Published in final edited form as:

Genet Epidemiol. 2009 May ; 33(4): 281–289. doi:10.1002/gepi.20386.

Ecogeographic Genetic Epidemiology

Chantel D. Sloan¹, Eric J. Duell, Ph.D.^{2,4}, Xun Shi, Ph.D.⁵, Rebecca Irwin, Ph.D.⁶, Angeline S. Andrew, Ph.D.^{2,3}, Scott M. Williams, Ph.D.⁷, and Jason H. Moore, Ph.D.^{1,2,3}

¹Computational Genetics Laboratory, Department of Genetics, Dartmouth Medical School, Lebanon, NH

²Departments of Community and Family Medicine, Norris-Cotton Cancer Center, Dartmouth Medical School, Lebanon, NH

³Department of Computer Science, University of New Hampshire, Durham, NH, Department of Computer Science, University of Vermont, Burlington, VT, Translational Genomics Research Institute, Phoenix, AZ

⁴Lifestyle, Environment and Cancer Group, International Agency for Research on Cancer, 69372 Lyon Cedex 08, France.

⁵Department of Geography, Dartmouth College, Hanover, NH

⁶Department of Biological Sciences Dartmouth College, Hanover, NH

⁷Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN

Abstract

Complex diseases such as cancer and heart disease result from interactions between an individual's genetics and environment, i.e. their human ecology. Rates of complex diseases have consistently demonstrated geographic patterns of incidence, or spatial “clusters” of increased incidence relative to the general population. Likewise, genetic subpopulations and environmental influences are not evenly distributed across space. Merging appropriate methods from genetic epidemiology, ecology and geography will provide a more complete understanding of the spatial interactions between genetics and environment that result in spatial patterning of disease rates. Geographic Information Systems (GIS), which are tools designed specifically for dealing with geographic data and performing spatial analyses to determine their relationship, are key to this kind of data integration. Here the authors introduce a new interdisciplinary paradigm, ecogeographic genetic epidemiology, which uses GIS and spatial statistical analyses to layer genetic subpopulation and environmental data with disease rates and thereby discern the complex gene-environment interactions which result in spatial patterns of incidence.

Keywords

Geographic Information Systems; Environmental Health; Population Genetics; Spatial Genetics; Medical Geography; Landscape Genetics

Spatial Disease Patterns

Gene-environment interactions are at the root of both the incidence and severity of common complex diseases such as heart disease, schizophrenia and cancer. Though genetic epidemiologists typically use a family-based or case-control study design to determine associations between genetic variants and disease, studies that incorporate environmental information provide a much richer understanding and interpretation of disease etiology. Typical genetic epidemiologic approaches also fail to stratify by geography, and thus they are unable to take advantage of vast amounts of spatial environmental data available from government, academic and business institutions. In such studies where environmental data are available, geography is not merely a surrogate for local environment, but can inform the discovery of the complex interactions involved not only in individual susceptibility, but the spatial variation of incidence.

The first law of geography as stated by Waldo Tobler is that “everything is related to everything else, but near things are more related than distant things [Tobler 1970].” The interactions that result in disease are no exception to this rule. Environmental influences, which are known to be unevenly distributed across space, are likely to have the most influence on disease within local geographical realms. The genetic subpopulations that environment (or local ecology) influence are also not evenly geographically distributed. The field of geographical genetics uses statistical measures to study patterns of spatial genetic variation and the processes that create it [Epperson 2003]. Recent investigations of human geographical genetics (examples of which will be cited hereafter) have discerned genetic subpopulations that vary across both geographic and political boundaries. The presence of genetically differentiated subpopulations is referred to as genetic structure, and results from populations having distinct ancestry [Rosenberg et al., 2002; Marchini et al., 2004; Falush et al., 2003; Epperson et al., 1996; Doligez et al., 1998]. Structure is commonly measured with a statistic known as F_{ST} , a measure of the difference in genetic variation between populations versus the total population [Wright 1950].

Integrating spatial genetic structure data with local environmental exposures may substantially increase our understanding of spatial patterns of human disease. Recent advances in gathering and analyzing environmental, genetic and disease rate data have provided substantially greater insights into etiology. Investigators are now in the position to initiate what is here referred to as ecogeographic genetic epidemiology, a paradigm for the study of the genetic basis of human disease that integrates spatial, environmental, and genetic data into models of geographic disease etiology.

The authors propose that using geographic information systems (GIS) to integrate multiple levels of available data will prove a valuable approach for epidemiologists, geneticists, geographers and ecologists who are attempting to determine the interactions that are at the root of common complex diseases. Here the philosophies, methodologies and applicable research from the fields of human ecology and landscape epidemiology, GIS, geographical genetics, and geographic genetic epidemiology are discussed. Geographic genetic epidemiology is a term that we will use loosely regarding any study of geographic genetic patterning in relation to disease. Finally, we will combine protocols and ideology from each field into ecogeographic genetic epidemiology.

Human Ecology and Landscape Epidemiology

Human ecology comprises those biological, cultural and environmental factors that influence the state of human populations. These factors were presented by Meade and Earickson as the “triangle of human ecology”, which posits that an individual's population, habitat and behavior together affect disease risk. Population is composed of genetics,

gender, and age, and is related to humans as biological entities- the organisms that interact directly with the environment in the development of disease [Meade et al., 2000]. Environmental influences are theorized to operate in this interaction by moderating the phenotypic expression of genotypes [Sing et al., 1996]. As a result, many investigators are calling for studies of gene-environment interaction for common diseases such as heart disease, cancer and even alcoholism [Schwartz et al., 1996; Rebbeck et al., 2007; Enoch 2006].

Landscape epidemiology encompasses most of the environmental variables involved in human ecology. Eugene Pavlosky first coined the term “landscape epidemiology” when he integrated the fields of natural and cultural ecology of human disease that epidemiologists Helmut Juszatz and Jacques May pioneered in the early 1960's. Landscape epidemiology incorporates the study of interacting environmental factors such as cultural anthropology, environmental science, sociology, meteorology and ecology into epidemiology [Meade et al., 2000; May 1958; Juszatz 1966; Pavlovsky 1966].

The methods employed in landscape epidemiology may be used to address a question of fundamental interest to medical research: What is responsible for the differing effects that environment has on an individual's or population's disease state? Alternatively, why do genetic susceptibilities not always result in the same disease phenotype? For example, one study observed that Hawaiians and African Americans that smoke no more than 30 cigarettes a day have an increased risk of lung cancer relative to Latin Americans, Japanese Americans and Caucasians who smoked the same amount [Haiman et al., 2006]. Therefore, it can be argued that genetic studies of disease are incomplete if they fail to incorporate environmental influences that affect phenotype; nor are environmental studies complete if they do not incorporate genetic risk factors with which they can interact.

Exploring Genetic Landscapes

There is an easily drawn connection between the fields of landscape epidemiology (the effect of environment on disease) and landscape genetics (the effect of environment on genetic patterning) [Storfer et al., 2007]. Methods for integrating analysis of spatial genetic data in relation to landscape features have been and continue to be developed [Epperson 2003; Storfer et al., 2007; Banks et al., 2005; Coulon et al., 2006; Manel et al., 2003]. Though spatial analyses have become very sophisticated, many available tools are still not commonly used by the human genetics community. In fact, studies of human disease are usually mutually exclusive, either looking at only genetic variation or only environmental influences.

Spatial analysis methods already in use may prove useful when complete ascertainment of individuals in an area is not possible, as is the case with most genetic studies. Geographic data can be plotted as points with exact locations, grouped into regions or used to estimate a varying smoothed surface across an area. Choosing a method for displaying and analyzing data should be given careful consideration and is context dependent. In the ecogeographic genetic epidemiology studies proposed here, the preferred or optimal level of smoothing is open to exploration. For instance, analyzing patterns regarding point locations (features that occur at exact points rather than varying continuously, such as pollution sources) can be accomplished with a variety of spatial point analyses. Spatial point analyses use methods such as Monte Carlo simulations to detect departures from complete spatial randomness (CSR) and hence, evidence of clustering [Ripley et al., 1977; Besag et al., 1993; Haggstrom et al., 1999; Waller et al., 2004]. Another option is to use kriging methods which are linear algorithms used to interpolate a continuous surface of data from point events by making quantitative predictions about an area based on surrounding values. In other words, kriging

methods estimate the value at a location that was not sampled based on the values of nearby samples [Waller et al., 2004; Matheron 1963].

For allele frequency maps, representing variation as a smooth surface created through spatial interpolation may be more reliable than a display of point data [Barbujani 2000]. Such smooth surfaces are created using a kind of spatial regression where a grid of cells with a certain window size is placed over an area and the average frequency is estimated at the grids' intersections, with the data points usually weighted by distance from the intersection [Piazza et al., 1981]. When creating smoothed maps of genetic structure investigators will inevitably encounter some challenging issues such as obtaining appropriate sample sizes and avoiding visual patterns that are not actually true reflections of the data. These are issues that have long been studied in geostatistics [Isaaks et al., 1989]. Examples of environmental data layers displayed in GIS as well as different methods for displaying data are shown in Figure 1.

GIS are information systems specifically designed for handling georeferenced data (data that can be associated to specific geographical locations). They include functionalities of integrating multiple data “layers” that represent different spatial features or phenomena, and performing analyses on them. Some statistical and spatial statistical functions have started to be incorporated into commercial GIS packages. To date, spatial studies of genetic patterns in relation to disease have mostly involved simple mapping techniques that are useful for visualizing patterns of variation, but not as a tool for further quantitative analysis, or integration with environmental data. GIS will be an important tool in these kinds of investigations.

Genetic structure may be discerned using a non-geographic method and then mapped using GIS. Measuring Wright's F_{ST} statistics is one of the most common methods for finding structure, while others include Principal Components Analysis and Bayesian methods implemented in the software packages STRUCTURE, Bayesian Analysis of Population Structure (BAPS) and Frequentist Estimation of Individual Ancestry Proportion (FRAPPE) [Rosenberg et al., 2002; Pritchard et al., 2000; Tang et al., 2005]. The STRUCTURE program is among the most popular computational methods, reporting a matrix of Q values indicating the proportion of an individuals' genome that originated from a certain subpopulation (admixture model) or the probability that an individual is from a certain subpopulation (no admixture model) as well as F_{ST} 's and graphical outputs. Though each method has its advantages and disadvantages, BAPS has the capability to do spatial analysis if the user supplies sampling coordinates [Corander et al., 2004]. It is important to note that the results of most of these methods are improved by the inclusion of ancestral or pseudo-ancestral individuals in the study population [Tang et al., 2005].

Many of the statistical methods developed for geographical genetics have come from the field of ecology, which has long dealt with issues of landscape genetics and studied the genetic structure of a variety of floral and faunal populations [Banks et al., 2005; Coltman et al., 2003; Vekemans et al., 2004; Clauss et al., 2006; Jump et al., 2006; Keeling 1999]. As a result of the increasing interest in both ecological and human spatial genetic patterning, several groups have developed specific analytical software tools (Table 1) [Cercueil et al., 2007; Bohonak 2002; Nason 1997; Peakall et al., 2006; Guillot et al., 2005; Takahashi 2003; Wartenberg 1989; Degen et al., 2001; Hardy et al., 2002]. The software packages were developed for different purposes and therefore, employ a variety of statistical measures. Some of these programs have taken into consideration the importance of visual displays of genetic structure patterning. The Geneland package creates a smooth map of genetic structure using Poisson-Voronoi tessellation, a method of randomly creating spatial tiles or polygons within a region and investigating the structure within each [Guillot et al., 2005].

Two other programs, GENECLUST and TESS also use a Bayesian method, and have been suggested to be perform better than Geneland, especially when used in conjunction with STRUCTURE [Francois et al., 2006; Chen et al., 2007].

An important consideration when assessing geographic genetic variation is that human geographical genetics is often measured in clusters, though it actually occurs in *clines* or gradual gradients of change. Indeed, most of the variation in studies of selection and landscape genetics can be explained by measuring clines [Handley et al., 2007]. Approaches include using model based or model free methods to investigate clines; others suggest looking for discrete boundaries (abrupt changes) in the spatial variation of genetic structure. The wombling method has been used for this purpose. This method samples across surfaces to determine if higher than expected shifts in allele frequencies, and thereby genetic boundaries, are present in the regions under analysis [Barbujani 2000; Barbujani et al., 1990a; Barbujani et al., 1990b]. Using both simulation studies and real data sets, one group was able to find genetic structure boundaries by calculating what they refer to as “genetic bandwidths” even at F_{ST} 's between 0.05-0.06, and using only 20 loci. This method has been implemented in a freely available software package called GENBMAP [Cercueil et al., 2007].

Geographical Genetics

The field of geographical genetics was established in 1943 with Sewall Wright's seminal paper on Isolation by Distance (IBD), and was further advanced in 1978 when spatial autocorrelation techniques were first applied to population genetics [Epperson 2003; Wright 1943; Sokal et al., 1978a; Sokal et al., 1978b]. IBD is a measure where two matrices, one measuring genetic and the other, geographic distance are created and the Mantel statistic is calculated in order to quantify the geographic distance between the genetic subpopulations. Afterward, Markov chains are often used to determine significance based on the Mantel distribution [Epperson 2003; Wright 1943; Bohonak 2002]. Spatial autocorrelation most commonly employs one of two statistics: Moran's I or Geary's c to measure the similarity between two locations and weight that similarity according to their spatial relationship (measured by adjacency or distance).

Using the aforementioned statistical measures, evidence of spatial human genetic patterns have been found in both macrogeographic and microgeographic studies. Genetic stratification is sometimes associated with locally constituted groups such as ethnicities or tribes, as well as with geographic location. In a landmark study of populations using the STRUCTURE program, Rosenberg and Pritchard were able to demonstrate that genetic structure is clustered by major geographic regions [Rosenberg et al., 2002]. In fact, in a study of 51 populations typed at 377 loci it was determined that geographic distance accounted for all the significant genetic variability in the sample before ethnicity was taken into account [Manica et al., 2005; Conrad et al., 2006]. This is largely due to the fact that human genetics are distinctly tied to human evolutionary and migration history on a global scale [Liu et al., 2006]. Therefore, the next logical question is to discern whether genetic markers can be used to distinguish structure within highly admixed populations.

Within human populations, geographic genetic structure is usually investigated relative to the genetic anthropology of a region or to prevent confounding issues in genetic epidemiology studies, as having an unbalanced proportion of cases or controls from a genetic subgroup could lead to spurious associations [Marchini et al., 2004; Thomas et al., 2002; Wacholder et al., 2002; Cardon et al., 2001; Setakis et al., 2006]. Europe has been a popular target for genetic anthropology studies, demonstrating a complex genetic history and multiple migrations from Northern Africa and the Mediterranean, though the continent

shows general geographic genetic patterns being stratified on north to southeast and west to east axes [Lefevre-Witier et al., 2006; Bosch et al., 2000; Crawford 2007]. These and many other exciting studies have helped to increase and clarify our understanding of historical events in conjunction with other forms of anthropological data. Notably missing is research of smaller geographic regions or those focused on more urban, highly admixed populations.

Some genetic epidemiology studies have examined more subtle differences in geographic genetic structure. The Wellcome Trust Case Control Consortium (WTCCC) conducted a genome-wide association study of 14,000 cases and 3,000 controls which were mostly Caucasians living in the British Isles. They found 13 genomic regions with allele frequencies that varied significantly over 12 geographic regions, most showing a northwest to southeast axis [Wellcome Trust Case Control Consortium 2007]. Another study conducted in Iceland found differences in geographic ancestry between age-defined cohorts, with the older cohorts demonstrating more local ancestry, as well as low levels of genetic structure (F_{ST} 's around 0.00017-0.00338)[Helgason et al., 2005]. Therefore, even in well-characterized, presumably homogeneous regions, there may still be structure that can be linked to geography. That structure may also be related to the age of the population, as well as the rurality of the region. Consideration should also be taken regarding the type of genetic data collected when looking for structure; for example, in a study of Finland, genetic differences were detected between Eastern and Western Finns only when Y chromosomal data was used, possibly due to regionally different sources of male gene flow [Lappalainen et al., 2006].

Another interesting consideration is that of time scale. Over evolutionary time, due to a variety of pressures, localized variation in allele frequencies results in the formation of subpopulations. These pressures generally happen in a certain geographic region over a certain period of time, whether during the advent of humans, or perhaps even more recent subdivisions within a new burgeoning city. Therefore, genetic structure is inherently tied to both geography and time (i.e. history). Many environmental influences are also tied to geography, and can be stable or fluctuate rapidly over time, but interact with susceptibility genes that were actually generated over long periods of time. Therefore, the genetic component of an ecogeographic epidemiology study mixes susceptibility genes that may be ancient in origin but were placed in geographical regions by more recent historical events, and therefore interact with environmental influences that they may or may not be well-adapted for.

Geographic Genetic Epidemiology

Although human geographic genetic patterning has been demonstrated in numerous instances, the connection between genetic landscape and spatial disease patterning is only beginning to be established. More inroads are being made regarding geographical genetics of pathogens than human disease susceptibility, which is not unexpected given the long ecological interest in geographic genetics and landscape epidemiology of parasite/environment interactions. Landscape can affect the genetics of pathogens, and the need to include geography in the analysis of infectious disease has been previously recognized [Real et al., 2007]. Recent examples include the geographic genetic epidemiology of human H5N1 avian flu infections. One study analyzed the virus's mutation pattern and reported local cases over the course of a decade [Janies et al., 2007]. One of the most innovative parts of this study was that it used interactive maps with Google Earth to investigate the spread of the virus and its mutation patterns over space. The group identified more virulent genotypes of the virus while demonstrating that reasons for the spread of avian flu were geographically and temporally context dependent (Figure 2). Malaria and Leishmania are also popular targets since they have such a complex life cycle, and effect such large populations in the

developing world [Antonio-Nkondjio et al., 2008; Al-Jawabreh et al., 2008; Gardella et al., 2008]. Recently, the genetic population structure of *Anopheles gambiae* and *Plasmodium falciparum* were investigated not only across time, but across space as well. The study found little difference between the groups and suggested that this may be a reason that antimalarial resistance spreads so quickly through Sub-Saharan Africa [Prugnolle et al., 2008]. Although these studies do not fully incorporate the interplay between geography, host, and pathogen, they do demonstrate that geography and genetics considered together are important in studies of human infectious disease. Such studies would be more powerful if they incorporated human geographical genetics because infectious disease mutation and virulence may be tied to the susceptibility of the host population.

Despite the few examples to date, human studies of the geographic genetic epidemiology of complex diseases are advancing. One such study occurred in Finland, where researchers have observed a geographic difference in multiple sclerosis (MS), with higher rates in the region of Ostrobothnia (located on the Southwest coast). Genetic markers present in those with ancestry from Ostrobothnia associated with higher susceptibility to MS in two datasets. This may be due to a founder effect as Vikings historically invaded southern Finland [Lappalainen et al., 2006]. Another example focused on the geographic genetic epidemiology of cancer in central Europe. Sokal et al. have published three papers investigating and developing methods for combining ethnohistoric, genetic and geographic data with incidence and mortality rates into a model for 31-42 cancer types. They used distance calculations followed with path analyses and subsequent k-means clustering to determine the pairwise correlations between variables, or the effect of ethnohistory, genes and geography on incidence and mortality [Sokal et al., 1997; Sokal et al., 2004; Sokal et al., 2000]. These represent some of the largest steps toward true ecogeographic genetic epidemiology research.

Ecogeographic Genetic Epidemiology

Full studies of the spatial interaction between genes and environment will substantially increase our understanding of the root causes of human disease and thus lead to the development of effective preventative measures and treatment. One can imagine the analogy of a movie projection, where three colors (red, green, and blue) are required to give the full picture, though one or two colors can provide the figures' basic outline. Each information source (genetics, environment and spatial disease patterns) can be considered a color in our projection that each provide useful information, but provide a much more comprehensive view of the total etiology when displayed together. With the variety and power of tools currently available in spatial statistics, geographical genetics, and landscape epidemiology, researchers can undertake these synergistic and interdisciplinary projects with increased efficacy. The final goal will be to extract both genetic and environmental information from a landscape and look for interactions that result in local disease rates. The major steps in reaching that goal will be to layer maps of genetic information and environment to perform subsequent statistical analyses and estimate the contribution (parameters) of each data layer to the disease outcome (Figure 3). For instance, smoothed layers showing pockets of genetic structure could be layered with data showing bedrock concentrations of uranium, power plants, farming communities, and sites of potential groundwater pollution that could then be related to spatial incidence of bladder cancer.

Environmental data are collected by a multitude of businesses and government agencies as well as academic departments in geography, epidemiology, geology, public health and medicine. Data can be and are collected in many different forms, such as complete lists of the locations of superfund sites or farms, or data such as air quality that result from sampling. One promising technology is remote sensing, in which measurements of

environmental variables are taken over geographic areas via airborne sensors [De La Rocque et al., 2004].

How one combines these data is still not completely clear, but a promising method for integrating data layers is geographically weighted regression (GWR). GWR is a set of methods in which a variables' contribution to an outcome can be estimated relative to its location, i.e. that parameters are allowed to vary, rather than determining and using a single global estimate for each parameter. A typical regression equation used to calculate global parameter estimates is as follows:

$$(1) \quad y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i$$

where y is the dependent variable (disease rate at i), β is the parameter being estimated for the independent variable x (environmental or genetic layer) and ε is a measure of error. Such analyses can make the estimates of disease risk much more meaningful. For instance, high disease rates may be due to genetic susceptibility in one location and an environmental variable in another location. Therefore, we can modify the regression equation to allow for local variation in the parameters as follows:

$$(2) \quad y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i$$

where (u_i, v_i) is the geographic location of point i . This is the fundamental equation for GWR. There are several methods for assigning weights to location, ranging in usage and complexity [Fotheringham et al., 2002].

In order to understand the development of chronic disease it will be necessary to collect data not only reflecting spatial variation, but across time as well. Human populations and environmental exposures are both highly dynamic. Most (if not all) current human geographical genetics studies do not measure time, but are “snapshots” of the genetic makeup of countries or continents at a single point in time. As mechanisms for genotyping improve and become less expensive, however, this could change. Rather than treating a population as a cohort, investigators will be able to treat a geographic region as a cohort; beginning at a single time point and sampling genetic structure and environmental variables over time to see what patterns consistently result in high incidence. This would be particularly useful if the onset of the disease occurs significantly after the environmental exposure, as is the case with most complex diseases, or if the population of an area is in constant flux. For instance, exposure to pollution may change with seasons, wind currents and local industry. Therefore, comparing maps that represent pollution levels, genetic structure and asthma rates at different time points would give a more complete picture of what gene-environment combinations are likely to result in asthma than a set of maps from a single time point. Estimation of lifetime exposure to certain spatial environmental variables is possible, even taking into account differences in time spent at home and work. Meliker et al. introduced a space-time information system (STIS) and applied it to individual exposure to arsenic based on GIS-estimated arsenic at former residences and places of employment for 440 individuals from a case-control study. They determined that there was a great deal of fluctuation for most individuals during their adult lives, and that their model was both highly flexible and effective [2007]. Of course the limitation to this is that investigators must become accustomed to collecting far more geographical data than they may be accustomed to doing, but this could be as simple as altering a survey instrument which is far easier than beginning a prospective cohort study.

It is intuitive to genetic epidemiologists to collect genetic data on candidate genes specifically related to the disease being studied. Genetic substructure that is discerned based on known susceptibility genes would more likely yield strong hypotheses regarding specific diseases. If clustering is based on known susceptibility genes, then investigators could more directly study not only that structure groups show different responses to environment, but begin to narrow down which alleles specifically are interacting. Along with these genetic data, researchers should collect individual ancestry information that might help identify different genetic groups and may substantiate structure findings.

Merging environmental and genetic information in multiple layers remains largely untested and there are several analytical challenges to overcome. The major challenge remains the determination of the spatial variation of genetic structure. Additional statistical challenges include environmental data layers that are likely to be measured on a different scale than genetic information, but still must be integrated into disease models with genetic layers. We suggest that forming business, government and academic collaborations will be key to obtaining as many appropriate environmental data layers as possible. Data that are thorough and especially those taken at many time points may be difficult to come by, though certain government data such as cancer registries and environmental databases such as superfund sites are fairly comprehensive. Naturally, these are already ongoing issues in traditional epidemiologic studies.

The above long term vision for ecogeographic genetic epidemiology may be fulfilled by meeting several more short term goals such as investigations of the most appropriate methods for mapping genetic information and integrating environmental and genetic data layers. Certainly the infrastructure for such an undertaking is now in place, but resources must be turned in this direction. The statistical and display capabilities of GIS as well as tools like Google Earth have the power to be the primary venues for revolutionizing spatial genetic epidemiology. Highly collaborative work between ecologists, genetic epidemiologists, geographers and statisticians will yield successful applications of current methods as well as the development of new ones that will certainly be critical to this undertaking.

In the future there will hopefully be more studies integrating human geographical genetics and landscape epidemiology, making full use of existing methods and computational tools while also exploring new approaches. Environmental influences can then be studied in relation to the actual biological entities on which they are acting, and potentially yield dramatic new insights into the causes of human disease.

Acknowledgments

The authors would like to thank Daniel Janies for his assistance in supplying Figure 2 along with a caption. This work was funded in part by National Institute of Health grants LM009012, AI59694 and RR018787.

References

- Al-Jawabreh A, Diezmann S, Muller M, Wirth T, Schnur LF, Strelkova MV, Kovalenko DA, Razakov SA, Schwenkenbecher J, Kuhls K, Schonian G. Identification of geographically distributed sub-populations of *Leishmania* (*Leishmania*) major by microsatellite analysis. *BMC Evol.Biol.* 2008; 8:183. [PubMed: 18577226]
- Antonio-Nkondjio C, Ndo C, Kengne P, Mukwaya L, Awono-Ambene P, Fontenille D, Simard F. Population structure of the malaria vector *Anopheles moucheti* in the equatorial forest region of Africa. *Malar J.* 2008; 7:120. [PubMed: 18601716]

- Banks SC, Lindenmayer DB, Ward SJ, Taylor AC. The effects of habitat fragmentation via forestry plantation establishment on spatial genotypic structure in the small marsupial carnivore, *Antechinus agilis*. *Mol.Ecol.* 2005; 14:1667–1680. [PubMed: 15836641]
- Barbujani G. Geographic patterns: how to identify them and why. *Hum.Biol.* 2000; 72:133–153. [PubMed: 10721615]
- Barbujani G, Jacquez GM, Ligi L. Diversity of some gene frequencies in European and Asian populations. V. Steep multilocus clines. *Am.J.Hum.Genet.* 1990a; 47:867–875. [PubMed: 2220827]
- Barbujani G, Sokal RR. Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc.Natl.Acad.Sci.U.S.A.* 1990b; 87:1816–1819. [PubMed: 2308939]
- Besag J, Green PJ. Spatial Statistics and Bayesian Computation. *J Royal Stat Soc.* 1993; 55:25–37.
- Bohonak AJ. IBD (Isolation by Distance): a program for analyses of isolation by distance. *J.Hered.* 2002; 93:153–154. [PubMed: 12140277]
- Bosch E, Calafell F, Perez-Lezaun A, Clarimon J, Comas D, Mateu E, Martinez-Arias R, Morera B, Brakez Z, Akhayat O, Sefiani A, Hariti G, Cambon-Thomsen A, Bertranpetit J. Genetic structure of north-west Africa revealed by STR analysis. *Eur.J.Hum.Genet.* 2000; 8:360–366. [PubMed: 10854096]
- Cardon LR, Bell JI. Association study designs for complex diseases. *Nat.Rev.Genet.* 2001; 2:91–99. [PubMed: 11253062]
- Cercueil A, Francois O, Manel S. The genetical bandwidth mapping: a spatial and graphical representation of population genetic structure based on the Wombling method. *Theor.Popul.Biol.* 2007; 71:332–341. [PubMed: 17353024]
- Chen C, Durand E, Forbes F, Francois O. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Mol.Ecol.Notes.* 2007; 7:747–756.
- Clauss MJ, Mitchell-Olds T. Population genetic structure of *Arabidopsis lyrata* in Europe. *Mol.Ecol.* 2006; 15:2753–2766. [PubMed: 16911198]
- Coltman DW, Pilkington JG, Pemberton JM. Fine-scale genetic structure in a free-living ungulate population. *Mol.Ecol.* 2003; 12:733–742. [PubMed: 12675828]
- Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat.Genet.* 2006; 38:1251–1260. [PubMed: 17057719]
- Corander J, Waldmann P, Martinen P, Sillanpaa MJ. BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics.* 2004; 20:2363–2369. [PubMed: 15073024]
- Coulon A, Guillot G, Cosson JF, Angibault JM, Aulagnier S, Cargnelutti B, Galan M, Hewison AJ. Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. *Mol.Ecol.* 2006; 15:1669–1679. [PubMed: 16629819]
- Crawford MH. Genetic structure of circumpolar populations: a synthesis. *Am.J.Hum.Biol.* 2007; 19:203–217. [PubMed: 17286261]
- De La Rocque S, Michel V, Plazenet D, Pin R. Remote sensing and epidemiology: examples of applications for two vector-borne diseases. *Comp.Immunol.Microbiol.Infect.Dis.* 2004; 27:331–341. [PubMed: 15225983]
- Degen B, Petit R, Kremer A. SGS--Spatial Genetic Software: a computer program for analysis of spatial genetic and phenotypic structures of individuals and populations. *J.Hered.* 2001; 92:447–449. [PubMed: 11773257]
- Doligez A, Baril C, Joly HI. Fine-scale spatial genetic structure with nonuniform distribution of individuals. *Genetics.* 1998; 148:905–919. [PubMed: 9504936]
- Enoch MA. Genetic and environmental influences on the development of alcoholism: resilience vs. risk. *Ann.N.Y.Acad.Sci.* 2006; 1094:193–201. [PubMed: 17347351]
- Epperson, BK. *Geographical Genetics*. Princeton, New Jersey; Princeton University Press; 2003.
- Epperson BK, Li T. Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proc.Natl.Acad.Sci.U.S.A.* 1996; 93:10528–10532. [PubMed: 8927628]

- Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*. 2003; 164:1567–1587. [PubMed: 12930761]
- Fotheringham, AS.; Brunsdon, C.; Charlton, M. Geographically Weighted Regression, the analysis of spatially varying relationships. John Wiley & Sons, LTD.; University of Newcastle, UK: 2002.
- François O, Ancelet S, Guillot G. Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*. 2006; 174:805–816. [PubMed: 16888334]
- Gardella F, Assi S, Simon F, Bogreau H, Eggelte T, Ba F, Foumane V, Henry MC, Kientega PT, Basco L, Trape JF, Lalou R, Martelloni M, Desbordes M, Baragatti M, Briolant S, Almeras L, Pradines B, Fusai T, Rogier C. Antimalarial drug use in general populations of tropical Africa. *Malar J*. 2008; 7:124. [PubMed: 18611279]
- Guillot G, Mortier F, Estoup A. Geneland: a computer package for landscape genetics. *Mol.Ecol.Notes*. 2005; 5:712–715.
- Haggstrom O, Van Lieshout, Marie-Colette NM, Moller J. Characterization results and Markov chain Monte Carlo algorithms including exact simulation for some spatial point processes. *Bernoulli*. 1999; 5:641–658.
- Haiman CA, Stram DO, Wilkens LR, Pike MC, Kolonel LN, Henderson BE, Le Marchand L. Ethnic and racial differences in the smoking-related risk of lung cancer. *N.Engl.J.Med*. 2006; 354:333–342. [PubMed: 16436765]
- Handley LJ, Manica A, Goudet J, Balloux F. Going the distance: human population genetics in a clinal world. *Trends Genet*. 2007; 23:432–439. [PubMed: 17655965]
- Hardy OJ, Vekemans X. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. 2002; 2:618–620.
- Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K. An Icelandic example of the impact of population structure on association studies. *Nat.Genet*. 2005; 37:90–95. [PubMed: 15608637]
- Isaaks, EH.; Srivastava, RM. An Introduction to Applied Geostatistics. Oxford University Press; New York: 1989.
- Janies D, Hill AW, Guralnick R, Habib F, Waltari E, Wheeler WC. Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). *Syst.Biol*. 2007; 56:321–329. [PubMed: 17464886]
- Jump AS, Penuelas J. Genetic effects of chronic habitat fragmentation in a wind-pollinated tree. *Proc.Natl.Acad.Sci.U.S.A*. 2006; 103:8096–8100. [PubMed: 16698935]
- Jusatz HJ. The importance of biometeorological and geomedical aspects in human ecology. *Int.J.Biometeorol*. 1966; 10:323–334. [PubMed: 5978010]
- Keeling MJ. The effects of local spatial structure on epidemiological invasions. *Proc.Biol.Sci*. 1999; 266:859–867. [PubMed: 10343409]
- Lappalainen T, Koivumaki S, Salmela E, Huoponen K, Sistonen P, Savontaus ML, Lahermo P. Regional differences among the Finns: a Y-chromosomal perspective. *Gene*. 2006; 376:207–215. [PubMed: 16644145]
- Lefevre-Witier P, Aireche H, Benabadji M, Darlu P, Melvin K, Sevin A, Crawford MH. Genetic structure of Algerian populations. *Am.J.Hum.Biol*. 2006; 18:492–501. [PubMed: 16788894]
- Liu H, Prugnolle F, Manica A, Balloux F. A geographically explicit genetic model of worldwide human-settlement history. *Am.J.Hum.Genet*. 2006; 79:230–237. [PubMed: 16826514]
- Manel S, Schwartz MK, Luikart G, Taberlet P. Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol.Evol*. 2003; 18:189–197.
- Manica A, Prugnolle F, Balloux F. Geography is a better determinant of human genetic differentiation than ethnicity. *Hum.Genet*. 2005; 118:366–371. [PubMed: 16189711]
- Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat.Genet*. 2004; 36:512–517. [PubMed: 15052271]
- Matheron G. Principles of Geostatistics. *Econ Geol*. 1963; 58:1246–1256.
- May, JM. The Ecology of Human Disease. MD Publications Inc.; New York: 1958.
- Meade, MS.; Erickson, RJ. Medical Geography. The Guilford Press; New York: 2000.

- Meliker JR, Slotnic MJ, Avruskin GA, Kaufmann A, Fedewa SA, Goovaerts P, Jacquez GJ, Nriagu JO. Individual lifetime exposure to inorganic arsenic using a space-time information system. *Int Arch Occup Environ Health*. 2007; 80:184–197. [PubMed: 16897097]
- Nason J. FijAnal. A Computer program for the analysis of spatial autocorrelation. 1997
- Pavlovsky, E. *Natural Nidality of Transmissible Diseases, with special reference to the landscape epidemiology of zoonanthroponoses*. University of Illinois Press; Urbana: 1966.
- Peakall R, Smouse PE. GENALEX6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol.Ecol.Notes*. 2006; 6:288–295.
- Piazza A, Menozzi P, Cavalli-Sforza L. The Making and Testing of Geographic Gene-Frequency Maps. 1981; 37:635–659.
- Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000; 155:945–959. [PubMed: 10835412]
- Prugnolle F, Durand P, Jacob K, Razakandrainibe F, Arnathau C, Villarreal D, Rousset F, de Meeus T, Renaud F. A comparison of *Anopheles gambiae* and *Plasmodium falciparum* genetic structure over space and time. *Microbes Infect*. 2008; 10:269–275. [PubMed: 18321750]
- Real LA, Biek R. Spatial dynamics and genetics of infectious diseases on heterogeneous landscapes. *J.R.Soc.Interface*. 2007
- Rebbeck TR, Khoury MJ, Potter JD. Genetic association studies of cancer: where do we go from here? *Cancer Epidemiol.Biomarkers Prev*. 2007; 16:864–865. [PubMed: 17507606]
- Ripley BD, Kelly FP. Markov Point Processes. *J. London Math. Soc*. 1977; 15:188–192.
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. *Science*. 2002; 298:2381–2385. [PubMed: 12493913]
- Schwartz GL, Turner ST, Sing CF. Association of genetic variation with interindividual variation in ambulatory blood pressure. *J.Hypertens*. 1996; 14:251–258. [PubMed: 8728304]
- Setakis E, Stirnadel H, Balding DJ. Logistic regression protects against population structure in genetic association studies. *Genome Res*. 2006; 16:290–296. [PubMed: 16354752]
- Sing CF, Haviland MB, Reilly SL. Genetic architecture of common multifactorial diseases. *Ciba Found.Symp*. 1996; 197:211–29. discussion 229–32. [PubMed: 8827376]
- Sokal RR, Oden NL. Spatial autocorrelation in biology, 1: methodology. *Biol. J. Linn. Soc*. 1978a; 10:199–228.
- Sokal RR, Oden NL. Spatial autocorrelation in biology, 2: some biological implications and four applications of evolutionary and ecological interest. *Biol. J. Linn. Soc*. 1978b; 10:229–249.
- Sokal RR, Oden NL, Rosenberg MS, DiGiovanni D. Ethnohistory, genetics, and cancer mortality in Europeans. *Proc.Natl.Acad.Sci.U.S.A*. 1997; 94:12728–12731. [PubMed: 9356518]
- Sokal RR, Oden NL, Rosenberg MS, Thomson BA. A new protocol for evaluating putative causes for multiple variables in a spatial setting, illustrated by its application to European cancer rates. *Am.J.Hum.Biol*. 2004; 16:1–16. [PubMed: 14689511]
- Sokal RR, Oden NL, Rosenberg MS, Thomson BA. Cancer incidences in Europe related to mortalities, and ethnohistoric, genetic, and geographic distances. *Proc.Natl.Acad.Sci.U.S.A*. 2000; 97:6067–6072. [PubMed: 10823950]
- Storfer A, Murphy MA, Evans JS, Goldberg CS, Robinson S, Spear SF, Dezzani R, Delmelle E, Vierling L, Waits LP. Putting the “landscape” in landscape genetics. *Heredity*. 2007; 98:128–142. [PubMed: 17080024]
- Takahashi M. PSAwinD version 1.1.1: a program for calculating spatial indices. *J.Hered*. 2003; 94:267–270. [PubMed: 12816970]
- Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. *Genet.Epidemiol*. 2005; 28:289–301. [PubMed: 15712363]
- Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol.Biomarkers Prev*. 2002; 11:505–512. [PubMed: 12050090]
- Tobler WR. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ Geog*. 1970; 46:234–240.
- Vekemans X, Hardy OJ. New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol.Ecol*. 2004; 13:921–935. [PubMed: 15012766]

- Wacholder S, Rothman N, Caporaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol.Biomarkers Prev.* 2002; 11:513–520. [PubMed: 12050091]
- Waller, LA.; Gotway, CA. *Applied Spatial Statistics for Public Health Data.* John Wiley & Sons, Inc.; Hoboken, NJ: 2004.
- Wartenberg D. SAAP- a spatial autocorrelation analysis program. 1989
- Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007; 447:661–678. [PubMed: 17554300]
- Wright S. Genetical structure of populations. *Nature.* 1950; 166:247–249. [PubMed: 15439261]
- Wright S. Isolation By Distance. *Genetics.* 1943; 28:114–138. [PubMed: 17247074]

Figure 1.

GIS data can be displayed in several different formats. The most appropriate format for any particular dataset is dependent on how the data are measured and the analyses conducted. There are a variety of different types of environmental layers that can be overlain and integrated with GIS. Pictured above are several examples of data layers and display methods in the state of NH. Namely Superfund Sites on the final National Priorities List (point data), Uranium levels (proportional symbols), groundwater locations (polygons), and two examples using lung cancer rates (one displayed regionally, i.e. by county, and one spatially interpolated). Lung cancer rate data were obtained from the Cancer Registry of New Hampshire, and environmental data from the New Hampshire Department of Environmental Services.

Figure 2.

Screenshot from Janies et al., (2007) depicting the spread of the genotype Lys-627 in the Polymerase Basic 2 protein of H5N1 in 351 Influenza strains [Janies et al., 2007]. The phylogenetic tree is based on full genome analysis of 351 isolates of H5N1 collected between 1996 and 2006. The genotype increases the ability of the virus to replicate in mammalian hosts. Earth background image sources: Google, TerraMetrics, and NASA.

Figure 3.

The field of ecogeographic epidemiology seeks to explore a landscape, including the genetic susceptibilities in hosts, their local environmental exposures, and the spatial patterning of disease rates.

Software Tools for Detecting Spatial Genetic Structure

Table 1

There are several tools available for investigating geographical genetics. They each have their focus and thereby their advantages and disadvantages. Researchers should consider carefully before choosing software appropriate to their needs.

	Notes on Operation	F _{ST}	Isolation by Distance	Spatial Autocorrelation	PCA	Geographic Information	Other Features
FijAnal (Nason 1997)	Macintosh			(see other features)			Kinship coefficients instead of Moran's I, Geary's C
GENALEX6 (Peakall & Smouse 2006)	Windows Excel		Mantel test	Several varieties for different sample sizes	X		Large variety of tests, including AMOVA, Nei's genetic distance etc.
GENBMAP (Cercueil 2007)						Wombling with output maps	
Geneland (Guillot 2005)	Linux, Macintosh, Windows	X				Poisson-Voronoi tessellation geographic display	Graphical displays of run information
IBD (Bohonak 2002)	Macintosh and Windows	X	Mantel test				Allele count, heterozygosity, RMA analysis
PSAwinD (Takahashi 2003)	Windows			Moran's I (for different layers of data)			NAC, SND, distance classes
SAAP (Wartenberg 1989)				Moran's I, Geary's C			Euclidian or spherical distances
SGS (Degen 2001)	Windows	X		Moran's I, Geary's C			Genetic, city-block and Tanimoto distances
SPAGeDi (Hardy & Vekemans, 2004)	Windows	X				1-3 spatial coordinates in input	Kinship and relatedness coefficients, F and R-statistics, etc.
TESS (Chen et al., 2007)	Windows, Command Line					Assigns spatial clusters	Conducts similar analyses to GENECLUST

NAC= number of alleles in common

SND= standard normal deviate