

Properties of Haplotype Structure of a Non-coding Region

Jian-hong Sun^{1,2}, Hai-cheng Xu¹

¹ Engineering College of Honghe University, Yunnan Mengzi, 661100

² Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Yunnan Kunming 650091, People's Republic of China.
sparkhonghe@foxmail.com

Abstract: Haplotype structures are important in the study of human evolutionary history. Here, based on 1,092 individuals' data from the 1000 Genomes Project, a haplotype block (Solid Spine of LD method defined) located on 22q12.3 (Position: 34876961- 34890077), a non-coding region of chromosome 22 was analyzed from the view of haplotype structure. Statistical tests suggested the region is selectively neutral with a very low recombination rate, and the genetic diversity analysis shows that the African population (AFR) has higher diversity than non-AFR populations in this region. As a haplotype block, the results also show that only 52 of 2,184 possible haplotypes are observed in this region. The haplotypes distribution concentrate on 4~5 common haplotypes in non-African populations (i.e., East Asian population (ASN), European population (EUR) and American population (AMR)). In contrast with non-AFR, the haplotypes are more evenly and more widely expressed in AFR. Furthermore, most of SNPs of ASN are wild type which reveals the bottleneck effect on ASN.

[Jian-hong Sun. **Properties of Haplotype Structure of a Non-coding Region.** *Life Sci J* 2015;12(4):90-96]. (ISSN:1097-8135). <http://www.lifesciencesite.com>. 11

Keywords: Haplotype; single-nucleotide polymorphisms (SNPs); linkage disequilibrium (LD); 1000 Genomes Project;

1. Introduction

In genetics, a haplotype is a combination of alleles at adjacent locations on a chromosome that are inherited together. Since 2001, Daly and his colleagues have reported haplotype structures across 500 kilobases on chromosome 5q31 using 103 single-nucleotide polymorphisms (SNPs) in a European-derived population (Daly et al. 2001). Various studies showed that many regions in the human genome can be organized into haplotype blocks of 5–100 kb. Most variation is accounted for by two to four haplotypes (Jeffreys et al., 2001; Patil et al., 2001; Gabriel et al., 2002; Anderson & Slatkin, 2004). These haplotype blocks are separated by a recombination hotspot, an area in which there is little or no evidence of historical recombination. With the advent of high-throughput genomics, biologists are joining the big-data club (Marx, 2013). In 2012, the 1000 Genomes Project Consortium reported a validated haplotype map of 38 million single nucleotide polymorphisms (SNPs), 1.4 million short insertions and deletions, and more than 14,000 larger deletions (1000 Genomes Project Consortium, 2012). Facing such massive data sets, life scientists are challenged with handling and further analyzing these data. SNPs are the most common type of genetic variation and can be divided into haplotype blocks for analysis and study, effectively reducing workload and complexity.

In this article, I applied neutrality tests, linkage disequilibrium (LD) analysis, nucleotide/haplotype diversity, recombination detection and haplotype

network analysis, etc., to analyze a haplotype block. My goal is to genetically characterize the commonalities and differences among different populations. Information on human evolutionary history can be mined from these analyses. Furthermore, I analyzed the generic causes of the nucleus strong LD structure of this block and investigated why the LD structure is not found in ASN. The Genome-wide association study (GWAS) research reports related to the LD structure were also discussed.

2. Material and Methods

The data (Version: Phase1_release_v3) was downloaded from the website of the 1000 Genomes Project. Then, 1092 individuals of phase1 data were collected worldwide from 14 populations in four common ancestry groups: Europe (EUR); Africa (AFR); East Asia (ASN); Americas (AMR). Only biallelic polymorphism sites were selected to study in this paper.

The definitions of haplotype blocks are not uniform. Wall and Pritchard cataloged these definition methods into two groups: defined blocks of regions with limited haplotype diversity and transition zones in which there are evidence for extensive pair-wise disequilibrium and historical recombination (Wall & Pritchard, 2003). The commonly used software, Haploview (Barrett et al., 2005), offers three different methods to generate haplotype blocks. They are as follows: Confidence Intervals (Gabriel et al., 2002), the Four Gamete Rule (Wang et al., 2002) and the Solid Spine of LD, which was developed by the Haploview

team internally. All three methods belong to the group previously mentioned. In my research, I selected the Solid Spine of LD similarity method to generate haplotype blocks. In my work, a "solid spine" of perfect LD ($r^2 = 1$) constructs a haplotype block. The specific steps are as follows: 1) search for a SNP group across a whole chromosome, in which any two SNPs are in perfect LD; and 2) determine the range of the longest distance between two SNPs of the previous SNPs' group. This step creates a haplotype block. The team of the 1000 Genomes Project identified 99.7% SNPs with a frequency of 0.05 (1000 Genomes Project Consortium, 2012). To maintain high accuracy, only the SNPs with $MAF \geq 0.05$ were selected for haplotype block analysis. On chromosome 22, 475,362 SNPs have been reported, and among of them, 95,602 SNPs have $MAFs \geq 0.05$. In my results, 5,235 perfect LD groups have been found in these SNPs with $MAFs \geq 0.05$. Among them, there are only 3 perfect LD groups out of more than 20 SNPs. In this article, the one with the maximum number of SNPs was selected for analysis. The group is located at position 34,876,961-34,890,077. The region works as a haplotype block of Solid Spine in the LD method, whose "spine" consists of 43 SNPs. To date, no gene has been found around of this region; no related reports have been published. For convenience, this region is named 22q12.3 (13K) in the following description.

The subsequent analysis used the following common software: DnaSP v5 (Librado & Rozas, 2009, University of Barcelona, 645 Barcelona 08071, Spain), Haploview 4.2 (Barrett et al., 2005, Daly Lab at the Broad Institute, Cambridge, MA 02141, USA) and NETWORK 4.6.12 (Fluxus Technology Ltd: 4 Market Hill | Clare | Suffolk CO10 8NN | England Company Number: 3790136, | VAT (Tax) number: GB740626544, <http://www.fluxus-engineering.com>).

The purpose of the neutrality test is to determine whether sequences fit the neutral theory model. Here, I used Tajima's D test (Tajima, 1989) and Fu & Li's D* and F* test for the analysis (Fu & Li, 1993). The LD calculation is an essential prerequisite for haplotype block division. Two commonly used measures, r^2 (Hill & Weir, 1994) and D' (Lewontin, 1964), were employed to analyze LD in the focused region. Nucleotide diversity is a concept in molecular genetics that is used to measure the degree of polymorphism within a population. In Nei and Li's model (Nei & Li, 1979), this measure is defined as the average number of nucleotide differences per site between any two DNA sequences chosen randomly from the sample population. This measure is denoted by π (Watterson, 1975). Haplotype diversity (Hd) was analyzed for four common ancestry populations groups (AFR, AMR, ASN and EUR). The analysis used the measure method

proposed by Nei (Nei, 1987). In population genetics, gene flow is the transfer of alleles or genes from one population to another. Here, I used estimators F_{st} and N_m (Hudson et al., 1992) to analyze gene flow from AFR to other non-AFR population groups. Recombination is one of the major evolutionary forces that shape genetic diversity (Bafna & Bansal, 2004). Many methods have been proposed to detect recombination from DNA sequences (Hudson, 1987; Salminen et al., 1996; Martin & Rybicki, 2000; Worobey, 2001). In this study, I used the classic estimator of recombination proposed by Hudson, which is based on the number of pairwise differences. The estimator shows as $R=4Nr$, where N is the population size (for autosomal loci of diploid organisms) and r is the recombination rate per sequence (per gene). To estimate R between adjacent sites, I used the following equation: $R=R(\text{per gene}) / D$, where D is the average nucleotide distance in the base pairs of the analyzed region. Another parameter is the minimum number of recombination events, RM (Hudson & Kaplan, 1985). The haplotype network allows us to visualize relationships between haplotypes. Reconstructing phylogenetic networks through a large number of haplotype samples allows us to trace potential evolutionary paths and mine historical information on human population differentiation. This study used the popular software NETWORK Version 4.6.1.2 to draw the haplotype network, in which the software offers the Reduced Median (RM) network algorithm (Bandelt et al., 1995) and a Median-Joining (MJ) network algorithm (Bandelt et al., 1999). It also offers the Maximum Parsimony (MP) algorithm (Polzin & Daneshmand, 2003) to identify unnecessary median vectors and links for optimizing the network. However, these algorithms are designed for non-recombining bio-molecules. Recombining bio-molecules will deliver high-dimensional networks, which are difficult to interpret. Consequently, analysis of the haplotype network depends on recombination.

3. Results

For the four population groups, the neutrality tests suggested that the region is selectively neutral. There are 272 SNPs reported in this region, 22q12.3 (13k). Ninety-two SNPs and their MAF (Minor Allele Frequency) are ≥ 0.05 , but the HWpval (Hardy-Weinberg equilibrium p-value) of 8 SNPs are so small that the 8 SNPs will be ignored in some analyses. A threshold value of 0.001 was set to cut off unqualified SNPs. The HWpval of a SNP is defined as the probability that its deviation from the H-W equilibrium could be explained by chance. In this study, the 84 screened SNPs are the focus of the analysis and discussion. As a reference, all 272 SNPs are used for analysis in some cases.

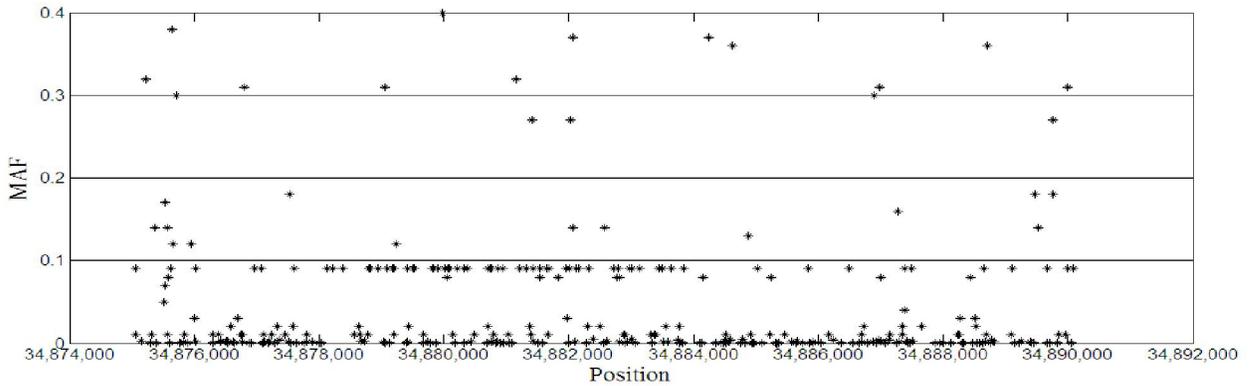


Figure 1. Scatter based on the minor frequency of SNPs. The strip close to the line of MAF = 0.1 is the “spine” of this block.

Table 1 SNP Information and Nucleotide Diversity

Pop Group	Nsi = 84 SNPs (MAF ≥ 5%, HWpval ≥ 0.001)						Nsi=272 SNPs (All SNPs)					
	η	Nse	h	π	Hd	σ ²	η	Nse	h	π	Hd	σ ²
AFR	84	492	29	0.30393	0.887	0.00003	196	492	95	0.12641	0.935	0.00003
AMR	84	362	21	0.15122	0.689	0.00029	165	362	56	0.0577	0.812	0.00025
ASN	17	572	16	0.0584	0.656	0.0001	59	572	58	0.02572	0.776	0.00015
EUR	84	758	15	0.20496	0.694	0.00007	137	758	62	0.07528	0.839	0.00007

Nsi is the number of sites; η is the total number of mutations; Nse is the number of sequences of the sample; h is the number of haplotypes observed in the sample; π is the nucleotide diversity; Hd is the Haplotype (gene) diversity and its sampling variance is shown as σ². The results are calculated in software DnaSP v5.

Among of the chosen SNPs, 43 SNPs are in perfect LD relationship ($r^2 = 1$ each other), working as a “spine” of the haplotype block. Another 27 SNPs are in strong LD relationship ($r^2 > 0.33$ and $|D'| > 0.67$) with previous 43 SNPs. A strong LD structure presents a strip near the line of MAF equal to 0.1, and as the backbone of this structure acts as the “solid spine” of the haplotype blocks (Figure 1). Another commonly used measure of LD show that all non-spine SNPs (41 SNPs) are in complete LD ($|D'| \approx 1$) relationship with every SNP of the “spine”. The LD of all SNPs (MAF ≥ 0.05) of region 22q12.3 (13K) shows that almost 84 SNPs (MAF ≥ 0.05, and 8 SNPs with very small HWpval are ignored.) are in strong LD (Measure of LD: $|D'|$) with each other. Notably, the SNPs in the strong LD structure are almost wild type, but the same SNPs in the other population (AFR, AMR and EUR) present as mutation type with high frequency (AF>0.05). Reducing the strength requirements of LD to $r^2 > 0.33$ and $|D'| > 0.67$ enables us to search all of chromosome 22. The furthest extension of the strong LD structure has been determined; it includes 174 SNPs (HWpval ≥ 0.001). The range of the strong LD structure extends to 113.56 Kbp across 34,864,640 ~ 34,978,291. In this region, 2,109 SNPs are reported. In

the extended region, 208 SNPs are ignored because their HWpval < 0.001 and their MAF ≥ 0.01. Among these SNPs on the strong structure, only 6 SNPs present as mutation type in ASN. For convenience, the Strong LD Structure is named 22q12.3 (SLDS) in the following description.

AFR has the highest nucleotide diversity, π, which is more than five times higher than the nucleotide diversity of ASN (Table 1). Many common SNPs of the other three populations cannot be observed mutation type in ASN. The data source consists of 1,092 individuals, which means there are a total of 2,184 possible haplotypes in theory. However, only 52 haplotypes are observed in this haplotype block. Among them, 8 haplotypes are most commonly observed and account for 93.55% of the total (Table 2). However, the common haplotypes show strong differentiation distribution in four ancestry-based groups (Table 2). In non-AFR populations groups, the haplotype distribution concentrates on fewer common haplotypes: a) In EUR, the most common haplotypes, H1, H2, H3 and H4, account for 97.76% of the total; b) In ASN, the most common haplotypes, H1, H2 and H3, account for 95.1% of the total; c) In AMR, H1, H2, H3, H4 and H5 account for 91.71% of the total. However, in AFR, the distribution of the common haplotypes is

more balanced. Twenty-nine haplotypes are observed in AFR; 21 haplotypes are observed in AMR; 16 haplotypes are observed in ASN and 15 haplotypes are

observed in EUR (Table 1). The haplotype diversity, H_d , is highest in the AFR and, unsurprisingly, the lowest in the ASN (Table 1).

Table 2. Commonly observed haplotype in the haplotype block 22Q12.3 (Position: 34876961- 34890077)

Haplotype	All (%)	EUR (%)	ASN (%)	AMR (%)	AFR (%)
H1	35.62%	36.54	42.48	45.86	18.70
H2	30.68%	38.13	37.76	29.83	11.59
H3	11.40%	14.25	14.86	9.94	4.07
H4	6.91%	8.84	0	4.70	13.62
H5	4.26%	0	0.52	1.38	17.28
H6	2.11%	0	0	0.55	8.94
H7	1.33%	0	0	0.28	5.69
H8	1.24%	0	0	0.28	5.28
Others	6.45%	2.24	4.38	7.18	14.83

The “All (%)” shows the proportion of each common haplotype in the total sample. “EUR (%)”, “ASN (%)”, “AMR (%)”, “AFR (%)” represent the proportion of each common haplotype in each common ancestry group.

The results of recombination detection estimate that the R per gene is 0.001. The estimate of R between adjacent sites is 0.0000. The minimum number of recombination events, R_m , is estimated to be 40. The R_m decreased from 40 to 19 when a subset of 84 SNPs was selected to participate in the calculation instead of all 272 SNPs. The results from recombination detection also show that the recombination rate of this region is very low, suggesting that the required conditions of haplotype network analysis have been satisfied.

To reduce the complexity, only the 84 selected SNPs participate in the haplotype network analysis. The median-joining haplotype network (Figure 2) contains four star-like haplotype clusters centered separately on nodes H1, H2, H3 and H4. From the ancestry haplotype “Root,” the mutations partition the haplotypes into two main branches, M1 and M2 (Figure 2(a)). No haplotypes are observed in ASN on branch M2. In the Tajima-Nei model (Tajima & Nei, 1984), the farthest genetic distance between the ancestry haplotype “Root” and a main haplotypes node (H1~H8) is 0.951. That value represents the distance between the “Root” and H3. Similarly, the nearest genetic difference is 0.690; this value represents the distance between the “Root” and H2. The ASN population groups are short of some medium transition haplotypes on clades H40 → H19 → H52 → H8 → H12 → H47 → H3 and H51 → mv6 → H6 → H8 → H12 → H47 → H3 (Figure 3).

4. Discussions

The following features are used to define haplotype blocks in this study: 1) each haplotype block contains ≥ 2 SNPs; 2) some haplotype blocks may be related as block intersections or blocks included; and 3) not every SNP of a chromosome should belong to a haplotype block. Furthermore, the definition of

haplotype block inherits the main feature of the LD Solid Spine: the first and last markers in a block are of strong LD (i.e., $r^2 = 1$) with all intermediate markers. However, the intermediate markers are not necessarily in LD with each other. This type of haplotype block is conducive to finding a strong LD structure (the “solid spine” of the haplotype block is the backbone of the strong LD structure), which has been interrupted by mutations or recombination.

Progressively filtering the mutations with lower minor frequencies extends the length of the LD structure, which prefers to recover its initial state. The Strong LD Structure 22q12.3 (SLDS) in non-African populations extends over longer distances than in Africans, which might reflect a population bottleneck at the time modern humans first left Africa (Wall & Pritchard, 2003; Marotta et al., 2012).

It would be interesting to know how and when the 22q12.3 (SLDS) is generated. There are several main factors that can produce linkage disequilibrium in a population: natural selection, random drift, genetic hitchhiking, and gene flow. Which factor is the predominate cause of the 22q12.3 (SLDS)? Accordingly, two reasons, natural selection and genetic hitchhiking can almost be excluded as the cause of the 22q12.3 (SLDS). First, the range of the 22q12.3 (SLDS) is across a long non-coding region. LOC441996 is the nearest gene; it is 6,488 bp away from the boundary of the 22q12.3 (SLDS). LOC441996 is only a mitochondrial pseudogene. Second, the results of the neutrality tests (Tajima's D test, Fu & Li's D^* test and Fu & Li's F^* test) show that evolution of the region 22q12.3 (13K) in AFR, AMR and EUR does not deviate from the neutral model. In fact, the reported evidence of natural selection is sparse at the molecular level, especially in the non-coding region. Furthermore, gene flow estimates can be computed using information from haplotype data.

Region 22q12.3 (13K) undergoes little recombination, which suggests that the gene flow estimator, N_m , is more accurate than F_{ST} (Hudson et al., 1992). Therefore, I tend to measure the gene flow using N_m . The calculated values of N_m between AFR and the other populations groups (AMR, ASN and EUR) are 1.45, 0.97 and 1.73, respectively. These values are based on the haplotypes of 1,092 total individuals; the four population groups of these SNPs are of 22q12.3

(SLDS). The values show the low levels of migration from AFR to the other populations groups. Therefore, I suggest that the 22q12.3 (SLDS) is most likely produced by the random drift process and that it was generated before our ancestors migrated out of Africa. The founder effect observed on the haplotype network also supports this opinion. This strong LD structure cannot be observed in ASN, which might be due to a genetic bottleneck or a sampling problem.

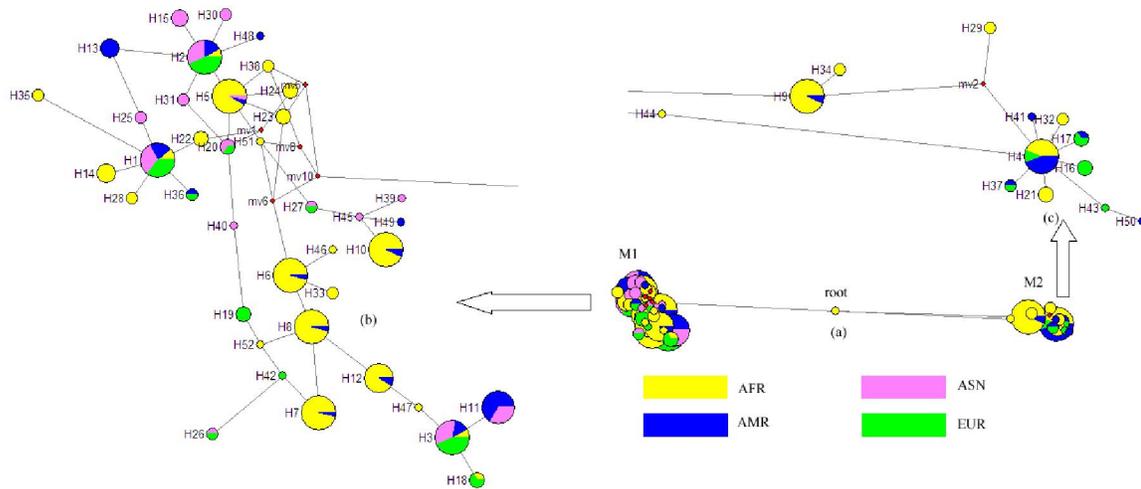


Figure 2. Median-joining network of 22q12.3 (13K) haplotypes: (a) Trunk figure; (b) enlarged figure of the main branch, M1; (c) enlarged figure of the main branch, M2. The tree was based on 52 haplotypes. Filled circles indicate the haplotypes; the numbers identify the haplotypes, with the size of each circle proportional to the observed frequency. The colors within the circles correspond to the different populations groups – yellow: AFR; blue: AMR; pink: ASN; green: EUR; black: ancestry haplotype (named “Root”); red: median vector. The size of the pie chart is proportional to the occurrence in the populations.

According to the published reports of GWAS (Hindorff et al., 2014), I found that one SNP of 22q12.3 (SLDS), rs738968, had been reported. (Another SNP, rs8141914, was excluded from the 22q12.3 (SLDS) because $HWP_{val} < 0.0001$.) (Kennedy et al., 2012). The SNP is located at the 3' end, downstream of LOC441996 (aconitase 2, mitochondrial pseudogene). The genetic distance from LOC441996 is 95,977 bp. In their results, rs738968 shows significant association with secreted TNF- α in the Caucasian cohort but not in the African-American cohort. However, the 22q12.3 (SLDS) was not mentioned, possibly because of the SNP's low-resolution ratio.

The possibility of recombination between a pair SNPs is small if they are in strong LD. Although previous LD analysis has indicated a low rate of recombination in 22q12.3 (13K), the results of the recombination estimator R are more accurate. Both the value of R per gene and the value of R between

adjacent sites are very low, which is consistent with the LD analysis. Note that R_m underestimates the total number of recombination events, which means recombination events might actually happen far more frequently than the results of the calculation suggest. Although the region has a very low recombination rate, which does not fully meet the requirements of algorithms in the haplotype network, the haplotype network analysis still has a reference value because the contribution of recombination is very limited.

In the haplotype network, star-like haplotype clusters usually indicate historic demographic expansion events. The center node of star-like clusters function as a founder haplotype node, such as H1, H2, H4 and H5 (Figure 2.(b) (c)). For haplotype nodes H1 and H5, the evidence of historic demographic expansion events in AFR populations is more obvious. For ASN and AMR populations, haplotype node H2 as a founder haplotype node might have led to the demographic expansion events throughout history

(Figure 2.(b)). The population diversity of the AMR cohort is closer to AFR than the ASN and EUR cohorts. As the birthplace of human ancestors in

Africa, the number of haplotypes, compared with the other three groups, is much richer.

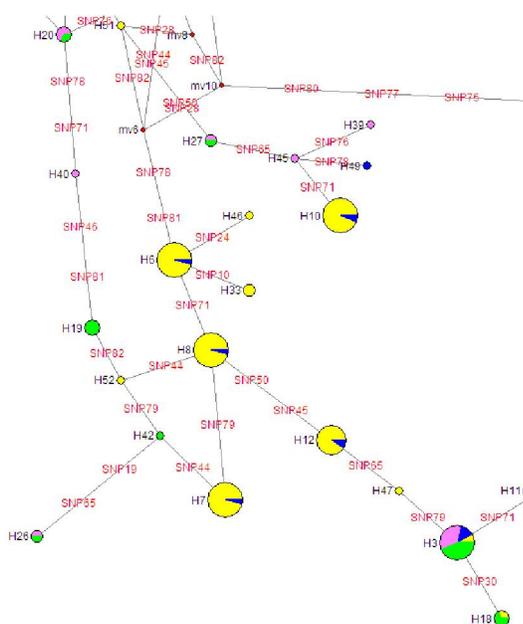


Figure 3. Median-joining local trees of 22q12.3 (13K) haplotypes. SNP names follow the supplementary table S2.

5. Conclusions

LD structure is important in evolutionary biology and human genetics. It provides information about past events and constrains the potential response to both natural and artificial selection (Slatkin, 2008). The most important feature is that it reduces the workload of genome-wide association studies (GWAS) and obviates the need to examine every mutation. Similarly, Chi and his partners used haplotypes instead of SNPs to measure breast cancer risk among African American women (Song et al., 2013). LD can also be used to improve genotype accuracy (1000 Genomes Project Consortium, 2012). In forensic DNA analysis, haplotype blocks are also proposed to transfer evidence, mixture, and kinship analyses (Ge et al., 2010). This paper focused on the 22q12.3 (SLDS) located in a non-coding region. It is still unknown whether the LD structure is involved in gene regulation. The related genome-wide association study requires further analysis to determine whether rs738968 associates with only secreted TNF- α or the whole 22q12.3 (SLDS). Although the 22q12.3 (SLDS) is observed in AFR, AMR and EUR population groups, it is still hard to explain why rs738968 shows significant association with secreted TNF- α in the Caucasian cohort but not in the African-American cohort. These questions direct our future research.

Acknowledgements:

This work was supported by the Innovation of the Training Mode of Applied Talents Foundation of Honghe University (MS1001).

Corresponding Author:

Jian-hong Sun
Engineering College of
Honghe University
Yunnan Mengzi, 661100, China
E-mail: sparkhonghe@foxmail.com

References

- Daly, M., Rioux, J. D., Schaffner, D. F., et al. High-resolution haplotype structure in the human genome. *Nature Genet.* 29, 229-232 (2001).
- Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 29:217-222.
- Patil N, Berno AJ, Hinds DA et al. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719-1723.
- Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B. & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576), 2225-2229.
- Anderson, E. C., & Slatkin, M. (2004). Population-genetic basis of haplotype blocks in the 5q31 region.

- The American Journal of Human Genetics, 74(1), 40-49.
6. Marx, V. (2013). Biology: The big challenges of big data. *Nature*, 498(7453), 255-260.
 7. 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), 56-65.
 8. Wall, J. D., & Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8), 587-597.
 9. Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps *Bioinformatics* 21: 263-265. Find this article online.
 10. Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., & Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *The American Journal of Human Genetics*, 71(5), 1227-1234.
 11. Librado, P., & Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11), 1451-1452.
 12. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585-595.
 13. Fu, Y. X., & Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133(3), 693-709.
 14. Hill, W. G., & Weir, B. S. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *American journal of human genetics*, 54(4), 705.
 15. Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, 49(1), 49.
 16. Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10), 5269-5273.
 17. Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2), 256-276.
 18. Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press.
 19. Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132(2), 583-589.
 20. Bafna, V., & Bansal, V. (2004). The number of recombination events in a sample history: conflict graph and lower bounds. *Computational Biology and Bioinformatics*, IEEE/ACM Transactions on, 1(2), 78-90.
 21. Hudson, R. R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetical research*, 50(03), 245-250.
 22. Salminen M, Carr J, Burke D, McCutchan F. (1996). Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. *AIDS Res Hum Retrovir* 11:1423-5.
 23. Martin, D., & Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics*, 16(6), 562-563.
 24. Worobey, M. (2001). A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. *Molecular Biology and Evolution*, 18(8), 1425-1434.
 25. Hudson, R. R., & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1), 147-164.
 26. Bandelt, H. J., Forster, P., Sykes, B. C., & Richards, M. B. (1995). Mitochondrial portraits of human populations using median networks. *Genetics*, 141(2), 743.
 27. Bandelt, H. J., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution*, 16(1), 37-48.
 28. Polzin, T., & Daneshmand, S. V. (2003). On Steiner trees and minimum spanning trees in hypergraphs. *Operations Research Letters*, 31(1), 12-20.
 29. Tajima, F., & Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. *Molecular biology and evolution*, 1(3), 269-285.
 30. Marotta, M., Chen, X., Inoshita, A., Stephens, R., Budd, G. T., Crowe, J. P., & Tanaka, H. (2012). A common copy-number breakpoint of ERBB2 amplification in breast cancer colocalizes with a complex block of segmental duplications. *Breast Cancer Res*, 14, R150.
 31. Hindorff LA, MacArthur J, Morales J, et al. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed May 31, 2014.
 32. Kennedy, R. B., Ovsyannikova, I. G., Pankratz, V. S., Haralambieva, I. H., Vierkant, R. A., & Poland, G. A. (2012). Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. *Human genetics*, 131(9), 1403-1421.
 33. Slatkin, M. (2008). Linkage disequilibrium-understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6), 477-485.
 34. Song C, Chen G K, Millikan R C, et al. (2013). A Genome-wide scan for breast cancer risk haplotypes among African American women. *PLoS one*, 8(2), e57298.
 35. Ge, J., Budowle, B., Planz, J. V., & Chakraborty, R. (2010). Haplotype block: a new type of forensic DNA markers. *International journal of legal medicine*, 124(5), 353-361.