032	0.388	0.094	0.025		0.444	(0.388, 0.508)	0.0090	(0.0076, 0.0106)		0.185	0.319
		Random intercept to	916.5	0.122	0.562	0.101	0.024		0.330	(0.280, 0.388)	0.0085	(0.0071, 0.0103)	0.087	0.164	0.25
		and from													
	log25th	Linear model	1034.55	0.049		0.041	0.025	0.783	0.603	(0.548, 0.664)	0.0081	(0.0077, 0.0086)			
		Random intercept to	978	0.061	0.305	0.039	0.025		0.582	(0.525,0.644)	0.0086	(0.0075, 0.0099)	0.112		0.319
		Random intercept from	979.5	0.048	0.399	0.097	0.025		0.591	(0.534, 0.655)	0.0088	(0.0074, 0.0103)		0.182	0.313
		Random intercept to	902.9	0.126	0.566	0.120	0.024		0.489	(0.436, 0.548)	0.0084	(0.0069, 0.0101)	0.084	0.163	0.243
		and from													
Validation	log25th	Random intercept to	451.3	0.097	0.539	0.218	0.009	0.698	0.576	(0.774, 0.879)	0.0088	(0.0075, 0.0103)	0.064	0.122	0.100
model*		and from													

*Incline model without migration rate ≥0.1

Figures

Figure 1. Scatter plots of logit of migration rates and cost function



Supplement tables

Supplement table 1. Summary of sampling counties, villages and snails

County	Village	Snail n (%)
AN	JG	40(4.8)
	BG1	36(4.3)
	BG2	42(5.0)
	JQ1	30(3.6)
	JQ3	44(5.3)
	HT3	37(4.4)
subtotal		229(27.5)
JI	GH3	14(1.7)
	GH7	40(4.8)
	GQ2	39(4.7)
	GQ3	14(1.7)
	SB	24(2.9)
	DS4	46(5.5)
	DS5	13(1.6)
	DC	27(3.2)
	GH	12(1.4)
	LJ	22(2.6)
	XQ3	32(3.8)
	XQ5	32(3.8)
	XQ7	13(1.6)
subtotal		328(39.4)
ZH	FX	11(1.3)

	HQ1	23(2.8)
	HQ2	37(4.4)
	HQ4	35(4.2)
	XD3	37(4.4)
	XD6	33(4.0)
	XZ	15(1.8)
	ZG	35(4.2)
	WX	25(3.0)
	XL	25(3.0)
subtotal		276(33.1)
Total		833(100)

Supplement table 2. Summary of Oncomelania hupensis snail microsatellite genotypes

OH08		**OH12**		**OH150**		**OH157**		**OH211**		**OH212**		**OH235**		**OH47**		**OH573**		**OH68**		**OH73**	
Repeats	N(%)	Repeats	N(%)	Repeats	N(%)	Repeats	N(%)	Repeats	N(%)	Repeats	N(%)	Repeats	N(%)	Repeats	N(%)	Repeats	N(%)	Repeats	N(%)	Repeats	N(%)
105	1(0.1)	180	3(0.2)	336	10 (0.7)	262	3(0.2)	140	14 (1.0)	219	7(0.6)	248	7(0.5)	208	7 (0.5)	172	1(0.1)	170	8 (0.5)	150	28 (1.8)
108	52 (4.8)	183	20 (1.3)	342	62 (4.4)	264	101 (6.6)	152	6(0.4)	222	9(0.7)	250	593 (44.2)	210	11 (0.8)	176	1010 (75.8)	172	28 (1.8)	153	94 (6.1)
111	146 (13.5)	186	16 (1.0)	345	399 (28.3)	266	2(0.1)	154	3(0.2)	225	37(3)	252	397 (29.6)	212	7 (0.5)	180	13 (1.0)	174	346 (22)	156	27 (1.8)
114	58 (5.4)	189	21 (1.4)	348	183 (13.0)	274	45 (2.9)	156	3(0.2)	228	72 (5.9)	254	248 (18.5)	214	83 (5.9)	182	280 (21.0)	176	226 (15)	159	133 (8.6)
117	28 (2.6)	192	30 (1.9)	351	351 (24.9)	276	92 (6.0)	158	1(0.1)	231	55 (4.5)	256	23 (1.7)	216	31 (2.2)	184	7(0.5)	178	684 (44)	162	208 (13.5)
120	29 (2.7)	195	82 (5.3)	354	179 (12.7)	278	935 (61.0)	160	156 (11.4)	234	76 (6.2)	258	12 (0.9)	218	537 (37.9)	186	2(0.2)	180	166 (11)	165	216 (14.0)
123	11 (1.0)	198	19 (1.2)	357	28 (2.0)	280	118 (7.7)	162	213 (15.5)	237	19 (1.6)	260	12 (0.9)	220	486 (34.3)	190	12 (0.9)	182	15 (1)	168	125 (8.1)
126	36 (3.3)	201	44 (2.9)	360	10 (0.7)	282	49 (3.2)	164	897 (65.4)	240	27 (2.2)	262	35 (2.6)	222	82 (5.8)	200	7(0.5)	184	1 (0.1)	171	154 (10.0)
129	41 (3.8)	204	117 (7.6)	363	12 (0.9)	284	3(0.2)	166	16 (1.2)	243	55 (4.5)	264	9(0.7)	224	151 (10.7)			188	1 (0.1)	174	211 (13.7)

OH08		OH12		OH150		OH157		OH211		OH212		OH235		OH47		OH573		OH68		OH73	
132	56	207	217	366	7(0.5)	286	13	168	26	246	160	266	3(0.2)	226	2(0.1)			190	70	177	87
	(5.2)		(14.0)				(0.9)		(1.9)		(13)								(4.5)		(5.6)
135	251	210	65	369	1(0.1)	288	56	178	8(0.6)	249	82	270	3(0.2)	238	19			198	1	180	154
	(23.2)		(4.2)				(3.7)				(6.7)				(1.3)				(0.1)		(10.0)
138	43	213	177	372	39	290	9(0.6)	188	26	252	59									183	77
	(4.0)		(11.5)		(2.8)				(1.9)		(4.8)										(5.0)
141	8(0.7)	216	16	375	7(0.5)	292	8(0.5)	190	2(0.2)	255	42									186	12
			(1.0)								(3.4)										(0.8)
144	8(0.7)	219	7(0.5)	378	17	294	14	194	1(0.1)	258	17									189	2(0.1)
					(1.2)		(0.9)				(1.4)										
147	20	222	4(0.3)	381	15	296	8(0.5)			261	13									198	11
	(1.9)				(1.1)						(1.1)										(0.7)
150	6(0.6)	225	1(0.1)	384	14	298	20			264	11									204	1(0.1)
					(1.0)		(1.3)				(0.9)										
153	22	279	1(0.1)	387	3(0.2)	300	8			267	18									207	4(0.3)
	(2.0)						(0.5)				(1.5)										
156	22	387	2(0.1)	390	16	302	15			270	14										
	(2.0)				(1.1)		(1.0)				(1.2)										
159	24	393	16	393	18	304	16			273	41										
	(2.2)		(1.0)		(1.3)		(1.0)				(3.4)										
162	18	396	537	396	9(0.6)	306	9(0.6)			276	81										

OH08		OH12		OH150		OH157		OH211	OH212		OH235	OH47	OH573	OH68	OH73
	(1.7)		(34.7)							(6.6)					
165	23	399	57	399	10	310	2(0.1)		279	34					
	(2.1)		(3.7)		(0.7)					(2.8)					
168	17	408	5(0.3)	405	2(0.1)	312	3(0.2)		282	56					
	(1.6)									(4.6)					
171	6(0.6)	411	37	408	2(0.1)	318	3(0.2)		285	27					
			(2.4)							(2.2)					
174	6(0.6)	414	52	411	9(0.6)				288	76					
			(3.4)							(6.2)					
177	13			414	7(0.5)				291	67					
	(1.2)									(5.5)					
180	6(0.6)			417	1(0.1)				294	30					
										(2.5)					
183	2(0.2)			423	1(0.1)				297	22					
										(1.8)					
186	4(0.4)								300	5(0.4)					
189	6(0.6)								303	10					
										(0.8)					
192	36														
	(3.3)														
195	11														

OH08		OH12	OH150	OH157	OH211	OH212	OH235	OH47	OH573	OH68	OH73
	(1.0)										
198	5(0.5)										
201	33										
	(3.1)										
204	5(0.5)										
207	3(0.3)										
219	3(0.3)										
225	2(0.2)										
228	10										
	(0.9)										
231	3(0.3)										
234	2(0.2)										
243	1(0.1)										
246	3(0.3)										
Tot	1080	1546	1412	1532	1372	1222	1342	1416	1332	1546	1554
al	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)

Supplement table 3. Summary of cost-path values for different cost models

Model	Parameters of EDG distribution	N	Maximum	Median	Minimum	IQR
Euclidean	Median	496	67213	20986	6129	21829
	25th percentile	496	67146	17947	3064	23343
topography	Median	496	67373	21028	6257	21703
	25th percentile	496	67275	17941	3160	23358
Incline	Median	992	67556	21095	6264	21870
	25th percentile	992	67457	18115	3159	23420
land use	Median	992	792110	185965	52631	149434
	25th percentile	992	750338	155263	27015	153913
distance from watershed	Median	496	93168	34317	10524	25226
	25th percentile	496	90316	30675	4792	27064
wetness	Median	496	80357	26891	6741	23826
	25th percentile	496	78086	24232	3311	25163
Stream only	Median	992	5328716	1099047	376972	814078
	25th percentile	992	4920968	927570	203168	809812
stream velocity	Median	992	5288993	1057957	369528	812309
	25th percentile	992	4851660	910861	221901	798687
streams and channels	Median	992	943752	293955	131763	270226
	25th percentile	992	860484	267931	95599	292998

Supplement table 4. Correlation coefficients of median of EGD between different environmental models

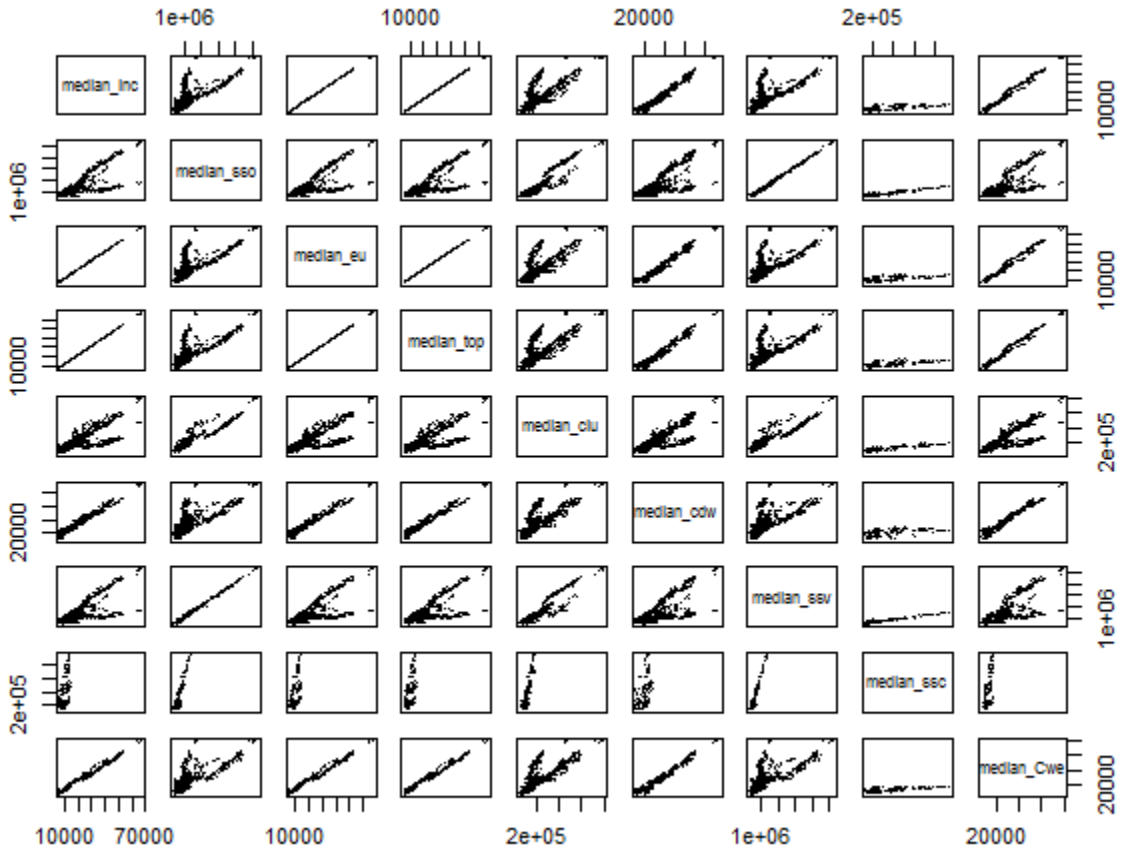
	Euclidean	topography	Incline	land use	distance from watershed	wetness	Stream only	stream velocity	streams and channels
Euclidean	1	>0.999	0.999	0.616	0.737	0.929	0.592	0.582	0.545
topography	>0.999	1	>0.999	0.617	0.738	0.927	0.591	0.580	0.543
Incline	0.999	>0.999	1	0.603	0.740	0.928	.577	0.565	0.528
land use	0.616	0.617	0.603	1	0.567	0.553	0.878	0.877	0.865
distance from watershed	0.737	0.738	0.740	0.568	1	0.627	0.512	0.462	0.444
wetness	0.929	0.927	0.928	0.553	0.627	1	0.587	0.594	0.557
Stream only	0.592	0.591	0.577	0.878	0.512	0.587	1	0.964	0.921
stream velocity	0.581	0.580	0.565	0.878	0.462	0.594	0.964	1	0.959
streams and channels	0.545	0.543	0.528	0.865	0.444	0.557	0.920	0.959	1

Supplement table 5. Correlation coefficients of 25th percentile of EGD between different environmental models

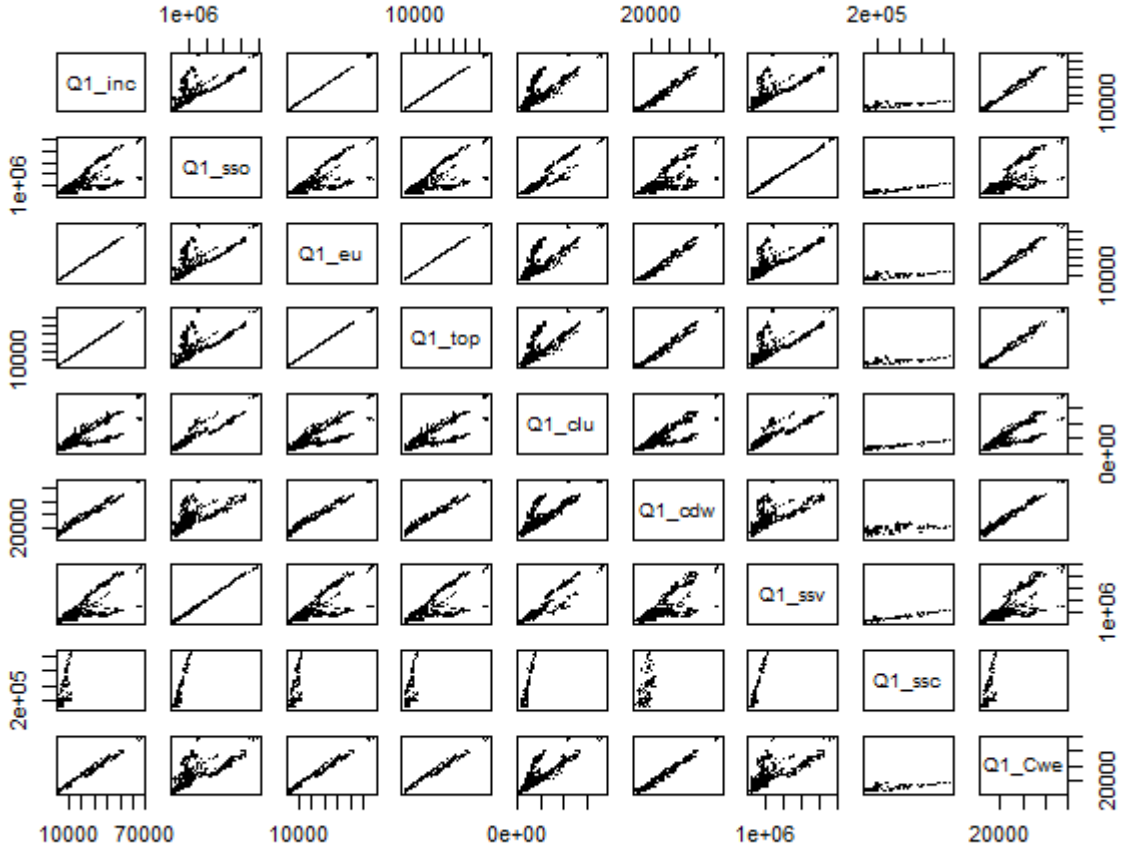
	Euclidean	topography	Incline	land use	distance from watershed	wetness	Stream only	stream velocity	streams and channels
Euclidean	1	>0.999	0.999	0.724	0.783	0.983	0.772	0.734	0.620
topography	>0.999	1	>0.999	0.723	0.781	0.984	0.781	0.734	0.619
Incline	0.999	>0.999	1	0.710	0.782	0.984	.758	0.720	0.603
land use	0.724	0.723	0.710	1	0.673	0.714	0.963	0.961	0.917
distance from watershed	0.783	0.781	0.782	0.673	1	0.735	0.664	0.623	0.513
wetness	0.983	0.984	0.984	0.714	0.737	1	0.761	0.731	0.618
Stream only	0.772	0.772	0.758	0.963	0.664	0.761	1	0.925	0.925
stream velocity	0.734	0.734	0.720	0.961	0.623	0.731	0.986	1	0.956
streams and channels	0.620	0.619	0.603	0.917	0.513	0.618	0.925	0.956	1

Supplement figures

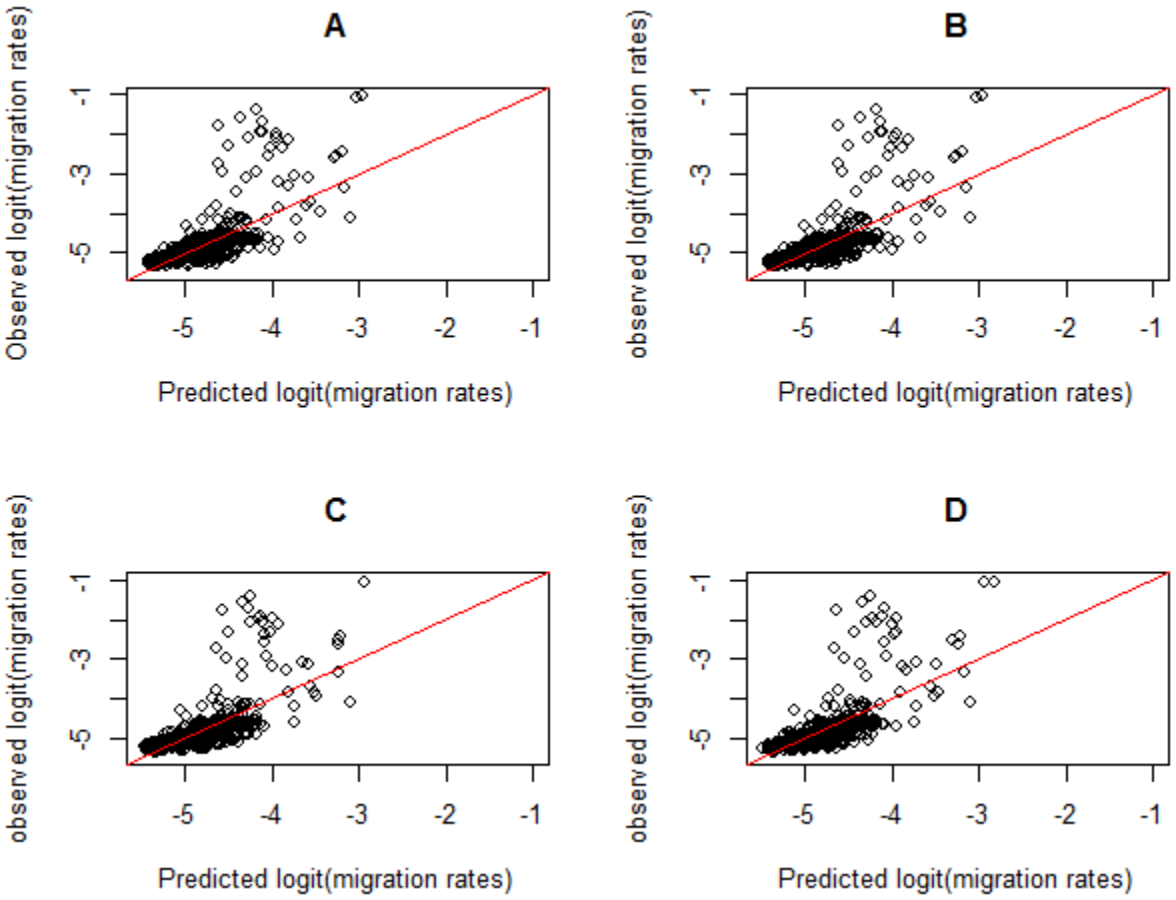
Supplement figure 1. Correlation of median of EGD between different environmental models



Supplement figure 2. Correlation of 25th percentile of EGD between different environmental models



Supplement figure 3. Observed versus predicted logit of migration rates from "to and from" random effect model for EGD log25th



Panel A represents predicted versus observed logit(migration rates) from Euclidean model log25th random effect "to and from" model. Panel B represents predicted versus observed logit(migration rates) from Incline model log25th random effect "to and from" model. Panel C represents predicted versus observed logit(migration rates) from Stream only model log25th random effect "to and from" model. Panel D represents predicted versus observed logit(migration rates) from Land use model log25th random effect "to and from" model.

References

1. Bayne, C.J., *Successful parasitism of vector snail Biomphalaria glabrata by the human blood fluke (trematode) Schistosoma mansoni: a 2009 assessment.* Mol Biochem Parasitol, 2009. **165**(1): p. 8-18.
2. Liang, S., et al., *Re-emerging schistosomiasis in hilly and mountainous areas of Sichuan, China.* Bull World Health Organ, 2006. **84**(2): p. 139-44.
3. Carlton, E.J., et al., *Evaluation of mammalian and intermediate host surveillance methods for detecting schistosomiasis reemergence in southwest China.* PLoS Negl Trop Dis, 2011. **5**(3): p. e987.
4. WHO. 1997; Available from: http://www.who.int/water_sanitation_health/resources/vector337to356.pdf.
5. Remais, J., *Modelling environmentally-mediated infectious diseases of humans: transmission dynamics of schistosomiasis in China.* Adv Exp Med Biol, 2010. **673**: p. 79-98.
6. Institute, I.R.R., *Vector-borne Disease Control in Humans Through Rice Agroecosystem Management*. 1987.
7. Akullian, A.N., et al., *Modeling the combined influence of host dispersal and waterborne fate and transport on pathogen spread in complex landscapes.* Water Qual Expo Health, 2012. **4**(3): p. 159-168.
8. Killeen, G.F., B.G. Knols, and W. Gu, *Taking malaria transmission out of the bottle: implications of mosquito dispersal for vector-control interventions.* Lancet Infect Dis, 2003. **3**(5): p. 297-303.
9. Cameron, R.A.D.a.W.P., *Estimating migration and the effects of disturbance in mark-recapture studies on the snail cepaea nemoralis L.* . J. Anim. Ecol., 1977(46): p. 173-179.
10. N., S.M.a.B., *A comparison of three indirect methods for estimating average levels of gene flow.* International Journal of organic evolution 1989. **43**(7): p. 1349-1368.
11. Tavare, S., *Line-of-descent and genealogical processes, and their applications in population genetics models.* Theor Popul Biol, 1984. **26**(2): p. 119-64.
12. Beerli, P. and J. Felsenstein, *Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach.* Genetics, 1999. **152**(2): p. 763-73.
13. Beerli, P. and J. Felsenstein, *Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach.* Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4563-8.
14. Cornuet, J.M., et al., *New methods employing multilocus genotypes to select or exclude populations as origins of individuals.* Genetics, 1999. **153**(4): p. 1989-2000.
15. Rannala, B. and J.L. Mountain, *Detecting immigration by using multilocus genotypes.* Proc Natl Acad Sci U S A, 1997. **94**(17): p. 9197-201.
16. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.
17. Wilson, G.A. and B. Rannala, *Bayesian inference of recent migration rates using multilocus genotypes.* Genetics, 2003. **163**(3): p. 1177-91.
18. Faubet, P., R.S. Waples, and O.E. Gaggiotti, *Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates.* Mol Ecol, 2007. **16**(6): p. 1149-66.
19. Anderson, C.S.M., D.B., *Genetic estimates of immigration and emigration rates in relation to population density and forest patch area in Peromyscus leucopus.* Conserv Genet, 2010(11): p. 1593–1605.
20. Bertrand, J.A., et al., *Extremely reduced dispersal and gene flow in an island bird.* Heredity (Edinb), 2014. **112**(2): p. 190-6.
21. Howes, B.J.B., J.W.; Gibbs, H.L.; Herman, T.B.; Mockford, S.W.; Prior, K.A.; Weatherhead, P.J., *Directional gene flow patterns in disjunct populations of the black ratsnake (Pantheropis obsoletus) and the Blanding's turtle (Emydoidea blandingii).* Conserv Genet, 2008.
22. Miller, H.J.W., E.A., *Representation and Spatial Analysis in Geographic Information Systems.* Annals of the Association of American Geographers, 2003. **93**(3): p. 574–594.

23. Remais, J., et al., *Analytical methods for quantifying environmental connectivity for the control and surveillance of infectious disease spread.* J R Soc Interface, 2010. **7**(49): p. 1181-93.
24. Gurarie, D. and E.Y. Seto, *Connectivity sustains disease transmission in environments with low potential for endemicity: modelling schistosomiasis with hydrologic and social connectivities.* J R Soc Interface, 2009. **6**(35): p. 495-508.
25. Liang, S., D. Maszle, and R.C. Spear, *A quantitative framework for a multi-group model of Schistosomiasis japonicum transmission dynamics and control in Sichuan, China.* Acta Trop, 2002. **82**(2): p. 263-77.
26. Remais, J., Hubbard, A., Zisong, W., Spear, R.C., *Weather-driven dynamics of an intermediate host:mechanistic and statistical population modelling of Oncomelania hupensis.* Journal of Applied Ecology, 2007(44): p. 781-791.
27. Pamilo, P., *Genotypic correlation and regression in social groups: multiple alleles, multiple loci and subdivided populations.* Genetics, 1984. **107**(2): p. 307-20.
28. Guo, J.G., et al., *A geographic information and remote sensing based model for prediction of Oncomelania hupensis habitats in the Poyang Lake area, China.* Acta Trop, 2005. **96**(2-3): p. 213-22.

APPENDIX

BAYESASS CODE

AN COUNTY
Input file: county_an.txt
Random seed=2345 MCMC iterations=10000000 Burn-in=2000000 Sampling interval=100
Mixing parameters: (dM=0.3,dA=0.3,dF=0.3) Output file=an_county.txt
Individuals: 229 Populations: 6 Loci: 11

JI_ZH COUNTY
Input file: ji_zh.txt
Random seed=12345 MCMC iterations=15000000 Burn-in=5000000 Sampling interval=100
Mixing parameters: (dM=0.5,dA=0.5,dF=0.5) Output file=output_ji_zh.txt
Individuals: 604 Populations: 23 Loci: 11