Scientific Research Publishing

Advances in Bioscience and Biotechnology, 2018, 9, 520-533
http://www.scirp.org/journal/abb
ISSN Online: 2156-8502
ISSN Print: 2156-8456

Mathematical Modeling the Biology of Single Nucleotide Polymorphisms (SNPs) in Whole Genome Adaptation

James Lindesay[1,2,3], Tshela E. Mason[1,2,3], William Hercules[1,2,3], Georgia M. Dunston[1,2,3]

[1]Department of Physics & Astronomy, Howard University, Washington, USA
[2]National Human Genome Center, Howard University, Washington, USA
[3]Department of Microbiology, Howard University, Washington, USA
Email: jlindesay@howard.edu, tmason@gmail.com, wmhercules@yahoo.com, gdunston@howard.edu

How to cite this paper: Lindesay, J., Mason, T.E., Hercules, W. and Dunston, G.M. (2018) Mathematical Modeling the Biology of Single Nucleotide Polymorphisms (SNPs) in Whole Genome Adaptation. *Advances in Bioscience and Biotechnology*, **9**, 520-533.
https://doi.org/10.4236/abb.2018.910036

Received: September 3, 2018
Accepted: October 19, 2018
Published: October 22, 2018

Abstract

As a living information and communications system, the genome encodes patterns in single nucleotide polymorphisms (SNPs) reflecting human adaptation that optimizes population survival in differing environments. This paper mathematically models environmentally induced adaptive forces that quantify changes in the distribution of SNP frequencies between populations. We make direct connections between biophysical methods (e.g. minimizing genomic free energy) and concepts in population genetics. Our unbiased computer program scanned a large set of SNPs in the major histocompatibility complex region and flagged an altitude dependency on a SNP associated with response to oxygen deprivation. The statistical power of our double-blind approach is demonstrated in the flagging of mathematical functional correlations of SNP information-based potentials in multiple populations with specific environmental parameters. Furthermore, our approach provides insights for new discoveries on the biology of common variants. This paper demonstrates the power of biophysical modeling of population diversity for better understanding genome-environment interactions in biological phenomenon.

Keywords

Genome-Environment Interactions, Genomic Adaptation, SNP Functional Correlations

1. Introduction

As a complex, dynamic information system, the human genome encodes and perpetuates the principles of life. The information is incorporated within a mostly fixed template, as well as within the structure of human genome sequence variation. Of the approximately 3 billion nucleotides of the human genome, only about 0.1% consist of bi-allelic single nucleotide polymorphisms (SNPs) distributed throughout the genome [1]. Once the statistical distribution of variation reaches homeostasis in a given environment, a human *population* can be described in terms of the maintained order and patterns of polymorphisms in the whole genome. We define the environment not just in terms of geophysical parameters, but rather as the complete interface of the population to biologic and evolutionary influences. We assert that the stability of whole genome adaptation is reflected in the frequencies of maintained diversity in these common variants (SNPs) for a population in its environment. As dynamic sites in the human genome, SNPs are often highly correlated into combinations referred to as *haploblocks* whose *haplotypes* are maintained throughout generations with fixed frequencies within a given population. Such combinations of SNPs are said to be in *linkage disequilibrium* (LD). This reflects that certain SNP allelic combinations never appear within the population, implying that only certain haplotypes are biologically viable and generationally maintained. In population dynamics, *viability* manifests as maintained survivability and functionality. *The formation of haploblocks is an emergent property of genomic information that cannot be characterized in the absence of the environmental influences that compel such phase transitions among populations.* Therefore, the dynamically independent statistical genomic units we use are SNP haplotypes together with alleles within SNP sites that are not in contiguous LD with any other SNPs. In particular, changes in the distribution of allelic and haplotypic responses to the environment directly reflect adaptive forces on the population. The resilience of living humans as embodiments of the genome allows for the adaptation of groups to new or changing environments. Differing human populations have emerged as a consequence of the various past migratory groups remaining within specific environments and developing the collective coping mechanisms that have allowed the groups to function effectively in their surroundings. We consider adaptation to be the dynamic process of modifying expressions of the genome towards optimizing the survivability of a group that remains in a particular environment. Using measures of genomic information that reflect the interplay of statistical variations due to the environmental baths within which stable populations exist motivates the development of "genodynamics" as an analog to macro-physical "thermodynamics" [2]. *This approach offers a novel way of thinking about population diversity, through the discovery of relationships between the environment and genome variation underlying biology.* In this paper, we mathematically model genome-environment interactions and demonstrate straightforward environmental influences upon common genomic variants.

2. Materials and Methods

2.1. Population Variation and Information

We begin by developing expressions that relate the genomic information measures of human groups whose diversity profile is stable over generations, to additive dynamic state variables that depend upon the environment occupied by that group. Most common informatic measures in the physical and communication sciences are related to the entropy of the statistical system being described. In order to develop entropy measures for a genomic population, the dynamic units of relevance must first be ascertained. Within a given environment, the statistical distributions of certain sets of SNPs become highly correlated as emergent units. This means that the genomic information dynamics in a specific environment is an *emergent phase of expression* of the human genome. The *specificentropy*$s^{(S)}$ (or per capita entropy) of a single SNP location (S) that is not in (contiguous) linkage disequilibrium will take the form of a canonical ensemble state variable in an environmental bath given by

$$s^{(S)} \equiv -\sum_{a=1}^{2} p_a^{(S)} \log_2 p_a^{(S)} , \tag{1}$$

where $p_a^{(S)}$ represents the probability (frequency) that allele a occurs in the population. It should be noted that the entropy so defined is a dimensionless measure of disorder without biophysical units. Likewise, the specific entropy of a SNP haploblock (H) consisting of a set of strongly correlated bi-allelic SNPs is taken to be

$$s^{(H)} \equiv -\sum_{h=1}^{2^{n^{(H)}}} p_h^{(H)} \log_2 p_h^{(H)} , \tag{2}$$

where $n^{(H)}$ is the number of SNP locations in haploblock (H), and $p_h^{(H)}$ represents the probability (frequency) that haplotype h occurs in the population. The upper limit in this sum represents the number of mathematically possible bi-allelic combinations of alleles within the haploblock. Commonly available tools were used to construct the haploblock structures [3].

Since entropy is a measure of the disorder of a distribution, a system with maximum disorder (equal statistical distribution of all mathematically possible combinations) is one of maximum entropy S_{max}. The *information content* (*IC*) of a maintained statistical distribution is measured by the degree of *order* that the distribution has relative to a completely disordered one, *i.e.*, the difference between the entropy of a completely disordered distribution and that of the given distribution; $IC = S_{\max} - S$ [4]. Such an information measure is likewise additive due to the additive nature of the entropy [5]. Thus, both entropy and information content are extensive state variables whose values increase proportionate with the population size. The *normalized information content* (*NIC*) for a given SNP haploblock (H) is a (non-additive) intrinsic measure defined by

$$NIC^{(H)} \equiv \frac{S_{\max}^{(H)} - S^{(H)}}{S_{\max}^{(H)}} = \frac{s_{\max}^{(H)} - s^{(H)}}{s_{\max}^{(H)}} = \frac{n^{(H)} - s^{(H)}}{n^{(H)}} , \tag{3}$$

where, as previously stated the specific entropy of the haploblock $s^{(H)}$ is just the entropy per population member $S^{(H)} = N_{\text{population}} s^{(H)}$. This normalized measure of information ranges between 0 and 1. Such a dimensionless measure allows one to explore the informatic spectra of regions of the genomes of individuals as well as populations [4].

To best parameterize environmental influences, only Phase 3 HapMap data were utilized [6]. These data include populations with African Ancestry in the Southwest USA (ASW), Utah residents with ancestry from Northern and Western Europe (CEU), Han Chinese in Beijing China (CHB), Chinese in Metropolitan Denver Colorado USA (CHD), Gujarati Indians in Houston Texas USA (GIH), Japanese in Tokyo Japan (JPT), Luhya in Webuye Kenya (LWK), Mexican Ancestry in Los Angeles California USA (MXL), Massai in Kinyawa Kenya (MKK), Toscani in Italia (TSI), and Yoruba in Ibadan Nigeria (YRI). Of the Phase 3 populations, the NIC of ASW is 0.52, CEU is 0.76, CHB is 0.76, GIH is 0.73, JPT is 0.77, LWK is 0.59, MXL is 0.71, MKK is 0.63, TSI is 0.74, and YRI is 0.63. It should be noted that for the so-called *founder* populations where the genotyping was more complete for Phase I, II, and III of HapMap data, the NIC values for CEU, CHB, and JPT are somewhat higher (0.88) than the NIC of 0.77 for YRI.

2.2. Information Dynamics of the Human Genome

We next develop dimensional scales and units that can quantify the relative pliability and elasticity of information dynamics between various populations and regions of the genome of the same population, analogous to the additive energy units in the physical sciences. In contrast to the fundamental particles of micro-physics, fundamental life units cannot maintain in the absence of the environments that support them. Therefore, the least complicated description of genomic dynamics should develop genomic *free energy* variables F_{Genome} as more fundamental than environmentally independent energetic measures.

The genomic free energy F_{genome} has been developed as a state variable that balances between the conservation and variation of SNPs and haplotypes within a given environmental bath. Minimizing the genomic free energy optimizes the population's survivability under the complete set of environmental stimuli and stressors, establishing the balance between conservation and variation of alleles and traits in the dynamics of the population distribution. A dimensional environmental potential T_E (which is an intensive state variable that is independent of the size of the population) will parameterize the intrinsic, pervasive agitation of the population due to stochastic environmental stimuli (analogous to how the *temperature* parameterizes agitation of fundamental physical units in a thermal bath). Similarly, dimensional allelic and haplotype potentials, $\mu_a^{(S)}$ and $\mu_h^{(H)}$, will parameterize the genomic free energy change in a population from the addition of one individual of allele *a* or haplotype *h*. For a given haploblock (*H*), the differential genomic free energy takes the form

$$dF^{(H)} = -S^{(H)} dT_E + \sum_h \mu_h^{(H)} dN_h^{(H)}, \tag{4}$$

where $N_h^{(H)}$ represents the number of individuals in the population with haplotype *h*. This form neglects any influence of the population upon the environment. The total genomic free energy is a sum over all SNP haploblocks and non-linked SNPs given by

$$F_{\text{genome}} = \sum_{H} F^{(H)} + \sum_{S} F^{(S)} . \tag{5}$$

As is the case in thermodynamics, the additive allelic potentials $\mu_h^{(H)}$ are expected to scale relative to the environmental potential T_E, and allelic or haplotypic potential differences should directly reflect in the ratio of the frequencies of occurrence of those dynamic units within the population. We assert that such properties are encompassed in the functional form

$$\frac{\mu_{h_2}^{(H)} - \mu_{h_1}^{(H)}}{T_E} = -\log_2 \frac{p_{h_2}^{(H)}}{p_{h_1}^{(H)}} \tag{6}$$

Defining a single human Genomic Energy Unit ($\tilde{\mu} \equiv 1GEU$) to be the allelic energy necessary to induce maximal variation within a single non-linked bi-allelic SNP location ($p_{a_1} = \frac{1}{2} = p_{a_2}$), the potential of the haplotype *h* or allele *a* in an environmental bath characterized by the environmental potential T_E that bathes the whole genome can be expressed as

$$\begin{aligned} \mu_h^{(H)} &= \left(\tilde{\mu} - T_E\right) n^{(H)} - T_E \log_2 p_h^{(H)} \\ \mu_a^{(S)} &= \left(\tilde{\mu} - T_E\right) - T_E \log_2 p_a^{(S)} \end{aligned} . \tag{7}$$

If only one allele is present at a SNP location for a given population, the allelic potential of that allele is defined to be at the *fixing potential* μ_{fixed} for that environment $\mu_{a_1}^{(S)} = \mu_{fixing} \equiv \left(\tilde{\mu} - T_E\right)$.

We will assume that the population is homeostatic (or at least quasi-homeostatic, which means that any changes occurring in the population distribution requires many generations to become significant). Population homeostasis is equivalent to the Hardy-Weinberg condition used in population biology that the statistical distribution is independent of any sub-divisions of the population data, including those associated with differing generations or ages. Our *population stability condition* will require that the genomic free energy be a (stable) minimum under changes in the population within the local environment when the population is in homeostasis with its environment, *i.e.*, $\left(\frac{\partial F_{\text{Genome}}}{\partial N_{\text{Population}}}\right) = 0$. By substituting the forms of the allelic potentials $\mu_h^{(H)}$ and $\mu_a^{(S)}$ expressed in terms of the probabilities in Equation (7) into the population stability condition and summing over all haploblocks and SNPs, an explicit expression of the environmental potential can be obtained:

$$T_E = \frac{\tilde{\mu} n_{SNPs}}{n_{SNPs} - s_{\text{Genome}}} = \frac{\tilde{\mu}}{NIC_{\text{Genome}}} . \tag{8}$$

This inversely relates the environmental potential to the intrinsic normalized information content characterizing the variation of the *whole* genome of the population, demonstrating that the whole genome is uniformly bathed in this particular environmental parameter. The population stability condition can be expressed in terms of the population averaged haplotype and allelic potentials. We refer to the average haplotype potential within a SNP haploblock $\sum_h \mu_h^{(H)} p_h^{(H)} = \left\langle \mu^{(H)} \right\rangle$ as the *block potential* for haploblock (*H*), and the average allelic potential at a non-linked SNP location $\sum_a \mu_a^{(S)} p_a^{(S)} = \left\langle \mu^{(S)} \right\rangle$ as the *SNP potential* for location (*S*). The population stability condition then requires that the sum of all block and SNP potentials for a given population vanishes:

$$\left(\frac{\partial F_{\text{Genome}}}{\partial N_{\text{Population}}} \right) = 0 \Rightarrow \sum_H \left\langle \mu^{(H)} \right\rangle + \sum_S \left\langle \mu^{(S)} \right\rangle = 0 . \tag{9}$$

This condition demonstrates that balance is established between diversity and conservation in a population to optimize its survivability within the given environment. One should note that the environmental potential T_E, the block potentials $\left\langle \mu^{(H)} \right\rangle$ and the SNP potentials $\left\langle \mu^{(S)} \right\rangle$ can only be constructed for a population. In addition, the individual allelic potentials $\mu_h^{(H)}$ and $\mu_a^{(S)}$ characterize an overall allelic potential for each individual in the population,

$$\mu_{\text{individual}} = \sum_H \mu_h^{(H)} + \sum_S \mu_a^{(S)} , \tag{10}$$

where the set of SNP haplotypes *h* and alleles *a* are unique to the individual. An individual's overall allelic potential is not a universal parameter, but rather depends strongly upon the environment.

To illustrate population dependent spectra of genomic block potentials, the genomic free energies of blocks in the major histocompatibility complex (MHC) region on chromosome 6 are displayed for a few founder populations using phase I, II, and III data from HapMap in **Figure 1**.

The MHC region encodes genes for the human immune response. This region of the genome is particularly relevant in host response to environmental stressors and is known to display straightforward biological correlations with environmental parameters. The emergent differences in the haploblock structure of the populations are immediately apparent. The *block binding potential* (which parameterizes the stability of an emergent haploblock) will be defined as the difference in the block potential from the sum of the individual SNP potentials that make up that block if they were not in linkage disequilibrium (LD). The corresponding spectra of binding potentials (per SNP) are demonstrated in **Figure 2**.

Those SNPs in haploblocks with more negative binding potential per SNP have enhanced biologic favorability for maintaining their correlated statistics throughout generations of the populations in the given environments. SNPs in haploblocks with nearly zero binding potential per SNP are nearly independent, indicative of the environmental transition point of the emergent genomic phase. Stated precisely, an emergent genomic phase indicated by the formation of a haploblock of statistically correlated SNPs on the genome of a population in homeostasis with a particular environment results in a non-vanishing binding potential for the SNPs in that haploblock. *The strength of the binding block*

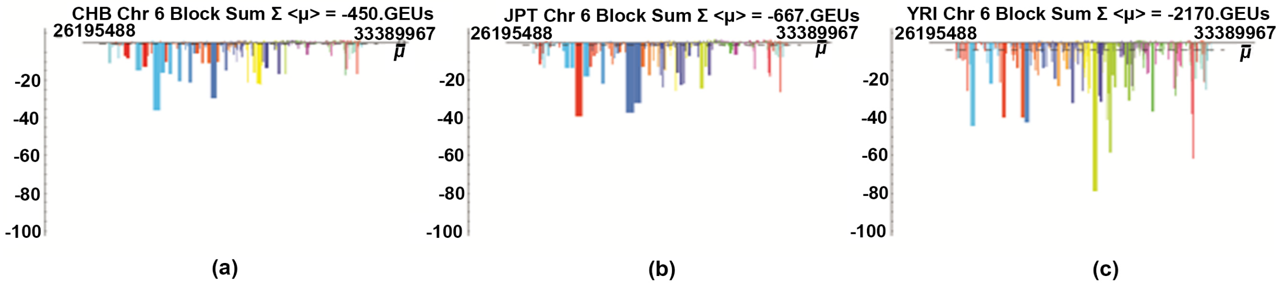

Figure 1. Block potentials for MHC region on Chromosome 6 as a function of location: (a) represents the block potentials for the Han Chinese in Beijing China; (b) represents the MHC region block potentials for the Japanese in Tokyo Japan; (c) represents the MHC region block potentials for the Yoruba in Ibadan Nigeria. Average values are demonstrated as the horizontal dashed lines. Regions of lower potential are indicative of a greater degree of conservation, and stronger binding of the correlated SNPs.

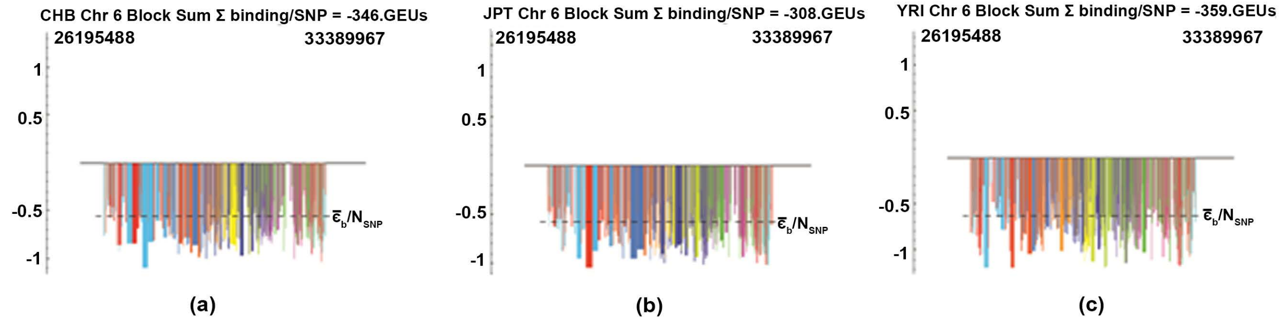

Figure 2. Binding block potential per SNP for the MHC region on Chromosome 6: (a) represents the binding block potentials for the MHC region for the Han Chinese in Beijing China; (b) represents the binding block potentials for the MHC region for the Japanese in Tokyo Japan; (c) represents the binding block potentials for the MHC region for the Yoruba in Ibadan Nigeria.

potential per SNP indicates the degree to which the SNP variation must be correlated in order to maintain a biologically viable population.

2.3. Distributive Genodynamics

The formulation of the information dynamics of the human genome in terms of genomic free energies directly results in well-defined forms for the SNP potentials for SNPs that are not in LD and for block potentials for correlated SNPs that are in LD. Since the SNP haploblock structure has an emergent form that differs between populations, meaningfully defined distributed potentials will reflect the biology underlying the participation of individual SNPs in the informatics architecture of its correlation with other SNPs in the haploblock. We will next develop *distributed SNP potentials* $\mu_S^{(H)}$ within a haploblock (*H*) such that they satisfy the following conditions:

If the SNP is occupied by an allele that is fixed in the given population, then its distributed SNP potential is the fixing potential μ_{fixed};

The sum of the distributed SNP potentials should be the same as the block potential $\mu^{(H)}$, *i.e.* $\left\langle \mu^{(H)} \right\rangle = \sum_{S=1}^{n^{(H)}} \mu_S^{(H)}$;

The block potential should be linearly distributed amongst the constituent SNPs in accordance with occurrences of the SNP alleles.

The first bullet insures that if the SNP is not variant within the population, its genomic energy is not modified from that of a SNP that is not in LD, and the second bullet requires that the distributed potentials should reconstruct the block potential in an additive way. The third bullet represents a simple mechanism for relating the distributed potentials to the degree of variation in the SNP. The mathematical form that satisfies these conditions is given by

$$\mu_S^{(H)} \equiv \mu_{\text{fixed}} + \left[\left\langle \mu^{(H)} \right\rangle - n^{(H)} \mu_{\text{fixed}} \right] \left(\frac{\bar{p}_S}{\sum_{S'} \bar{p}_{S'}} \right), \tag{11}$$

where $\bar{p}_S = 1 - p_S$ is the minor allele frequency of the SNP labeled (*S*). Using this form, the distribution of the haploblock potential to any constituent SNP is proportionate to the occurrence of the minor allele in the population in a manner that increases the SNP's genomic free energy as the SNP has higher variation (*i.e.*, becomes less conserved).

The degree of stability of the participation of the SNP in the biology of the emergent haploblock can be quantified in terms of its binding potential defined by

$$\varepsilon_{\text{binding}}^{(S)} \equiv \mu_S^{(H)} - \left\langle \mu^{(S)} \right\rangle, \tag{12}$$

where $\left\langle \mu^{(S)} \right\rangle$ would be the SNP potential of the genomic variant were it not in LD. As defined, this metric of SNP binding within the haploblock is always negative, reflecting the increased genomic conservation inherent in LD.

We can furthermore assign allelic measures from the distributed SNP potentials in a manner that constructs the SNP potentials as population averages of derived *distributed allelic potentials* $\mu_{a_S}^{(H)}$, *i.e.*, $\mu_S^{(H)} = \sum_{a_S} p_{a_S}^{(H)} \mu_{a_S}^{(H)}$. The most straightforward form that uniformly assigns the distributed SNP potential within a haploblock, and maintains the expected correlation that increased genomic potential reflects increased variation, results by simply adjusting the non-linked allelic potentials using the SNP binding potential, *i.e.*,

$$\mu_{a_S}^{(H)} \equiv \mu_{a_S}^{(S)} + \varepsilon_{\text{binding}}^{(S)}. \tag{13}$$

It should be noted that all distributed potentials are only defined at the *population* level and cannot be ascribed to *individuals*. Only the emergent haplotype potentials $\mu_h^{(H)}$ can be ascribed to individuals within the population. However, since distributed potentials are defined for the population as a whole, they can bequite useful for parameterizing the environmental influences upon that population. *Distributed potentials are particularly useful for describing the adaptation of the population to stimuli and stressors with known biological correspondence to particular alleles or SNPs.* The description of genomic variants using distributed potentials inherently includes any presently unknown *whole genome* response to specific stressors.

2.4. Adaptive Forces

Once genomic free energy measures have been developed for individual alleles and genomic regions, environmentally induced adaptive forces can be characterized using gradients of those additive measures down the slope of environmental parameters. For a given allele *a* on the genome that is biologically connected to a definable environmental parameter λ (such as UV light, lactose in diet, prevalence of malarial plasmodia, etc.), we define the environmentally induced *adaptive force* on that allele by

$$f_a \equiv -\frac{\partial \mu_a}{\partial \lambda}, \tag{14}$$

with analogously defined adaptive forces on potentials characterizing SNPs, haploblocks, haplotypes, genes, and even perhaps whole chromosomes. Such an expression is only meaningful if there is a functional relationship between the biology of the genomic unit and the particular environmental parameter λ. In such cases, positive adaptive forces drive the conservation of the given genomic unit down the slope of the genomic potential. Increased *survivability* might drive the genomic unit towards more diversity, or more conservation, depending on the nature of the environmental influence upon the homeostatic population. Quantifying such forces inherently involves comparisons between differing environments.

To explore environmental impacts on adaptation, we will confine our investigation to phase III data of HapMap, since this represents the broadest set of populations with somewhat uniform genotyping. We have chosen to exclude ASW, CEU, CHD, GIH and MXL from our parameterization of adaptive forces, since these populations do not reside in their geographical origin. In this paper, the genomic potentials of the set of SNPs in the MHC region on chromosome 6 were chosen to conduct a double-blind exploration for possible correlations with three particularly straightforward environmental parameters: annual exposure to UV-B radiation, altitude above sea level, and exposure to malarial vectors. In order to simplify the analysis of any results, the set of all SNPs in this region that are *not* in LD for most of the populations were pre-selected out for the computational search. The algorithm examines whether the genomic potentials for the SNPs and alleles can be fitted to simple functional forms (curves) singly dependent on a given environmental parameter. If the root-mean-squared (RMS) deviation of the data points from the curves, as compared to the maximum variation of the data, falls within 10%, the SNP is flagged by the program, and adaptive forces are calculated for the curves.

The averaged ancestral annual UV-B radiation exposure used was expressed in units of Joules per square meter (UV radiance) as estimated from the following cited source [7]. In these units, estimates of annual UV radiance for the CHB population averaged 2180 (ranging from 1500 to 2600), for the JPT population averaged 2400 (ranging from 2300-2500), for the LWK population averaged 5764 (ranging from 5450 to 6500), for the MKK population averaged 5624 (ranging from 5000 to 6125), for the TSI population averaged 1507 (ranging from 950 to 2500), and for the YRI population averaged 5129 (ranging from

3500 to 6300). The altitude values used are averaged estimates of elevations of populated regions for ancestral homelands in units of meters using data from [8]. In units of meters, estimates of population elevation for the CHB population averaged 22 (ranging from 3 to 48), for the JPT population averaged 107 (ranging from 5 to 287), for the LWK population averaged 1711 (ranging from 1203 to 2486), for the MKK population averaged 1507 (ranging from 712 to 2383), for the TSI population averaged 74 (ranging from 1.3 to 143), and for the YRI population averaged 211 (ranging from 12 to 337). The parasite data were based upon the *Plasmodium falciparum* parasite rate (PfPR), used by the World Health Organization [9]. We expect that all of the examined populations had higher malarial exposure in ancestry than at present. In particular, the TSI population likely had significantly higher malarial exposure in ancestry than in present time, since relatively recent developments have significantly reduced the prevalence of the insects and treatment of the disease. In units of parasite reproductive rate, estimates of PfPR for the CHB population averaged 0.01 (ranging from 0 to 0.05), for the JPT populations averaged 0.0002 (ranging from 0 to 0.001), for the LWK population averaged 12 (ranging from 2 to 35), for the MKK population averaged 8 (ranging from 1 to 25), for the TSI population averaged 0.8 (ranging from 0 to 5), and for the YRI population averaged 70 (ranging from 20 to 95).

In the following plots, if there is a best fit curve plotted with the points, then the data was flagged by the computer program. Blue points represent populations with the flagged SNP not in linkage disequilibrium. The thickness of the curves in the plots represents the degree of correlation of the data with the fitting curve, with bolder curves indicating stronger correlations.

3. Results and Discussion

Our program flagged functional dependencies on altitude of phase III HapMap data for the SNP rs1109771 in the MHC region for the populations CHB, LWK, MKK, TSI and YRI. The curves are plotted in **Figure 3**.

The relative RMS deviation for the SNP potential was 0.03, for the *G* allelic potential was 0.008, and for the *A* allelic potential was 0.001. A significant adaptive force of about +1.5 GEUs/kilometer at lower altitudes on allele *A* towards increased conservation is apparent. At higher altitudes, significant variation is maintained, as indicated by the SNP potential remaining very near the maximum value of 1 GEU (maximal variation). This implies that the *G* allele continues a significant presence in the population in order to optimize its survivability in the higher altitudes available in the HapMap data.

High Altitude and NOTCH4

Over the course of human history, adaptation to challenging environments has necessitated modulation of biological pathways at the genomic level to combat the toxic effects present in said environments. High altitude is an excellent example of how humans have adapted to an environmental stressor (e.g., low

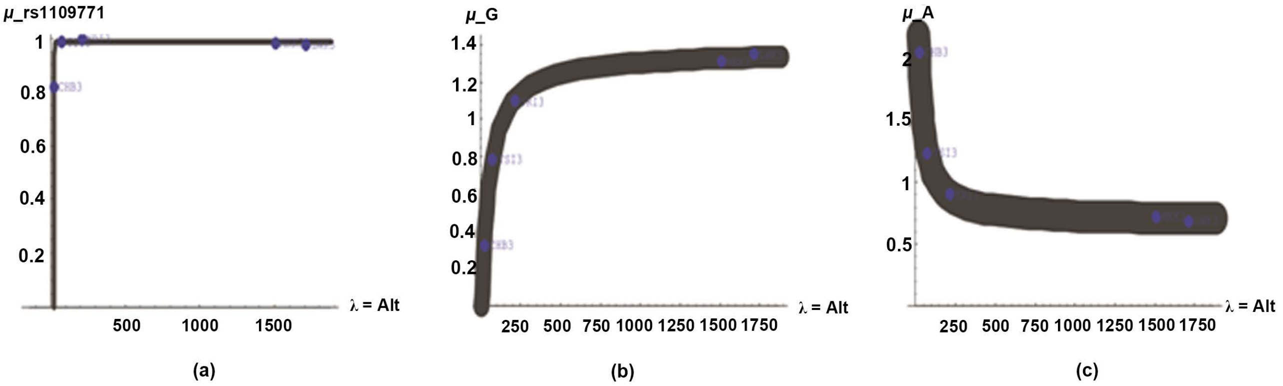

Figure 3. SNP rs1109771 in MHC region on Chromosome 6. The horizontal axis labeled by the environmental parameter λ is altitude in units of kilometers. The vertical axis gives the SNP (μ_rs1109771) and allelic (μ_G, μ_A) potentials in genomic energy units (GEUs): (a) illustrates the functional correlation of SNP rs1109771 with altitude in the Chinese, Kenyan, Tuscan and Yoruban populations; (b) and (c) illustrate the allelic correlations (G and A respectively) with altitude in the aforementioned populations.

oxygen content). The body's response to chronic exposure to alveolar hypoxia is to hyperventilate, thereby increasing resting heart rate and stimulating the production of red blood cells to maintain the oxygen content of arterial blood at or above sea level values [10]. Moreover, an insufficient supply of oxygen prompts the formation of new vessels from the walls of existing ones, *i.e.* angiogenic sprouting [11]. Growth factors and chemokines are secreted from hypoxic tissues, stimulating endothelial cells to break away from vessel walls. These angiogenic factors then coordinate sprouting, branching, and new lumenized network formation until the oxygen content rises and normoxia can be re-established [12]. The Notch signaling pathway plays a key role in shaping the formation and remodeling of the vascular network under hypoxic conditions [11]. This pathway is an evolutionarily conserved intracellular signaling pathway that was originally identified in Drosophila. Notch has four transmembrane receptors, with Notch 1 and Notch 4 being expressed by endothelial cells [13] [14] [15]. It has been shown that targeted deletion of Notch 4 in mice results in the deregulation of arterial and venous specification of endothelial cells as well as the deformation of arteries and veins [16] [17]. In addition, overexpression of the intracellular domain of Notch 4 in endothelial cells results in a β1 integrin-mediated increase in adhesion to collagen resulting in cells that show a reduced sprouting response to vascular endothelial growth factor both *in vitro* and *in vivo* [18]. Thus, it appears that Notch signaling promotes cellular responses in endothelial cells that help to alleviate the harmful effects of hypoxia in the human body. Consequently, population differences in allelic frequencies in this pathway could effectively provide an adaptive advantage for survival in response to this environmental stressor.

As a demonstration of the potential guidance offered by this formulation towards future discovery in the biology of whole genome adaptation, our program flagged functional dependencies on plasmodium parasite load from HapMap data for rs430620 in the MHC region for the populations CHB, LWK, MKK, TSI and YRI. The curves plotted in Figure 4 represent a strong flag for parasite dependency of a SNP in the intervening sequence of the genome with no known association to any gene. The relative RMS deviation for the SNP potential was 0.007, for the *G* allelic potential was 0.02, and for the *A* allelic potential was 0.008. A significant adaptive force of about +3 GEUs/unit PfPR for initial parasite loads on allele *A* towards increased conservation is apparent. The *A* allele has very low occurrence within populations with no parasite load, and the SNP approaches fixation towards allele *G*. Once again, for higher parasite loads, significant variation is maintained, as indicated by the SNP potential approaching the maximum of 1 GEU, indicative of the importance of maintaining a significant occurrence of the *G* allele in the population. The possibility of an association of the *A* allele with increased survivability under an environmental stressor that parallels this parasite load is intriguing. Furthermore, this correlation exemplifies how genodynamics can be utilized in the discovery of genomic variants with previously unknown clear functional dependencies on simple environmental factors, thereby motivating other researchers to contextualize these functional dependencies relative to their own laboratory/clinical findings.

We have demonstrated the utility of associating genomic free energy measures with environmental influences on whole genome adaptation. Double-blind smooth mathematical functions flagged relationships between altitude and the allelic energies of a SNP associated with oxygen deprivation. From these functional relationships, genomic energy gradients quantify adaptive forces in a manner analogous to corresponding concepts in the physical sciences. Our formulation of genomic information dynamics optimizes the survivability of a population in a given environment. Specifically, whole genome SNP distributions represent an environmentally influenced balance between genome sequence variation and conservation. Furthermore, double-blind smooth mathematical

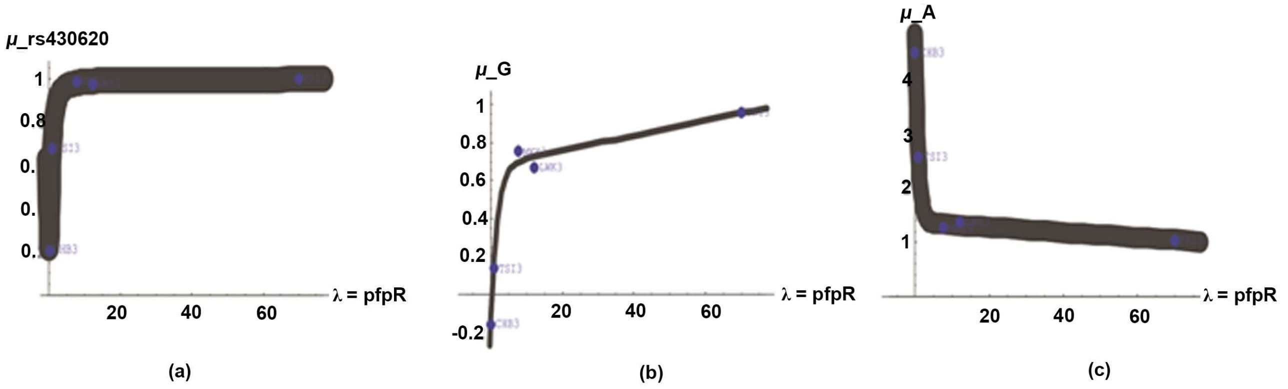

Figure 4. rs430620 in MHC region on Chromosome 6. The horizontal axis labeled by the environmental parameter λ is plasmodium parasite load in units of *Plasmodium falciparum* parasite rate. The vertical axis gives the SNP (μ_rs430620) and allelic (μ_G, μ_A) potentials in genomic energy units (GEUs): (a) illustrates the functional correlation of SNP rs430620 with PfPR in the Chinese, Kenyan, Tuscan and Yoruban populations; (b) and (c) illustrate the allelic correlations (G and A respectively) with PfPR in the aforementioned populations.

functions flagged relationships between parasite load and the allelic energies of a SNP with no known association to a gene. This provides an intriguing opportunity and direction for future discovery of the biology associated with this SNP.

Moreover, population diversity in genome-wide common variants, such as SNPs that are non-randomly embedded in the human genome, represent a "quintessential experiment of nature" in whole genome adaptation to environmental stimuli and stressors associated with population diversity in health outcomes. SNPs associated with common diseases not only reveal mechanisms underlying the complex biology of common diseases, but also the "genomic cost" to populations in whole genome adaptation to environmental stimuli and stressors. By parameterizing the information dynamics of SNPs in HapMap populations, we developed a mathematical model of environmentally induced adaptive forces as drivers of population health and diversity in health outcomes. Our model provides new lenses through which SNP data can be explored to solve problems in population-based patterns of genome variation in common complex diseases which we submit is significant in clinical translation.

Acknowledgements

The authors would like to acknowledge the continuing support of the National Human Genome Center, and the Computational Physics Laboratory, at Howard University. This research was supported in part by NIH Grant NCRR 2 G12 RR003048 from the RCMI Program, Division of Research Infrastructure. The authors hereby certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., *et al*. (2001) A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms. *Nature*, **409**, 928-933. https://doi.org/10.1038/35057149

[2] Lindesay, J., Mason, T.E., Hercules, W. and Dunston, G.M. (2014) Development of Genodynamic Metrics Forexploring the Biophysics of DNA Polymorphisms. *Journal of Computational Biology and Bioinformatics Research*, **6**, 1-14.

[3] Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: Analysis and Visualization of LD and Haplotype Maps. *Bioinformatics*, **21**, 263-265. https://doi.org/10.1093/bioinformatics/bth457

[4] Lindesay, J., Mason, T.E., Ricks-Santi, L., Hercules, W., Kurian, P. and Dunston, G.M. (2012) A New Biophysical Metric for Interrogating the Information Content in Human Genome Sequence Variation: Proof of Concept. *Journal of Computational Biology and Bioinformatics Research*, **4**, 15-22.

[5] Lindesay, J. (2013) Foundations of Quantum Gravity. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511919909

[6] International HapMap3 Consortium (2010) Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature*, **467**, 52-58.

[7] Herman, J.R., Krotkov, N., Celarier, N.E., Larko, D. and Labow, G. (1999) Distribution of UV Radiation at the Earth's Surface from TOMS-Measured UV-Backscattered Radiances. *Journal of Geophysical Research*: *Atmospheres*, **104**, 12059-12076. https://doi.org/10.1029/1999JD900062

[8] Globe Task Team (1999) The Global Land One-Kilometer Base Elevation (GLOBE) Digital Elevation Model, Version 1.0. https://www.mgdc.noaa.gov/mgg/topo/globe.html

[9] World Health Organization (2008) World Malaria Report 2008. Geneva World Health Organization, Switzerland. http://www.who.int/malaria/publications/atoz/9789241563697/en/

[10] Grocott, M.P.W., Martin, D.S., Levett, D.Z., McMorrow, R., Windsor, J. and Montgomery, H.E. (2009) Arterial Blood Gases and Oxygen Content in Climbers on Mount Everest. *The New England Journal of Medicine*, **360**, 140-149. https://doi.org/10.1056/NEJMoa0801581

[11] Phng, L.K. and Gerhardt, H. (2009) Angiogenesis: A Team Effort Coordinated by Notch. *Developmental Cell*, **16**, 196-208. https://doi.org/10.1016/j.devcel.2009.01.015

[12] Gerhardt, H., Golding, M., Fruttiger, M., Ruhrberg, C., Lundkvist, A., Abramson, A., *et al.* (2003) VEGF Guides Angiogenic Sprouting Utilizing Endothelial Tip Cell Filopedia. *The Journal of Cell Biology*, **161**, 1163-1177. https://doi.org/10.1083/jcb.200302047

[13] Hoffman, J.J. and Iruela-Arispe, M.L. (2007) Notch Signaling in Blood Vessels: Who Is Talking to Whom about What? *Circulation Research*, **100**, 1556-1568. https://1161/01.RES.0000266408.42939.e4

[14] Farve, C.J., Mancuso, M., Maas, K., McLean, J.W., Baluk, P. and McDonald, D.M. (2003) Expression of Genes Involved in Vascular Development and Angiogenesis in Endothelial Cells of Adult Lung. *American Journal of Physiology-Heart and Circulatory Physiology*, **285**, H1917-H1938. https://doi.org/10.1152/ajpheart.00983.2002

[15] Villa, N., Walker, L., Lindsell, C.E., Gasson, J., Iruela-Arispe, M.L. and Weinmaster, G. (2003) Vascular Expression of Notch Pathway Receptors and Ligands Is Restricted to Arterial Vessels. *Mechanisms of Development*, **108**, 161-164. https://doi.org/10.1016/S0925-4773(01)00469-5

[16] Carlson, T.R., Yan, Y., Wu, X., Lam, M.T., Tang, G.L., Beverly, L.J., *et al.* (2005) Endothelial Expression of Constitutively Active Notch 4 Elicits Reversible Arteriovenousmal Formations in Adult Mice. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 9884-9889. https://doi.org/10.1073/pnas.0504391102

[17] Kim, Y.H., Hu, H., Guevara-Gallardo, S., Lam, M.T., Fong, S.Y. and Wang, R.A. (2008) Artery and Vein Size Is Balanced by Notch and Ephrin B2/EphB4 during Angiogenesis. *Development*, **135**, 3755-3764. https://doi.org/10.1242/dev.022475

[18] Leong, K.G., Hu, X., Li, L., Noseda, M., Larrivee, B., Hull, C., Hood, L., Wong, F.A. and Karsan, A. (2002) Activated Notch 4 Inhibits Angiogenesis: Role of 1-Integrin Activation. *Molecular and Cellular Biology*, **22**, 2830-2841. https://doi.org/10.1128/MCB.22.8.2830-2841.2002