Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Article

Genetic Flux Between *H1* and *H2* Haplotypes of the 17q21.31 Inversion in European Population

Libin Deng[1,2,3#], Xiaoli Tang[1#], Xiangwen Hao[2], Wei Chen[2], Jiari Lin[3], Yangyu Yu[3], Dake Zhang[2], and Changqing Zeng[2*]

[1]*Faculty of Basic Medical Science, Nanchang University, Nanchang 330006, China;*
[2]*Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China;*
[3]*Institute of Translational Medicine, Nanchang University, Nanchang 330006, China.*

Genomics Proteomics Bioinformatics 2011 Jun; 9(3): 113-118 DOI: 10.1016/S1672-0229(11)60014-4

Received: Dec 06, 2010 Accepted: May 16, 2011

Abstract

The chromosome 17q21.31 inversion is a 900-kb common structural polymorphism found primarily in European population. Although the genetic flux within inversion region was assumed to be considerable suppressed, it is still unclear about the details of genetic exchange between the *H1* (non-inverted sequence) and *H2* (inverted sequence) haplotypes of this inversion. Here we describe a refined map of genetic exchanges between pairs of gene arrangements within the 17q21.31 region. Using HapMap phase II data of 1,546 single nucleotide polymorphisms, we successfully deduced 96 *H1* and 24 *H2* haplotypes in European samples by neighbor-joining tree reconstruction. Furthermore, we identified 15 and 26 candidate tracts with reciprocal and non-reciprocal genetic exchanges, respectively. In all 15 regions harboring reciprocal exchange, haplotypes reconstructed by clone sequencing did not support these exchange events, suggesting that such signals of exchange between two sister chromosomes in certain heterozygous individual were caused by phasing error regions. On the other hand, the finished clone sequencing across 4 of 26 tracts with non-reciprocal genetic flux confirmed that this kind of genetic exchange was caused by gene conversion. In summary, as crossover between pairs of gene arrangements had been considerably suppressed, gene conversion might be the most important mechanism for genetic exchange at 17q21.31.

Key words: genetic flux, inversion, haplotype, polymorphism

Introduction

Genetic flux plays an important role at genome evolution. Crossover and gene conversion are important forms of genetic information exchange. Although recombination has been and is being studied across human genome at variable levels (high and low resolution), studies of gene conversion have only covered several specific regions (such as major histocompatibility complex regions and globin genes), or duplication genes distributed at different locus (such as human endogenous retroviral sequence in human Y chromosome) (*1-3*). It is less known about gene conversions across autosomes, since it is hard to distinguish gene conversion events from double crossovers on mammal meiotic products (*4*).

As one kind of chromosomal rearrangements, inversion is believed to be an important model to dis-

# Equal contribution.
*Corresponding author.
E-mail: czeng@big.ac.cn

cover gene conversions. Since gene flow among different oriented chromosome segments is inhibited, these segments would accumulate mutations independently and evolve towards different directions (*5*). According to previous nucleotide diversity researches in *Drosophila* genus, gene conversion rate would be uniform along inversion loop, whereas double-crossover rate would be higher in central part of inversion loop when compared with inversion breakpoints (*5*).

In previous research, human chromosome 17q21.31 region has been identified as a 900-kb inversion polymorphism by assembly of chromosome-specific BAC clones (*6*). Two orientations of this inverted segment represent distinct lineages, *H1* (non-inverted sequence) and *H2* (inverted sequence). The inverted configuration is linked to *H2*, which is relatively common in Europen samples but nearly absent in other HapMap populations. Recent studies have suggested no evidence of having recombined between two haplotype groups. Therefore, such inversion region provides a good sample to discover the gene conversions within autosomal regions.

In this study, we used phase II haplotype data across the human chromosome 17q21.31 inversion polymorphism region to detect conversion events in HapMap CEU samples. Based on the assumption that no recombination events had taken place between *H1* and *H2* haplotypes, 26 non-reciprocal gene conversions were detected. Meanwhile, several conversion events (reciprocal genetic flux) were false due to haplotype phasing error in HapMap dataset. Our study shows that haplotype-based methods can be applied to scan for genetic flux at inversion polymorphism regions.

Results

Haplotype substructures in 17q21.31 region

To determine the segmental orientations, we carried out cluster analyses of haplotypes by considering this 900-kb inversion segment as a specific polymorphism frame. In total, 120 haplotypes were constructed based on 1,546 single nucleotide polymorphisms (SNPs) from HapMap phase II data in all unrelated individuals. Further haplotype tree analysis showed that the CEU population was composed of two distinct subgroups (**Figure 1**). Using two SNPs, namely rs1800547 and rs9468, for orientation determination, the larger clade containing 96 haplotypes corresponded to the *H1* lineage as in the study of Stefansson *et al* (*6*), and the other clade containing 24 haplotypes corresponded to the *H2* lineage as inversion group (Table S1). These haplotypes were consistent with Stefansson's result by combining microsatellites and SNPs.

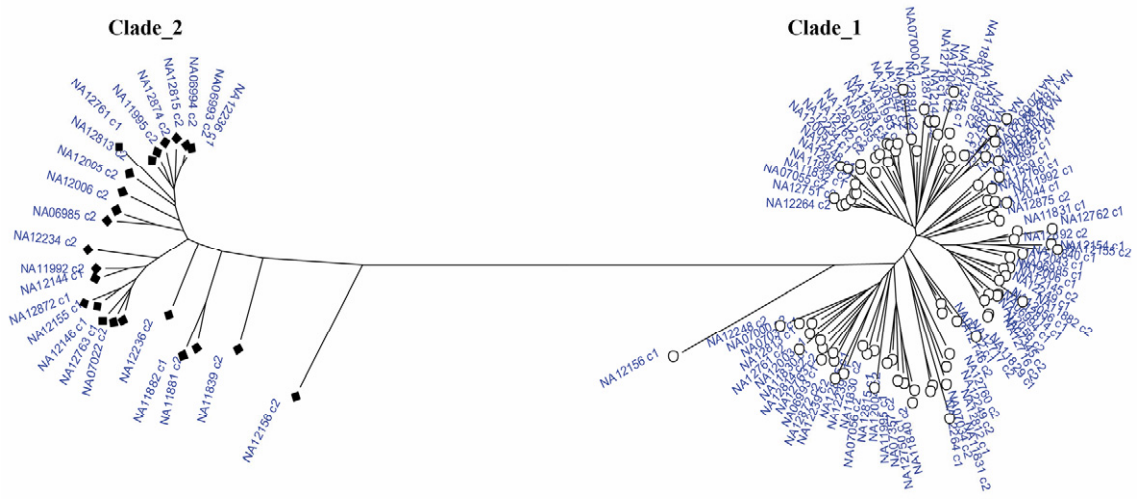

Figure 1 Neighbor-joining tree of long-range haplotypes from CEU population samples. With each branch representing a haplotype, clusters are constructed from data of 120 HapMap chromosomes at 17q21.31. In the two distinct clades, "○" represents *H1*, and "♦" represents *H2*.

Candidate conversion events within inversion region

Using phased haplotypes in *H1* and *H2* lineages, we predicted 41 candidate conversion regions (Table S2) with Betran's algorithm (*7*). The average length of them was 14,118 bp, longer than the length of other known conversion tracts. The longest one was 66,767 bp, containing 157 SNPs. Within these predicted conversions, 15 candidate regions had the corresponding conversion events observed in the same position of the sister chromosome from same individuals. We defined these paired predicted regions as reciprocal genetic exchange (**Table 1**). As shown in **Figure 2**, group 1 illustrated these reciprocal exchanges, and individuals with same candidate regions came from the same family, such as NA11881 and NA11882, NA12155 and NA12156. Moreover, non-reciprocal genetic exchanges could be observed in both *H1* and *H2* lineages (group 2 in Figure 2). In all, 26 regions of

Table 1 Predicted and observed haplotypes in three reciprocal conversion regions

Region	Sample	HapMap haplotype	Observed haplotype
41,163,838 - 41,182,076	NA11881:C1	T-A-**A-T...G-G-A**-G-C	T-A-***G-C...A-A-G***-G-C
	NA11881:C2	C-C-***G-C...A-A-G***-A-C	C-C-**A-T...G-G-A**-A-C
	NA11882:C1	C-C-**G-C...A-G-A**-G-T	C-C-***A-T...A-A-G***-G-C
	NA11882:C2	T-A-***A-T...A-A-G***-A-C	T-A-**G-C...G-G-A**-A-C
41,458,711 - 41,460,355	NA11881:C1	T-G-**C-C-C...C-C-A-A**-C-T	T-G-***A-T-T...T-T-C-G***-C-T
	NA11881:C2	C-C-***A-T-T...T-T-C-G***-T-G	C-C-C-**C-C...C-C-A-A**-T-G
	NA11882:C2	T-G-**C-C-C...T-T-C-G**-T-G	T-G-***A-T-T...C-C-A-?***-?-G
	NA11882:C1	C-C-***A-T-T...T-C-A-A***-C-T	C-C-**C-C-C...T-T-C-G**-C-T
41,578,112 - 41,644,878	NA12156:C1	C-T-A-**C-G-G**-A-C-G	C-T-A-***T-A-A***-A-C-G
	NA12156:C2	T-C-G-***T-A-A***-C-T-T	T-C-G-**C-G-G**-C-T-T

Note: C1, one sister chromosome; C2, the other sister chromosome.

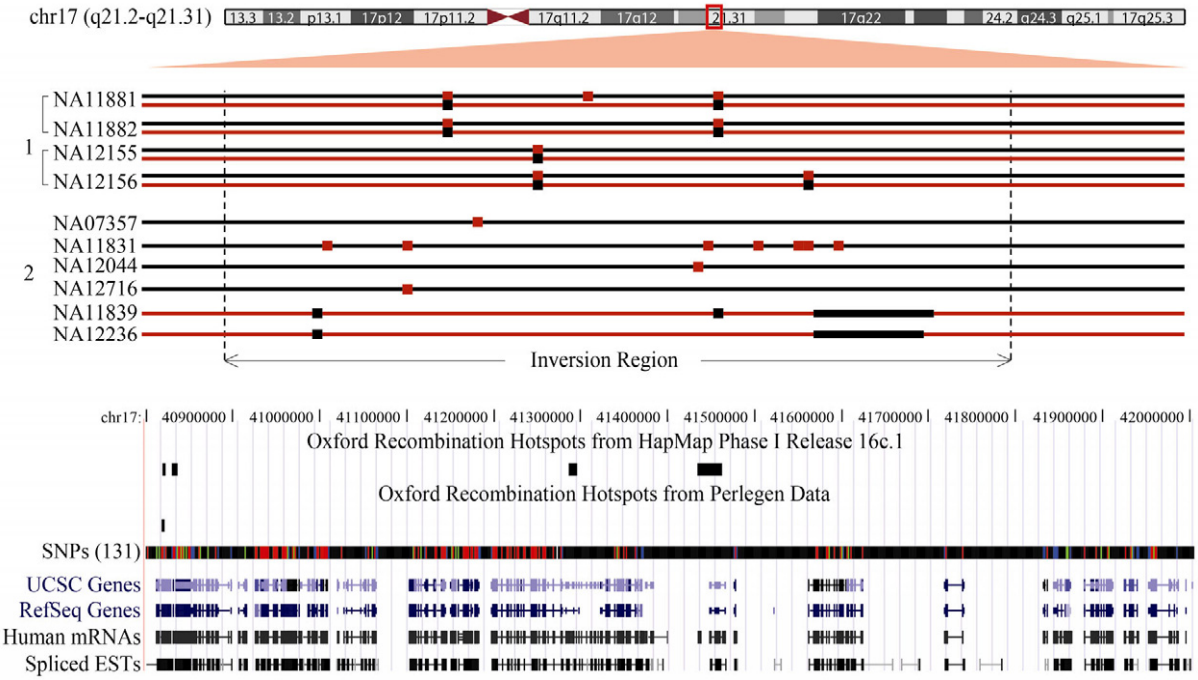

Figure 2 Distribution of predicted candidate conversion regions. In the upper panel, the left column shows the individuals. Black lines represent *H1* and red lines represent *H2*; red bars in *H1* and black bars in *H2* represent predicted candidate conversion regions. Two dash lines demonstrated the inversion region. The bottom panel shows the recombination hotspots within this inversion region (retrieved from the HapMap official website at http://www.hapmap.org/).

this kind were identified to have diverse length. These regions may be introduced from other population by gene conversion.

For the 15 reciprocal genetic exchanges, candidate regions were observed in samples from the same family, implying that this kind of conversion may be the result of phasing error. To test this possibility, we obtained the real haplotypes of these candidate regions by clone sequencing. As shown in Table 1, comparison between the HapMap-predicted haplotypes and the observed ones demonstrated the existence of phasing errors. Factors contributed to phasing errors include conversion events occurred in parent generation and inaccuracy in algorithms adopted by software for phasing. We further sequenced candidate regions in NA10859 and in her parents (NA11881 and NA11882), and found that haplotypes were identical in both generations. Therefore, these errors were most likely due to algorithms in haplotype deduction.

Verification of conversion events

To validate the conversion event of the 26 non-reciprocal genetic exchanges, we selected four regions for sequencing verification (**Table 2**). Most of the candidate conversion regions were confirmed by clone sequencing, and small conversion regions seemed to be easier for verification. For instance, a 15-bp predicted conversion region, chr17: 41,111,640-41,111,654, was successfully validated in both NA11831 and NA12716. Large candidate regions may contain more SNP phasing errors, and haplotype deduction may be inaccurate. The candidate region chr17: 41581663-41634438 was longer than all known conversion tracts. It was predicted in both NA12236 and NA11839 samples, but was only confirmed in the latter one. The longest candidate region was predicted in NA12156, spanning 66,767 bp. In our attempt for its haplotype construction, we only obtained one haplotype at each breakpoint using clone sequencing. In addition, we found that both upstream and downstream haplotypes had phasing error SNPs, precluding us to identify this candidate as a conversion region.

Discussion

In 17q21.31 inversion region, haplotype subgroups corresponded to the two alleles with opposite segmental orientations, *H1* and *H2*, respectively (*6*). Considering the high genetic divergence between these two arrangements (F_{ST}=0.81), the crossover between two oriented chromosome segments has been significantly inhibited. In this study, we used phased haplotype data from CEU samples to detect gene conversion within 17q21.31 inversion region. After excluding the phasing error from HapMap haplotype construction, we identified 26 confident regions subjected to historic gene conversions. As conversion event can weak the strength of linkage disequilibrium (LD) (*4*), it should be concerned at searching related disease genes within this region.

The goal of International HapMap Project was to

Table 2 Verification of four "interspersed" conversion regions

Region	41,012,678-41,015,758		41,111,640-41,111,654		41,195,723-41,195,778		41,581,663-41,634,438	
	Sample	Haplotype	Sample	Haplotype	Sample	Haplotype	Sample	Haplotype
	NA07000:C1	**C-G-T-C-T**	**NA07000:C1**	**G-G-A-C**	**NA07000:C1**	**A-T-G-C**	**NA07000:C1**	**G-G-A-T…T-A-G-A**
	NA07000:C2	**C-G-T-C-T**	**NA07000:C2**	**G-G-A-C**	**NA07000:C2**	**A-T-G-C**	**NA07000:C2**	**G-G-A-T…T-A-G-A**
	NA11839:C1	**C-G-T-T-T**	**NA11831:C1**	**G-G-A-C**	**NA07357:C1**	**T-G-G-C**	**NA07055:C1**	**G-G-A-T…T-A-G-A**
	NA11839:C2	C-G-T-C-T	**NA11831:C2**	**A-G-C-A**	**NA07357:C2**	**A-T-G-C**	**NA07055:C2**	**G-G-A-T…T-A-G-A**
	NA12236:C1	C-A-T-C-C	**NA12716:C1**	**A-G-C-A**	NA12236:C1	T-G-A-A	**NA12236:C1**	**A-A-T-C…C-G-A-G**
	NA12236:C2	C-A-T-C-C	**NA12716:C2**	**G-G-A-C**	NA12236:C2	T-G-A-A	**NA12236:C2**	**G-A-T-C…C-G-G-G**
			NA12236:C1	G-G-C-A				
			NA12236:C2	G-G-C-A				

Note: Bold, H1 haplotype; Unbolad, H2 haplotype; Underlined, conversion region; C1, one sister chromosome; C2, the other sister chromosome.

help researchers to mapping genes related to diseases using LD among markers. Although Bayesian statistical method was more accurate than other algorithms, such as expectation-maximization algorithm (*8*), the genotyping error and structure variation could still affect the data qualification of HapMap. It reminds that researchers should sequence the target regions and determine the correct associated markers. For most applications, although it is easy to estimate the switch error rate of phasing based on computer simulation, the identification of phasing errors in particular loci is inefficient. In this study, we did verify several haplotype phasing error sites compared to real autosomal haplotypes from clone sequencing. However, the predicted method based on detection of "reciprocal genetic exchange" cannot apply to autosomal region without chromosome rearrangement.

Conclusion

Here we described a refined map of genetic exchanges between pairs of gene arrangements within the 17q21.31 region. Using HapMap phase II data of 1,546 SNPs, we successfully deduced 96 *H1* and 24 *H2* haplotypes in European samples by neighbor-joining (NJ) tree reconstruction. In particular, the large genetic differentiation between these two haplotype clades might be caused by the suppression of gene flux between these two arrangements. Furthermore, we applied methods of Betran *et al* (*7*) to scan for genetic flux between *H1* and *H2* in the 120 CEU haplotypes, and identified 15 and 26 candidate tracts with reciprocal and non-reciprocal genetic exchange, respectively. In all 15 regions harboring reciprocal exchange, haplotypes reconstructed by clone sequencing did not support these exchange event, suggesting that such signals of exchange between two sister chromosomes in certain heterozygous individual were caused by phasing error regions. On the other hand, the finished clone sequencing across 4 of 26 tracts with non-reciprocal genetic flux confirmed that this kind of genetic exchange was caused by gene conversion. In summary, as crossover between pairs of gene arrangements had been considerably suppressed, gene conversion might be the most important mechanism for genetic exchange at 17q21.31.

Material and Methods

Haplotype analysis and conversion prediction

Phasing haplotype data of 60 unrelated CEU samples (Utah residents with ancestry from northern and western Europe) were downloaded from the HapMap website (HapMap Phase II/rel#21a data files; http://www.hapmap.org/) (*9*). In total, 1,546 autosomal markers across chr17q21.31 (41.02-41.98 Mb), which were segregating in CEU HapMap samples, were applied for analysis.

To infer the evolutionary relationships of haplotypes, NJ tree algorithm was carried out for haplotype clustering (*10*). Haplotype genetic distances were estimated by the allele sharing distance method (*11*). The haplotype tree was then constructed by means of the NJ algorithm implemented in the MEGA3 (*12*). Next, we used Betran's method (*7*) to detect gene conversion tracts from haplotype data between arrangements (or haplotype subgroups). The algorithm was provided by the DnaSP software (version 4.10.9) (*13*). Based on the parameter φ (Equation A4 in Betran *et al*) (*7*), which measures the probability of detecting a converted site, we estimated the number and the length distribution of conversion tracts across 17q21.31 region.

Clone sequencing and conversion verification

DNA samples of the International HapMap Project were obtained from Coriell Repositories (Camden, NJ, USA). To verify the candidate conversions, seven regions within inversion region were selected for clone sequencing (Tables 1 and 2). All the PCR reactions were performed with TaKaRa LA Taq kit. PCR products were electrophoresed on gel of 1% agrose and were retrieved using AxyPrep DNA Gel Extraction kit (AxyGen Company) according to the recommendations of the manufacturer. In order to get the single haplotype, we cloned the products with pGEM-T vector (pGEM-T vector system, Promega) under the instructions of manufacturer. For each sample, we selected five clones.

All the clones were cultured at 37°C for 12 h, and plasmids were extracted using the AxyPrep Plasmid

kit according to the manufacturer instruction. DNA sequencing was performed on ABI 3730 DNA sequencer system using Big Dye chemistry. SNPs were found by aligning with reference sequence according to assembly 37 using Lasergene software.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 30871348, 30700470, 30890030 and 30890031) and Educational Department of Jiangxi Province (No. GJJ10303).

Authors' contributions

LD, XT and XH performed major data analyses and drafted the manuscript. XH, WC, JL and YY collected the dataset and performed clone sequencing. LD, DZ and CZ designed the study, supervised the project and co-wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Papadakis, M.N. and Patrinos, G.P. 1999. Contribution of gene conversion in the evolution of the human beta-like globin gene family. *Hum. Genet.* 104: 117-125.
2 Rozen, S., *et al.* 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423: 873-876.
3 Zangenberg, G., *et al.* 1995. New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. *Nat. Genet.* 10: 407-414.
4 Jeffreys, A.J. and May, C.A. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.* 36: 151-156.
5 Rozas, J., *et al.* 1999. Molecular population genetics of the rp49 gene region in different chromosomal inversions of *Drosophila subobscura*. *Genetics* 151: 189-202.
6 Stefansson, H., *et al.* 2005. A common inversion under selection in Europeans. *Nat. Genet.* 37: 129-137.
7 Betran, E., *et al.* 1997. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. *Genetics* 146: 89-99.
8 Hellenthal, G. and Stephens, M. 2006. Insights into recombination from population genetic variation. *Curr. Opin. Genet. Dev.* 16: 565-572.
9 Frazer, K.A., *et al.* 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.
10 Desper, R. and Gascuel, O. 2006. Getting a tree fast: Neighbor Joining, FastME, and distance-based methods. *Curr. Protoc. Bioinformatics* Chapter 6: Unit 6.3.
11 Gao, X. and Martin, E.R. 2009. Using allele sharing distance for detecting human population stratification. *Hum. Hered.* 68: 182-191.
12 Kumar, S., *et al.* 2004. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* 5: 150-163.
13 Rozas, J., *et al.* 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496-2497.

Supplementary Material

Tables S1 and S2

DOI: 10.1016/S1672-0229(11)60014-4