# Genetic Variations in Human Populations: A Continuum Approach

Amir Levy
*Departments of Physics, Massachusetts Institute of Technology,Cambridge, Massachusetts 02139 USA*

(Dated: May 18, 2018)

Small genetic variations between different human populations are often described by Wright's "infinite island" model, where populations are assumed to be isolated. In reality, genetic differences are related to geographical distances, and are not captured in Wright's island model. We extend the island model to a continuum model, where local diffusion and long-ranged migration lead to a generalized Langevin equation. We find closed-form expressions for genetic variations in arbitrary dimensions, as well as analytical solutions in 1D. Finally, we compare our against publicly available allele frequency data. Fitting our model to the data, we estimate the range of diffusion as well as the rate of migration, and obtain reasonable results.

## INTRODUCTION

A common measure of genetic distance between groups is called $F_{st}$, or Fixation Index[1, 2]. The fixation index is defined as the ratio of the variance of allele frequency between the two sub-populations, and the variance of the allele frequency in the entire population. This measure was introduced by Wright in the context of animal inbreeding, but is widely used today for human population as well [3]. If sub-populations are almost isolated, with a small probability of migration ($m$), the fixation index has a simple formula: $F_{\rm st} = 1/(2mN + 1)$, where $N$ is the sub-population size (for a diploid population, the $N$ should be replaced by $2N$). An important assumption of the model is that the population is infinite, which guarantees that the allele frequency remains constant over time, and neither allele takes over the population.

Wright's infinite island model does not take into account any geographical aspect - the tendency to migrate to nearby areas. In a commonly used extension, the "stepping stone" model, sub-populations (or colonies) are placed on a lattice, and higher rates of migrations are allowed between neighbouring sites [4]. The stepping-stone model predicts not only the variance between populations, but also show how correlations decay with distance. Indeed, studies of human allele frequencies in different populations agree with this trend: genetic differences increase with geographic distance [3].

Yet, the interpretation of the genetic data remains controversial [5]. There are many factors that contribute to genetic differences and are not accounted for in the simplified models. For example, ecological and climate variations can favor different allele in different places[6, 7] (sickle cell anemia is a prominent example). Mass migrations are another leading cause for genetic diversity [8, 9]. Nevertheless, random migrations is still a major force in shaping the spatial pattern we observe.

The simplifications of the popular "infinite island" and "stepping stone" models makes it difficult to directly compare them to data, and to extract meaningful and quantitative information. For example, studies show that genetic differences are varying smoothly in large areas[5, 9, 10]. A discrete set of "sub-population" is therefor somewhat artificial. Moreover, the "populations" defined in research today ("Mexicans","Italians","Druze", etc...) may not correspond to the "stepping stone" idea of a sub-populations.

To address this problem, we modify the stepping-stone model to a continuum description of individuals, as oppose to colonies. We solve for the distribution of alleles in time and space, and then artificially construct "sub-populations" by grouping individuals that share the same area. Interestingly, we find that these artificial sub-populations behave as isolated islands, if this area is sufficiently large. Closed form expressions are obtained for the genetic variations and spatial correlations at any radius and any dimension. It worth noting that similar approaches appear in the literature[11], but they are not commonly used, and seem to be less general.

We compare our model to both simulations and to publicly available genetic data from 50 sub-populations from Africa, Asia and Europe[12]. By fitting our model to real human data, we extract several parameters (such as migration rate) that can potentially shed light on migration patterns in our history.

## WRIGHT'S INFINITE ISLAND MODEL

Genetic drift in small isolated populations leads to variations in allele frequencies. Let us consider only random drift, without mutation or selection, in a single locus with two possible alleles. Eventually, random selection will lead one of the alleles to fixate in the population. However, if the small population (the "island") is not completely isolated, but weakly connected to a much larger population by migration, new mutations are constantly introduced into the population. At equilibrium, the mean allele frequency in the island has to equal the larger population allele frequency. The variance of the allele frequency depends on the migration rate, and can serve as a measure of the isolation of the island.

Wright proposed this model in 1931 [13]. We will present here a different derivation, based on the Langevin equation, that will be useful in the extension for a more realistic continuum model. Let us assume that one of the alleles ($A$) appears in the island population at time $t$ with probability $p(t)$. In the general population (outside of the island), the $A$ allele appears with probability $p_0$. In the spirit of the Moran process [14], we assume generations are discrete, and at each generation individuals are randomly replaced by an individual from the previous generation. At the new generation, the fraction of $A$s can change either due to migration (individuals replaced by immigrants that carry $A$ with probability $p_0$) or genetic drift (noise):

$$p(t+dt) = (1-m)p(t) + mp_0 + \eta(t), \tag{1}$$

where $m$ is the fraction of the population that is replaced by migrants, and $\eta$ is a noise term. To estimate the noise term, we neglect variations in immigrants population. For random selection the number of $A$ alleles is distributed binomially. Hence, the variance in the fraction of $A$s in the new generation equals $p(1-p)/N$, and the noise term is characterized by:

$$\langle \eta(t)\eta(t')\rangle = \delta(t-t')\frac{p(t)(1-p(t))}{N} \tag{2}$$

We can recast Eq. 1 as a Langevin equation for $\delta p = p - p_0$:

$$\frac{d\delta p}{dt} = -m\delta p + \eta(t). \tag{3}$$

The Langevin equation is a simple 1D ODE, which has a closed form solution[15]. For convenience, let us assume the island original population had the same allele frequency as the general population, $\delta p(0) = 0$:

$$\delta p(t) = \int_0^t \mathrm{e}^{-m(t-\tau)}\eta(\tau)d\tau. \tag{4}$$

The variance in the allele frequency of the population reads:

$$\begin{aligned}\langle \delta p(t)^2\rangle &= \int_0^t\int_0^t \mathrm{e}^{-m(2t-\tau-\tau')}\langle\eta(\tau)\eta(\tau')\rangle d\tau\tau' \\ &= \frac{1}{N}\int_0^t \mathrm{e}^{-2m(t-\tau)}\langle p(\tau)(1-p(\tau))\rangle d\tau \\ &= \frac{p_0(1-p_0)}{2mN}\left[1-\mathrm{e}^{-2mt}\right] \\ &- \frac{1}{N}\int_0^t \mathrm{e}^{-2m(t-\tau)}\langle\delta p(\tau)^2\rangle d\tau. \end{aligned} \tag{5}$$

This implicit equation is easily solved in the $t \to \infty$ limit. The second term averages over values of $\langle\delta p(\tau)^2\rangle$ from the last $1/2m$ generations. After many generations the variance is fixed and we can take $\langle\delta p^2\rangle$ out of the integration. Solving the the variance yields:

$$\langle\delta p(t)^2\rangle = \frac{p_0(1-p_0)}{2mN+1} \tag{6}$$

The fixation index is the ratio of the sub-population variance the total variance, $p_0(1-p_0)$. Hence, we recover Wright's well-known result [1]:

$$F_{\mathrm{st}} = \frac{1}{2mN+1}. \tag{7}$$

## THE DIFFUSION-MIGRATION MODEL

We now extend the infinite island model to a account for local "diffusion" (replacement by nearby individuals) as well as long-range migration. Correspondingly, the allele frequency function, $p(\mathbf{x},t)$ has an $\mathbf{x}$ dependence. At each generation, each individual is either replaced by one individual from it's environment, or by a migrant from the entire population. Similarly to the island model, the overall population has a fixed allele frequency $p_0$. Hence, the allele frequency at the new generation reads:

$$\begin{aligned}p(\mathbf{x},t+dt) = (1-m)A_L^{-1}\int_{|\mathbf{x}-\mathbf{x}'|<L} p(\mathbf{x}-\mathbf{x}',t)\mathrm{d}^d\mathbf{x}' \\ + mp_0 + \eta(\mathbf{x},t).\end{aligned} \tag{8}$$

Where $L$ and $A_L$ are the distance and area for local replacements, correspondingly, and where $d$ is the dimensionaly of the system.. Assuming $p(x,t)$ is a slowly varying function in space, we can Taylor expand $p(\mathbf{x}-\mathbf{x}',t)$ around $\mathbf{x}$, and get:

$$\begin{aligned}&A_L^{-1}\int_{|\mathbf{x}-\mathbf{x}'|<L} p(\mathbf{x}-\mathbf{x}',t)dx' \approx \\ &\approx p(\mathbf{x},t) + \sum \frac{\partial^2 p(\mathbf{x},t)}{\partial x_i \partial x_j} A_L^{-1}\int_{|\mathbf{x}-\mathbf{x}'|<L} x_i' x_j' \\ &= p(\mathbf{x}) + \frac{L^2}{d+2}\nabla^2 p(\mathbf{x},t)'.\end{aligned} \tag{9}$$

We again define $\delta p = p - p_0$, and find the generalized Langevin equation:

$$\frac{\partial\delta p(x,t)}{\partial t} = R^2\nabla^2\delta p(x,t) - m\delta p(x,t) + \eta(x,t), \tag{10}$$

where $R = \sqrt{(1-m)/(d+2)}L$. The noise term, $\eta(\mathbf{x},t)$ describes the genetic drift at each point in time and space. If each individual takes an area of $a^d$ the noise correlation function is:

$$\begin{aligned}\langle\eta(x,t)\eta(x',t')\rangle &= a^d\delta(t-t')\delta(x-x')p(x,t)(1-p(x,t)) \\ &\approx a^d\delta(t-t')\delta(x-x')p_0(1-p_0),\end{aligned} \tag{11}$$

where we make the simplifying assumption that the allele frequency $p(x,t)$ is close enough to $p_0$ so we can ignore the small variations. In the island model, for example, this simplification corresponds to $mN \gg 1$. The Fourier transform of the noise, $\eta(\mathbf{q},t) = \int \mathrm{d}^d\mathbf{x}\eta(\mathbf{r},t)\exp(i\mathbf{q}\cdot\mathbf{r})$, has the following correlation function:

$$\langle\eta(\mathbf{q},t)\eta(\mathbf{q}',t')\rangle = a^d\delta(t-t')(2\pi)^d\delta^d(\mathbf{q}+\mathbf{q}')p_0(1-p_0). \tag{12}$$

Taking the Fourier transform of the Langevin equation, we find a decoupled set of equations, each similar to a 1D Langevin equation:

$$\frac{\partial \delta p(\mathbf{q},t)}{\partial t} = -[m + (Rq)^2]\delta p(\mathbf{q},t) + \eta(\mathbf{q},t). \quad (13)$$

The solution of this Langevin equation is the same as Eq. 4, and the correlation function is given by:

$$\begin{aligned} &\langle \delta p(\mathbf{q},t)\delta p(\mathbf{q}',t)\rangle = \\ &= \delta^d(\mathbf{q}+\mathbf{q}')\frac{a^d(2\pi)^d p_0(1-p_0)}{2(m+R^2q^2)}\left[1-\exp(-(m+R^2q^2)t)\right]. \end{aligned} \quad (14)$$

The time-scale for equilibrium is depends on the $q = 0$ mode, and is proportional to $1/m$ just like in the infinite island model. For shorter wavelength, equilibrium is reached much faster. Wavelengths that are shorter than $R$, for example, are negligible after a single generation. This is expected since we are mixing at each generation individuals from that distance. To find the spatial correlation function, $\langle \delta p(\mathbf{x},t)\delta p(\mathbf{x}',t)\rangle$, we use the the inverse Fourier transform: $\delta p(\mathbf{x},t) = \int \mathrm{d}^d\mathbf{q}/(2\pi)^d \delta p(\mathbf{q},t)\exp(i\mathbf{q}\cdot\mathbf{x})$ and obtain:

$$\begin{aligned} &\langle \delta p(\mathbf{x},t)\delta p(\mathbf{x}',t)\rangle = \\ &= \frac{a^d p_0(1-p_0)}{2(2\pi)^d R^2}\int \mathrm{d}^d\mathbf{q}\frac{1-\exp(-(m+R^2q^2)t}{q^2+\xi^-2}\exp(i\mathbf{q}\cdot\mathbf{x}), \end{aligned} \quad (15)$$

where the correlation length $\xi$ is defined as $R/\sqrt{m}$. For $t \gg m^{-1}$ the system reaches equilibrium, and the correlation function simplifies to:

$$\langle \delta p(0,t)\delta p(\mathbf{x},t)\rangle = \frac{a^d p_0(1-p_0)}{2(2\pi)^d R^2}\int \mathrm{d}^d\mathbf{q}\frac{\exp(i\mathbf{q}\cdot\mathbf{x})}{q^2+\xi^-2} \quad (16)$$

## EFFECTIVE FIXATION INDEX: GENETIC VARIANCE OF SUB-POPULATIONS

Now that we have found the correlation function between individuals, we can go back and explore how sub-populations behave under this model. In the island model sub-population had a very natural definition, albeit not a very realistic one. Now we need to construct a sub-population artificially. To be consistent with the island model, we build the sub-population from $N$ individuals. This adds another length-scale to the problem, the radius of the population $\lambda_{sp}^d \propto Na^d$. If the population is centered around $\mathbf{x}_{sp}$, the average allele frequency reads:

$$\bar{p}(\mathbf{x}_{sp}) = A_{sp}^{-1}\int_{|x|<\lambda_{sp}} \mathrm{d}^d\mathbf{x} p(|\mathbf{x}-\mathbf{x}_{sp}|), \quad (17)$$

where $A_{sp}$ is the area of the sub-population. The variance of sub-population allele frequency is:

$$\langle \bar{\delta p}^2\rangle = \frac{1}{A_{sp}^2}\int \mathrm{d}^d\mathbf{x}\mathrm{d}^d\mathbf{x}'\langle \delta p(\mathbf{x}-\mathbf{x}_{sp})\delta p(\mathbf{x}'-\mathbf{x}_{sp})\rangle. \quad (18)$$

Plugging in the individual correlation function (Eq. 16), we get:

$$\langle \bar{\delta p}^2\rangle = \frac{a^d p_0(1-p_0)}{2(2\pi)^d A_{sp}^{2d} R^2}\int\int \mathrm{d}^d\mathbf{x}\mathrm{d}^d\mathbf{x}'\mathrm{d}^d\mathbf{q}\frac{\exp(i\mathbf{q}\cdot(\mathbf{x}-\mathbf{x}'))}{q^2+\xi^-2} \quad (19)$$

The spatial integrals are Fourier transform of a $d$ dimensional ball, which are known. We further exploit the spherical symmetry to simplify the final expression:

$$\begin{aligned} \langle \bar{\delta p}^2\rangle &= \frac{p_0(1-p_0)}{2mN}\left(\frac{\lambda_{sp}}{\xi}\right)^2\int \frac{J_{d/2}^2(y)\mathrm{d}^d\mathbf{y}}{y^d(y^2+(\lambda/\xi)^2)}. \\ &= \frac{p_0(1-p_0)}{2mN}g_d\left(\frac{\lambda_{sp}}{\xi}\right). \end{aligned} \quad (20)$$

where $J_n(x)$ is the $n$-th Bessel function, and $g_d(\alpha)$ is defined as:

$$g_d(\alpha) = d\int_0^\infty \frac{J_{d/2}^2(y)dy}{y(1+(y/\alpha)^2)} \quad (21)$$

Consequently, the effect of diffusion on the fixation index is:

$$F_{\mathrm{st}} = F_{\mathrm{st}}^0 g_d\left(\frac{\lambda_{sp}}{\xi}\right). \quad (22)$$

The function $g_d(\alpha)$ goes from 0 to 1 for any dimension. If $\xi \ll \lambda_{sp}$ ($\alpha \to \infty$) we get back the island model fixation index (assuming also $2mN \gg 1$). In this limit, the function $g_d(\alpha)$ equals 1, following a general identity of Bessel functions.

$$\int_0^\infty \frac{J_{d/2}^2(y)dy}{y} = \frac{1}{d}. \quad (23)$$

In the opposite limit, where the correlation length is very large ($\alpha \to 0$) we find a very different result compare to the island model. In this limit the correlation length is a better measure for the effective size of the population.

### Spatial Correlations

So far we only considered the variance in allele frequency across different sub-populations. A more interesting feature we can predict from our model is spatial correlations. Exending Eq. 20 to populations that are located at different places, the correlation between the average allele frequency of populations that are located around $\mathbf{x}_{sp}$ and $\mathbf{x}'_{sp}$ is:

$$\langle \bar{\delta p}(\mathbf{x}_{sp})\bar{\delta p}(\mathbf{x}'_{sp})\rangle = \int \mathrm{d}^d\mathbf{x}\mathrm{d}^d\mathbf{x}' \langle \delta p(\mathbf{x}-\mathbf{x}_{sp})\delta p(\mathbf{x}'-\mathbf{x}'_{sp})\rangle \tag{24}$$

Substituting the individual pair correlation function (Eq. 16), the only difference compare with the variance calculation, is an additional $\exp[i\mathbf{q}\cdot(\mathbf{x}_{sp}-\mathbf{x}'_{sp}]$ term. The variance is only the 0th component of a Fourier transform. The Fourier transform can be simplified using the spherical symmetry to include a single integration. Finally, we get:

$$\langle \bar{\delta p}(\mathbf{x}_{sp})\bar{\delta p}(\mathbf{x}'_{sp})\rangle = \frac{p_0(1-p_0)}{2mN} h_d\left(\frac{r}{\lambda_{sp}};\frac{\xi}{\lambda_{xp}}\right) \tag{25}$$

where the function $h_d\left(\frac{r}{\lambda_{sp}};\frac{\xi}{\lambda_{xp}}\right)$ is defined as:

$$h_d\left(\frac{r}{\lambda_{sp}};\frac{\xi}{\lambda_{xp}}\right) =$$

$$= \Gamma\left(\frac{d}{2}\right) d \left(\frac{r}{2\lambda_{sp}}\right)^{\frac{2-d}{2}} \int_0^\infty \mathrm{d}y \frac{J_{\frac{d-2}{2}}\left(\frac{ry}{\lambda_{sp}}\right) J^2_{\frac{d}{2}}(y)}{y^d\left[1+\left(\frac{y\lambda_{sp}}{\xi}\right)^2\right]} \tag{26}$$

The function is plotted in Fig.1. for 1 and 2 dimensions. For $d=1$, the Bessel functions $J_{1/2}$ and $J_{-1/2}$ take a simple form and the integrals can be evaluated analytically. Alternatively, one can directly evaluate the Fourier transform of the correlation function for individuals (Eq. 16), and carry out the remaining integrals. The individuals correlations function is an exponentially decaying function, with decay length $\xi$. Correspondingly, the sub-populations correlation function would also decay with same exponent. If the sub-populations are not intersecting ($r > 2\lambda_{sp}$), the correlation function is particularly simple:

$$\langle \bar{\delta p}(\mathbf{x})\bar{\delta p}(\mathbf{x}')\rangle = \frac{p(1-p)}{2mN}\frac{\xi}{\lambda_{sp}}\sinh^2\left(\frac{\lambda_{sp}}{\xi}\right)\mathrm{e}^{-|\mathbf{x}-\mathbf{x}'|/\xi}. \tag{27}$$

In case there is an overlap between the populations, the expression more cumbersome:

$$\langle \bar{\delta p}(\mathbf{x})\bar{\delta p}(\mathbf{x}')\rangle = \frac{p(1-p)}{2mN}\left[1-\frac{r}{2\lambda_{sp}} - \frac{\xi\mathrm{e}^{-\lambda_{sp}/\xi}}{\lambda_{sp}}\left(\cosh\frac{r}{2\xi}\sinh\frac{r/2-\lambda_{sp}}{\xi}+\sinh\frac{\lambda_{sp}}{\xi}\sinh\frac{r}{\xi}\right)\right]. \tag{28}$$

So far we haven't specified the dimensionality of the system. Clearly, the dimension cannot be greater than 2, as human diffusion and migrations physically takes place on a two-dimensional sphere. Yet, if we consider for example a population that is located along a river bank, it might be better described as 1 dimensional, rather than 2. The general function, $h_d$, allows us to naturally interpolate between 1 and 2 dimensions and is the main result of our work. Despite its unpleasant form, $h_d$ can be easily evaluated numerically, and thus can be useful in comparison to real genetic data. Before we apply it to human populations, we test its validity in the next section by comparing it against a simulation of the Moran process in 1D.

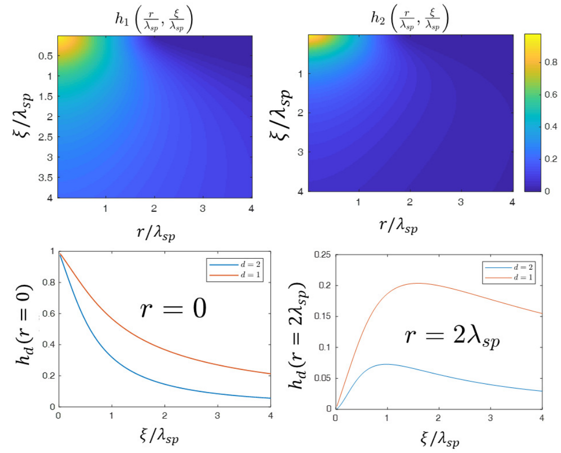

FIG. 1. The sub-population correlation function, $h_d$ (Eq. 25) illustrated for $d=1$ (top left) and $d=2$ (top right), as a function of their two variables, distance ($r$) and correlation length ($\xi$), normalized by the sub-population size ($\lambda_{sp}$). Examples for $r=0$ and $r=2\lambda_{sp}$ are shown in the bottom left and right, correspondingly. The $r=0$ limit measures the sub-populations variance, and approaches the classical wright formula at small correlation length. The $r=2\lambda_{sp}$ measures the correlations between neighbouring populations, and has a maximum when the correlation length matches the sub-population size.

## COMPARISON TO SIMULATIONS

The generalized Langevin equation (Eq. 10), which we solved in order to derive the pair correlation function for sub-populations, was based on a modified Moran process (without selection). In a Moran process, individuals in the population can carry one of two alleles. At each generation, each individual is replaced randomly by some individual from the previous generation. To include the spatial dependence, we restricted this replacement only to near by individuals. However, we also allowed for some small probability ($m$) for individuals to be replaced by the general population. In the derivation of the corresponding Langevin equation, several assumptions were made. Most notably is the existence of an equilibrium allele frequency, with only small local deviations. In general, the Moran process would eventually lead the entire population to fixate on one of the two alleles. Other assumptions included truncating the Taylor expansion of $p(t,x)$ after the second order term, and fixing the variance of the noise.

To test the validity of these assumptions, we compare our theoretical result (Eq. 25) with a simulation of Moran process. We use a system of $5 \cdot 10^5$ individuals, with different migrations rates and replacement radius of $L = 25$. The initial allele assignment is random, with probability of $p_0 = 0.5$. The data was collected after 2000 generations. Results are shown in Fig. 2, for both fixed population size as a function of distance, and varying population size at $r = 0$. While there are clear deviations between simulation data and the theoretical predictions, the approximated continuum models is able to capture most of the behavior of the full simulation.

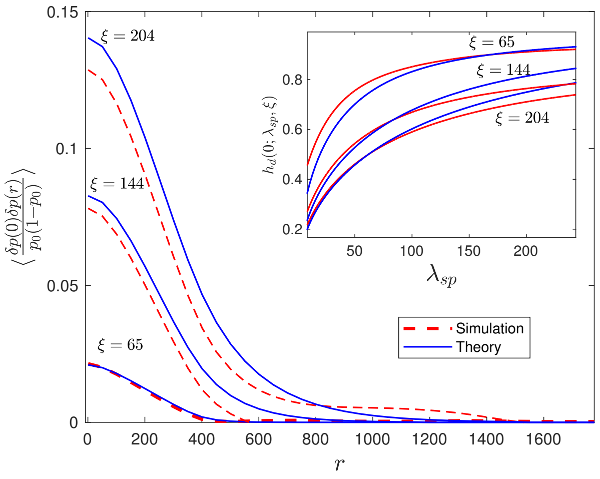

FIG. 2. Comparing between the sub-populations correlation function in a Moran process simulation and analytic formula (Eq. 25). Results are shown for replacement radius of 25, population size of 400 and 3 values of migration probabilities - 0.05,0.01 and 0.005. Inset - Relative Fixation index ($g_d = h_d(0) = F_{\rm st}/F^0_{\rm st}$) as a function of population size.

## APPLICATION TO HUMAN POPULATION

We conclude by looking at human population data. Using publicly available data, we compare allele frequencies in 50 human populations from Asia, Africa and Europe to our continuous model ([12], See Fig. 3b for locations). The data is based on 128 *microhaplotypes*, regions of the DNA that have single nucleotide polymorphisms (SNPs) in close proximity. Overall, we extracted 394 SNPs with two possible mutations. Though different alleles in the same haplotype are very much correlated, we regard all SNPs as uncorrelated. This is the first of many simplifying assumptions, as we only expect to get a rough estimation of the models parameters.

For each pair of populations, located at $\mathbf{x}$ and $\mathbf{x}'$, we estimate the correlation function $\langle \delta p(\mathbf{x}) \delta p(\mathbf{x}') \rangle$ by averaging over the different SNPs. In order to average over SNPs with different mean allele frequency $(p_0)$, we normalize each population allele frequency by $\sqrt{p_0(1-p_0)}$. As we can see in Fig. 3, after this normalization the data appears to be distributed normally. The mean frequency $p_0$ is averaged over all populations. The distance between pair is the great circle distance.

There are 4 parameters in our model (Eq. 25): The correlation length $\xi$, the sub-population radius $\lambda_{sp}$, the total number of migrants $mN$ and dimensionality $d$. Knowing the total number of individuals in the population, $N$, allows us to extract the migration probability, $m$, and from there the replacement radius, $L$. Since we are dealing with human population, we replace $N$ with $2N$.

It is not clear how to choose the dimensionality of the human network. A reasonable assumption is something between 1 and 2 dimensions. We might draw some intuition on the dimension of human connectivity networks by looking at transportation networks nowadays. Recent studies show they have a fractal dimension between 1 and 1.5 [16]. Transportation in the neolithic period was very different, but it perhaps share some of the underlying structure. For simplicity, we will use 1D in our comparison. The average population size $N$ is taken to be $10^4$.

Fig.3a show the correlation function, for distances up to 8000km, compared with the best fit. The fit show that the sub-population radius is $\approx$ 500km. This is in the order of magnitude of sub-population sizes as they appear in the database, yet slightly larger. The correlation length is quite large, and is almost 2000km. The migration rate is very small, and is estimated to be about $10^{-4}$, leading to a replacement radius $L =\approx$ 30km.

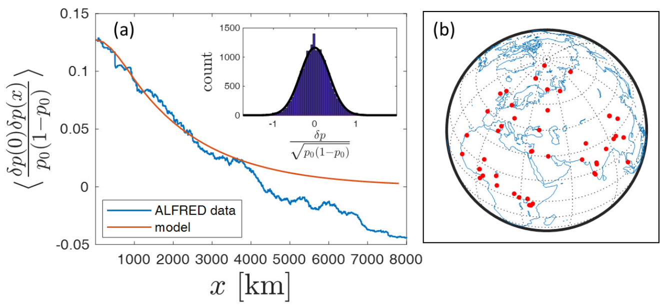

FIG. 3. Genetic variations as a function of distance in human populations. (a) Correlation between the normalized allele frequencies as a function of distance (blue line) fits our predictions (red line) with the following parameters: $\lambda_{sp} = 520$km, $L = 32$km, $m = 10^{-4}$ and $N = 10^4$. Inset- the distribution of normalized allele frequencies is indeed normal. (b) Location of the populations we consider

## CONCLUSIONS

"Isolation by distance" is a term coined by Wright in 1942 to describe how genetic difference increase with distance, making distant populations effectively behave like "islands" [17]. This behavior gained growing support over the years, and our small analysis of the ALFRED dataset clearly show that correlations indeed decay with distance.

We find two important length-scales that determine the behavior of the correaltion function: the correlation length ($\xi$) and is the sub-population radius ($\lambda_{sp}$). In the limit of small correlation length, different population are indeed "isolated by distance", and are described well by Wright infinite island model. However, many of the populations that are studied in genetic surveys today are probably small compared to the correlation length. Hence, when interpreting genetic data, it is important to extend the island model to include more realistic migration patterns.

We obtained a general result that relates the correlation function to the two length-scales. Our expression is applicable to any finite dimension, and offers natural extrapolation for fractional dimensions as well. By comparing our model to genetic data, we found a good fit with reasonable parameters.

The relatively large variations in the data-set, which correspond to small migration rates, limits the applicability of our model. Long range migration acts to smooth variations across the entire population, which is an important assumption of our model. If migrations rates are small, there is a stronger dependence on initial conditions, which we did not take into account.

While we tried to take a small step into a more realistic model, we neglected many important factors that might actually dominate spatial correlations. Mutation, selection (especially geographic dependent selection), and mass migration are some of the ingredients a more realistic model should include. Hopefully, the flood of available genetic data that continues to accumulate will aid the development of more realistic population genetics models.

[1] S. Wright, Annals of Human Genetics **15**, 323 (1949).

[2] M. Nei, Proceedings of the National Academy of Sciences **70**, 3321 (1973).

[3] L. J. L. Handley, A. Manica, J. Goudet, and F. Balloux, TRENDS in Genetics **23**, 432 (2007).

[4] M. Kimura and G. H. Weiss, Genetics **49**, 561 (1964).

[5] J. Novembre and M. Stephens, Nature genetics **40**, 646 (2008).

[6] D. P. Kwiatkowski, The American Journal of Human Genetics **77**, 171 (2005).

[7] S. Manel, M. K. Schwartz, G. Luikart, and P. Taberlet, Trends in ecology & evolution **18**, 189 (2003).

[8] M. E. Weale, D. A. Weiss, R. F. Jager, N. Bradman, and M. G. Thomas, Molecular Biology and Evolution **19**, 1008 (2002).

[9] P. Menozzi, A. Piazza, and L. Cavalli-Sforza, Science **201**, 786 (1978).

[10] T. Jombart, S. Devillard, and F. Balloux, BMC genetics **11**, 94 (2010).

[11] R. Lande, Genetics **128**, 443 (1991).

[12] H. Rajeevan, M. V. Osier, K.-H. Cheung, H. Deng, L. Druskin, R. Heinzen, J. R. Kidd, S. Stein, A. J. Pakstis, and N. P. Tosches, Nucleic Acids Research **31**, 270 (2003).

[13] S. Wright, Genetics **16**, 97 (1931).

[14] P. A. P. Moran, in *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 54 (Cambridge University Press, 1958) pp. 60–71.

[15] M. Kardar, *Statistical physics of fields* (Cambridge University Press, 2007).

[16] Y. Lu and J. Tang, Environment and Planning B: Planning and Design **31**, 895 (2004).

[17] S. Wright, Genetics **28**, 114 (1943).