Protein Cell 2014, 5(2):99–112
DOI 10.1007/s13238-013-0004-1

Protein & Cell

REVIEW

Genome-scale analysis of demographic history and adaptive selection

Qi Wu, Pingping Zheng, Yibu Hu, Fuwen Wei✉

Key Lab of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
✉ Correspondence: weifw@ioz.ac.cn (F. Wei)
Received July 23, 2013 Accepted November 4, 2013

ABSTRACT

One of the main topics in population genetics is identification of adaptive selection among populations. For this purpose, population history should be correctly inferred to evaluate the effect of random drift and exclude it in selection identification. With the rapid progress in genomics in the past decade, vast genome-scale variations are available for population genetic analysis, which however requires more sophisticated models to infer species' demographic history and robust methods to detect local adaptation. Here we aim to review what have been achieved in the fields of demographic modeling and selection detection. We summarize their rationales, implementations, and some classical applications. We also propose that some widely-used methods can be improved in both theoretical and practical aspects in near future.

KEYWORDS genomics, demographic history, local adaptation, natural selection

INTRODUCTION

Identifying adaptive selection has been a central issue in the study of molecular evolution, since Kimura Motoo (1968) argued that it is neutrality instead of selection driving the majority of variations in DNA. There has also been long interest in understanding the nature of selection in study of domestication since Charles Darwin (1859), during which artificial selection leads to phenotypic and genetic variation distinguishing domesticated organisms from their wild ancestors (Mannion, 1999). The inference of demographic history of related population(s) plays a vital role for these aims, for the reason that a proper inferred model could offer a null hypothesis for expectation of neutrality (Nielsen et al., 2007). To distinguish selective traits from those caused by bottleneck effects, understanding of the population history that the first population captured from wild became domestic population has also been a vital task (Axelsson et al., 2013). Besides, demographic models inferred from genetic data complement archeological evidence in understanding prehistorical events, such as number and timing of major continental fluctuations of population size as well as migration. Therefore, the research of demographic history as well as adaptive selection play essential role in evolutionary biology.

In the past decade, there has been an explosive progress in genomics (International Human Genome Sequencing Consortium, 2001; Li et al., 2010; Huo et al., 2012; NCBI Resource Coordinators, 2013). The explosion started from the revolution of sequencing technique and stimulated accumulation of genomic data, which subsequently pushed the improvement of analysis methods. Today, the data accumulation rate is hundreds of times higher than that when the Human Genome Project was first stated (NCBI Resource Coordinators, 2013). The available genomic data have been extended for various species, from the initial goal of human and key lab model species to primates and domesticated animals and plants, and presently to endangered organisms with special scientific or cultural values (NCBI Resource Coordinators, 2013; Grigoriev et al., 2013). Such a data flood made population genetics approaches widely applied in numerous organisms, which in turn stimulated the development of population genetic approaches, for example, to infer more detailed demographic history, to identify adaptive selection more accurately and sensitively, and to perform the computation more rapidly with more data and less constraints (Nielsen et al., 2007; Crisci et al., 2012).

For demographic inference, the most straightforward and simplest approach was based on polymorphic data organized

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

in a Site Frequency Spectrum (SFS). A coalescent process or diffusion process could be applied to trace the history of species. In that way a series of parameters describing the history could be inferred by maximum likelihood, Bayesian approximation or Markov Chain Mento Carlo method (MCMC) (Crisci et al., 2012). As for the identification of local adaptation, population genetic statistic methods were applied to seek the outliers of genetic variation and differentiation across genomes within and between species driven by selection forces (Sabeti et al., 2006). In this review we focused on some recent advances on demographic inference as well as identification of adaptive selection. We also provided some perspectives related to the improvement of those mentioned approaches. We suggested that the genome point of view might contribute to the future progress of population genomics, in both theoretical and applicable aspects.

APPROACHES ON DEMOGRAPHIC HISTORY WITH GENOME-SCALE DATA

To infer demographic history is to estimate population events in the past with population data at present. With the development of the next-generation genome sequencing technique, the present population data could be either mass of genome-scale traditional molecular polymorphism data from multiple individuals or heterozygosity data obtained from one whole genome sequence. We will discuss them respectively.

Methods with polymorphism dataset

In order to make inference about the population history, two steps are needed. One must firstly formulize the history of the population with certain mathematical model, in which the evolutionary affair during the history was described as a set of parameters. Secondly statistical inference methods could be implied on the parameters. In the past decade, the coalescent process was the most widely used model (Nielsen and Wakeley, 2001; Crisci et al., 2012). Recently, a diffusion approach was adopted as well. These two approaches are discussed in the section of "Coalescent process versus diffusion process". The statistical inference methods are discussed in the section of "Methods to infer demographic parameters".

Coalescent process versus diffusion process

The isolation-with-migration model (Nielsen and Wakeley, 2001) was one of the most common models to infer demographic scenarios, under which different methods could be used to trace the evolutionary history of the genetic variation. Straightforwardly, the genealogy of alleles could be traced backward in time under the process of coalescence, during which the parameters of demographic model could be derived, including the change of effective population size and the time point on events of bottleneck and exponential growth. When two or more populations were considered, the divergence time could also be derived. If not limited to the consideration of the Wright-Fisher model, coalescent method could consider migration rate and recombination rate in the gene tree as well. Actually, coalescent method is the most widely used method and has been applied in numerous demographic inference programs (Wooding and Rogers, 2002; Adams and Hudson, 2004; Hey and Nielsen, 2004; Thornton and Andolfatto, 2006; Becquet and Przeworski 2007; Lopes et al., 2009; Hey, 2010).

The polymorphism dataset could be organized as the Site Frequency Spectrum (SFS), which is the distribution of allele frequencies in a sampled dataset. In the case of multiple populations, a joint SFS (JSFS) could be used, and the evolution of genetic polymorphism among populations could be described as the change over time of allele distribution in the SFS/JSFS. In the context of neutral theory, the change could be approximated with a diffusion process. The Kolmogorov forward equation for diffusion approximation of neutrality could be introduced to approximate the distribution of allele frequencies at given time (Hartl and Clark, 2007). Different from the methods based on coalescent process, the method based on diffusion process could provide more flexible demographic history model with acceptable computational performance (Gutenkunst et al., 2009), and it has been used to deal with complicated demographic model including three populations with migration and recombination based on genome-scale SNP dataset (Gutenkunst et al., 2009; Zhao et al., 2013).

We would like to review some more about the implementation using the method of diffusion process, not only because the diffusion process has been a classic model in population genetics (Kimura, 1955), but also because it is a distinctive and fairly novel method in demographic inference. A comprehensive implementation named ∂a∂i has been developed until recently (Gutenkunst et al., 2009). Based on the joint distribution of allele frequencies across biallelic variants from multiple populations, the program uses compositional likelihood method to infer expected SFS under a specific demographic model via an evolution process simulated with diffusion. With the assumption of infinite-sites model and Wright-Fisher model, the evolution of density distribution of derived mutations $\Phi(x_1, x_2, \ldots, x_P; t)$ in P populations could be formulated as

$$\frac{\partial}{\partial \tau}\Phi = \frac{1}{2}\sum_{i=1}^{P}\frac{\partial^2}{\partial x^2}\frac{x_i(1-x_i)}{v_i}\Phi - \sum_{i=1}^{P}\frac{\partial}{\partial x_i}\left[\gamma_i x_i(1-x_i) + \sum_{j=1}^{P} M_{i\leftarrow j}(x_j - x_i)\right]\Phi \quad (1)$$

Here $\tau = \frac{t}{2N_{ref}}$ is the time unit, where t is the time in generations and N_{ref} is the reference effective population size. x means the population frequencies runs from 0 to 1. $v_i = \frac{N_i}{N_{ref}}$ is the relative effective size of population i. $M_{i\leftarrow j} = 2N_{ref}m_{i\leftarrow j}$ is the scaled migration rate, where

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

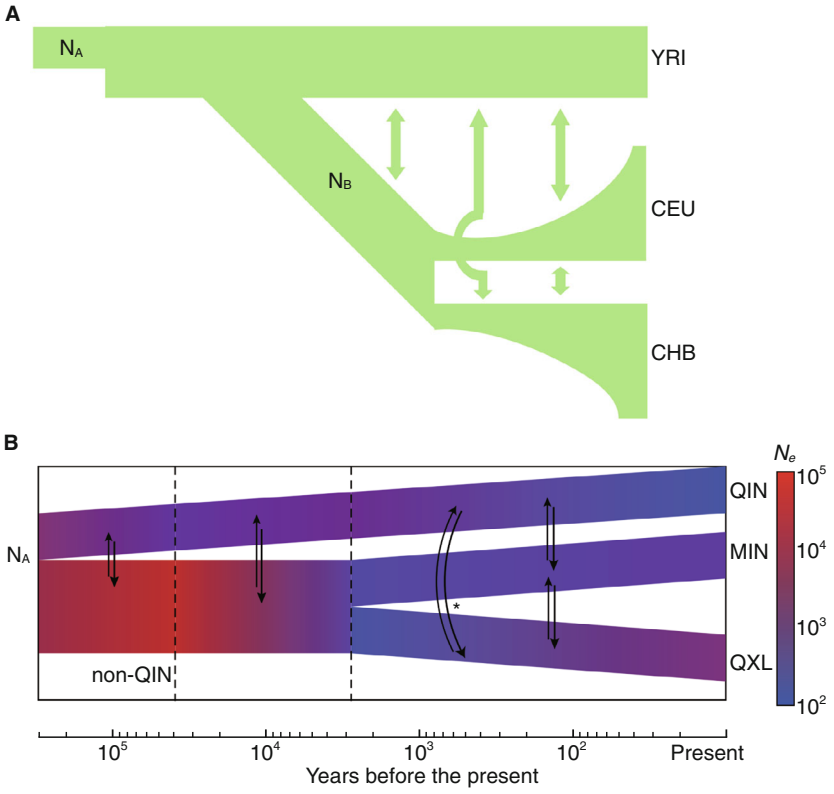


Figure 1. Human and giant panda demographic history inferred by ∂a∂i (See details in Gutenkunst et al. (2009) and Zhao et al. (2013)). Here we focus on the complexity of the two demographic models. Both involve the evolution of three populations (two split from the ancestral population) and changes of the effective population size. (A) The demographic history of three human populations. Population A is the ancestral population whose effective size is shown as N_A. Population B is the population out of the Africa with effective size N_B. YRI is the Yoruba individuals from Ibadan, Nigeria. CHB means Han Chinese in Beijing, China. CEU means CEPH Utah residents with European ancestries. (B) The demographic history of three giant panda populations. A is the ancestral population with effective size N_A. The non-QIN and QIN are divergent populations of population A, and MIN and QXL are divergent populations of non-QIN. QIN and MIN represent the panda population in the Qinling Mountains and Minshan Mountains, respectively. QXL represents a combined population from the Qionglai, Daxiangling, Xiaoxiangling, and Liangshan Mountains. The asterisk in the figure shows the asymmetric migration from QIN to QXL.

$m_{i \leftarrow j}$ is the proportion of "chromosomes" per generation in population i that are new migrants from population *j*. And $\gamma_j = 2N_{ref}s_i$ is the scaled selection, in which s_i is the relative selective advantage of variants in population *i*.

The program uses single nucleotide polymorphism (SNP) data in a given genomic region as input dataset, the region of which would be as large as the whole genome obtained by genome resequencing. If outgroup is used, a statistical correction is needed for ancestral state misidentification (Hernandez et al., 2007). Such variations, which is caused by varying mutation rates across sites and over time, violate the parsimony assumption that the ancestral state of each SNP matches the orthologous allele in the outgroup locus (Hwang and Green, 2004). In the original paper of ∂a∂i which inferred a population history of human, a tri-nucleotide transition rate matrix for primate lineage was used for the correction of the misidentification. Since the tri-nucleotide transition rate matrix varies among different mammalian lineages, a customized matrix should be inferred accordingly.

The usage of diffusion approximation offers several advantages of ∂a∂i (Gutenkunst et al., 2009). It considers multiple populations in historical time with population size fluctuations and asymmetric migrations. It goes beyond the assumption of independent non-recombining regions. It uses the full dataset instead of a restricted summary to guarantee the statistical power. Finally and most importantly, ∂a∂i offers a great flexibility to the demographic model design and can be used to model complicated demographic scenarios. With the rapid falling of sequencing cost, ∂a∂i program has been used in several model and non-model species for the whole genome resequencing data. It has been used in soybean to infer domestication history, usually including a bottleneck and a following effective population size fluctuation in the time span of thousand years (Lam et al., 2010). Besides two-

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

population models, ∂a∂i also works well in constructing three-population models as shown in human and giant panda (Gutenkunst et al., 2009; Zhao et al., 2013). In human, a SNP data set of 68 individuals from four populations (YRI, CHB, CEU, and MXL) are used to model human expansion out of Africa and settlement of the New World. To take the out-of-Africa model as an example, three populations, YRI, CHB, CEU, are involved. Population A is the ancestral population. After a population expansion, population B diverges from population A, then split into CEU and CHB with following increases of effective population size respectively (Gutenkunst et al., 2009). In giant panda, a SNP data set of 34 wild individuals from three populations, QIN, MIN, and QXL, are used. The first stage was that ancestral population split into two populations of QIN and non-QIN. Then non-QIN experienced an increase and a following decrease in population size. MIN and QXL split after the decrease of non-QIN, whereas the QIN showed small fluctuations after its split from the ancestor. Besides, a significant asymmetric migration was found from QIN to QXL (Zhao et al., 2013). The brief illustration of the models is shown in Fig. 1.

Methods to infer demographic parameters

Demographic history is the population events in the past, while the genetic diversity data available is contemporary. Therefore, one has to find out proper estimation of historical parameters which give the best fit to the present polymorphism dataset. Several alternative statistical inference procedures could be used for these purposes as discussed below. Here we discuss maximum likelihood method and Bayesian approximation method. We also discuss some about Markov Chain Mento Carlo (MCMC) method, which is widely used in Bayesian computation.

In essence, inference of demographic history could be a statistical procedure. It looks for the most possible distribution pattern of SFS under the constraints of given demographic parameter set that fits the real dataset sampled from the population. Therefore, it would be natural to introduce the Maximum Likelihood (ML) method, which estimates the most possible measure of probability in the probability space that fits the known sample with the highest likelihood.

If the sites in the observed dataset are unlike, they could be regarded as statistically independent. Thereby the log-likelihood with the condition of a hypothetical population history model is

$$L(D\,|\,H) = \sum_{k=1}^{n-1} S_k \ln \sigma_k \tag{2}$$

in which $L(D\,|\,H)$ means the log-likelihood of dataset D under the condition of population history model, S_k is the number of site occurring k times in the sample and σ_k is the probability of a polymorphism sites occurring k times in the sample (Wooding and Rogers, 2002). The method has been used to test the fluctuation of effective population size in human population history. In ∂a∂i, the composite likelihood scheme was used (Gutenkunst et al., 2009). With a given distribution of polymorphism of P populations, the expected value of each entry of the SFS was defined as $M[d_1, d_2, \ldots, d_P]$, where the d is the same as the k in Wooding and Rogers (2002). So the likelihood equation was written as

$$L(\Theta\,|\,S) = \prod_{i=1}^{P} \prod_{d_i=0}^{n_i} \frac{e^{-M[d_1, d_2, \ldots, d_P]} M[d_1, d_2, \ldots, d_P]^{S[d_1, d_2, \ldots, d_P]}}{S[d_1, d_2, \ldots, d_P]!} \tag{3}$$

In which $S[d_1, d_2, \ldots, d_P]$ is the joint SFS of the P populations, $L(\Theta|S)$ is the likelihood function of the joint SFS under the diffusion model with the parameter set of Θ.

The Approximation Bayesian Computation (ABC) is another method to simulate the parameter values from demographic models that could have given rise to the observed dataset. Suppose the observed dataset x and the joint density of parameter values θ that defined the population history model, the probability of θ with given x could be considered as a posterior of $p(\theta|x)$ according to the Bayesian formula. Thus the essential of the computations become the integral of a certain function of the posterior distribution. The method does work in the case that the posterior distribution was simple or low dimensional, for example, the fluctuation of effective population size on single population (Thornton and Andolfatto, 2006). When the complicated model is considered, the computation becomes a complicated high-dimensional integral which is intractable. Therefore a summary statistics could be used to a restricted set of data to simplify the computation. popABC (Lopes et al., 2009) was following this way, which makes it possible to consider both recombination and migration with genomic data. The program as well as its successive implementations has been used in the detection of rapid radiation in spiny lobsters (Palero et al., 2009), the inference of Africa pygmies demographic history (Batini et al., 2011), as well as the recombination rate variation and the speciation study in rodents (Nachman and Payseur, 2012).

Summary statistics only uses part of information in dataset, which may reduce the statistical power. To use full information of dataset and avoid complicated computation of posterior, the Markov Chain Monte Carlo (MCMC) method was introduced to overcome the difficulty of the complicated high-dimension integral of the posterior in ABC. The method uses a series of sampling based on constructing a Markov chain to get a reliable inference to the probability distribution of the total, instead of computation of high-dimension integral. It starts from settling an initial distribution of the total as prior. Then a series of sampling from the total are performed. For each sampling a distribution could be calculated as the posterior distribution and used to correct the prior one. When sampling times are large enough and the posterior distribution tends to be stable, the stable distribution of samples

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

Table 1. Comparison of some recent demographic models

Citations	Simulation method	Inference procedure	Summary statistics	Migration	Recombination
Wooding and Rogers, 2002	Coalescent process	ML	No	No	No
Adams and Hudson, 2004	Coalescent process	ML	No	No	No
Hey and Nielsen, 2004	Coalescent process	MCMC	No	Yes	No
Hey, 2010	Coalescent process	MCMC	No	Yes	No
Thornton and Andolfatto, 2006	Coalescent process	ABC	No	No	Yes
Becquet and Przeworski, 2007	Coalescent process	MCMC	Yes	Yes	Yes
Lopes et al., 2009	Coalescent process	ABC	Yes	Yes	No
Gutenkunst et al., 2009	Diffusion approximation	ML	No	Yes	Yes

(exactly the equilibrium distribution of Markov chain) could be considered as the distribution of the total (Beaumont, 2010). The MCMC method was powerful to give a simulation result for the posterior distribution of Bayesian computation particularly for complicated demographic pattern.

The methods for demographic history inference using polymorphism dataset mentioned above were compared in Table 1. From the table it can be seen that majority of the simulation method is coalescent process, that nearly half of the inference procedures is maximum likelihood approach, others is approximation Bayesian computation, and three of the five ABC procedures uses MCMC method. Of the two methods using ML in coalescent process (the first two lines in Table 1), Wooding and Rogers' method (2002) considered only one population. While Adams and Hudson's considered two populations. Comparing the two methods using MCMC with full statistics (the line 3 and 4 in Table 1), Hey and Nielsen's method (2004) considered just two populations while Hey's extensive method (2010) considered more than two populations.

Model with heterozygosity dataset

Beside above mentioned approaches, another novel approach using heterozygotes of one genome is applied to obtain information of the population parameters (McVean and Cardin, 2005). Such a method, named as the Pairwise Sequentially Markovian Coalescent (PSMC) model, was used to infer human population history (Li and Durbin, 2011). The PSMC model considered the local density of heterozygous sites along chromosomes which reflects how the constant Most Recent Common Ancestors (MRCA, or TMRCA abbreviated by Li and Durbin (2011)) were separated by historical recombination events. Therefore the population parameters, such as the past effective population size and the recombination rate could be inferred.

PSMC deals with heterozygotes in diploid genome. The free parameters of the PSMC model include the scaled mutation rate, the recombination rate, and piecewise constant ancestral population sizes. Indeed, the accuracy and variance of the results depend on the number of recombination events, and very small number of recombination events would increase the variance and reduce the power of the model. Thus too ancient or too recent recombination events would not make a precise inference, because few of these events could be detected in genome. In the original paper of PSMC which inferred a detailed population history from human reference genome, the time span was from 20 Kya to 2 Mya for human demographic history (Li and Durbin, 2011). In the case of bears and giant panda, the time range was 5 Kya to 1 Mya (Miller et al., 2012), and 10 Kya to 9 Mya (Zhao et al., 2013), respectively. For the same species from different populations, the pattern of PSMC fluctuated in a similar way, but skewed a little due to population substructure (Li and Durbin, 2011). Fig. 2 showed the results of population history inference in giant panda and human.

GENOME-WIDE SCANNING FOR LOCAL ADAPTATION

From "Survival of the fittest" to "Neutral or near neutral mutations"

The concepts of selection, adaptation, and evolution were first described in the famous Origin of Species by Darwin (1859). Natural selection changes the fitness by accumulating tiny variations from generation to generation to adapt to environments. Phenotypic adaptation is the result of mutation and selection during evolution. A few decades later, Ronald Aylmer Fisher (1930) proposed the idea of "Fisher's fundamental theorem", where mathematical approaches were applied to define the fitness rate as a function of its genetic variation. Natural selection could then be measured not only by fitness changes but also by genetic variations.

Later, principles of population genetics were extended to sequence data with the development of molecular biology. The neutral theory considered that genetic variations were accumulation of neutral mutation, and were removed by genetic drift (Kimura, 1968). Successive theories argued that with a broader definition of neutrality, majority of genetic

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

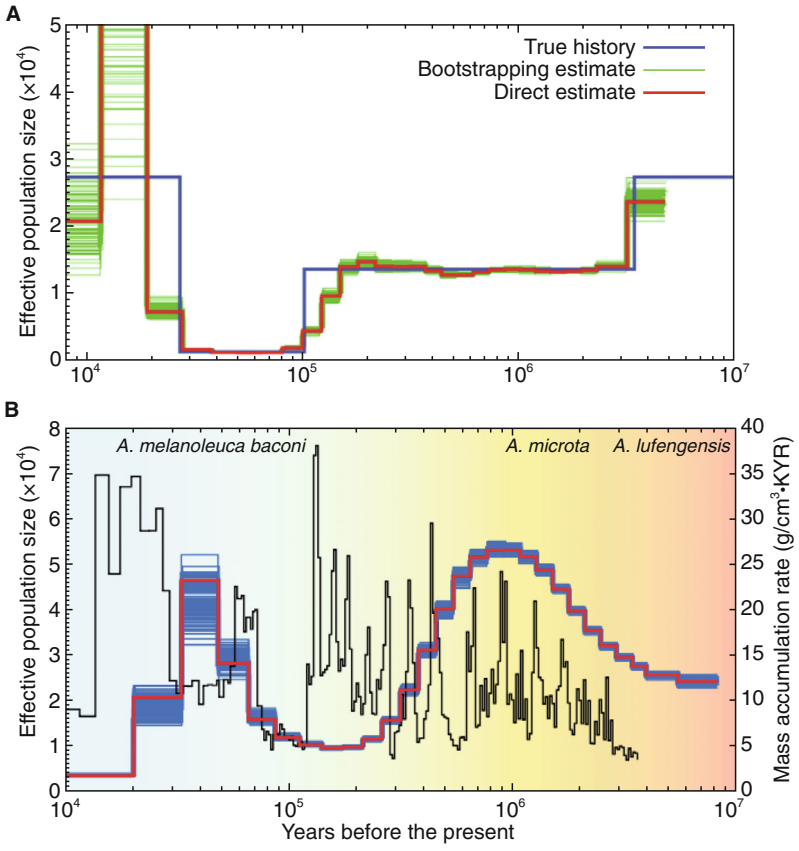


Figure 2. Human and giant panda demographic history inferred by PSMC (See details in Li and Durbin (2011) and Zhao et al. (2013)). (A) The history recovered using human reference genome, which shows a bottle neck during 3–10 Kya following an explosive growth. The red curve is the PSMC estimate on the originally simulated sequence; the 100 thin green curves are the PSMC estimates on 100 sequences randomly resampled from the original sequence. The blue curve is the population-size history of human used in simulation. (B) The history recovered using giant panda genome, in which two peaks of population growth and two bottle necks are detected. The red line represents the estimated effective population size (Ne); the 100 thin blue curves represent the PSMC estimates for 100 sequences randomly resampled from the original sequence; the brown line shows the MAR of Chinese loess which represents the climate changes during the history.

variations could be attributed to random drift during evolution instead of adaptation to local environment (Ohta, 1992; Nei, 2005). Therefore it becomes a challenge to detect the signatures under selection within genome via statistical approaches, which has been one of the central tasks of evolutionary genetics for the recent two decades (Kreitman and Akashi, 1995).

Genome is shaped by two evolutionary forces: neutrality during demographic history and natural selection. Generally, genetic drift, population growth, migration, and other demographic events affect the whole genome; but natural selection by local environment changes make imprint on episodes or a part of structures of the genome, which could have effects on phenotypes and fitness. Natural selection also changes the frequency of mutations across populations. The advantageous mutations approach to genetic fixation under directional selection; the advantageous heterozygotes are maintained by balancing selection; and purifying selection removes the deleterious mutants. The genetic signatures in genome sequence open a door to detect natural selection. Application of mathematical methods and statistical tests throw light on interpreting the imprints of evolution and adaptation.

Seeking genome-wide signatures of adaptation

Statistical methods are developed to distinguish causal genetic variations subject to selection from neutral genetic variations (Hartl and Clark, 2007). Whole genome sequencing provides large-scale genetic variation data for this purpose. Two strategies are applied. The first one is based on genome-wide selection scans (GWSS) to detect outliers or structure violations as the signatures of selection. The second is based on genome-wide association approaches, in which a prior

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

phenotype or environmental information is required to obtain associated genetic variation loci as the potential candidates for selection. Ultimately, the function of related genes should be investigated to associate the candidate genetic variation with the phenotype variation. For the first strategy, it is a challenge to distinguish selective effect from that caused by neutral events in past. For the second, there are less fine-scaled phenotypic records in non-model species than those in model species or domestic creatures. No matter what strategy is applied, an experimental validation would be highly persuasive for candidate variations, but in most cases it would be very difficult.

Comparative genomic data (between/among species)

The simple ratio of non-synonymous (*dN*) to synonymous (*dS*) substitution of coding regions is often used to identify adaptive loci deviation from neutral state between species. Under neutral theory, $\omega = dN/dS$ is expected to reflect selection types of species: $\omega = 1$, >1, and <1, indicating the sites tested under neutral state, positive selection, and purifying selection, respectively. Comparison of average ω indicator across gene or DNA segments will present the most conservative results because the effects of real sites under selection might be weakened by neutral sites unless the whole region was under selection. Yang (1997, 1998) set varying ω values across lineages or among protein sites to estimate whether a specific lineage was subjected to Darwinian natural selection among protein sites or evolutionary lineages. A likelihood test was applied to determine the null and the alternative hypotheses (Yang, 1997, 1998). This approach was incorporated in PAML software, which included branch model, site model, and branch-site model, allowing ω varying across branches, or sites among proteins, and both across branch and among proteins, respectively (Yang, 1997, 1998; Yang and Nielsen, 2002; Zhang et al., 2005; Yang, 2007; Yang and Nielsen, 2008; Yang and dos Reis, 2011).

These methods have been widely applied in comparative genomic analyses. In mammals, several studies using extensive genomes in mammals have been performed to detect the loci or lineages under selections (Clark et al., 2003; Kosiol et al., 2008; Nielsen et al., 2005; Li et al., 2010). For example, the results of Li et al. (2010) showed that the positively selected genes were significantly enriched in the functional categories of blood circulation and gas exchange activity in mammals. Zhang et al. (2013) compared two related bat genomes and identified genes responsible for DNA damage checkpoint and NF-κB pathways subjected to strong selection, implying possible adaptation to flight. In addition, comparative analysis of two falcons, peregrine and saker falcon, showed an accelerated evolution rate on homeostasis-related genes responsible for circulation (Zhan et al., 2013).

Comparative population genomic data (within and between species)

Under the neutral theory, polymorphic and divergence data are expected to be the accumulation of neutral mutations within and between species respectively. Hudson et al. (1987) established a statistical test to examine whether the DNA sequence with higher/lower evolutionary rate between species also presented higher/lower polymorphic rate within species. Although Wright and Charlesworth (2004) extended HKA tests by incorporating maximum likelihood tests, the HKA has been rarely used in genome-wide analysis for its constant effective population size assumption (Nei and Kumar, 2000). McDonald and Kreitman (1991) applied the idea underlying HKA tests and proposed that the ratio of dN:dS between species is equal to the ratio of dN:dS within species when sequences are selective neutrality. They found an excess of the ratio in divergence than that in polymorphism across the ADH sequences among three species of fruit flies, suggesting that beneficial alleles were maintained by positive selection. Bustamante et al. (2005) also applied this approach to compare the divergence and polymorphism data of human and chimpanzee to identify signatures under positive and purifying selection across human genome. The positively selected genes are predicted to be enriched in defense/immunity, transcription, sensory perception and so on, while negative selection are expected to affect processes of cell structure and motility, ectoderm development, general vesicle transport, and intracellular protein traffic.

Population genomic data within species

Evolutionary forces will skew genetic diversity and frequency of loci across genomes. Various methods based on estimators of genetic diversity within species have been developed to detect selection signatures. Here we mention tests based on heterozygosity, tests based on F_{ST}, tests based on Allele frequency, and tests based on haplotype. We also discuss composite approaches combining multiple tests.

Local adaptation signals have two manifestations. Within one population, they are expected to be with low heterozygosity. Between populations they show higher genetic differentiation measured by F_{ST}. These two indicators, heterozygosity and F_{ST}, and their derived formations are widely-used in population genomic analysis.

Rubin et al. (2010) applied Z-transformation to pooled heterozygosity $Z(H_p)$ across sliding windows of the genomes in domesticated chicken, and identified that the loci for thyroid stimulating hormone receptor (*TSHR*) were under selective sweeps. They further tested this gene across 271 birds from 36 geographic populations and proved that the mutant allele is a domestication locus in chicken. The formula they used is as follows.

$$H_p = \frac{2 \times \sum n_{MAX} \times \sum n_{MIN}}{(\sum n_{MAX} + \sum n_{MIN})^2} \quad (4)$$

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

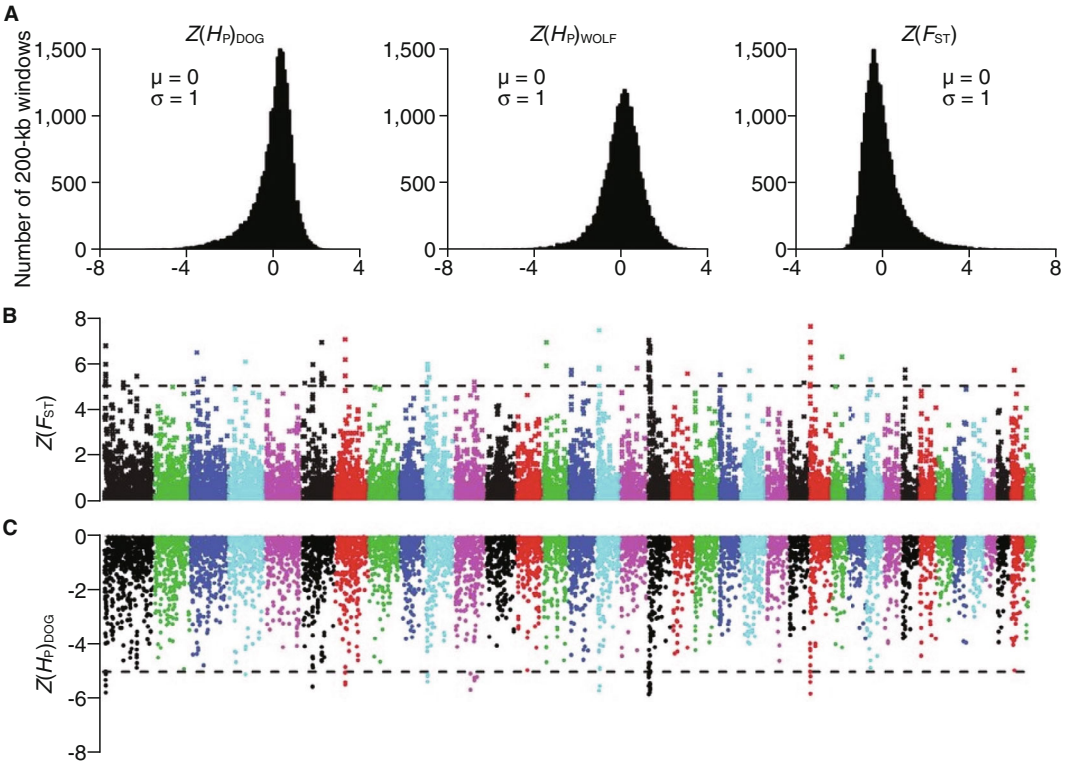


Figure 3. Selection analyses identified 36 candidate domestication regions (cited from Axelsson et al., 2013). Neighbored different colors in (B) and (C) distinguish different chromosomes. (A) Distribution of Z(H_p) of dog and wolf, Z(F_{ST}) between dog and wolf. (B) Distribution of Z(F_{ST}) across genomes. The loci with Z(F_{ST}) over the dashed line cut-off are candidates for outliers. (C) Distribution of Z(H_p) of dogs across genomes. The loci with Z(H_p) value lower than the dashed line are candidate for outliers. The loci with higher Z(F_{ST}) and lower Z(H_p) values in dogs will be considered as the signatures across genomes under domestication of dogs.

and

$$Z(H_p) = \frac{H_P - \mu H_P}{\sigma H_P} \quad (5)$$

Genome-wide scanning of adaptation signatures during dog domestication applied the same idea to transform heterozygosity and genetic differentiation index F_{ST} into $Z(H_p)$ and $Z(F_{ST})$ between dogs and wolves (Fig. 3), identifying a series of genes in dogs with low $Z(H_p)$ and high $Z(F_{ST})$, which attributed to adaptation to starch-rich food during domestication (Axelsson et al., 2013).

Genetic differentiation between populations is usually measured with genetic fixation index F_{ST}. Generally, positive selection gives rise to less heterozygosity within populations and higher genetic differentiation of loci between populations. F_{ST} statistic has been developed to estimate selection for decades (Wright, 1943; Weir and Cockerham, 1984; Slatkin and Voelm, 1991; Cockerham and Weir, 1993). Pairwise F_{ST} between populations is compared to detect the differentiated signals under positive selection directly (Akey et al., 2002; Barreiro et al., 2008; Lam et al., 2010; Zhao et al., 2013). The integrated approach based on different estimation of F_{ST} has been successfully applied into detection of local adaptation signals in giant pandas population genomics (Zhao et al., 2013).

Yi et al. (2010a) log-transformed F_{ST} into T value:

$$T = -\log(1 - F_{ST}) \quad (6)$$

They compared the braches of T values of Tibetan, Han, and Denes populations to identify signals of Tibetan adapting to high altitudes:

$$PBS = \frac{T^{TH} + T^{TD} - T^{HD}}{2} \quad (7)$$

Branch length indicated the genetic differentiation levels. The outliers with larger branches than the average genomic branch in Tibetan were considered to be candidates. To further validate the power of the transformed methods, a larger sample of Tibetan was genotyped, and the frequency of advantageous allele of the outlier loci EPAS1 was examined. The results showed that the SNP in EPAS1 was associated with erythrocyte count and hemoglobin quantity (Yi et al., 2010a).

The proportion of segregating sites in all sites of aligned sequences $\hat{\theta}_S$ and the average proportion of pairwise

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

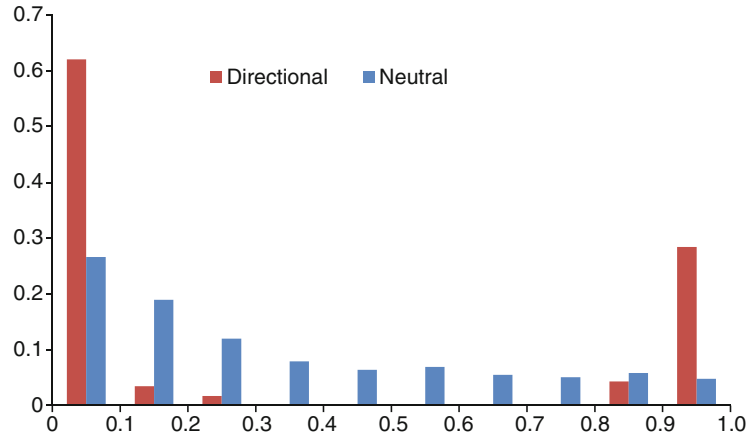


Figure 4. Derived allele frequency spectrum of SNPs under positive selected and neutral selections (modified from Zhao et al,. 2013). Positive selection maintained an excess of lower and higher frequency of alleles within populations.

mismatches over all compared sequences $\hat{\theta}_\pi$ could be used to measure nucleotide variation in populations. Tajima (1989) incorporated the two estimators and proposed a statistic test as

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{SE(\hat{\theta}_\pi - \hat{\theta}_S)} \quad (8)$$

in which *SE* means standard error. Under neutral state, the excess of low-frequency alleles will increase $\hat{\theta}_\pi$ and that of intermediate-frequency alleles would affect $\hat{\theta}_S$. A minus *D* value might indicate population under purifying selection or population growth; otherwise, a positive *D* suggests balancing selection, or diversifying selection, and recent admixture of different populations (Hartl and Clark, 2007). Fu and Li (1993) incorporated mutations of external and internal branch to estimate the nucleotide variation, and developed a G statistic test to check the selection signals caused by increased singleton polymorphism. Fay and Wu (2000) proposed an H statistic test to focus on signals from the higher-frequency and intermediate-frequency data. A significant minus H means neutral alleles approached to be fixed for genetic hitchhiking. Among the three statistic methods, Tajima's D is the most widely used, while the latter two tests incorporated out-group information such as related species to provide ancestral state.

Strong positive selection maintains excess high frequency of beneficial alleles in populations, as shown in Fig. 4. The new advantageous allele relative to its ancestral allele exhibits higher frequency in populations under positive selection. The allele from the related species was taken as the ancestral allele, and the derived alleles were considered the new ones after their divergence. Derived allele of each locus across genomes was compared by simple frequency or relative ratio of frequency among populations to localize the selected new advantageous alleles and its adapted populations, e.g. ΔDAF and DAF tests (Sabeti et al., 2006, 2007; Grossman et al., 2010). The DAF-based tests are expected to identify recent selection events since species divergence. In the example of human and chimpanzee, the derived allele of human SLC24A5 gene shows higher frequency across Europe populations but is absent in most of Asian populations, which is suggested to be responsible to skin pigment difference (Sabeti et al., 2007).

Haplotype provides integrative information of a set of neighboring SNPs rather than a sum of individual SNPs. Diversity and frequency of haplotype in populations will be distorted under very recent positive selection. Since the phased high-resolution human HapMap was available, statistic methods based on variance in heterozygosity, length and frequency of haplotype were developed to detect very recent adaptation within and between populations (Sabeti et al., 2002, 2006; Voight et al., 2006). Extended haplotype homozygosity (EHH test) was applied to identify long-range haplotype with reduced heterozygosity as signatures under recent selective sweeps (Sabeti et al., 2002). The ancestral and derived haplotypes were compared using integrated Haplotype Score (iHS test) to measure the selection strength of each locus (Sabeti et al., 2006; Voight et al., 2006). Differential frequency of haplotype between populations was measured and compared in Cross Population EHH (XP-EHH test) to localize population adaptive to local environment (Sabeti et al., 2007; Grossman et al., 2010). The haplotype-based methods were also widely used in domesticated species to detect signals of local adaptation to new environments under artificial selection. Vonholdt et al. (2010) used F_{ST} and XP-EHH methods to identify several SNPs involving memory formation and behavioral sensitization during dog's domestication. Toomajian et al. (2006) proposed a haplotype-sharing statistical analysis and used another haplotype-based method EHH as well to identify the early-flowering alleles in *Arabidopsis thaliana.*

To get a reliable result, it is persuasive to combine multiple tests based on different assumptions for the dataset, and only

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

the outliers supported by multiple tests are considered as reliable candidates (Simonson et al., 2010; Zhao et al., 2013). An alternative approach is to seek the genomic regions with multiple contiguous outlier loci as the candidates under selective sweep (Li et al., 2010). However, Grossman et al. (2010) proposed a composite likelihood test (CMS) incorporating as many as six independent statistic methods (F_{ST}, ΔDAF, DAF, XP-EHH, iHS, IHH) to identify truly selected variants. A SNP locus was estimated to be with a probability of selected / unselected of each independent test (*Si*). Bayesian factor and CMS were calculated as follows:

$$BF = \prod_{i=1}^{n} \frac{P(S_i \mid selected)}{P(S_i \mid neutral)} \tag{9}$$

$$CMS = \prod_{i=1}^{n} \frac{P(S_i \mid selected)}{P(S_i \mid selected) \times \pi + P(S_i \mid neutral) \times (1-\pi)} \tag{10}$$

in which the Greek π was the prior probability of selection, being estimated from prior information of the data. The distribution of posterior probability of CMS scores was used to estimate the confidence intervals. A modified CMS test was applied to detect region within genomes for recent adaptation in human populations (Grossman et al., 2013). The authors applied statistical methods to prove that the CMS tests have more power in detection of adaptive signals than independent test (Grossman et al., 2010, 2013).

Functional analysis of candidates

Generally, three approaches are used to investigate biological function of outlier candidates. One is the prediction of potential gene function. For protein within the candidate selective locus, the spatial structure of the protein will be simulated to infer its potential conformation change influenced by the variant locus (Sabeti et al., 2007; Grossman et al., 2013). For all the outlier data, candidate loci will be annotated by gene ontology database and the functional categories enriched with candidate loci would be considered as potentially selected. An alternative precise functional analysis associates studies of outlier variations with phenotype variation or environmental change. Some researchers have made good examples on genome-wide selection scanning as well as genome-wide association analysis (Yi et al., 2010; Grossman et al., 2013).

However, the prediction analysis and statistical association analysis could only tell us about the general information about potential biological function, it's important to examine whether the genetic variations with adaptive selection signals successfully express the phenotype with high fitness into new environment or not. In dog, expression and functional experiments were performed after the GWWS analysis to test adaptive signatures under domestication, which proved that domesticated dogs have more powers in enzymatic activity for digestion of starch-rich diets than their ancestor wolves (Axelsson et al., 2013).However, few cases could examine the functions of polymorphic loci in inter-gene under selections because of no protein expression.

PERSPECTIVES

A diffusion approximation model with recombination

The model used in ∂a∂i could give very complicated demographic model with the least constraints on dataset. The Kolmogorov forward equation, based on Fokker-Planck equation, was very solid in mathematics as well as in population genetics (Kimura, 1955; Haken, 1983). One possible improvement of the model might include the consideration of recombination. Here we suggest a concept of "loci group" to discuss a possible method considering recombination in the diffusion process of polymorphism distribution among populations. In ∂a∂i, all derived mutations from one genome are put into one pool or group to get a density distribution, which could be used to infer demographic history. Thus a prerequisite of unlinked loci within the whole group is needed and the recombination events between any two mutations could not be considered. Here we suggest an improvement that considers two such groups, group A and group B. Both of them satisfy the condition of performing demographic inference following the approach of ∂a∂i. Besides, each SNP in one group has one and only one close-linked locus in the other group. We could say the two groups are to some extent "linked". For group A, the evolution of the density could be modeled following ∂a∂i's approach as

$$\frac{\partial}{\partial \tau}\Phi_A = \frac{1}{2}\sum_{i=1}^{P}\frac{\partial^2}{\partial x_A^2}\frac{x_{Ai}(1-x_{Ai})}{v_{Ai}}\Phi_A - \sum_{i=1}^{P}\frac{\partial}{\partial x_{Ai}}\left[\gamma_{Ai}x_{Ai}(1-x_{Ai}) + \sum_{j=1}^{P} M_{A:i\leftarrow j}(x_{Aj}-x_{Ai})\right]\Phi_A \tag{11}$$

Similarly for group B, one has

$$\frac{\partial}{\partial \tau}\Phi_B = \frac{1}{2}\sum_{i=1}^{P}\frac{\partial^2}{\partial x_B^2}\frac{x_{Bi}(1-x_{Bi})}{v_{Bi}}\Phi_B - \sum_{i=1}^{P}\frac{\partial}{\partial x_{Bi}}\left[\gamma_{Bi}x_{Bi}(1-x_{Bi}) + \sum_{j=1}^{P} M_{B:i\leftarrow j}(x_{Bj}-x_{Bi})\right]\Phi_B \tag{12}$$

Since the group A and B are from the same genomes and experienced the same history for the same populations. Then we have

$$\frac{\partial}{\partial \tau}\Phi_A = \frac{\partial}{\partial \tau}\Phi_B \tag{13}$$

The Φ_A and Φ_B should be influenced by the same migration affairs in history. Suppose there is no selection ($\gamma_i = 0$), we should have

$$\sum_{j=1}^{P} M_{B:i\leftarrow j}(x_{Bj}-x_{Bi}) = \sum_{j=1}^{P} M_{A:i\leftarrow j}(x_{Aj}-x_{Ai}) \tag{14}$$

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

Equations (13) and (14) determine that the drift terms in (11) and (12) should also be equal

$$\frac{x_{Ai}(1-x_{Ai})}{v_{Ai}} = \frac{x_{Bi}(1-x_{Bi})}{v_{Bi}} \quad (15)$$

That means, if there is a certain proportion of chromosomes transformed from population *i* to *j* per generation detected in group A, there could be expected the same proportion in group B. However, if recombination exists, the linkage between A and B might be broken in a probability $r_{i\leftarrow j}$, the recombination rate for one migration event from *j* population to *i* population. In this case, although one allele from group A was migrated, its linked allele in group B was left in the original population in the probability of $r_{i\leftarrow j}$. Therefore one has

$$\sum_{j=1}^{P} M_{B:i\leftarrow j}(x_{Bj}-x_{Bi}) = \sum_{j=1}^{P} M_{A:i\leftarrow j}(1-r_{i\leftarrow j})(x_{Bj}-x_{Bi}) \quad (16)$$

Then the recombination rate could be included in the extensive diffusion process of the density distribution Φ as

$$\begin{aligned}\frac{\partial}{\partial \tau}\Phi = {} & \frac{1}{2}\sum_{i=1}^{P}\frac{\partial^2}{\partial x_A^2}\frac{x_{Ai}(1-x_{Ai})}{v_{Ai}}\Phi + \frac{1}{2}\sum_{i=1}^{P}\frac{\partial^2}{\partial x_B^2}\frac{x_{Ai}(1-x_{Ai})}{v_{Ai}}\Phi \\ & + \sum_{i=1}^{P}\frac{\partial^2}{\partial x_A \partial x_B}\left[\sum_{j=1}^{P} M_{A:i\leftarrow j}(x_{Aj}-x_{Ai})\right] \\ & \times \left[\sum_{j=1}^{P} M_{A:i\leftarrow j}(1-r_{i\leftarrow j})(x_{Bj}-x_{Bi})\right]\Phi \\ & - \sum_{i=1}^{P}\frac{\partial}{\partial x_{Ai}}\sum_{j=1}^{P} M_{A:i\leftarrow j}(x_{Aj}-x_{Ai})\Phi \\ & - \sum_{i=1}^{P}\frac{\partial}{\partial x_{Bi}}\sum_{j=1}^{P} M_{A:i\leftarrow j}(1-r_{i\leftarrow j})(x_{Bj}-x_{Bi})\Phi \end{aligned} \quad (17)$$

For total *P* populations there will be P × (1 – P) number of *r* values, their mean value could represent the recombination rate of the P populations within one generation. In the whole process, the most urgent task is to determine the SNP pairs. One pair of SNPs should be tightly linked so that there is only one recombination event could happen between the SNPs, as the same time every two pairs of SNPs should be distant enough so that the linkage could be ignored within groups. However, there is an arbitrary assumption which may be violated in real data set in the model. We suppose that to make sure equation (13) is equal, every term in the diffusion function (11) and (12) are equal. Therefore we obtain equation (14) and (15), which mean migration and drift are equal between groups. However, there does be possibility that drift and migration have similar effects, therefore equation (14) and (15) are not equal but equation (13) is still satisfied. If so, a deviation may appear in the estimation of recombination rate. Future studies are needed to access the deviation by either simulation or real dataset.

Challenges for identifying genomic signature of selection

In genome era, selection tests for recent positive selection were well developed as shown above. However, the test for the balancing and negative selection within species is rare. Although some approaches and softwares made efforts to test balancing selection (Excoffier et al., 2009; Excoffier and Lischer, 2010) and negative selection (Tajima, 1989; Yang, 1998), it could not meet the needs for GWWS, e.g. few balancing and purifying selection tests based on haplotypes. Signatures for balancing and negative selection should be interpreted and mined from large-scale genome data, which calls for diversifying methods in the future.

Genomic signatures for selection detected by GWWS could be distorted by demographic history and population structure. Tajima's D could not distinguish the signatures caused by selection force from those caused by demographic fluctuation (Yang, 2006, Hartl and Clark, 2007). Other estimators are also sensitive to population growth and contraction (Hartl and Clark, 2007), and invisible population structure and admixture will influence the heterozygosity and haplotype diversity (Hartl and Clark, 2007). Hence it is important to infer population genetic structure and demographic background to distinguish the effect of selection from neutrality. Williamson et al. (2005) set a good example on this idea. He used presumed neutral loci to construct a population history. And then he used the simulation results as the null hypothesis to infer selection signals. The more genomic data available, the more intensive and integrative methods are needed in the future to incorporate both selection forces and demographic history influences.

Last but not least, it is still a barricade to test the selective effect on certain loci using experimental approaches, either for cellular and molecular approach or high-throughput technique. The difficulties come from two aspects. Firstly, it is still a hard task to avoid false negative results. The theoretical assumption may be violated in the populations of certain species, genetic structure or demographic history may not be totally excluded. Selection signals may be flooded by noise. All these may introduce false negative results. Secondly, present methods could give evidence of whether one locus was under selection, but could not give clues that in which levels the locus shows the selection: at molecular level, or cellular level, or tissue/organ level, or individual level. It is impossible to test every possible level for each loci, let alone the permutation and combination of all loci. In general, although it is easy to consider some possibilities that the loci in non-synonymous sites would cause variations in protein sequence and possibly cause the change of protein structure, to verify the biological effects of the identified selective loci will lag behind the progress of identifying selective loci in the next several years.

ACKNOWLEDGEMENTS

We are grateful to the Supercomputing Center of Chinese Academy of Science (CAS) for the supercomputing resource for demographic history simulation. Studies in our laboratory were supported by grants from the National Natural Science Foundation of China (Grant

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

No. 31230011) and Knowledge Innovation Program of Chinese Academy of Sciences (KSCX2-EW-Z-4).

ABBREVIATIONS

CMS, composite likelihood test; EHH, Extended haplotype homozygosity; GWSS, genome-wide selection scans; iHS, integrated Haplotype Score; MCMC, Markov Chain Mento Carlo method; ML, maximum likelihood; MRCA, Most Recent Common Ancestors; PSMC, Pairwise Sequentially Markovian Coalescent; SFS, Site Frequency Spectrum; SNP, single nucleotide polymorphism; TSHR, thyroid stimulating hormone receptor.

COMPLIANCE WITH ETHICS GUIDELINES

The authors declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

OPEN ACCESS

This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

REFERENCES

Adams AM, Hudson RR (2004) Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics 168:1699–1712

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814

Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495:360–364

Batini C, Lopes J, Behar DM, Calafell F, Jorde LB, VanderVeen L, Quintana-Murci L, Spedini G, Destro-Bisol G, Comas D (2011) Insights into the demographic history of African Pygmies from complete mitochondrial genomes. Mol Biol Evol 28:1099–1110

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. Annu Rev Ecol Evol Syst 41:379–406

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. Genome Res 17:1505–1519

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B et al (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963

Cockerham CC, Weir BS (1993) Estimation of gene flow from F-statistics. Evolution 47:855–863

Crisci JL, Poh Y, Bean A, Simkin A, Jensen JD (2012) Recent progress in polymorphism-based population genetic inference. J Hered 103:287–296

Darwin C (1859) On the origin of species. John Murray, London

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity (Edinb) 103:285–298

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA et al (2013) The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res 40(Database issue):D26–D32

Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O et al (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883–886

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH et al (2013) Identifying recent adaptations in large-scale genomic data. Cell 152:703–713

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695

Haken H (1983) Synergetics. Springer, Berlin

Hartl DL, Clark AG (2007) Principles of population genetics, 4th edn. Sinauer Associates Inc, Sunderland

Hernandez RD, Williamson SH, Bustamante CD (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24:1792–1800

Hey J (2010) Isolation with migration models for more than two populations. Mol Biol Evol 27:905–920

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167:747–760

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159

Huo T, Zhang Y, Lin J (2012) Functional annotation from the genome sequence of the giant panda. Protein Cell 3:602–608

Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci USA 101:13994–14001

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

Kimura M (1955) Solution of a process of random genetic drift with a continuous model. Proc Natl Acad Sci USA 41:144–150

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six Mammalian genomes. PLoS Genet 4:e1000144

Kreitman M, Akashi H (1995) Molecular evidence for natural selection. Annu Rev Ecol Syst 26:403–422

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1059

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493–496

Li MH, Iso-Touru T, Lauren H, Kantanen J (2010) A microsatellite-based analysis for the detection of selection on BTA1 and BTA20 in northern Eurasian cattle (*Bostaurus*) populations. Genet Sel Evol 42:32

Lopes JS, Balding D, Beaumont MA (2009) PopABC: a program to infer historical demographic parameters. Bioinformatics 25:2747–2749

Mannion AM (1999) Domestication and the origins of agriculture: an appraisal. Prog Phys Geogr 23:37–56

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654

McVean GA, Cardin NJ (2005) Approximating the coalescent with recombination. Philos Trans R Soc Lond B Biol Sci 360:1387–1393

Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, Kim HL, Burhans RC, Drautz DI, Wittekindt NE et al (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proc Natl Acad Sci USA 109:E2382–E2390

Nachman MW, Payseur BA (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Trans R Soc Lond B Biol Sci 367:409–421

NCBI Resource Coordinators (2013) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 41(Database issue):D8–D20

Nei M (2005) Selectionism and neutralism in molecular evolution. Mol Biol Evol 22:2318–2342

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, Oxford

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158:885–896

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ et al (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3:e170

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8:857–868

Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–286

Palero F, Lopes J, Abello P, Macpherson E, Pascual M, Beaumont MA (2009) Rapid radiation in spiny lobsters, *Palinurus* spp) as revealed by classic and ABC methods using mtDNA and microsatellite data. BMC Evol Biol 9:263

Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S et al (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464:587–591

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ et al (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive natural selection in the human lineage. Science 312:1614–1620

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R et al (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo FR, Xing J, Jorde LB et al (2010) Genetic evidence for high-altitude adaptation in Tibet. Science 329:72–75

Slatkin M, Voelm L (1991) FST in a hierarchical island model. Genetics 127:627–629

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Thornton K, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics 172:1607–1619

Toomajian C, Hu TT, Aranzana MJ, Lister C, Tang C, Zheng H, Zhao K, Calabrese P, Dean C, Nordborg M (2006) A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome. PLoS Biol 4:e137

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72

Vonholdt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, Degenhardt JD, Boyko AR, Earl DA, Auton A et al (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464:898–902

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci USA 102:7882–7887

Wooding S, Rogers A (2002) The matrix coalescent and an application to human single-nucleotide polymorphisms. Genetics 161:1641–1650

Wright S (1943) Isolation by distance. Genetics 28:114–138

Wright SI, Charlesworth B (2004) The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics 168:1071–1076

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z (2006) Computational molecular evolution. Oxford University Press, Oxford

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang Z, dos Reis M (2011) Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28:1217–1228

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908–917

Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25:568–579

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS et al (2010a) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75–78

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS et al (2010b) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75–78

Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, Ni P, Hu L, Liu Y, Hou H et al (2013) Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. Nat Genet 45:563–566

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–2479

Zhang G, Cowled C, Shi Z, Huang Z, Bishop-Lilly KA, Fang X, Wynne JW, Xiong Z, Baker ML, Zhao W et al (2013) Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science 339:456–460

Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, Hu Y, He W, Zhang S, Fan W et al (2013) Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nat Genet 45:67–71

© The Author(s) 2014. This article is published with open access at Springerlink.com and journal.hep.com.cn