Title

Genetic diversity and disease spread: a protocol for meta-analysis

Authors

Amanda K Gibson; *corresponding author: akg5nq@virginia.edu

Affiliation: Department of Biology, University of Virginia, 229 Gilmer Hall, 485 McCormick Road, Charlottesville, Virginia 22904

AKG is the guarantor. AKG developed and will conduct the study in its entirety.

Amendments

If amendments to this protocol are needed, I will describe the change, explain the rationale for the change, and identify the date at which the change was made.

Support

Sources: Funding for AKG comes from the Wissenschaftskolleg zu Berlin. There is no direct funding for this project.

INTRODUCTION

Rationale

Why does infectious disease sweep through some host populations and not others? One prominent hypothesis proposes that disease can spread more readily through host populations with low genetic diversity. The hypothesis stems from the idea that genetically similar hosts are likely to be susceptible to the same infections. Hence the rate of transmission of an infection will be higher in groups of genetically similar hosts relative to groups of genetically dissimilar hosts.

This hypothesis has an intuitive appeal, and large epidemics of crop monocultures provide dramatic examples of the vulnerability of low diversity host populations. The term "monoculture effect" describes this association of high infection prevalence and low genetic diversity observed in agricultural systems. King and Lively (2012) noted that the protective effect of genetic diversity has taken on the status of "conventional wisdom." Yet it remains unclear if the effect of genetic diversity on disease spread is general.

Host diversity may also reduce variance in disease spread between populations. I will explain this point with a simple example: imagine that host genotypes are completely resistant or completely susceptible to a parasite. At any given point in time, the infection rate may be 0% or 100% in host populations of a single genotype. In contrast, host populations with a mixture of host genotypes are unlikely to be disease-free, as most possible mixtures include at least one susceptible host genotype. Mixtures are, however, also unlikely to experience infection rates of 100%, as most possible mixtures also include at least one resistant host genotype. Hence we would observe increased variance in infection rate for low diversity host populations relative to high diversity populations. This difference in variation should be detectable in experimental or agricultural populations, where coevolutionary feedbacks are limited. The idea predicts that low diversity populations experience more lows but also more devastating highs in infection rate than do high diversity host populations. Such variation is crucial to the study and control of infectious disease: a single large epidemic could decimate a crop or drive local host extinctions. Yet variation has been largely overlooked in tests of the association between genetic diversity and disease.

Objectives

I will address these current gaps in the field using meta-analysis. I will use a systematic review to collect studies from a variety of systems (agricultural and non-agricultural) and a variety of approaches (observational and experimental). I will then quantitatively synthesize the results of these studies. My primary questions are as follows:

1) Does genetic diversity reduce the spread of disease?

2) Does genetic diversity reduce the variation in disease spread across populations?

3) Are these effects context-dependent? For example, are effect sizes on average greater in agricultural vs. non-agricultural systems?

METHODS

Eligibility Criteria

I will include studies that provide 1) some measure of population-level genetic variation of the host species and 2) some measure of disease risk across host populations. These studies may be observational or experimental. All host-parasite systems are eligible for inclusion. Parasites and pathogens will be defined as in Lafferty and Kuris (2002). For inclusion, a study must compare multiple populations and provide data necessary for the calculation of effect sizes and their variance. The study must also address host populations rather than individual hosts within a population. Table 1 provides further details on inclusion/exclusion criteria, with some examples. I will only consider studies written in English.

Table 1: Inclusion and exclusion criteria for study selection

Criteria		Include	Exclude
Availability	Language	English	Other languages
	Data availability	Required: • Data for calculating effect size: means, correlation coefficients, risk ratios, etc. • Sample size • Data for calculating standard deviations I will obtain these from the paper, publicly available data sets associated with the paper, or communications with authors when possible	No relevant data Error or replication not available or cannot be deciphered
	Original	Study includes original data	Study duplicates data from another study (in this case, the study is excluded and data are drawn from the original study)
Basic design	Focal player	Genetic diversity for the **host**	Genetic diversity for the **parasite.** Studies in which parasite genetic diversity was the focus will largely be excluded by title. I will, however, consider studies that make use of a variety of parasite genotypes and species, as long as the data can be used to address the stated research question.
	Approach	Compares multiple host populations via observation or experimental manipulation For experimental studies, host populations vary in the degree of genetic diversity, through, for example, combining different numbers of distinct host genotypes or controlling mating such that the resulting broods vary in their genetic diversity.	Only a single population is considered. Replication is at the level of the individual, not the population. For example, the probability of an individual being infected is linked to individual-level genomic heterozygosity. Addresses individual-level outcomes (e.g. individual

		Addresses population-level outcomes (e.g. prevalence of infection, mean parasite load)	immune responses or parasite load)
Genetic diversity	Approach	Explicitly controlled or measured. In cases of experimental studies, genetic diversity is typically sufficiently controlled such that measurement is not necessary (e.g. if researchers control degree of outbreeding during mating or create mixtures of distinct genotypes). In cases of observational studies, genetic diversity must be measured, and the metric must be at the level of the population (e.g. genotype richness, evenness, population-level heterozygosity)	Not controlled, not measured, or not defined at the level of the population (e.g. sexual and asexual populations, lineages or species are compared, without measuring or controlling genetic diversity; individual-level heterozygosity)
	Defining diversity	Acceptable types of diversity include: Mixtures of distinct genotypes, cultivars, or lineages Broods or populations derived from single vs. multiple mating (polyandry/polygyny) Broods or populations derived from outcrossing vs. inbreeding/cloning	
	Divergence level	Intraspecific	Interspecific; hybrids between species
	Mixtures	True mixtures (e.g. mixtures of genotypes); coexistence within a population	Mixing of genotypes is in time, e.g. by crop rotation
Disease	Antagonist	Parasites, pathogens, defined as in Lafferty and Kuris (2002)	Herbivores
	Disease outcomes	Measured outcome is relevant to disease transmission (e.g. prevalence of infection; mean parasite load; R_0)	Measured outcome is not clearly relevant to transmission (e.g. immune response, yield) Though data for multiple parasite species may be independently extracted from a single study, parasite richness will not be treated as a relevant outcome, as the hypothesis being tested concerns transmission of an individual parasite species (rev. in Johnson et al. 2015)

Information Sources

I will conduct literature searches using the following electronic databases: Web of Science (1900-2018), PubMed, OpenGrey, ProQuest Theses and Dissertations Global, and the USDA National Agricultural Library. I will find additional studies by 1) scanning the citations of included studies and relevant reviews and 2) scanning the studies that cite these included studies and reviews in Google Scholar.

Search Strategy:

I will divide my literature search into three phases. First, I will locate studies through electronic databases. I will search Web of Science with the following search terms:

TI = (("genetic diversity" OR "genetic variation" OR monoculture OR polyculture OR inbred OR inbreeding OR mixtur* OR heterogen* OR polyandr* OR polygyn* OR promisc* OR outcros*) AND (parasit* OR infect* OR pathogen OR disease OR susceptib* OR epidemic OR resist*))

These terms were identified using a pilot literature search in Web of Science in the summer of 2018. After the initial search, I will continually screen for new studies through a weekly citation alert with the above search terms. I then will search PubMed using a large subset of these search terms. I will search OpenGrey, ProQuest Dissertations and Theses Global, and the USDA National Agricultural Library with a smaller subset of these search terms. Second, I will locate studies through the citation lists and citing articles of relevant reviews. Third, I will locate studies through the citation lists and citing articles of included studies.

I began searching for relevant literature on September 5, 2018.

Study records

Data Management: All references will be collected and organized in Endnote. Excel sheets will be used as the pool of eligible studies shrinks to a manageable size.

Selection Process: I will first use the inclusion criteria in Table 1 to screen titles. I began this process on September 5, 2018.

For titles that appear to meet the inclusion criteria or are ambiguous, I will screen abstracts. For abstracts that appear to meet the inclusion criteria or are ambiguous, I will screen the full texts of the articles. If possible, I will contact authors of studies for which eligibility remains ambiguous. For all steps excepting the title screen, I will record reasons for excluding an article. All records of studies, included and excluded, will be retained in an Endnote library. During screening, I will be blind to study authors and institutions.

Data Collection Process: I will develop an Excel sheet for data entry and an instruction manual for guidelines on how to extract and code data from each included study. For a subset of the studies, I will repeat data extraction to ensure that the results are replicable. When possible, I will contact authors to resolve ambiguities.

In the event that additional researchers are recruited to assist in this step, I will train them using the instruction manual. Data extraction would then be conducted in duplicate for all studies. Disagreements will be resolved via discussion.

Data Items

I will manually extract data from text and tables. Data contained in figures will be extracted using an image analysis program (e.g. ImageJ or PlotDigitizer). I will collect data necessary for calculating effect sizes; in most cases, these will be means, estimates of variance, and sample size. I will collect study meta-data: title, year, authors, and journal. I will collect data regarding the biology of the study: study system (host and parasite taxa; species and relevant higher levels, e.g. Insecta or Mammalia) and system context (agricultural or non-agricultural). I will collect data regarding the methodology of the study: approach (observational or experimental), assembly of experimental host populations (random, non-random), metric for host diversity, method for introducing parasites in experiments, metric for disease risk, diversity range (e.g. maximum number of host genotypes; max frequency of any host genotype) and estimates of host density (if possible). I will document and justify any additional data items collected after development of this protocol. Where possible, ambiguities will be resolved by contacting authors.

Outcomes and Prioritization

The first main outcome will be the effect of genetic diversity on a measure of disease risk. Depending on the study, this measure of disease risk may be population prevalence of a parasite, mean parasite load across individuals in the population, area under the disease progress curve, or estimates of R_0. For each effect size, we will record the specific measurement used. The second main outcome will be the effect of genetic diversity on variance in that measure of disease risk.

Risk of Bias in Individual Studies

The studies included in this meta-analysis will employ a variety of experimental designs. These experimental designs vary in their ability to directly test the contribution of genetic diversity itself to variation in disease risk. I will evaluate the experimental design and keep records of elements that influence my confidence that genetic diversity itself is being testing.

First, I will evaluate the approach to assembling diverse experimental populations of hosts. Experimental host populations may be assembled randomly, by selecting a random set of host genotypes to mix or by mating females with a random selection of males to create diverse broods. By random, I mean to indicate that the researchers do not display prior knowledge of a genotype's resistance level when assembling experimental populations. Such random assemblages allow researchers to test the role of genetic diversity itself in limiting disease spread. Experimental host populations may also be assembled non-randomly, by combining a host genotype that is known to be susceptible to infection with genotypes that are known to be resistant. I would also include in this category studies in which genotypes are known to vary at a single, or very few, loci. In such a design, the contribution of genetic diversity to disease risk is confounded with the contribution of resistance or of a single locus. I will evaluate the significance of this aspect of the design by running data analyses with and without studies that use non-random designs.

Second, the number of host genotypes included in experimental populations influences our confidence in a study's test of the role of genetic diversity. I will include studies that use two or more host genotypes. However, I have more confidence that genetic diversity itself is being tested if a study uses three or more host genotypes (or outcrosses females with three or more males). To deal with this issue, I will run data analysis with and without studies limited to two host genotypes.

Finally, all experimental studies must include appropriate controls. For experimental studies, this means that a study must include treatments that are identical to one another in all ways except their diversity level.

Data Synthesis

I will use statistical models to identify the general trend across studies, quantify heterogeneity between studies, and explain the observed heterogeneity between studies via moderators. Separate models will be built for observational and experimental studies; the design of observational vs. experimental studies differ in fundamental ways that are predicted to influence the strength and direction of the relationship between genetic diversity and disease risk. The specific factors included in each model will depend on the completeness of the data I am able to collect for each factor and the anticipated power of the analysis. The factors included will vary between observational and experimental studies (e.g. factors related to experimental design will be limited to analysis of experimental studies).

For comparison of means, all effect sizes will be converted to a common metric (e.g. Hedges' *g*) (Borenstein et al. 2009). I will analyze these data using a multilevel meta-analytic, or meta-regression, approach. With this approach, a single model can include categorical (e.g. system context - agricultural or non-agricultural; metric of disease risk; design - random or non-random assembly) and continuous predictors (e.g. host density; number of host genotypes). I will draw multiple effects sizes from a single host-parasite system and a single study. To account for non-independence amongst these effect sizes, I will specify system and study identifiers as random effects. The R packages MCMCglmm and metafor offer functions for carrying out such analyses (Hadfield 2010; Viechtbauer 2010, respectively). The best-fitting subset of models will be identified using Akaike information criteria (for example, using the R package glmulti (Calcagno and de Mazancourt 2010)).

To measure heterogeneity, I will estimate the fraction of the total variance accounted for by each random factor. Nakagawa and Santos (2012) outline this extension of the I_2 statistic for multilevel meta-analysis.

For comparison of variance, I will modify the approach outlined above using methods introduced by Nakagawa et al. (2014). They advise either 1) the use of the effect size lnCVR, the natural logarithm between the coefficients of variation of two groups and/or 2) a hierarchical model in which the mean effect size and host diversity are predictors of standard deviation. This approach is designed for comparing the variance between groups within a study; therefore, I will restrict this analysis to experimental studies. Observational studies lack treatment groups (i.e. high and low genetic diversity host populations); rather populations vary continuously in some metric of genetic diversity.

Meta-biases

Many tests of publication bias require independence of effect sizes and hence cannot be used when multiple effect sizes are drawn from the same study or the same system. To address this problem, I will follow the recommendations of Nakagawa and Santos (2012) and test for publication bias using meta-analytic residuals rather than the effect sizes themselves. Tests will include Egger's test to quantify funnel asymmetry and the trim and fill method to adjust for missing studies.

List of Works Cited

Borenstein, M., L. Hedges, J. Higgins, and H. Rothstein. 2009. Converting among effect sizes, Pages 45-49 Introduction to Meta-Analysis. West Sussex, UK, John Wiley & Sons, Ltd.

Calcagno, V., and C. de Mazancourt. 2010. glmulti: an R package for easy automated model selection with (generalized) linear models. Journal of statistical software 34:1-29.

Hadfield, J. D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software 33:1-22.

Johnson, P. T. J., R. S. Ostfeld, and F. Keesing. 2015. Frontiers in research on biodiversity and disease. Ecology letters 18:1119-1133.

King, K. C., and C. M. Lively. 2012. Does genetic diversity limit disease spread in natural host populations? Heredity 109:199-203.

Lafferty, K. D., and A. M. Kuris. 2002. Trophic strategies, animal diversity, and body size. Trends in Ecology & Evolution 17:507-513.

Nakagawa, S., R. Poulin, K. Mengersen, K. Reinhold, L. Engqvist, M. Lagisz, and A. M. Senior. 2014. Meta-analysis of variation: ecological and evolutionary applications and beyond. Methods in Ecology and Evolution 6:143-152.

Nakagawa, S., and E. S. Santos. 2012. Methodological issues and advances in biological meta-analysis. Evolutionary Ecology 26:1253-1274.

Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software 36.