Journal of Computational Biology and Bioinformatics Research Vol. 4(2), pp. 15-22, February 2012
Available online http://www.academicjournals.org/jcbbr
DOI: 10.5897/JCBBR11.026
ISSN-2141-2227 ©2012 Academic Journals

Full Length Research Paper

A new biophysical metric for interrogating the information content in human genome sequence variation: Proof of concept

James Lindesay[1], Tshela E. Mason[2], Luisel Ricks-Santi[3], William Hercules[1], Philip Kurian[1] and Georgia M Dunston[2,4]*

[1]Computational Physics Laboratory, Howard University, Washington, DC, 20060, U.S.
[2]National Human Genome Center, Howard University, Washington, DC, 20060, U.S.
[3]Cancer Center, Howard University, Washington, DC, 20060, U.S.
[4]Department of Microbiology, Howard University, Washington, DC, 20060, U.S.

Accepted 21 November, 2011

The 21st century emergence of genomic medicine is shifting the paradigm in biomedical science from the population phenotype to the individual genotype. In characterizing the biology of disease and health disparities in population genetics, human populations are often defined by the most common alleles in the group. This definition poses difficulties when categorizing individuals in the population who do not have the most common allele(s). Various epidemiological studies have shown an association between common genomic variation, such as single nucleotide polymorphisms (SNPs), and common diseases. We hypothesize that information encoded in the structure of SNP haploblock variation in the human leukocyte antigen-disease related (HLA-DR) region of the genome illumines molecular pathways and cellular mechanisms involved in the regulation of host adaptation to the environment. In this paper we describe the development and application of the normalized information content (NIC) as a novel metric based on SNP haploblock variation. The NIC facilitates translation of biochemical DNA sequence variation into a biophysical quantity derived from Boltzmann's canonical ensemble in statistical physics and used widely in information theory. Our normalization of this information metric allows for comparisons of unlike, or even unrelated, regions of the genome. We report here NIC values calculated for HLA-DR SNP haploblocks constructed by Haploview, a product of the International Haplotype Map Project. These haploblocks were scanned for potential regulatory elements using ConSite and miRBase, publicly available bioinformatics tools. We found that all of the haploblocks with statistically low NIC values contained putative transcription factor binding sites and microRNA motifs, suggesting correlation with genomic regulation. Thus, we were able to relate a mathematical measure of information content in HLA-DR SNP haploblocks to biologically relevant functional knowledge embedded in the structure of DNA sequence variation. We submit that NIC may be useful in analyzing the regulation of molecular pathways involved in host adaptation to environmental pathogens and in decoding the functional significance of common variation in the human genome.

Key words: Information theory, entropy, genomic variation, biological information.

INTRODUCTION

The human genome is arguably the most sophisticated knowledge system ever discovered, as evidenced by the exquisite information it encodes and communicates via the structure of its DNA sequence. Such information underpins the structure, function, and regulation of complex molecular pathways and network systems transmitted from cell to cell, individual to individual, and

*Corresponding author. E-mail: gdunston@howard.edu. Tel: 202-806-7372. Fax: 202-986-3972.

generation to generation via the genome. New knowledge derived from sequencing the human genome (International Human Genome Sequencing Consortium, 2001) and researching genome variation (International Human Genome Sequencing Consortium, 2003) challenges traditional views of biological identity and how biology works at the molecular level. The integration of this new knowledge into theoretical models of living systems demands a more complete and comprehensive understanding of the life sciences in general and the science of the human genome in particular. In many respects, the human genome displays features of communication systems based in information theory, such as pattern recognition, data compression, signal processing, and regulation that are also seen from a biophysical perspective of life and living systems.

The concept of information as a basic conserved property of the universe has been successfully demonstrated in the physical sciences from cosmology (Susskind and Lindesay, 2005) to telecommunications (Shannon, 1948). In both classical and quantum physics there is a sense that information is conserved. The information content (IC) of an isolated system can be quantified using the fundamental thermodynamic concept of entropy (Susskind and Lindesay, 2005). Complex biophysical systems, like the human genome, are not isolated but rather evolve within external environments which can be assumed to have quasi-static properties. This environmental contact leads to fluctuations of the entropy associated with the system. IC is measured as a difference between the maximum possible entropy and the entropy of a coherently maintained population distribution. In the biological realm, a coherent system maintains its characteristics over generations until perturbed or modified by external forces from the environment.

When applied to whole genome sequence-based biology, we hypothesize that genomic IC can be measured by identifying the “dynamic sites” of the genome and examining variation within and among populations. We identify dynamic sites as single nucleotide polymorphisms (SNPs) for statistical analysis of the genome (Mason et al., 2009). Almost all common SNPs have only two alleles and are therefore bi-allelic. SNP haplotype blocks (haploblocks) can be identified, within which the variability of nucleotides can be interrogated and the occurrence of particular combinations can be determined. Analysis of the frequencies of these SNP combinations leads to a statistical distribution of haplotypes within the population. Since the sizes of SNP haploblocks within the genome are of variable lengths, it is nontrivial to directly compare the IC associated with different haploblocks in a meaningful manner. In order to compare the IC among different haploblocks across the genome, as well as among different human populations, a normalized IC (NIC) was developed. The concept for NIC we developed from statistical physics was discovered to be similar to Shannon’s concept of *redundancy* (Shannon, 1948). Because our focus is the genome, NIC values here apply to the transmission of information in this biological system. If the NIC value of a SNP haploblock for a population is high compared to 50%, we can deduce that there are likely environmental factors skewing the distribution of haplotype frequencies in the population. Similarly, a low NIC value implies high variability and substantially fewer external factors biasing the haplotype frequency distribution in the population. In particular, a SNP haploblock that is completely homogeneous for a population has identical nucleotides at all dynamic sites for all members, thereby exhibiting no variability in the alleles encoded in that haploblock. Such a *monomorphic* haploblock has a NIC value of unity. Likewise, a population with maximum variation in the alleles will have a NIC value of zero. We assert that populations maintain themselves by establishing coherent SNP haplotype frequencies.

In this paper, we seek to explore the biophysical underpinnings of common variation in the genome. This perspective makes the physics of DNA sequence variation in the human genome relevant in new ways to concepts in biology and biomedical science. In so doing, the intent is to connect the genomic frontiers of biology and the health sciences with the biophysical frontiers of information theory and quantum physics.

MATERIALS AND METHODS

Derivation of the normalized information content equation

The degree of variability within a SNP haploblock population can provide a measure of the maintained order associated with that haploblock. SNP haplotype diversity will vary across different SNP haploblocks. Each population group is defined by the maintained order of its SNP haplotype diversity within the SNP haploblock structure; however the latter might be defined. Thus, haplotype diversity is herein reflected in the frequencies with which the SNP haplotypes occur within a given haploblock structure.

In order to provide a meaningful comparison of the information content among different regions of the genome as well as amongst different populations, the normalized information content (NIC) parameter was developed. NIC measures the difference between the entropy and the maximum possible entropy of a SNP haploblock within a given population. Since we expect that the external environment will significantly influence the state of the genome, we choose a form for the entropy measure as illustrated in equation 1.

$$S_{A,coherent} = -k \sum_{j}^{N_A} P^A{}_j \log_2 P^A{}_j \ . \quad (1)$$

where $P^A{}_j$ represents the probability or frequency with which a particular SNP haplotype j occurs within the particular haploblock A, and $N_A = 2^{N^A_{SNPs}}$ represents the number of mathematically possible SNP combinations for N^A_{SNPs} active biallelic SNP sites. For our purposes, $S_{coherent}$ has the potential of being an additive genostatic parameter that can be used to quantify the information in a system. Since all probabilities are non-negative, the minimum

value this entropy can take occurs when one of the probabilities itself is unity. This defines a *homogeneous* population yielding $S_{A,min}$=0. The maximum value this entropy can take occurs when all the probabilities are equal, $P^A{}_j = \frac{1}{N_A}$, defining an informationally *gray population*. In this case, one obtains the result for the entropy as $S_{A,\max} = k \log_2 N_A = k N^A_{SNPs}$. This represents the mathematical maximum of entropy for the SNP haploblock A across all human populations, giving a universal upper bound for this value. A given population will generally not have all possible SNP combinations expressed as viable SNP haplotypes, so that a population made up of equal frequencies for just the expressed haplotypes would not represent a universal maximum across populations. However, some of the SNP combinations that are not expressed in the populations *do* contribute to the informa-tion content of the haploblock. For these reasons, the chosen form for the maximum entropy is both universal and complete. Since the maximum entropy represents the upper limit that any entropy can attain, genomic information can be expected to relate to the difference between the maximum entropy of a gray population and the coherent frequency distribution maintained by a given popula-tion. However, the number of SNPs haplotypes that completely describe all populations varies for different regions of the genome. The information measure is most useful when it represents a dimensionless parameter that can be used to compare the informa-tion content of different SNP haploblocks as well as the same block amongst different populations. We have therefore chosen to normalize our measure of information content in a manner that gives zero for a gray population with no information content and unity for a homogeneous population that has a single maintained SNP haplotype. This can be mathematically expressed for haploblock A as follows:

$$NIC_A = \frac{S_{A,\max} - S_{A,coherent}}{S_{A,\max}} = \frac{N^A_{SNPs} + \sum_{j=1}^{2^{N^A_{SNPs}}} P^A{}_j \log_2 P^A{}_j}{N^A_{SNPs}} \quad (2)$$

This information measure is bounded by $0 \le NIC \le 1$, allowing for an informational comparison of apples to oranges. Normalization of the information metric therefore provides a means to interrogate genomic information from different regions of the genome. Besides limiting the range of the *NIC*, the maximum entropy state of specified block A is common to all populations; only the frequency distribution varies amongst the populations. This dimensionless form allows one to gain considerable insights into genomic informa-tion without a need for detailed information about the dynamics and history of the genome or knowledge of the form of all genomic parameters. The NIC is similar to an independently derived from (Nothnagel et al., 2004) where an entropic measure was used to construct haploblocks rather than interrogate the information content of SNP haploblock variation.

Analysis of NIC for the human leukocyte antigen-disease related (HLA-DR) region of the major histocompatibility complex (MHC)

The human MHC encoding the HLA system is the most highly expressed polymorphic system in the genome, and it plays an essential role in regulation of the immune response in host adaptation to environmental stimuli. It is a strong genomic marker of historical changes in environmental conditions. The NIC values were calculated for the HLA-DR region located on chromosome 6 between positions 415,611 and 3,908,995 (HLA-DRA1, HLA-DRB1, HLA-DRB5) and between positions 32,515,990 and 32,663,637 (HLA-DRB2, HLA-DRB3, HLA-DRB4). SNP haploblocks were constructed using the confidence interval algorithm (Gabriel et al., 2002) in Haploview v 4.2 from HapMap phase III data on the African American population from the southwest United States (N=98). Haploview uses a two marker expectation-maximization algorithm with a partition-ligation approach which creates highly accurate population frequency estimates of the phased haplotypes based on the maximum-likelihood as determined from the unphased input (Barrett et al., 2005; Haploview, 2003). The confidence interval algorithm defines the haploblock as a region over which a very small proportion (<5%) of comparisons among "informative SNP pairs" shows strong evidence of historical recombination and within which independent measures of pairwise linkage disequilibrium (LD) did not decline substantially with distance. Informative SNP pairs were in "strong LD" if the one-sided upper 95% confidence bound on the pairwise correlation factor D' was > 0.98 and the lower bound above 0.7. Conversely, "strong evidence of historical recombination" was defined by an upper confidence bound on D' less than 0.9 (Gabriel et al., 2002). The NIC values for the various blocks were collectively plotted in order to assess the statistical features of the distribution. We defined outliers on the distribution by identifying those values proximal to or beyond two root mean squared (RMS) deviations (2σ) of the mean, 0.61 (RMS (σ) =0.13; standard error=0.01). Root mean squared deviation is defined by $\sigma = \sqrt{< (p - < p >)^2 >}$, where <*p*> represents the mean of the distribution with discrete elements p_j. Lastly, all of these regions were scanned for potential regulatory elements using the publicly available bioinformatics tools ConSite and miRBase.

RESULTS

Haploview generated 189 haploblocks for the HLA-DR region for the ASW population. As illustrated in Figure 1, the NIC values were distributed between zero and one. There were ten blocks with values proximal to the lower bound (between 0.28 and 0.39), and those blocks were comprised of twenty-three SNPs. All twenty-three SNPs proximal to the lower bound were located in intergenic regions (Table 1). Conversely, there were no blocks with NIC values beyond the upper bound (0.86). The sequences of the ten blocks located proximal to the lower bound were scanned for potential regulatory elements, such as transcription factor binding sites (TFBS) and micro RNA (miRNA) motifs. Putative TFBS were identified for the ten blocks proximal to the lower bound which are listed in Table 2. Of the TFBS found, FOXI1, SOX5, and SOX17 sites were present in all ten blocks. Also, there were five SNPs that had TFBS changes when their minor alleles were present (Table 3). Lastly, we found that all ten blocks had miRNA motifs (Table 4).

DISCUSSION

The advent of geographically-defined, population-based, genome-wide variation resources such as the haplotype

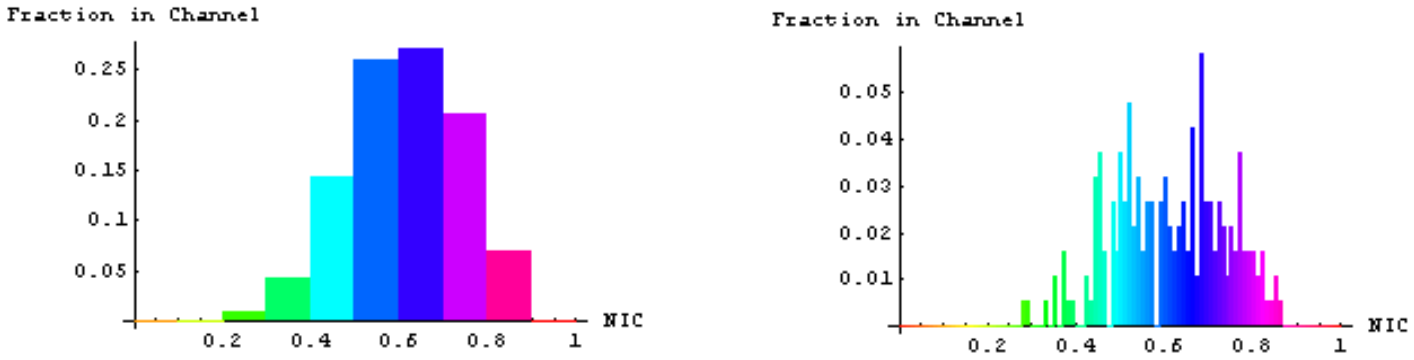

Figure 1. NIC Values for HLA-DR in an African American Population from the Southwest, US. The height of each bar is the fraction of all NIC values that fall within the channel given by the width of each bar.

Table 1. Location of SNPs proximal to the Lower Bound. In order to determine the location of each SNP in the region proximal to the lower bound, each reference sequence (rs) number was searched through the SNP database (dbSNP) maintained by the National Center for Biotechnology Information (NCBI).

Block ID	NIC Value	SNPs	SNP Location
DRB234-174	0.28	rs9378385	Intergenic Region
		rs9503746	Intergenic Region
DRB234-63	0.29	rs1028380	Intergenic Region
		rs7774941	Intergenic Region
DRB234-17	0.33	rs1890366	Intergenic Region
		rs2788212	Intergenic Region
DRB234-182	0.35	rs9405676	Intergenic Region
		rs9378389	Intergenic Region
DRB234-23	0.35	rs7751939	Intergenic Region
		rs6597267	Intergenic Region
		rs2317217	Intergenic Region
DRB234-38	0.37	rs11970370	Intergenic Region
		rs845896	Intergenic Region
DRB234-132	0.37	rs6924630	Intergenic Region
		rs9378763	Intergenic Region
		rs6596945	Intergenic Region
DBR234-22	0.37	rs9392155	Intergenic Region
		rs9505192	Intergenic Region
		rs9505153	Intergenic Region
DRB234-108	0.38	rs2449447	Intergenic Region
		rs1773015	Intergenic Region
DRB234-40	0.39	rs1764136	Intergenic Region
		rs845883	Intergenic Region

Table 2. Putative TFBS for the Haploblocks proximal to the Lower Bound. The sequence for the ten haploblocks located proximal to the lower bound region were scanned for putative TFBS using ConSite, which is a web-based tool that finds cis-regulatory elements in genomic sequences via high quality transcription factor models and cross species filters.

Block ID	ARNT	c-FOS	CHOP	COUP-TF	CREB1	C-REL	E2F	E4BP4	FOXA2	FOXD1	FOXD3	FOXF2	FOXI1	FOXQ1	EVI-1	HEN1	HLF	IRF-1	MAX	MEF2	MYC-MAX	MYF	n-MYC	NF-κB	NRF2	PAX6	PBX	P65	ROR-α1	ROR-α2	RUNX1	RXR-VDR	SAP1	SOX5	SOX17	SPZ1	STAF	TEF1	USF
DRB234-174		X				X		X	X	X	X	X	X	X		X							X									X	X	X	X			X
DRB234-63	X	X				X	X	X	X	X	X	X	X	X		X	X		X				X	X			X		X	X	X	X	X	X	X			X
DRB234-17	X	X				X	X			X			X				X		X				X	X	X			X		X		X	X	X	X		X	X
DRB234-182	X	X				X	X		X	X	X		X	X			X		X	X			X					X		X			X	X				X
DRB234-23		X		X	X				X	X	X	X	X	X					X				X			X				X			X	X				X
DRB234-38		X	X				X		X				X				X					X								X			X	X	X			X
DRB234-132	X	X				X	X		X	X	X	X	X	X			X		X			X	X		X			X		X			X	X			X	X
DRB234-22									X	X	X		X		X													X					X	X	X			
DRB234-108	X	X	X					X		X	X		X	X		X	X		X	X	X	X	X							X			X	X	X			X
DRB234-40	X	X				X			X	X	X		X	X		X			X	X		X	X	X		X	X			X		X	X	X	X	X		X

Table 3. SNPs proximal to the Lower Bound with TFBS changes when their minor allele is present. Of the twenty three SNPs interrogated for putative TFBS, these five SNPs were found to have a gain or loss of a TFBS when scanned using ConSite.

SNPs	Alleles	TFBS Change
rs9378763	A>C	Gain of FOXI1 site
rs17464136	C>G	Gain of SOX-17 site
rs11970370	A>C	Loss of SOX-5 site
rs6924630	A>T	Loss of TEF-1 site
rs7751939	C>A	Loss of IRF-1 and gain of FOXA2, FOXD3 & FOXI1 sites

map (i.e. HapMap) has opened a new era in human population genetics. Single nucleotide polymorphisms (SNPs), the most common type of natural variation in the human genome, offer an unprecedented opportunity to investigate evolutionary forces that have shaped human genome variation in natural populations. Information content (IC), as a new metric grounded in the biophysical matrix of DNA-sequence based biology, has been used to explore the biomedical significance of natural variation in the human genome. As the most polymorphically expressed biological system, the HLA region was examined for proof of concept that natural variation encodes fundamental information about the biology of host adaptive mechanisms in response to environmental pathogens. The data on the normalized information content (NIC) of SNP haploblocks in the HLA-DR region relate natural variation to pathways of innate immunity. The molecules and cells of the innate immune system are the first

Table 4. Characterized miRNAs located proximal to the lower bound. The sequence for the ten blocks proximal to the lower bound was scanned for miRNA motifs via the web-based tool miRBase.

Block ID	miRNA
DRB234-23	let-7e
DRB234-17	miR-16
	miR-93
	miR-222
DRB234-63	miR-29
DRB234-132	miR-142-p
	miR-548
DRB234-38	miR-21

responders to environmental disturbances of homeostasis. This system can initiate the inflammatory response by activating the cellular release of cytokines, chemokines, reactive oxygen (ROS), and reactive nitrogen (RNS) species. However, when the inflammatory response is sustained at a chronic level, these molecules can inflict a variety of damaging effects.

In our analysis, the transcription factor binding sites (TFBS) present in all ten of the blocks proximal to the lower bound have been reported to be regulated by the p38 mitogen activated protein kinase (MAPK) and Wnt (wingless*)* pathways (Gazel et al., 2008; Perreault et al., 2001; Zorn et al., 1999). These two pathways are activated in response to oxidative stress and inflammation (Maiese et al., 2008; Nagata et al., 1998). p38 is a kinase that is localized to the cytoplasm until activated, when it translocates to the nucleus. Its activity is critical for normal immune and inflammatory responses. p38 is activated by macrophages, neutrophils, and T cells in response to cytokines, chemokines, and bacterial lipopolysaccharide (LPS) (Roux and Blenis, 2004). It is known to phosphorylate its cellular targets such as the following transcription factors: ATF-1 and -2, MEF2A, Sap-1, Elk-1, NF-κB, Ets-1, and p53 (Roux and Blenis, 2004). An interesting feature of the p38 pathway is that it can regulate the Wnt pathway when activated by ROS and other stressors. Wnt signaling controls a variety of signal transduction pathways that involve protein kinases, caspases, NF-κB, GSK-3β, iNOS and FOXes (Maiese et al., 2008; Du et al., 2006; Savage et al., 2010).

Thus, it is possible that cross-talk between the p38 and Wnt pathways represents a network that has formed in response to oxidative stress. This proposed network would be advantageous to a population under constant challenge in a tropical environment with a plethora of pathogenic agents, such as *Schistosoma mansoni* (*S. mansoni)*. In innate immunity, dendritic cells (DCs) exposed to helminth products (such as Lacto-N-Fucopentaose III (LNFPIII), a milk sugar containing the Lewis^{x} trisaccharide found in the schistosome egg antigen (SEA)), have been reported to activate NF-□B by stimulating its nuclear translocation (Carvalho et al., 2008). It is worth noting that Lewis structures can occur on both N-glycan and mucin-type O-glycan cores, and these fucosylated glycans have been involved in many functions, like selecting recognition (Haltiwanger, 2009). Selectins are known for mediating extravasation of leukocytes and lymphocytes, pathogen adhesion, and modulation of signal transduction pathways (Becker and Lowe, 2003). The hallmark of *S. mansoni* infection is the switching of the host immune response from Th1 to Th2, resulting in the persistent survival of the parasite. One of the key components involved in modulating the host immune response from Th1 to Th2 is NF-□B. Studies conducted by Goodridge et al. (2007) and Thomas et al. (2005), found that neither SEA nor LNFPIII-dextran pulsed NF-κB $^{-/-}$ DCs were able to induce a Th2 response. Interestingly, increased levels of IL-4, a Th2 cytokine, have been demonstrated in murine schistosomiasis to control the generation of reactive oxygen and nitrogen intermediates in the liver (LaFlamme et al., 2001).

Additionally, our analysis showed that all the blocks proximal to the lower bound contained putative miRNA structures. This is consistent with findings of miRNAs in intronic and intergenic regions (Bartel, 2004). There were four blocks that contained well characterized miRNAs (Table 4). These miRNAs have been identified as tumor suppressors in a multitude of cancers (Calin and Croce, 2006; Esquela-Kerscher and Slack, 2006). Furthermore, miRNA expression has been shown to regulate the inflammatory response. For example, O'Connell et al. (2010) reported that an increase in miR-155 and a decrease in let-7e levels enhanced the response of Akt$^{-/-}$ macrophages to LPS.

Particularly noteworthy was the SNP variation proximal to the lower bound, which reflected that five SNPs had TFBS changes when their minor alleles were present (Table 3). rs9378763 and rs1764136 had a gain of FOXI1 and SOX-17 sites, respectively. rs11970970 and rs6924630 had a loss of SOX-5 and TEF-1 sites, respectively. rs7751939 was the only SNP whose minor allele resulted in loss of an IRF-1 site and gain of FOXA2, FOXD3 and FOXI1 sites. These FOX transcription factors have been shown to activate and be activated by the Wnt pathway (Pohl and Knochel, 2001; Zorn et al., 1999). In the study by Zhang et al. (2008), SOX-17 was shown to negatively regulate the Wnt pathway via suppression of β-catenin/T-cell factor-regulated transcription. IRF-1 is constitutively expressed in various cell types and induces the expression of pro-inflammatory cytokines in response to the activation of pathogen recognition receptors, such as the Toll-like receptor and nucleotide-binding oligomerization domain (NOD)-like receptor families (Tamura et al., 2008). Also, IRF1 transcriptionally targets a number of

genes, and is required for Th1 differentiation of interferon (IFN)-stimulated macrophages. When IRF-1 is absent, the induction of Th2-type immune responses occurs (Taki et al., 1997). In addition, when p38 is activated by IFN, it contributes to the phosphorylation of NF-□B, AP-1, IRF-1, IRF-4, IRF-8, and PU.1 (van Boxel-Dezaire et al., 2006). It has been reported that transcription enhancer factor 1 (TEF-1) directly binds to poly (ADP-ribose) polymerase 1(PARP-1) which is known to participate in DNA repair processes (Ha et al., 2002). Under conditions of oxidative stress, the activation of PARP1 results in greater expression of AP-1 and NF-□B-dependent genes (Virag and Szabo, 2002). Also, in a study conducted by Braam et al. (2005) using endothelial cells, SOX-5 had more pronounced representation in genes regulated by nitrous oxide (NO) than the other transcription factors studied. Interestingly, in *falciparum* malaria NO inhibits the adhesion of parasitized red cells to vascular endothelium (Clark et al., 2004). Hence, it is possible that not only has schistosomiasis acted as an environmental stressor in shaping the allelic variation in the proposed p38-Wnt compensatory network, but so has malaria. It is intriguing that independent of the potential gains and losses of transcription factor binding sites, there is continued regulation of the oxidative stress process irrespective of specific allele selection.

We also assessed the performance of the measure by comparing blocks with NIC values most proximal to the upper bound with those most proximal to the lower bound. It is noteworthy that more than 96% of SNPs in the blocks proximal to the upper bound were located in genic regions, in contrast to the SNPs in the blocks proximal to the lower bound, *none* of which were found in genic regions. The identification of genic haploblocks with high information content and a more in-depth assessment of the biological significance of the entire NIC distribution are being investigated.

This paper has introduced a biophysical metric for analyzing the information content of SNP haploblock variation. NIC values in the HLA-DR region highlighted common variants involved in regulation of host immunity to environmental stressors. This supports our hypothesis that information encoded in the structure of SNP haploblock variation can elucidate molecular pathways and cellular mechanisms involved in the regulation of host adaptation to the environment. Using our analysis, p38 and Wnt, are proposed as a communication network connected by transcription factors and miRNAs in population adaptation to pathogens. Since the genetic variation highlighted by NIC values is in a representative sample of the population, its relevance in disease association studies remains to be determined. Because our motivation has been to use common variation in a reference population to interrogate the biology of health, the further understanding of disease using this approach would be a by-product. Finally, NIC values derived from common variation in the HLA-DR region suggest its involvement with regulation of innate immune mechanisms.

ACKNOWLEDGMENTS

The research is supported in part by NIH Grants NCRR 2 G12 RR003048 from the RCMI Program, Division of Research Infrastructure; NIGMS S06 GM08016, and NCI 5U54 CA091431.

Abbreviations: IC, Information content; **SNPs,** single nucleotide polymorphisms; **Haploblocks,** SNP haplotype blocks; **NIC,** normalized information content; **HLA-DR,** human leukocyte antigen-disease related; **MHC,** major histocompatibility complex; **TFBS,** transcription factor binding sites; **miRNA,** MicroRNA; **ARNT,** aryl hydrocarbon receptor nuclear translocator; **c-FOS,** proto-oncogene c-FOS; **CHOP,** DNA damage-inducible transcript 3 protein; **COUP-TF,** COUP transcription factor; **CREB1,** Cyclic AMP responsive element binding protein 1; **c-REL,** proto-oncogene c-REL; **E2F,** transcription factor E2F; **E4BP4,** nuclear factor interleukin 3 related protein; **FOXA2,** Hepatocyte nuclear factor 3-beta; **FOXD1,** forkhead box protein D1; **FOXD3,** forkhead box protein D3; **FOXF2,** Forkhead box protein F2; **FOXI1,** Forkhead box protein I1; **FOXQ1,** Forkhead box protein Q1; **EVI-1,** MDS1 and EV-1 complex locus protein EV1; **HEN1,** helix-loop-helix protein 1; **HLF,** endothelial PAS domain containing protein 1; **IRF-1,** interferon regulatory factor 1; **MAX,** protein Max; **MEF2,** myocyte-specific enhancer factor 2A; **MYC-MAX,** MYC proto-oncogene protein-protein MAX; **MYF,** myogenic factor; **n-MYC,** N-myc proto-oncogene protein; **NF-□□,** NF-kappa beta; **NRF2,** Nuclear factor erythroid 2-related factor 2; **PAX6,** paired box protein PAX6; **PBX,** Pre-B cell leukemia transcription factor; **p65,** Transcription factor p65; **ROR-α1,** Nuclear receptor ROR- **α**1; **ROR- α2,** nuclear receptor ROR-**α**2; **RUNX1,** runt related transcription factor 1; **RXR-VDR,** retinoic acid receptor RXR-vitamin D receptor complex; **SAP1,** Receptor type tyrosine protein phosphatase H; **SOX5,** transcription factor SOX5; **SOX17,** transcription factor SOX17; **SPZ1,** Spermatogenic leucine zipper protein 1; **STAF,** zinc finger protein 143; **TEF1,** transcriptional enhancer factor TEF-1; **USF,** upstream stimulatory factor; **HapMap,** haplotype map project; **ROS,** reactive oxygen species; **MAPK,** mitogen activated protein kinase; **Wnt,** Wingless; **LPS,** lipopolysaccharide; **ATF1,** cyclic AMP dependent transcription factor ATF-1; **ATF2,** cyclic AMP dependent transcription factor ATF-2; **Elk1,** ETS domain containing protein Elk1; **Ets1,** protein C-ets-1; **p53,** cellular tumor antigen p53; **GSK-3β,** glycogen synthase kinase 3 beta; **iNOS,** inducible nitric oxide synthase; ***S. mansoni*,** *Schistosoma mansoni*; **LNFPIII,** lacto-N-Fucopentaose **III; SEA,** schistosome egg antigen; **Th1,** T-helper 1 cells;

Th2, T-helper 2 cells; **DCs,** dendritic cells; **IL-4,** interleukin 4; **NOD,** nucleotide binding oligomerization domain; **IFN,** interferon; **AP1,** transcription factor AP1; **PARP1,** poly(ADP-ribose) polymerase 1.

REFERENCES

Barrett JC, Fry B, Maller J, Daly MJ (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformation, 21: 263-265.

Bartel D (2004). MicroRNAs: genomics, biogenesis, mechanism and function. Cell, 116: 281-297.

Becker D, Lowe J (2003). Fucose: biosynthesis and biological function in mammals. Glycobiology, 13: 41R-53R.

Braam B, de Roos R, Bluyssen H, Kemmeren P, Holstege F, Joles JA, Koomans H (2005). Nitric oxide-dependent and nitric oxide-independent transcriptional responses to high shear stress in endothelial cells. Hypertension, 45: 1-9.

Calin G, Croce C (2006). MircoRNA signatures in human cancers. Nat. Rev. Cancer, 6: 857-866.

Carvalho L, Sun J, Kane C, Marshall F, Krawczyk C, Pearce EJ (2008). Review series on helminths, immune modulation and the hygiene hypothesis: Mechanisms underlying helminth modulation of dendritic cell function. Immunol., 126: 28-34.

Clark I, Alleva L, Mills AC, Cowden W (2004). Pathogenesis of malaria and clinically similar conditions. Clin. Microbiol. Rev., 17: 509-539.

Du Q, Park KS, Guo Z, He P, Nagashima M, Shao L, Sahai R, Geller DA, Perwez Hussain S (2006). Regulation of human nitric oxide synthase 2 expression by Wnt □-catenin signalling. Cancer Res., 66:7024-7031.

Esquela-Kerscher A, Slack F (2006). Oncomirs-microRNAs with a role in cancer. Nat. Rev. Cancer 6: 259-269.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart I, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ., Altshuler D (2002). The Structure Blocks in the Human Genome. Sciences, 296: 2225-2229.

Gazel A, Nijhawan RI, Walsh R, Blumenberg M (2008). Transcriptional profiling defines the roles of ERK and p38 kinases in epidermal keratinocytes. J. Cell. Physiol., 215: 292-308.

Goodridge H, McGuiness S, Houston KM, Egan CA, Al-Riyami L, Alcocer MJC, Harnett MM, Harnett W (2007). Phosphorylcholine mimics the effects of ES-62 on macrophages and dendrictic cells. Parasite Immunol., 29: 127-137.

Ha H, Hester L, Snyder S (2002). Poly (ADP ribose) polymerase-1 dependence of stress-induced transcription factors and associated gene expression in glia. PNAS, 99: 3270-3275.

Haltiwanger R (2009). Fucose is on the TRAIL of colon cancer. Gastroenterol., 137: 36-39.

Haploview (2003). http://www.broadinstitute.org/scientificcommunity/science/programs/medical-and population-genetics/Haploview/Haploview

International Human Genome Sequencing Consortium (2003). The International HapMap project. Nat., 426: 789-794.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nat., 409: 860-921.

LaFlamme AC, Patton EA, Bauman B, Pearce EJ (2001). IL-4 plays a crucial role in regulating oxidative damage in the liver during schistosomiasis**.** J. Immunol., 166: 1903-1911.

Maiese Y, Li F, Chong ZZ, Shang YC (2008). The Wnt signalling pathway: Aging gracefully as a protectionist. Pharmacol. Ther., 118: 58-81.

Mason TE, Ricks-Santi L, Lindesay J, Kurian P, Hercules W, Dunston GM (2009). Mining the Information Content of Natural Variation in Health Disparity Research. The American Society of Human Genetics, 59th Annual Meeting, Honolulu, Hawaii, US, October 2009; 1942/T/Poster Board #491.

Nagata Y, Takahashi N, Davis RJ, Todokoro K (1998). Activation of p38 MAP kinase and JNK but not ERK is required for erythropoietin-induced erythroid differentiation. Blood, 92: 1859-1869.

Nothnagel M, Furst R, Rohde K (2004). Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum. Hered., 54: 186-198.

O'Connell R, Rao D, Chaudhuri A, Baltimore D (2010). Physiological and pathological roles for MircoRNAs I the immune system. Nat. Rev. Immunol., 10: 111-122.

Perreault N, Katz JP, Sackett SD, Kaestner KH (2001). FOXI1 controls the Wnt/β-catenin pathway by modulating the expression of proteoglycans in the gut. J. Biol. Chem., 276: 43328-43333.

Pohl B, Knochel W (2001). Overexpression of the transcriptional repressor FOXD3 prevents neural crest formation in *Xenopus* embryos. Mech. Dev., 103: 93-106.

Roux P, Blenis J (2004). ERK and p38 MAPK-activated protein kinases: a family of protein kinases with diverse biological functions. Microbiol. Mol. Biol. Rev., 68: 320-344.

Savage J, Voronova A, Mehta V, Sendi-Mukasa F, Skerjanc IS (2010). Canonical Wnt signalling regulated FOXC1/2 expression in P19 cells. Differentiation, 79: 31-40.

Shannon C (1948). A mathematical theory of communication. Bell Syst. Tech. J., 27: 379-423.

Susskind L, Lindesay J (2005). An Introduction to Black Holes, Information and the String Theory Revolution. World Scientific Publishing Company, New Jersey, US.

Taki S, Sato T, Ogasawara K, Fukuda T, Sato M, Hida S, Mitsuyama M, Shin EH, Kojima S, Taniguchi T, Asano Y (1997). Multistage regulation of Th1 type immune responses by the transcription factor IRF1. Immunity, 6: 673-679.

Tamura T, Yanai H, Savitsky D, Taniguchi T (2008). The IRF family transcription factors in immunity and oncogenesis. Annu. Rev. Immunol., 26: 535-584.

Thomas PG, Carter MR, Da'dara AA, DeSimone TM, Harn DA (2005). A helminth glycan induces APC maturation via alternative NF-kappa B activation independent of I kappa B alpha degradation. J. Immunol., 175: 2082-2090.

van Boxel-Dezaire A, Rani M, Stark G (2006). Complex modulation of cell type-specific signalling in response to Type 1 interferons. Immunity, 25: 361-372.

Virag L, Szabo C (2002). The therapeutic potential of Poly (ADP-ribose) polymerase inhibitors. Pharmacol. Rev., 54: 375-429.

Zhang W, Glockner S, Guo M, Machida EO, Wang DH, Easwaran H, Van Neste L, Herman JG, Schuebel KE, Watkins DN, Ahuja N, Baylin SB (2008). Epigenetic inactivation of the canonical Wnt antagonist SRY-Box containing gene 17 in colorectal cancer. Cancer Res., 68: 2764-2772.

Zorn AM, Barish GD, Williams BO, Lavender P, Klymkowsky MW, Varmus HE (1999). Regulation of Wnt signalling by SOX proteins: XSoxa/b and XSox3 physically interact with β-catenin. Mol. Cell., 4: 487-498.

URLs:

dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/)
HapMap (http://hapmap.ncbi.nlm.nih.gov/)
ConSite (http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite)
miRBase (http://www.mirbase.org/search.shtml)