Computational Biology and Chemistry 64 (2016) 94–98

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

A model for the clustered distribution of SNPs in the human genome

Chang-Yong Lee

The Department of Industrial and Systems Engineering, Kongju National University, Cheonan 330-717, South Korea

ARTICLE INFO

Article history:
Received 14 January 2016
Received in revised form 16 April 2016
Accepted 6 June 2016
Available online 8 June 2016

Keywords:
Single nucleotide polymorphism
Mutation
Human genome
Probability distribution
Hapmap

ABSTRACT

Motivated by a non-random but clustered distribution of SNPs, we introduce a phenomenological model to account for the clustering properties of SNPs in the human genome. The phenomenological model is based on a preferential mutation to the closer proximity of existing SNPs. With the Hapmap SNP data, we empirically demonstrate that the preferential model is better for illustrating the clustered distribution of SNPs than the random model. Moreover, the model is applicable not only to autosomes but also to the X chromosome, although the X chromosome has different characteristics from autosomes. The analysis of the estimated parameters in the model can explain the pronounced population structure and the low genetic diversity of the X chromosome. In addition, correlation between the parameters reveals the population-wise difference of the mutation probability. These results support the mutational non-independence hypothesis against random mutation.

1. Introduction

The most common type of genetic variants in the human genome is the single nucleotide polymorphism (SNP), which, as a result of mutation, has a difference in a single nucleotide within a population of samples (Barreiro et al., 2008). SNP data, together with gene expression and other biological information, are an important resource to answer various biological questions regarding the genetic variation, such as the mutational pattern of the genome, the phylogenetic classification, and the association with phenotype data.

In recent years, as the cost of genotyping has dropped dramatically due mainly to the advance in the genotyping technology (Metzker, 2010; LaFramboise, 2009), much effort has been put into the identification of SNPs in the human genome (International HapMap Consortium, 2005; The 1000 Genomes Project Consortium, 2010, 2012). Notably, the International HapMap project (International HapMap Consortium, 2005, 2007; The International HapMap 3 Consortium, 2010) (hereafter, Hapmap) is an international effort to identify the genetic variation in the human genome to develop a haplotype map. Although Hapmap includes some datasets on the copy number variation, SNP data are the main resource not only for understanding and characterizing the differences in genome structure but for association studies with diseases and/or environmental factors.

It has been known that SNPs in the human genome are not distributed randomly but clustered across the genome (Amos, 2010; Tenaillon et al., 2008; Koboldt et al., 2006; Hellmann et al., 2005; Lindblad-Toh et al., 2000). This clustering property suggests that mutations tend to occur not randomly but preferentially to the proximity of existing mutations. In addition to the interpretation of the clustering as the reflection of mutational hotspots (Rogozin and Pavlov, 2003), clustered SNPs can emerge in various ways. Natural and balancing selections can modulate local variability and tend to create regions of increased variability that results in non-randomness (Bubb et al., 2006). A high variance of genes within a population of samples in the time to the most recent common ancestor causes different recombination rates in a chromosome (Eriksson et al., 2002). It was also proposed that microsatellites can also generate mutational biases in their flanking regions by expansion and erosion from the perspective of microsatellite evolution (Vowles and Amos, 2004; Webster and Hagberg, 2007; Varela et al., 2008). Clustered SNPs can also arise from ascertainment biases in the SNP discovery process (Kuhner et al., 2000). Examples include the SNP identification based on maximally dissimilar sequences, the usage of not enough samples, and finding all possible SNPs not on a whole genome but on a given region of a chromosome.

However, when SNP clusters are found throughout a whole genome with a large number of samples from different global populations, it is unlikely that the observed clusters would be due to ascertainment biases. Thus, as pointed out in reference Amos (2010), the majority of SNP markers along a whole genome should reflect the underlying mutation pattern. In this respect, a non-random mutation process was proposed and tested against the

E-mail address: clee@kongju.ac.kr

http://dx.doi.org/10.1016/j.compbiolchem.2016.06.003

Table 1
The number N of SNPs in each of single population (left two columns) and in each chromosome averaged over 11 populations (right four columns). The population names are abbreviated and the full names can be found in reference International HapMap Consortium (2005, 2007) and hap2. Note that the number of SNPs in each single population and the total number of SNPs of all chromosomes are less than 1,440,616 SNPs obtained from all global populations.

Pop. name	N	Chr.	N	Chr.	N
ASW	1,399,533	1	102,848	13	46,752
CEU	1,269,095	2	103,994	14	40,899
CHB	1,181,090	3	86,600	15	37,884
CHD	1,173,514	4	77,006	16	39,392
GIH	1,260,550	5	79,132	17	33,734
JPT	1,154,331	6	82,545	18	36,802
LWK	1,366,422	7	67,937	19	23,092
MEX	1,324,625	8	67,576	20	32,426
MKK	1,385,579	9	57,444	21	17,557
TSI	1,266,622	10	66,389	22	17,826
YRI	1,340,306	11	63,333	X	41,408
		12	61,202	Total	1,283,778

random mutation by generating a semi-realistic population of chromosomes from stochastic computer simulations that implements the concept of ‘the sphere of influence’ (Amos, 2010). As millions of SNPs on a whole genome are now available in public domains, the mutation pattern can be systematically investigated.

In this paper, we propose a probabilistic model for the clustered distribution of SNPs. The proposed model assumes non-independent mutations in which subsequent mutations occur not randomly but preferentially to near mutated sites. Within the model, SNP clusters could form mainly through a non-negligible tendency of the mutation process in the closer proximity of existing SNPs. The proposed model was tested against Hapmap SNP data and the proposed model was confirmed as suitable to explain empirical SNP distributions of the human genome. We also tested the proposed model against the random mutation model in which all mutations occur independently, and we confirm that the proposed model explains the distribution more appropriately than the random model.

As the X chromosome is a haploid in males, its SNP distribution may have characteristics different from the distributions of the autosomes. With the estimated parameters in the proposed model, we characterized the clustered SNP distributions obtained from different chromosomes, including the X chromosome. Whereas the proposed model is valid irrespective of the ploidy (i.e., either diploid or haploid), our analysis of estimated parameters accounts for the characteristics, such as the pronounced population structure and the low mutation rate, specific to the X chromosome.

2. Materials and methods

2.1. Data

We use the genome-wide Hapmap SNP data of Phase III, which consists of 1,440,616 SNPs in 1184 reference individuals from 11 global ancestry groups of three continental regions. The data are publicly available and can be downloaded at http://hapmap.ncbi.nlm.nih.gov/. To investigate the population-specific differences, we extracted SNPs that are polymorphic within each of a single population. For each of 11 global populations, we analyzed SNP distributions on 22 autosomes and the X chromosome. Generally, the number of SNPs depends on the number of samples, the chromosome, and the genetic diversity of the population. Thus, the number of identified SNPs may fluctuate with populations and chromosomes.

Table 1 shows the number of SNPs identified in each population summed over all chromosomes and each chromosome averaged over 11 global populations. The number of SNPs for each population in Table 1 roughly illustrates the regional difference in the genetic diversity. The populations that originated from Africa (ASW, LWK, MKK, YRI) have a larger number of SNPs than other continental regions, indicating a higher genetic diversity. On the other hand, the population from Asia (CHB, CHD, JPT) have a smaller number of SNPs and a lower genetic diversity than others.

The clustering property of SNPs can be represented by taking the ordered locations of all SNPs and examining how proximate they are located. We quantify the proximity of SNPs in terms of the SNP space, which is defined as the number of nucleotides between two adjacent SNPs in their ordered locations. Specifically, let ℓ_i be the location, in the number of nucleotides counting from 5′ of a sequence, of the ith SNP in a chromosome. Then, the ith SNP space s_i is defined as

$$s_i \equiv \ell_{i+1} - \ell_i, \quad i = 1, 2, \ldots. \tag{1}$$

2.2. Random model

An SNP is the result of a mutation and it is commonly assumed that the mutation arises randomly in DNA sequence. Under this assumption, we tested the hypothesis that SNPs are distributed randomly in a sequence. In the random mutation model, mutations occur independently of each other with a constant probability β of $0 < \beta < 1$. With the random model, the probability distribution of the SNP space defined in Eq. (1) can be derived as follows. Suppose that an SNP is found at the location ℓ_k. Then, the probability of finding a subsequent SNP at the location $\ell_{k+1} = \ell_k + s$ is given as

$$p(s\,|\,\beta) = (1-\beta)^{(s-1)}\beta, \quad \text{for } s = 1, 2, \ldots \tag{2}$$

Note that Eq. (2) is the probability mass function of a geometric distribution (Pitman, 1993). The geometric distribution is the discrete analog of the exponential distribution and has a property of being memoryless. The distribution is often used for modeling the number of trials until the first success, in our case, an SNP.

By taking a logarithm on the both sides of Eq. (2), we get

$$\ln p(s\,|\,\beta) = \ln(1-\beta)s + \ln\frac{\beta}{1-\beta}. \tag{3}$$

This illustrates that $\ln p(s\,|\,\beta)$ is linear in s with $\ln(1-\beta) < 0$ being the proportionality, or a slope. The parameter β can be estimated, for example, by the maximum likelihood estimation (MLE). Thus, if mutations occur randomly, we expect that $\ln p(s\,|\,\beta)$ should be a straight line in s with a negative slope of $\ln(1-\beta)$.

In Fig. 1, we plot the empirical distribution of the SNP spacing s for different chromosomes of ASW. A comparison with the random model reveals that the empirical distributions do not follow a geometric distribution (the dotted line), especially for small values of s in which the major deficiency of the model occurs. In particular, Fig. 1 shows that the probability increases sharply and non-linearly as s becomes small which suggests that the SNPs are clustered. This tendency of the distribution is more or less independent of the chromosome. This reflects that SNPs are distributed not randomly, but clustered. From these we can infer that the random mutation hypothesis is inadequate to explain the clustering property of SNPs.

2.3. Proposed model

The clustered SNPs imply that when a mutation occurs, another mutation is more likely to occur as they are closer in their locations. This suggests that the mutation probability is not independent of the location but dependent on how close the mutations are in their locations. This non-independent mutation can be modeled after the mutation probability being inversely proportional to some power of the separation in nucleotides between two consecutive mutations. Formally, given that a mutation occurs at the location