Chinese Science Bulletin

Positive correlation between recombination rate and nucleotide diversity is shown under domestication selection in the chicken genome

FANG Lin[2*], YE Jia[2*], LI Ning[1,2*], ZHANG Yong[1,2], LI SongGang[1,2], GANE Ka-Shu WONG[3] & WANG Jun[1,2,4,5†]

[1] Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China;
[2] Beijing Genomics Institute, Beijing 101300, China;
[3] University of Washington Genome Center, Department of Medicine, Fluke Hall, M/C 352145, Seattle, USA;
[4] Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230, Odense M, Denmark;
[5] The Institute of Human Genetics, University of Aarhus, DK-8000 Aarhus C, Denmark

Positive correlation between recombination rate and nucleotide diversity has been observed in a wide variety of eukaryotes on megabase scale. On the basis of genome-wide chicken genetic variation map generated by comparing three domestic breeds with wild ancestor and the positions of markers on the genetic linkage map, we found that SNPs rates were similar for all chromosomes while the recombination rates increased in micro chromosomes. In other words no correlation exists in chromosome size. Nevertheless, when we scanned the genome by calculating the values of each characteristic within non-overlapping windows, instead of single value for each chromosomes, the nucleotide diversity was found to be significantly correlated with the recombination rate (r=0.27, P<0.0005). Furthermore, the significant association not only existed between these two features, but also existed between all 6 pairwise combinations of nucleotide diversity, recombination rate, GC content and average gene length. This co-variation is very meaningful for the studies of sequence evolution.

recombination rate, nucleotide diversity, positive correlation, domestication selection, chicken

The positive correlation between recombination rate and nucleotide diversity was first observed in the *Drosophila*[1] and subsequently in a wide variety of eukaryotes at megabase-scale, such as nematode[2], maize[3], mouse[4,5] and human beings[6,7]. Begun and his colleagues[1] found that the regions of the *Drosophila* genome with low rates of recombination exhibit low levels of polymorphism within populations. Three hypotheses were supposed to account for the correlation[8]: a strictly neutral hypothesis, a hitchhiking with selective sweeps of advantageous mutations hypothesis, and a background selection of deleterious mutations hypothesis. The neutral theory suggests that recombination is a latent factor for mutation[9]. According to the theory, the crossover events are prone to stimulate mutations, which is the likely consequence of faulty repair of double strand breaks with recombination[7]. On the other hand, the hitchhiking with selective sweeps model gives the explanation for low levels of polymorphism in region of low recombination, that the hitchhiking effect makes selectively advantageous mutants sweep through the population and eliminats variation at tightly linked

Received September 5, 2007; accepted November 19, 2007
doi: 10.1007/s11434-008-0159-y
*Contributed equally to this work
†Corresponding author (email: wangj@genomics.org.cn)
Supported by the Chinese Academy of Sciences (Grant Nos. GJHZ0701-6 and KSCX2-YW-N-023), Ministry of Science and Technology under high-tech program 863 (Grant Nos. 2006AA10A121 and 2006AA02Z334), Chinese 973 Program (Grant Nos. 2007CB815703 and 2007CB815705), Ministry of Education (Grant No. XXBKYHT2006001), National Natural Science Foundation of China (Grant Nos. 30725008, 90608010, 90403130 and 90612019), and Chinese Municipal Science and Technology Commission (Grant No. D07030200740000)

sites[10]. In regard to the background selection model, as with the selective sweeps model, it proposes that the selection against deleterious alleles might induce the reduction in large amounts of genetic variability at linked neutral sites[11].

Chicken is a modern descendant of dinosaurs and a non-mammalian amniote; its genome provides a new perspective on vertebrate genome evolution. Furthermore, the genetic variations map of chicken nucleotide polymorphisms[12] is based on the comparison between wild ancestor and domestic breeds. Domestic chickens that have been selected for different purposes are useful models for studying the genetic basis of extensive phenotypic diversity. Domestication selection, as an artificial process, has the particular evolution character which was distinguished from natural selection. The recent availability of the complete genome sequence of the chicken[13] and the genetic variation map of chicken nucleotide polymorphisms[12], combined with publicly available genetic maps of chicken[13−16] provide opportunities for the analysis of the genetic characterization of the chicken genome.

1 Methods and data source

The chicken genome consists of 38 autosomes and 2 sex chromosomes. According to their highly different chromosome size, the autosomes were classified into three groups by the ICGSC (International Chicken Genome Sequencing Consortium): macrochromosomes (GGA 1−5), intermediate chromosomes (GGA 6−10) and microchromosomes (GGA 11−38). In our analysis, we chose the same chromosomes as the ICGSC genetic analysis used, ranging from GGA1 to GGA28. Particularly, GGA16 and GGA22 were excluded due to their insufficient sequence, as well as GGA23 and GGA25 were excluded due to their insufficient genetic markers. Spearman's rank correlation coefficient was adopted to explain the correlative level.

1.1 Recombination rate

We estimated recombination rate (cM/Mb) based onthe genetic and physical length. We calculated the rate for each 5 Mb non-overlapping window within macro and intermediate chromosomes, while for each microchromosome we obtained one global rate. The data of genetic markers that we adopted for recombination rate were provided by the International Chicken Genome Sequencing Consortium[13]. Nevertheless, it was inevitable that some markers were not available as their physical positions conflicted with genetics positions. A dynamic programming method was employed to find a path with maximum markers as the optimal dataset of genetics markers. The recombination rates were calculated by assuming the same crossover rates between neighboring genetic markers (or flanking two mapped loci at each boundary of a window).

1.2 SNPs data

The essential purpose of the chicken SNP project[12] is to increase the density of markers on the genetic map. In addition, the SNPs data allow researchers to construct detailed haplotypes that segregate in different QTL crosses. We compared the sequence reads from the three domestic breeds to the 6.6X RJF (red jungle fowl) reference genome sequence (Chicken Genome Consortium, 2004). For each domestic breed, we sequenced one-quarter coverage of the genome (approximate one million reads from each breed) by using automated capillary sequencers (Amersham MegaBACE 1000). We used more stringent thresholds to minimize sequencing errors, with Q>25[17,18] for variant site and Q>20 in both flanking 5-bp regions. All data have been housed in ChickVD[19].

We summed up all the variant sites and the effective length of each region. The SNPs rate was then calculated as the slope of them, within 5 Mb non-overlapping windows the same as calculating recombination rate. All the nucleotide diversity values discussed here are the average of the three domestic lines and are magnified as $\pi\times10^3$ (SNPs/kb).

1.3 Average gene length

We used Ensembl gene (Release 27) here. Gene length referred to the whole gene length containing introns and exons. We averaged all the genes that were fully embedded in 5 Mb size windows. Correspondingly, the genes were ignored, if they were mapped at the boundary.

There were 18355 genes and the average length of them was 23.63 kb. Separately, 7138, 2646 and 4524 genes were from the macrochromosomes, intermediate chromosomes and microchromosomes. For these three classifications, the average gene lengths were 28.16, 22.65 and 17.05 kb respectively.

2 Results and discussions

2.1 Recombination and nucleotide diversity

At first glance, the observation of chicken was unexpected, given the existing backdrop of results from other species. On the basis of chicken genetic variation map[12] and the positions of markers on the genetic linkage map, when calculating SNPs rate and recombination rate at chromosomal level, we found that the SNPs rates were similar for all chromosomes while the recombination rates increased in microchromosomes. Nevertheless, when we used non-overlapping windows to calculate the values of each characteristic, instead of single value for each chromosome, the nucleotide diversity was found to be significantly correlated with the recombination rate. Notably, recombination rate and nucleotide diversity indicated significant co-variation for a window size less than 10 Mb (Figure 1(a), Table 1). Our result was identical with similar investigation in other species, which was always based on megabase-sized windows in the calculation as we have done in this paper. In other words, the correlation exists but only on appropriate window

Table 1 The table lists the *P* value of correlation coefficient for different window sizes[a)]

	Window size (Mb)					
	RR vs. SNP	RR vs. GC	SNP vs. GC	RR vs. GL	SNP vs. GL	GC vs. GL
0.5	$2.23\times10^{-5***}$	$1.86\times10^{-44***}$	$3.49\times10^{-54***}$	$1.17\times10^{-2**}$	$7.58\times10^{-10***}$	$6.42\times10^{-28***}$
1	$4.36\times10^{-5***}$	$2.05\times10^{-33***}$	$6.28\times10^{-35***}$	$4.49\times10^{-2**}$	$2.04\times10^{-7***}$	$4.07\times10^{-19***}$
2	$4.96\times10^{-6***}$	$2.13\times10^{-21***}$	$1.34\times10^{-19***}$	$1.78\times10^{-3***}$	$1.21\times10^{-5***}$	$1.13\times10^{-19***}$
5	$5.34\times10^{-4***}$	$7.81\times10^{-22***}$	$2.11\times10^{-6***}$	$4.35\times10^{-10***}$	$3.55\times10^{-3***}$	$9.42\times10^{-19***}$
10	$4.15\times10^{-2**}$	$1.90\times10^{-17***}$	$8.40\times10^{-3***}$	$1.08\times10^{-10***}$	$7.58\times10^{-2*}$	$1.16\times10^{-16***}$
20	$7.22\times10^{-2*}$	$2.58\times10^{-11***}$	$9.57\times10^{-2*}$	$1.44\times10^{-9***}$	$2.90\times10^{-1*}$	$2.36\times10^{-16***}$
Chromosome size	$2.55\times10^{-1*}$	$1.74\times10^{-5***}$	$6.90\times10^{-1*}$	$2.17\times10^{-6***}$	$9.09\times10^{-1*}$	$2.35\times10^{-9***}$

a) ***, significant relationship at $P<0.01$; **, significant relationship at $0.01<P<0.05$; **, non-significant case. RR, recombination rate; GC, GC content; SNP, nucleotide diversity; GL, gene length.

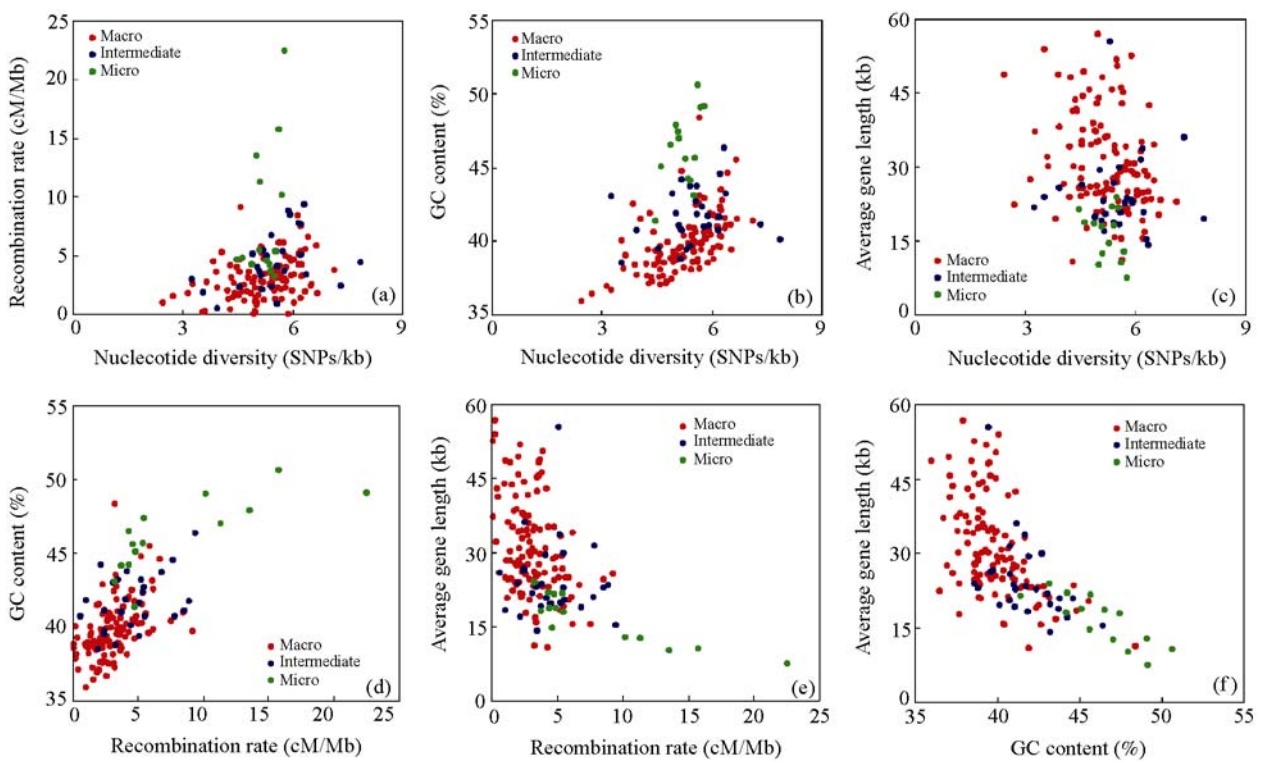

Figure 1 Scatterplots of all the 6 pairwise combinations of recombination rate, nucleotide diversity, gene length and GC content. Significant relationships were found in each combination. The macro and intermediate chromosomes are only given as one data point per chromosome.

Table 2 Correlation coefficient of all the 6 pairwise combinations of recombination rate, nucleotide diversity, gene length and GC content

	Recombination rate	GC content	Average gene length
Nucleotide diversity	0.27328 (5.34×10^{-4})	0.36807 (2.11×10^{-6})	–0.23137 (3.55×10^{-3})
Recombination rate		0.67036 (7.81×10^{-22})	–0.47204 (4.35×10^{-10})
GC content			–0.63012 (9.42×10^{-19})

a) The bracketed parts are the associated probabilities. A 5-Mb non-overlapping window was used for macro and intermediate chromosomes calculation, and one data point per chromosome was given in the estimation of microchromosomes.

scale. So, what happened when we jumped from chromosome level to megabase size?

It is well known that shorter chromosomes tend to have higher average recombination rates, both in human and mouse genomes. Given the obligatory chiasma per bivalent to ensure normal segregation during meiosis, the increased recombination rate of chicken's microchromosomes was expected. We calculated one average recombination rate for each chicken chromosome, and the waving interval of all the rates was approximately from 2 to 20 cM/Mb[13]. Correspondingly, when we calculated the rate for each 5 Mb non-overlapping window on whole genome, these rates fluctuated from 0.04 to 22.48 cM/Mb. Although 0.04 cM/Mb seems immoderate, it is consistent with the fact that lower recombination rates were observed near the centromere of macrochromosomes. Furthermore, within the given window size, insufficient genetic markers in some microchromosomes might be another important reason for the lower recombination rates. As regards to SNPs rates, they varied approximately from 4.4 to 6.7 SNPs/kb at chromosomal level, and from 2.43 to 7.84 SNPs/kb at 5 Mb window size. About the SNPs data, we have to note the little population attribute and the deficient coverage. When we did the correlation analysis at megabase-sized level, the interference should be weakened, which was due to the difference of data quality.

2.2 Correlations between all 6 pairwise combinations

All 6 pairwise combinations of SNPs rate, recombination rate, GC content and average gene length are found to be significantly correlated (Figure 1(a)—(e)). With the exception of the average gene length, the other 3 measures exhibit co-variation. The average gene length is negatively correlated with the other three features. Correlation coefficients and probabilities are summarized in Table 2.

GC content seems to be the most strongly correlated factor of the four considered characteristics. The compositional variation of GC content over large scales is referred to as an "isochore"[21,22]. It is a feature of warm-blooded vertebrates, including mammals and birds[22]. The GC content shows a broad distribution among individual chromosomes of the chicken genome[13]. The association between recombination rate and GC content can be explained by "biased gene conversion", which induces a high GC content in regions with high recombination rate[13,23].

The variation of gene size is mostly caused by intron size variation. The negative correlation between average gene size and GC content could partly be explained by the fact that GC-rich regions tend to be gene-dense with many compact genes[24].

Therefore, recombination rate is positively correlated with nucleotide diversity and GC content. The average gene length is negatively correlated with the GC content.

2.3 Domestication

As all the SNPs data were obtained by comparing domestic breeds with wild ancestor, to some extent, we have taken the initial step to understand the population genetic processes that produce the patterns of molecular polymorphism and the patterns of divergence observed during the long way of domestication. There is no further result because it is limited by the two problems: (1) little population attribute of the recent nucleotide diversity map for chicken, (2) insufficient positions of the markers on the genetic linkage map.

The authors thank the International Chicken Genome Consortium for the sharing of chicken genome data and collection of genetic markers, especially Prof. Martien A.M. Groenen and John Wallis. This work was further supported by the Danish Platform for Integrative Biology, the Ole RØmer grant from the Danish Natural Science Research council and the Danish Medical Research Council.

1 Begun D J, Aquadro C F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature, 1992, 356(6369): 519－520

2 Cutter A D, Payseur B A. Selection at linked sites in the partial selfer *Caenorhabditis elegans*. Mol Biol Evol, 2003, 20(5): 665－673[doi]

3 Tenaillon M I, Sawkins M C, Long A D, et al. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. mays L.). Proc Natl Acad Sci USA, 2001, 98(16): 9161－9166[doi]

4 Nachman M W. Patterns of DNA variability at X-linked loci in Mus domesticus. Genetics, 1997, 147(3): 1303－1316

5 Waterston R H, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. Nature, 2002, 420(6915): 520－562[doi]

6 Nachman M W. Single nucleotide polymorphisms and recombination rate in humans. Trends Genet, 2001, 17(9): 481－485[doi]

7 Lercher M J, Hurst L D. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet, 2002, 18(7): 337－340[doi]

8 Hudson R R. How can the low levels of DNA sequence variation in regions of the *Drosophila* genome with low recombination rates be explained? Proc Natl Acad Sci USA, 1994, 91(15): 6815－6818[doi]

9 Hellmann I, Ebersberger I, Ptak S E, et al. A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet, 2003, 72(6): 1527－1535[doi]

10 Wiehe T H, Stephan W. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol Biol Evol, 1993, 10(4): 842－854

11 Charlesworth B, Morgan M T, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics, 1993, 134(4): 1289－1303

12 Wong G K, Liu B, Wang J, et al. A genetic variation map for chicken with 2.8 million singlenucleotide polymorphisms. Nature, 2004, 432(7018): 717－722[doi]

13 Hillier L W, Miller W, Birney E, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature, 2004, 432(7018): 695－716[doi]

14 Groenen M A, Cheng H H, Bumstead N, et al. A consensus linkage map of the chicken genome. Genome Res, 2000, 10(1): 137－147

15 Schmid M, Nanda I, Guttenbach M, et al. First report on chicken genes and chromosomes 2000. Cytogenet Cell Genet, 2000, 90(3-4): 169－218[doi]

16 Romanov M N, Price J A, Dodgson J B. Integration of animal linkage and BAC contig maps using overgo hybridization. Cytogenet Genome Res, 2003, 102(1-4): 277－281[doi]

17 Ewing B, Hillier L, Wendl M C, et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res, 1998, 8: 175－185

18 Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res, 1998, 8: 186－194

19 Wang J, He X M, Ruan J, et al. ChickVD: A sequence variation database for the chicken genome. Nucleic Acids Res, 2005, 33(Database issue): 438－441[doi]

20 Axelsson E, Webser M T, Smith N G, et al. Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. Genome Res, 2005, 15(1): 120－125[doi]

21 Eyre-Walker A, Hurst L D. The evolution of isochores. Nat Rev Genet, 2001, 2(7): 549－555[doi]

22 Bernardi G. Isochores and the evolutionary genomics of vertebrates. Gene, 2000, 241(1): 3－17[doi]

23 Marais G. Biased gene conversion: implications for genome and sex evolution. Trends Genet, 2003, 19(6): 330－338[doi]

24 Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome. Nature, 2001, 409(6822): 860－921[doi]