Chinese Science Bulletin

REVIEW

Positive Darwinian selection in human population: A review

WU DongDong[1,3] & ZHANG YaPing[1,2†]

[1] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China;
[2] Laboratory for Conservation and Utilization of Bio-resource, Yunnan University, Kunming 650091, China;
[3] Graduate University of Chinese Academy of Sciences, Beijing 100049, China;

This paper reviews a large number of genes under positive Darwinian selection in modern human populations, such as brain development genes, immunity genes, reproductive related genes, perception receptors. The research on the evolutionary property of these genes will provide important insight into human evolution and disease mechanisms. With the increase of population genetics and comparative genomics data, more and more evidences indicate that positive Darwinian selection plays an indispensable role in the origin and evolution of human beings. This paper will also summarize the methods to detect positive selection, analyze the interference factors faced and make suggestions for further research on positive selection.

natural selection, positive selection, population genetics, comparative genomics

GENETICS

As one of the most important scientific discoveries of the 19th centenary, Darwinian natural selection theory articulated by Darwin and Wallace[1] in 1858 contains mainly excessive multiplication, struggle for existence, heritable variation, survival of the fittest, etc. Then, although the theory was extensively accepted, the role of natural selection in the evolutionary process was continuously debated. In 1925, Morgan introduced mutationism and argued that the most important factor of evolution was the occurrence of advantageous mutations, and natural selection was just to save advantageous mutations and to select out unfavourable ones (reviewed in ref. [2]). However, an overwhelming majority of mutations found in the subsequent experiments were disadvantageous, and the theory was gradually replaced by neo-Darwinism, advocated by Fisher, Wright and Dobzhansky et al. They thought of natural selection as a key factor in producing and maintaining genetic polymorphism; therefore, neo-Darwinism is also known as selectionism (reviewed in ref. [2]). However, in the 1960s, protein sequencing and electrophoretic techniques have been used widely in evolutionary analyses, e.g. large scale genetic variation was observed in human and *Drosophila melanogaster*; similar amino acid substitution rates in different vertebrate lineages by comparing haemoglobin and cytochrome sequences, which neo-Darwinism cannot explain[3]. In allusion to the problem, Kimura[4], King and Jukes[5] advocated the well-known neutral theory in 1968 and 1969, which challenged classical Darwinism. The theory states that most mutations are disadvantageous, and mutations are randomly fixed or deleted, dependent on gene drift. Additionally, the mutation rate is dependent on population size[4,5]. Over the following 30 years, the role of natural selection in molecular evolution has been extremely controversial and the key debate between neo-Darwinism and neutralism is about the role of natural selection and genetic drift on evolution. Detection of positive selection depends primarily on the differences in non-

Received September 6, 2007; accepted March 2, 2008
doi: 10.1007/s11434-008-0202-z
†Corresponding author (email: zhangyp1@263.net.cn or zhangyp@mail.kiz.ac.cn)
Supported by the National Natural Science Foundation of China (Grant Nos. 30621092 and 30430110), and Bureau of Science and Technology of Yunnan Province

synonymous and synonymous substitution rates, based on which more and more natural selection events are revealed in proteins[6]. For example, through analysis Smith and Eyre-Walker[7] found out that about 45% amino acid substitutions result from positive natural selection, and one favorable mutation will be fixed every 45 years. Another investigation indicated that the ratio of non-synonymous substitution rates to synonymous between *D. melanogaster* and *Drosophila simulans* was 3 times that of the former polymorphism[8]. The importance of natural selection on evolution is gaining growing acceptability. Accordingly, Ohta[9−11] amended neutral theory and put forward "near-neutrality" theory, in which, "mutation-drift-selection" plays important roles in molecular evolution but that of natural selection is elevated. In the present day, with the increase of population genetics and comparative genomics data, abundant genomic detections of positive selection suggest that natural selection events occur more frequently than previously in evolutionary population history[12−17]. It is still difficult to evaluate the extent to which natural selection contributes to evolution.

Recently, research on positive selection has been developed rapidly, which promotes the investigation on human origin and evolution, and provides pivotal information on the mechanism of various diseases. Here we review recent research progress on positive selection in human populations, common problems and corresponding solutions.

1 Genes under positive selection during human evolution

Darwinian natural selection has maintained high phenotypic diversity in modern human populations, including differences in appearance, resistance to disease and drug metabolism, in order to adapt to distinct environments. It is exact positive selection that accomplishes more advantaged *Homo sapiens* than other primates in many fields, such as the achievement of cognitive capacity. Many related genes, immunity and reproduction genes have been found to be targets of Darwinian positive selection.

1.1 Genes associated with brain development

The genomes of human beings and our close relative the chimpanzee are very similar, but the phenotypes differ greatly. Cognition is earmarked for distinguishing human and other higher primates, and the brain is the physiological basis. Brain size has developed rapidly in primates, particularly lineages related to humans; and some genes associated with brain development have evolved rapidly under positive selection[18−24], an important area of research on positive selection.

Microcephalin is one of the key genes regulating human brain development. Mutations of the gene cause primary microcephaly. The brain volume of patients of microcephaly is only about 1/4—1/3 of normal individuals, and equivalent to that of early anthropoid. ASPM (abnormal spindle-like microcephaly-associated), another primary microcephaly associated gene located at 1q31 spanning 65 kb, has 28 exons encoding 10434 bp coding sequence. The protein contains an N-terminal microtubule-binding domain, a calponin-homology domain, an IQ repeat domain, and a C-terminal region[18−24]. Analyses based on polymorphism data on human populations from all over the world indicated that these crucial brain development genes were still evolving under positive selection, and the effect is ongoing[18,24]. Speech and language are also unique characters of *H. sapiens*. *FOXP2* is a recently identified gene associated with human language expression and singing ability in birds. Analyses on comparative data of primates suggested that the evolutionary rate of *FOXP2* gene was accelerated in the human lineage by positive selection, and similarly to the aforementioned brain development genes, *FOXP2* was also affected by recent positive selection in human population[25−27]. It can be concluded that human intelligence is one product of adaptive evolution. PACAP (Pituitary adenylate cyclase-activating polypeptide) is a neuropeptide that plays an important role in the nervous system by mediating regulation and development of the nervous system in the cerebral cortex. Its chromosomal region identified was associated with holoprosencephaly in human populations. The importance of this function leads to a highly conserved peptide in mammals, but phylogenetic analysis indicated UD and PRP regions of precursor evolved under highly positive selective pressure in the human lineage[28]. Investigations of these genes show they are responsible for distinct human features. For instance, Clark et al.[27] separated positive selection genes unique to humans and chimpanzees from 7645 orthologies using maximum likelihood method. This will undoubtedly provide pivotal clues on the difference between the two species and human evolution.

1.2 Genes in the immune system

The major histocompatibility complex (MHC) locus is a classic example of positive selection. MHC acts mainly for antigen presentation in the immune system; and the ratio of non-synonymous substitution rates to synonymous (d*N*/d*S*) was found to be higher in the antigen-binding region[29]. The gene family evolves under a "birth-and-death" process among species and holds high polymorphism, in which MHC is also affected by balancing selection. There is an "arms race" process caused by the co-evolution between host and pathogen where the pathogen comes under positive selection for changing and updating its pathway to enter the host, and comparatively, the host evolves under positive selection to resist the pathogen. This is the Red Queen theory articulated by van Valen[30]. Many anti-virus genes like APOBEC3G[31,32], CCR5[33−38] support this hypothesis. Detection at the genomic level also demonstrated that the immune system experienced significant positive selection[12,13,15,17].

1.3 Genes associated with reproduction

In primates, many reproduction related genes (i.e. Female reproductive proteins: ZP2, ZP3, OGP; and Male reproductive proteins: PRM1, PRM2) have progressed with rapid adaptive evolution, due to sperm competition, sexual conflict and cryptic female choice[39−41]. Studies at the genomic level revealed that reproduction related genes evolved under recent positive selection and selected loci were significantly distinct between different human populations[13].

1.4 Perception receptor genes

In the long-term evolution in which animals choose self-favored food but avoid injurious in complex environment, gustatory sense (sweet, bitter, salty, sour and spicy) system plays indispensable roles. The mammalian bitter taste receptor (TAS2R or T2R) gene family experiences a "birth-and-death" process via adaptive evolution, namely positive selection and shows enormous variations[42].

TAS2R16 mediates sensitivity to some toxic components in nature. Individuals with K172N mutation in the human population have an even higher sensitivity to these components. Analysis on 997 individuals from 60 populations indicated that derived alleles at this polymorphism site rs846664 evolved before the expansion of early humans out of Africa and has had a very high frequency by positive selection all through human migration[43]. PTC (TAS2R28), its two haplotypes corresponding to "taster" and "nontaster" phenotypes (hsA and hsG), maintain a high frequency in African, Asian and European populations. Considering the effect of recent expansion of humans, Wooding et al.[44] detected balancing selection occurring in the PTC gene maintaining the two haplotypes after simulating the extent and time of the population expansion. However, it is not clear why the two "taster" and "nontaster" individuals occur at a high frequency in the population, when there is presumably a heterozygote advantage.

In the same manner, the olfactory receptor (OR) gene super-family also experiences adaptive evolution and "birth-and-death" process in mammals. Owing to distinct habitation, the sizes and ratios of intact to pseudo genes diverge extremely. Between humans and chimpanzees, genetic diversity of genes is lower. This results from positive selection in the human population, but purifying selection in the latter, because the non-synonymous substitution rate is significantly lower than the synonymous substitution rate in chimpanzees[45].

With the advent of enormous genomic and polymorphic data, detections at the genomic level suggest that positive selection events are more frequent than previously believed. For example, a recent paper reported abundant recent positive selection in Africa[13]. Discovery of the mechanism on positive selection will provide profound information on human origin and evolution, ability and mechanism of various viruses. Actually, many genetic disease associated genes have been detected under positive selection in human population.

1.5 Positively selected genetic disorder associated genes

Main pathogeny of spinocerebellar ataxia (SCA) type 2 (OMIM: 183090) is an expanded mutation of CAG repeats within the 3′ coding region of the SCA-2 gene. Research has indicated recent positive selection on haplotypes associated with the (CAG)8CAA(CAG)4CAA-(CAG)8 allele in modern Europeans[46]. Duchenne muscular dystrophy (DMD, OMIM: 300377) is a lethal X-linked excessive genetic disorder, the main clinical feature being progressive deterioration of muscle tissue, caused by a mutation of the DMD gene. Various genetic polymorphism patterns, e.g. low genetic diversity, significant population differentiation and long range linkage disequilibrium indicated that the 7th intron was af-

fected by positive selection[47]. OCA2, MYO5A, DTNBP1, TYRP1 and SLC24A5 are all albinism-associated genes, which were analyzed under positive selection in Europeans in a genomic scan[13]. Other positively selected genetic disease associated genes include BRCA1[48,49] and SRY[50]. Although the mechanism by which these genes are under positive selection is unclear, it will provide crucial insight into research on pathogenesis and therapy.

Malaria is one of the most important causes of child mortality worldwide, annually killing more than 1 million children in Africa[51]. Many genes resistant to malaria are positively selected[51]. HBB encoding a β-globin is the earliest identified selected gene, and later a variety of statistical detections validated the evolutionary patterns of HBB alleles in different populations[14]. Individuals with variant 202A of glucose-6-phosphate dehydrogenase (G6PD) have the ability to protect themselves from malaria[52]; as one variant in the promoter of CD40L (TNFSF5), an essential protein in the immune system, is associated with resistance to malaria[53]; long range haplotype tests suggested that they were recent targets of positive selection in the human population. GYPA, a gene encoding surface glycoprotein, was identified to be evolving under positive selection and balancing selection in the human population[54]. Duffy locus has been always thought of as a potential target of positive selection, considering the unique geographical distribution of the three alleles FY*A, FY*B and FY*O. For instance, FY*O has been fixed in sub-Saharan African populations; in contrast, Chinese populations have only FY*A; FY*B was present in their ancestors, but it was lost in most of the population now. Polymorphism features also concluded positive selection of FY*O and A[56,57], e.g. genetic diversity of FY*O in African populations was very low, and frequency of derived allele was high.

The question to consider is why genetic mutations that cause genetic diseases still evolve under positive selection. It is unclear but can be deduced from gene pleiotropy, where the mutation is advantageous in one instance but unfavorable in another[58]. As mentioned above, an advantageous mutation resistant to malaria in Africa causes sickle cell anemia. Another contributor is the changing environment; one mutation was advantageous in the past but might be disadvantageous after recent environmental changes. There is also the example of the relationship between genetic variation of dopamine receptor D4 (DRD4) and changes in human cultural structure[59]. There is a characteristic 48 bp variable number tandem repeat (VNTR) polymorphism in the 3rd exon, varying from 2 repeats (2R) to 11 repeats (11R), in which 7R has been known to be associated with attention deficient hyperactivity disorder (ADHD). However, Ding et al.[59] argued that 7R reached a high frequency in recent human evolution by positive selection according to population genetics. The distribution of 7R can be related to human migration considering relationship between 7R and activity. Additionally, in an early male-competitive society, 7R carriers enjoyed the advantages of mating partners, obtaining food and taking care of offspring. Recently, as human migration has declined, the importance of physical activity decreased to presumably induce genetic diseases[60,61]. However, on the other hand, research on genetic selection will provide important clues on pathogenesis of corresponding genetic diseases.

2 Detection of positive selection in genomic level

With the near-completion of various comparative genomic data, namely the genomes of humans, chimpanzees, mice and dogs, and release of SNP data of genetic variation, a great many studies of positive selection at the genomic level have been published. The advantage of research at the genomic level is exclusion of interference of population demographic history. However, many studies are based on SNP database, which are highly affected by ascertainment bias, even if using many different emendated methods. Comparative genomic analyses have high false positive rates using maximum likelihood methods for very few species. Additionally, the selection genes detected in different reports differ significantly, presumably because different methods used are suitable for selection at different times, and the data used are also different.

3 Methods to detect positive selection

Based on different data sets used, these methods can be categorized into three groups: methods based on intra-specific polymorphism data, methods based on inter-specific divergence data, and methods based on both intra- and interspecific data. We will give a brief intro-

duction on several commonly used methods in the following text.

3.1 Methods based on intra-specific polymorphism data

Natural selection can be simply classified into three kinds: positive selection, negative selection (or purifying selection) and balancing selection. For simplicity in explanation, a favorable mutation is fixed rapidly in a population under positive selection, negative selection is a process of deleting disadvantageous mutations in the population, and balancing selection can maintain two or more favorable alleles. Positive selection not only works on affected genes or regions but also induces similar genetic variation on adjacent regions of linkage genes, which is just genetic hitch-hiking, also named selective sweep. Specifically, genetic hitch-hiking indicates that a linked allele will be driven to a higher frequency when the frequency of the positively selected allele rises, and selective sweep defines that genetic diversity at linked loci is reduced when diversity of positively selected genes diminishes. Actually, the two concepts denote the same genetic phenomenon, and therefore they usually are not differentiated. This is the effect of positive selection on the gene frequency spectrum; frequency spectrum is simply noted as a fraction of the various alleles with different frequencies of the alleles. Alleles under positive selection possess long haplotypes, the linkage disequilibrium is also higher than average, and the frequency of derived alleles is higher. For the population genetic polymorphism patterns under positive selection, please refer to refs. [14,62]. These signatures deviating from neutrality are just bases of methods for detecting positive selection.

(i) Methods based on frequency spectrum. Ewens sampling formula[63] can be said to be a milestone of population genetics, and based on the formula, Watterson developed the famous Ewens-Watterson test[64]. Comparing the distribution of observed and predicted theoretical average gene diversity (average heterozygosity) which is calculated from the average of multi-loci heterozygosity as a null model, one is able to detect whether deviation from neutrality is occurring[3]. The detections are used only for infinite allele model data, like allozymes. Many other methods are developed for infinite site model data, e.g. nucleotide. Tajima's D test[65], similar to Ewenss-Watterson test in theory, is designed to detect the relation between a segregating site and nucleotide diversity. In neutrality, the two parameters indicate equal variation, and significant difference suggests deviation from neutrality. Both positive and negative selection will produce negative D values. And, when $D>0$, the gene or region evolves presumably under balancing selection[62,66]. For example, the Tajima's D observed values of PTC genes are 2.94 ($p(D>0)=0.01$), and 2.91 ($p(D>0)=0.01$) in Asian and European populations, respectively[44]. The D does not actually fit well with normal distribution, but fits well with β distribution. Besides, detection by stimulating D distribution using β also introduces errors[3].

The evolutionary process of sequences can be traced by coalescent theory in the population. Mutations can be classified into external, present recently, and internal. In positive and/or background selection, external mutations will increase relative to internal. According to the difference, Fu and Li[67−69] developed a series of similar tests, which can be used for different purposes. Unlike the aforementioned test, Fs is suitable for detecting deviation from neutrality by population expansion and/or positive selection[63]; comparatively, Fu and Li's D and F are more appropriate for detecting background selection. However, these methods are influenced significantly by population expansion.

Fay and Wu's H test[70] was designed based on the idea that genetic hitchhiking will produce new derived alleles at a high frequency, but background selection cannot. The method needs an outgroup, usually the orangutan. The ancestral allele of humans and chimpanzee is calculated by maximum likelihood or maximum parsimony method considering that one mutation occurs in one site, and then the other is a derived allele. For instance, three SNPs: rs2692396, rs846664 and rs1204014 at the human TAS2R16 loci have an extraordinarily high derived allele frequency. The fact that Fay and Wu's H value was significantly lower than 0 demonstrated that positive selection occurred in the gene[43]. In contrast with former methods, H test is influenced more by disturbance of population differentiation, but less by population expansion[70]. Under positive selection, high frequency derived alleles will be fixed in a very short time, and be suitable for detecting positive selection of alleles <~80000 years[14].

(ii) Population differentiation. In 1973, Lewontin and Krakauer[71] brought forward a selection detection method that took advantage of the fact that positive se-

lection within populations can promote differentiation among populations. However, the method is confounded by demographic history and population structure, as will be discussed later. Demographic history influences genome variation, but positive selection only affects a single gene and/or region. Therefore, selected loci will deviate from neutrally distributed loci when looking at a greater genomic level, which is the method to detect positive selection using population differentiation parameter F_{st} based on outlier approach. Taking advantage of this method, Akey et al.[72] obtained 174 candidates with positively selected genes according to Fst distribution of genomic SNPs, which was also the method for the first human positive selection gene map based on SNP data. There are still some examples, like IL4, which play an essential role in the cytokine signal pathway in the immune system. The polymorphism site rs2243250 in the promoter region differentiates significantly among many populations, which is the signature of positive selection to regulate IL4 expression: expression of IL4 with allele T is 3 times higher than C[73].

(iii) Detection based on linkage disequilibrium and haplotype structure. Linkage disequilibrium, known as correlation between two loci, can be broken down by recombination, and the haplotype structure will also be disrupted. However, in the positive selected loci and/or region, an advantageous mutation will be fixed so rapidly in the population that recombination does not substantially break down the haplotype, and the region possesses long range linkage disequilibrium and haplotype structure, which positively correlates with the intensity of positive selection[74]. Accordingly, haplotype length, similarity, and length and decay rate of linkage disequilibrium can be used for detecting deviation from neutrality. Relative to the methods based on frequency spectrum, analysis on haplotype is more powerful in detecting positive selection, but is more sensitive to recombination. Because of the limited time a haplotype is maintained, the corresponding methods are mainly used for detecting recent positive selection, like that unique to different human populations.

The parameter that long range haplotype (LRH) test uses is extended haplotype homozygosity (EHH), which is defined as the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by descent[74]. Recombination rate is highly heterogeneous and changes rapidly in the human genome. Therefore, the credible and convincing statistical test must take this factor into account. Another parameter rEHH (relative EHH) of LRH considers this. LRH is useful for detecting positive selection with allele frequencies as low as ~10%, and can identify accurately a single gene, and furthermore, is relatively robust to ascertainment bias[14]. Sabeti et al.[33] verified the selection of CCR5-Δ32 using this method. CCR5 is a chemokine receptor mediating entry of the HIV virus, and has many non-synonymous mutations in the human population for positive selection. The 32bp deletion allele in the coding region (CCR5-Δ32) occurs with a high frequency in the human population, the homozygote mutation can protect from HIV infection, and the heterozygote can delay infection[33]. Another classical example is the LCT locus; one of the haplotypes spreads to 1 Mb in Europeans[13]. As with LRH, there are also haplotype similarity (HS) tests[75], integrated EHH (iHS) test[13]and linkage disequilibrium decay test[12].

3.2 Detection based on interspecific divergence data

(i) Positive selection detection at the gene level. Ratio of non-synonymous substitution rate (dN) to synonymous substitution rate (dS) between two sequences is traditionally used to detect evolutionary patterns of genes between species (ω=dN/dS >,=,<1 indicates positive selection, neutral evolution and negative selection respectively). However, different sites among proteins evolve under different selective pressures for different functions; additionally, genes undergo distinct pressures during different evolutionary times. For these cases, Nielsen and Yang [76,77] developed site-specific models, branch-specific model[78,79], and branch-site model[80] based on maximum likelihood method. The models are all based on constructed nested model: one null model (ω values of all sites are lower than 1), and one corresponding general model (allowing for sites with ω>1) for detecting the significance of difference of likelihood values between the two models using likelihood ratio test (LRT). The null model will be rejected when chi square test is significant. In the "site-specific" model, Bayes methods are employed to detect potential positively selected sites, including two test : NEB (Native empirical Bayes)[76,81] and BEB (Bayes empirical Bayes)[82]. The former will generate a high false positive; therefore, it is advisable to use the latter method[83]. Simulation and actual data analyses indicate that the precision and power of many models are very low, and the commonly used nested model is M1a (Nearly Neu-

tral) vs. M2a (positive Selection) and M7 (β) and M8 (β&ω). In M7, ω fits with β (p, q) distribution, and ω values falls between 0 and 1, so no positive selection sites are allowed, and it is a null model. M8 has p0 fraction of sites fit with β (p, q) distribution and other p1 fraction has positive selection sites ($\omega > 1$), and is the general model. Many data analyses suggested that the power of M7 vs. M8 to detect positive selection sites is lower than M1a vs. M2a, as the former makes false positive data apparent. Although the model is a powerful tool to detect positive selection sites, the main problem is the false positive error generated. Anisimova et al.[84,85] gave the following suggestion after simulation: (1) high identity or few sequences (e.g., $S \leqslant 0.11$ or $T \leqslant 6$) detected will be less reliable; high diverse sequences will also be less accurate; detection efficiency will decrease as the codons of sequences drop below 100; (2) increasing number of sequences is the most efficient strategy to increase accuracy; (3) multiple models are employed. In the "Branch-specific" model, ω values in different lineages can be calculated, considering that genes evolve under different pressures at different times. For example, new genes appear by positive selection (or relaxation of selective constraints) after duplication, and non-synonymous mutations will not occur when the new gene is stabilized under purifying selection as others with $\omega<1$. "Branch-site" includes model A and B, which classifies phylogenetic lineages into foreground and background lineage. Foreground lineage is detected for selection and allows for positive selection sites, and others belong to background lineage which does not permit selective sites and lineage. Comparing likelihood values between Model A and M1a (Nearly Neutral) (free degree $df = 2$) is test 1; and it can also compare Model A and its null model (ω value is fixed 1 in foreground lineage), which is test 2. Computer simulation found that false positive rates were very high using test 1 (e.g. 20%—70%) and it was difficult to distinguish positive selection from relaxation of selective constraint, and accommodate it using test 2[83]. Additionally, reliability is high when calculating posterior probabilities using BEB (false positive rate is lower than 5% when significance is higher than 95%), but the method is too conservative to get a positive selection site[83]. The problem that dN/dS method has to solve is how to be able to distinguish between positive selection and relaxation of selective constraints. Normally, dN increases but dS does not change under positive selection and both dN and dS will increase when relaxation of selective constraint is occurring. Positive selection is so weak that likelihood ratio test does not yet support the event, which is also an issue with the dN/dS method. More sensitive methods to distinguish effectively positive and negative selection in protein are needed.

dN/dS is mostly scaled as protein evolutionary rate. For example, Wang et al.[86] found that evolutionary rate of brain-expressed genes in humans was not higher than in other primates, suggesting presumably that rapid evolution of human brain is attributed to several key genes.

(ii) Detection of positive selection in protein level. On the other hand, the aforementioned gene level detection methods of positive selection will produce errors when treating all amino acid changes equally. Accordingly, statistical models based on amino acid physicochemical properties have been developed, which compares distribution of amino acid changes deduced from phylogenetic trees and expected neutral distribution to detect sites deviating from neutrality[87,88]. TreeSAAP is the corresponding program to detect amino acid changes[89].

3.3 Detection based on intraspecific polymorphism and interspecific divergence data

The famous McDonald-Kreitman test[90] employs intraspecific polymorphism and interspecific divergence data. It classifies nucleotides into two kinds: fixed sites, in which species A holds one kind of nucleotide but the other species holds another nucleotide and all others are polymorphic sites. The two kinds are further grouped into synonymous and non-synonymous substitution sites. The four sites can be detected in a 2×2 contingency table[90]. The method is not influenced by such factors as demographic history and recombination rates; it only looks at coding sequences. HKA test is a statistical method for positive selection developed by Hudson et al.[91] based on the presumption that high rate loci have higher variation than low rate ones. However, HKA test is not commonly used because it requires observation of DNA sequences over long periods of time and constant population size which are not viable factors[3,91].

Theoretical bases of the methods vary greatly, and the results from different methods of the same data are usually different. For example, McDonald-Kreitman test compares non-synonymous and synonymous substitu-

tions from inter- and intra-specific data, and is suitable for detecting positive selection several million years ago[14]. The reality is that no single test can effectively determine all polymorphic data deviating from neutrality.

4 Influence of various factors on detection of positive selection

Wright-Fisher model is a standard neutral model, it uses a population in which mutations do not affect fitness and is panmictic and has constant size. Therefore, the precision will be influenced by these factors when detecting selection using the null model. For instance, population expansion or decline, variation in recombination rate and ascertainment bias of data will disrupt the model. Undoubtedly, these factors have to be excluded to verify selected genes or regions.

4.1 Demographic history

Actually, demographic history, including bottleneck, expansion, differentiation and fusion, can also produce similar genetic variation caused by selection. For example, both population expansion and positive selection can engender low frequency alleles; two or more high frequency alleles maintained in the population may be attributed to population subdivision rather than balancing selection. Consequently, detection of positive selection must take demographic history into account.

Both ASPM and Microcephalin are crucial regulating factors in brain development, and have been proven to be under strong positive selection in primates particular humans and anthropoid lineages by phylogenetic and maximum likelihood analyses. In modern human populations, the genes have high frequency haplotypes (ASPM haplotype 63 is 21%; Microcephalin haplotype 49 is 33%), which derive presumably from ongoing positive selection[18,24]. However, Currat et al.[92] argued that population differentiation following expansion can also produce similar patterns after coalescent simulation. But the parameters used should make clear whether they are consistent with normal human history, and the simulation should not consider the effect of recombination rate[93]. However, the differences between demographic history and natural selection are due to the fact that demographic history influences the overall genome, but the latter only affects one gene or region. Many recent detections of positive selection at the genomic level turn out to be advantageous. Additionally, demographic history will not be influenced by divergence data, McDonald-Kreitman test is accordingly not influenced by this factor. More credible and robust human history is also needed to exclude the disturbance of demographic history.

4.2 Recombination rate

In the human genome, nearly half of nucleotide variation results from recombination, e.g. genetic variation in the centromere, and telomere region has low recombination rates[94,95]; empirical data also demonstrated that positively selected genes are distributed mainly in lower recombination regions. Many methods have considered this factor; therefore, high-scale recombination rate maps are essential in detecting natural selection[94,95]. They are actually inter-effected in statistics when detecting demographic, recombination and selection using polymorphic data. In a similar pattern, positive selection will induce high false positive rates when detecting recombination hot-spots[96].

4.3 Ascertainment bias

Ascetainment bias is the biggest problem with SNP databases, which are concerned with the SNP acquiring process. Firstly, SNP sites are found in very few individuals and their frequencies are then obtained from a larger sample, which will lose many low frequency SNPs but obtain large numbers of mediate frequency SNPs. Ascertainment bias will have a substantial impact on the detection of demographic history, recombination rate and natural selection. Specific and detailed ascertainment bias and mathematical theory of calibration of SNP are suggested[97]. However, calibration still results in a large margin for error, and SNPs in databases should only be used as reference material for detecting natural selection of single genes. Detailed data are obtained once the overall gene region is sequenced in each individual from a large sample covering nearly the whole population, which will produce unbiased SNP data.

5 Prospect

Although the research on positive selection is still in its initial stages, it has already made much headway recently. It is only for a small fraction even if larger numbers of positively selected genes are detected today. Investigation has turned to polymorphic data from comparative genomic data, probably because people are now

more concerned with more recent human evolution. Besides the search for genes associated with brain, immunity and reproduction, enormous phenotypic variation in modern human populations will be emphasized in future research. Specifically, high diversity of social and cultural structure, particular variation in appearance, and food habits, have been observed among different human populations, such as expression of melanin in skin, hair related genes, genes for saccharide and amino acid metabolism. Recent studies have found important roles of large quantities of non-coding sequences, like cis-regulating sites[98], microRNA[99]; however, selection patterns on non-coding sequences have not been intensely studied, so further research is needed. Unbiased data is ideally required, but is not always possible. To minimize this issue, a sample should be large and have a wide-coverage in practice. Additionally, a variety of genomic detections of positive selection is influenced by ascertainment bias and thus requires further verification. Recently, genomic technologies like sequencing, SNP genotype, have been developed rapidly, making it possible to obtain large samples of data and research multi-genes in complex trait. In methodology, more rigorous statistical methods are needed in the treatment of various factors, e.g. demographic history, recombination rate, which have yet to be addressed. Presently a difficult point of research in positive selection is how to relate selection with specific functions of genes and to give directions for further functional studies, the ultimate objective of selection research. It is now an operable work to study genetic evolution and its function by positive selection in combination with disease associated studies.

The authors thank Ms Nishara Somasundaram, and Ms Shan Fang for helping to revise the manuscript.

1 Darwin C, Wallace A R. On the tendency of the species to form varieties and on the perpetuation of the species by natural means of selection. J Proc Linnaean Soc London (Zoology), 1858, 3: 45－62
2 Nei M. Selectionism and neutralism in molecular evolution. Mol Biol Evol, 2005, 22: 2318－2342
3 Nei M, Kumar S. Molecular Evolution and Phlogenetics. New York: Oxiford University, 2000
4 Kimura M. Evolutionary rate at the molecular level. Nature, 1968, 217: 624－626
5 King J L, Jukes T H. Non-Darwinian evolution. Science, 1969, 164: 788－798
6 Eyre-Walker A. The genomic rate of adaptive evolution. Trends Ecol Evol, 2006, 21(10): 569－575
7 Smith N G, Eyre-Walker A. Adaptive protein evolution in *Drosophila*. Nature, 2002, 415: 1022－1024
8 Fay J C, Wyckoff G J, Wu C I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature, 2002, 415: 1024－1026
9 Ohta T. Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci USA, 2002, 99: 16134－16137
10 Ohta T. The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst, 1992, 23: 263－286
11 Ohta T. Population size and rate of evolution. J Mol Evol, 1972, 1: 305－314
12 Wang E T, Kodama G, Baldi, P, et al. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc Natl Acad Sci USA, 2006, 103: 135－140
13 Voight B F, Kudaravalli S, Wen X, et al. A map of recent positive selection in the human genome. PLoS Biol, 2006, 4(3): e72
14 Sabeti P C, Schaffner S F, Fry B, et al. Positive natural selection in the human lineage. Science, 2006, 312: 1614－1620
15 Nielsen R, Bustamante C, Clark A G, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol, 2005, 3: e170
16 Nielsen R. Molecular signatures of natural selection. Annu Rev Genet, 2005, 39: 197－218
17 Bustamante C D, Fledel-Alon A, Williamson S, et al. Natural selection on protein-coding genes in the human genome. Nature, 2005, 437: 1153－1157
18 Mekel-Bobrov N, Gilbert S L, Evans P D, et al. Ongoing adaptive evolution of ASPM, a brain size determinant in *Homo sapiens*. Science, 2005, 309: 1720－1722
19 Evans P D, Anderson J R, Vallender E J, et al. Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. Hum Mol Genet, 2004, 13: 489－494
20 Zhang J. Evolution of the human ASPM gene, a major determinant of brain size. Genetics, 2003, 165: 2063－2070
21 Kouprina N, Pavlicek A, Mochida G H, et al. Accelerated evolution of the ASPM gene controlling brain size begins prior to human brain expansion. PLoS Biol, 2004, 2: e126
22 Wang Y, Su B. Molecular evolution of microcephalin, a gene determining human brain size. Hum Mol Genet, 2004, 13: 1131－1137
23 Evans P D, Anderson J R, Vallender E J, et al. Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size. Hum Mol Genet, 2004, 13: 1139－1145
24 Evans P D, Gilbert S L, Mekel-Bobrov N, et al. Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science, 2005, 309: 1717－1720
25 Zhang J, Webb D M, Podlaha O. Accelerated protein evolution and origins of human-specific features FOXP2 as an example. Genetics, 2002, 162: 1825－18352
26 Enard W, Przeworski M, Fisher SE, et al. Molecular evolution of

FOXP2, a gene involved in speech and language. Nature, 2002, 418: 869－872
27 Clark A G, Glanowski S, Nielsen R, et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science, 2003, 302: 1960－1963
28 Wang Y Q, Qian Y P, Yang S, et al. Accelerated evolution of the pituitary adenylate cyclase-activating polypeptide precursor gene during human origin. Genetics, 2005, 170: 801－806
29 Hughes A L, Nei M, Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature, 1988, 335: 167－170
30 van Valen L. A new evolutionary law. Evol Theory, 1973, 1: 1－30
31 Zhang J, Webb D M. Rapid evolution of primate antiviral enzyme APOBEC3G. Hum Mol Genet, 2004, 13: 1785－1791
32 Sawyer S L, Emerman M, Malik H S. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol, 2004, 2: e275
33 Sabeti P C, Walsh E, Schaffner S F, et al. The case for selection at CCR5-Δ32. PLoS Biol, 2005, 3: e378
34 Bamshad M J, Mummidi S, Gonzalez E, et al. A strong signature of balancing selection in the 5′ *cis*-regulatory region of CCR5. Proc Natl Acad Sci USA, 2002, 99: 10539－10544
35 Maayan S, Zhang L, Shinar E, et al. Evidence for recent selection of the CCR5-Δ32 deletion from differences in its frequency between Ashkenazi and Sephardi Jews. Genes Immun, 2000, 1: 358－361
36 Wooding S, Stone A C, Dunn D M, et al. Contrasting effects of natural selection on human and chimpanzee CC chemokine receptor 5. Am J Hum Genet, 2005, 76: 291－301
37 Zhang Y W, Oliver A, Ryder, et al. Intra- and interspecific variation of the CCR5 gene in higher primates. Mol Biol Evol, 2003, 20(10): 1722－1729
38 Li H P, Zhang Y W, Zhang Y P, et al. Neutrality tests using DNA polymorphism from multiple samples. Genetics, 2003, 163: 1147－1151
39 Swanson W J, Yang Z, Wolfner M F, et al. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci USA, 2001, 98: 2509－2514
40 Swanson W J, Vacquier V D. The rapid evolution of reproductive proteins. Nat Rev Genet, 2002, 3: 137－144
41 Wyckoff G J, Wang W, Wu C I. Rapid evolution of male reproductive genes in the descent of man. Nature, 2000, 403: 261－263
42 Shi P, Zhang J, Yang H, et al. Adaptive diversification of bitter taste receptor genes in mammalian evolution. Mol Biol Evol, 2003, 20: 805－814
43 Soranzo N, Bufe B, Sabeti P C, et al. Positive selection on a high-sensitivity allele of the human bitter-taste receptor TAS2R16. Curr Biol, 2005, 15: 1257－1265
44 Wooding S, Kim U K, Bamshad M J, et al. Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. Am J Hum Genet, 2004, 74: 637－646
45 Gilad Y, Bustamante C D, Lancet D, et al. Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am J Hum Genet, 2003, 73: 489－501
46 Yu F, Sabeti PC, Hardenbol P, et al. Positive selection of a pre-expansion CAG repeat of the human *SCA2* gene. PLoS Genet, 2005, 1(3): e41
47 Nachman M W, Crowell S L. Contrasting evolutionary histories of two introns of the duchenne muscular dystrophy gene, Dmd, in Humans. Genetics, 2000, 155: 1855－1864
48 Huttley G A, Easteal S, Southey M C, et al. Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Nat Genet, 2000, 25: 410－413
49 Fleming M A, Potter J D, Ramirez C J, et al. Understanding missense mutations in the BRCA1 gene: An evolutionary approach. Proc Natl Acad Sci USA, 2003, 100: 1151－1156
50 Wang X, Zhang J, Zhang Y P. Erratic evolution of SRY in higher primates. Mol Biol Evol, 2002, 19: 582－584
51 Kwiatkowski D P. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet, 2005, 77: 171－192
52 Ruwende C, Hill A. Glucose-6-phosphate dehydrogenase deficiency and malaria. J Mol Med, 1998, 76: 581－588
53 Sabeti P C, Usen S, Farhadian S, et al. CD40L association with protection from severe malaria. Genes Immun, 2002, 3: 286－291
54 Baum J, Ward R H, Conway D J. Natural selection on the erythrocyte surface. Mol Biol Evol, 2002, 19: 223－229
55 Zimmerman P A, Woolley I, Masinde G L, et al. Emergence of FY*A null in a plasmodium vivax-endemic region of Papua New Guinea. Proc Natl Acad Sci USA, 1999, 96: 13973－13977
56 Hamblin M T, Thompson E E, Rienzo A D. Complex signatures of natural selection at the duffy blood group locus. Am J Hum Genet, 2002, 70: 369－383
57 Hamblin M T, Rienzo A D. Detection of the signature of natural selection in humans: Evidence from the Duffy blood group Locus. Am J Hum Genet, 2000, 66: 1669－1679
58 Clark N L, Swanson W J. Pervasive adaptive evolution in primate seminal proteins. PLoS Genet, 2005 1(3): e35
59 Ding Y C, Chi H C, Grady D L, et al. Evidence of positive selection acting at the human dopamine receptor *D4* gene locus. Proc Natl Acad Sci USA, 2002, 99(1): 309－314
60 Harpending H, Gregory C. In our genes. Proc Natl Acad Sci USA, 2002, 99(1): 10－12
61 Hughes A L. Strength in numbers. Nature, 2002, 417: 795
62 Bamshad M, Wooding S. Signature of natural selection in the human genome. Nat Rev Genet, 2003, 4: 99－111
63 Ewens W J. The sampling theory of selectively neutral alleles. Theor Popul Biol, 1972, 3: 87－112
64 Watterson G A. Heterosis or neutrality? Genetics, 1977, 85: 789－814
65 Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 1989, 123: 585－595
66 Nielsen R, Williamson S, Kim Y, et al. Genomic scans for selective sweeps using SNP data. Genome Res, 2005, 15: 1566－1575
67 Fu Y X, Li W H. Statistical tests of neutrality of mutations. Genetics,

1993, 133: 693－709
68 Fu Y X. New statistical tests of neutrality for DNA samples from a population. Genetics, 1996, 143: 557－570
69 Fu Y X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics, 1997, 147: 915－925
70 Fay J C, Wu C I. Hitchhiking under positive darwinian selection. Genetics, 2000, 155: 1405－1413
71 Lewontin R C, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. Genetics, 1973, 74: 175－195
72 Akey J M, Zhang G, Zhang K, et al. Interrogating a high-density SNP Map for signatures of natural selection. Genome Res, 2002, 12: 1805－1814
73 Rockman M V, Hahn M W, Soranzo N, et al. Positive selection on a human-specific transcription factor binding site regulating IL4 expression. Curr Biol, 2003, 13: 2118－2123
74 Sabeti P C, Reich D E, Higgins J M, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature, 2002, 419: 832－837
75 Hanchard N A, Rockett K A, Spencer C, et al. Screening for recently selected alleles by analysis of human haplotype similarity. Am J Hum Genet, 2006, 78: 153－159
76 Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics, 1998, 148: 929－936
77 Yang Z. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol, 2000, 51: 423－432
78 Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol, 1998, 15: 568－573
79 Yang Z. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol, 1998, 46: 409－418
80 Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol, 2002, 19: 908－917
81 Yang Z, Nielsen R, Goldman N, et al. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics, 2000, 155: 431－449
82 Yang Z, Wong W, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol, 2005, 22: 1107－1118
83 Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol, 2005, 22: 2472－2479
84 Anisimova M, Bielawski J P, Yang Z. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol, 2002, 19: 950－958
85 Anisimova M, Bielawski J P, Yang Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol, 2001, 18: 1585－1592
86 Wang H Y, Chien H C, Osada N, et al. Rate of evolution in brain-expressed genes in humans and other primates. PLoS Biol, 2007, 5(2): e13
87 Xia X. What amino acid properties affect protein evolution? J Mol Evol, 1998, 47: 557－564
88 McClellan D A, McCracken K G. Estimating the influence of selection on the variable amino acid sites of the cytochrome *b* protein functional domains. Mol Biol Evol, 2001, 18: 917－925
89 Woolley S, Johnson J, Smith M J, et al. TreeSAAP: Selection on amino acid properties using phylogenetic trees. Bioinformatics, 2003, 19: 671－672
90 McDonald J, Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature, 1991, 351: 652－654
91 Hudson R, Kreitman M, Aguade M. A Test of neutral molecular evolution based on nucleotide data. Genetics, 1987, 116: 153－159
92 Currat M, Excoffier L, Maddison W, et al. Comment on "ongoing adaptive evolution of ASPM, a brain size determinant in homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". Science, 2006, 313: 172a
93 Mekel-Bobrov N, Posthuma D, Gilbert S L, et al. Response to comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in *Homo sapiens*" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". Science, 2006, 313: 172b
94 McVean G A, Myers S R, Hunt S, et al. The fine-scale structure of recombination rate variation in the human genome. Science, 2004, 304: 581－584
95 Crawford D C, Bhangale T, Li N, et al. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet, 2004, 36: 700－706
96 Reed F A, Tishkoff S A. Positive selection can create false hotspots of recombination. Genetics, 2006, 172: 2011－2014
97 Nielsen R. Population genetic analysis of ascertained SNP data. Hum Genomics, 2004, 1: 218－224
98 Haygood R, Fedrigo O, Hanson B, et al. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet, 2007, 39(9): 1140－1144
99 Chen K, Rajewsky N. Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet, 2006, 38: 1452－1456