European Journal of Human Genetics (2016) 24, 139–141

www.nature.com/ejhg

SHORT REPORT

Rare missense variants within a single gene form yin yang haplotypes

David Curtis*

Yin yang haplotype pairs differ at every SNP. They would not be accounted for by population models that incorporate sequential mutation, with or without recombination. Previous reports have claimed that there is a tendency for common SNPs to form yin yang haplotypes more often than would be expected by sequential mutation or by a random sample of all possible haplotypic arrangements of alleles. In the course of analysing next-generation sequencing data, instances of yin yang haplotypes being formed by very rare variants within a single gene were observed. As an example, this report describes a completely yin yang haplotype formed by eight rare missense variants in the *ABCA13* gene. Of 1000 genome subjects, 21 have a copy of the alternate allele at all eight of these positions and a single subject is homozygous for all of them. None of the other 1070 subjects possesses any of the altetrnates. Thus, the eight alternate alleles are always found together and never occur separately. The existence of such yin yang haplotypes has important implications for statistical methods for analysing rare variants. Also, they may be of use for gaining a better understanding of the history of human populations.
European Journal of Human Genetics (2016) **24,** 139–141; doi:10.1038/ejhg.2015.74; published online 22 April 2015

INTRODUCTION

Yin yang haplotypes, in which there is a partial or complete tendency for alleles to be different at every SNP, were first described using markers with a high minor allele frequency (MAF). Using a liberal definition based on haplotypes derived from small numbers of consecutive SNPs it was claimed that such haplotypes were consistent with a neutral evolutionary model incorporating mutation and recombination.[1] However, subsequent studies based on longer haplotypes derived from regions of homozygosity using SNPs with MAF >0.05 concluded that common haplotypes were in fact more divergent from each other than would be expected from a random sample of all possible haplotypes.[2,3] As haplotypes formed by processes of mutation and recombination are expected to be more similar to each other than would be a random sample, the observation of these highly divergent haplotypes did not seem to be consistent with such a model. Examples were given of 9-locus haplotypes where the 10 commonest haplotypes in the population included pairs that were completely or almost completely discordant at every SNP. These 9-SNP haplotypes in some cases extended over large genetic distances, the most extreme example being over 14 Mb on the X chromosome, between rs6638361 and rs241393. The mechanism whereby such haplotypes had arisen was unclear but it was proposed that they might represent ancient founder effects and that they might be useful for elucidating population history.

In contrast to these previously described haplotypes of common variants stretching over large genetic distances, the present report describes a similar although more extreme phenomenon occurring with rare coding variants within a single gene. Multiple examples of this phenomenon were observed in the process of developing a novel statistical method that aimed to detect recessive effects by identifying an excess of compound heterozygotes.[4] To do this, the method identifies subjects carrying two or more missense variants within the same gene to see whether there is an excess of such subjects among cases. It was noted that sometimes there was a tendency for variants at two or more different positions to tend to occur together in the same subjects. While this might possibly reflect the existence of compound heterozygotes, a likely alternative explanation would be that the variants were in linkage disequilibrium with each other. Obviously, this could occur if a mutation forming the second variant occurred in a haplotype bearing the first. As one would expect little recombination within a gene, such pairs might well then tend to be preserved in the population. What was more surprising was the observation that there sometimes seemed to be a complete co-occurrence of the alternate alleles, such that each was always found with the other and none was observed on its own. A particularly striking example of this was observed in the *ABCA13* gene.

MATERIALS AND METHODS

The general method of analysis, which seeks to identify genes exerting recessive effects, is described elsewhere and is implemented as an option in the SCOREASSOC program.[4] The method is designed to identify subjects in whom variants may occur together either as homozygotes or as compound heterozygotes. Thus, it begins by identifying all subjects who carry two or more missense, frameshift, splice site and nonsense variants having MAF <0.1. It was applied to the 1000-genome data.[5] It quickly became apparent that many variants within genes were in strong LD with each other, in that one variant would tend to occur in the presence of another. More strikingly, it was observed that some variants would form yin yang haplotypes, meaning that pairs of variants occurred together exclusively, that is, either both would be present or neither. Although the phenomenon was observed to a greater or lesser extent in many different genes, for purposes of illustration results are presented here only for the *ABCA13* gene.

RESULTS

It was noted that the following missense variants within ABCA13 always occurred together: chr7.hg19.g 48312490G>A, chr7.hg19.g

UCL Genetics Institute, UCL, Molecular Psychiatry Laboratory, London, UK
*Correspondence: Professor D Curtis, UCL Genetics Institute, UCL, Molecular Psychiatry Laboratory, Rockefeller Building, 21 University Street, London WC1E 6JJ, UK. Tel: +44 20 7791 5207; Fax: +44 20 7791 5201; E-mail: d.curtis@ucl.ac.uk
Received 19 August 2014; revised 19 February 2015; accepted 17 March 2015; published online 22 April 2015

Table 1 List of subjects carrying the haplotype consisting of eight alternate missense alleles of ABCA13NA18510 is homozygous for the haplotype

Sample	*Family ID*	*Population*	*Population description*
HG00126	HG00126	GBR	British in England and Scotland
HG00233	HG00233	GBR	British in England and Scotland
HG00243	HG00243	GBR	British in England and Scotland
HG00339	HG00339	FIN	Finnish in Finland
HG01488	CLM62	CLM	Colombian in Medellin, Colombia
NA18510	Y010a	YRI	Yoruba in Ibadan, Nigeria
NA18868	Y007	YRI	Yoruba in Ibadan, Nigeria
NA18912	Y028	YRI	Yoruba in Ibadan, Nigeria
NA18933	Y036	YRI	Yoruba in Ibadan, Nigeria
NA19099	Y105	YRI	Yoruba in Ibadan, Nigeria
NA19102	Y042	YRI	Yoruba in Ibadan, Nigeria
NA19107	Y006	YRI	Yoruba in Ibadan, Nigeria
NA19256	Y092	YRI	Yoruba in Ibadan, Nigeria
NA19313	LWK001	LWK	Luhya in Webuye, Kenya
NA19377	NA19377	LWK	Luhya in Webuye, Kenya
NA19394	NA19394	LWK	Luhya in Webuye, Kenya
NA19700	2367	ASW	African Ancestry in Southwest US
NA19701	2367	ASW	African Ancestry in Southwest US
NA19909	2427	ASW	African Ancestry in Southwest US
NA19923	2434	ASW	African Ancestry in Southwest US
NA20339	2486	ASW	African Ancestry in Southwest US
NA20799	NA20799	TSI	Toscani in Italy

48313256A>G, chr7.hg19.g 48313563A>G, chr7.hg19.g chr7.hg19.g 48315724C>T, chr7.hg19.g 48315898T>C, chr7.g19.g 48317836G>A, chr7.hg19.g 48318098A>G, chr7.hg19.g 48318400T>G. One subject was homozygous for all the alternate alleles, whereas 21 subjects were heterozygous at all positions. None of the other 1070 subjects carried an alternate allele at any of these positions. Other rare variants appeared to be in LD with this haplotype to a greater or lesser extent. The haplotype always occurred on the background of alternate alleles at chr7.hg19.g 48311602G>A and chr7.hg19.g 48314929T>A. However, four additional subjects were also heterozygous at chr7.hg19.g 48311602G>A and another four subjects were heterozygous at chr7.hg19.g 48314929T>A. Thus, the haplotype would have a D′ of 1 with each of these loci, but they did not exhibit their full yin yang effect because they also sometimes occurred in isolation. The subjects carrying the eight variant alleles are listed in Table 1. It can be seen that they have a wide geographical distribution and that, with the exception of NA19700 and NA19701, they are not known to be related to each other.

DISCUSSION

This brief report does not seek to present a systematic survey of the extent to which yin yang haplotypes occur between rare variants. Such a study would require rigorous definitions and also the ability to derive haplotypes reliably from sequence data. Nor does it attempt to assign an exact *P* value to the observation reported. The aim is simply to draw attention to the existence of this phenomenon and its implications. The yin yang haplotype reported here has also been observed in exomes sequenced for large case–control association studies currently submitted for publication. It appears to have a similar frequency in all samples studied and there is no suggestion that it has any functional consequence or that the fact that it occurs in ABCA13 is of any special relevance.

It is important to be aware that yin yang haplotypes can occur even among rare missense variants. Any form of LD between variants would invalidate a simple burden test, which assumed that variants occurred independently. As discussed elsewhere, such LD will also invalidate many tests for association that use Monte Carlo methods to permute genotypes rather than haplotypes.[6] Likewise, if one tests for an excess of subjects carrying two or more variant alleles, in an attempt to detect compound heterozygotes, it is difficult to distinguish compound heterozygotes from alleles co-occurring through LD. Under a simple coalescent model, one might not expect two very rare variants to occur in a single haplotype and if one observed a single affected subject with two novel variants in the same gene one might suspect that this represented a compound heterozygote disrupting both copies of a gene. However, the existence of yin yang haplotypes implies that one would need to consider the possibility that both variants lay within the same copy of the gene. Another implication is that if certain pairs of variants always occur together then it would never be possible to use population association methods to decide which was causative – one would need to rely on functional studies of artificial constructs in which only one variant was present.

The mechanism whereby such yin yang haplotypes occur is difficult to discern. It is hard to think of plausible biological mechanisms whereby such a rare haplotype might be distributed so widely without any of the constituent variants surviving in isolation. One would need to speculate that either all had arisen simultaneously or they were produced by some form of admixture or bottleneck. However, it is still difficult to visualise exactly what set of circumstances would produce this result. Alternatively, they might arise from some kind of chromosomal rearrangement or sequencing artefact, although again it is not exactly clear what the nature of this might be. It seems that it would be fruitful to study such

phenomena further in order to gain a clearer understanding of the mechanisms involved and implications for analytic methodologies.

CONFLICT OF INTEREST

The author declares no conflict of interest.

1 Zhang J, Rowe WL, Clark AG, Buetow KH: Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. *Am J Hum Genet* 2003; **73**: 1073–1081.

2 Curtis D, Vine AE, Knight J: Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Ann Hum Genet* 2008; **72**: 261–278.

3 Curtis D, Vine AE: Yin yang haplotypes revisited - long, disparate haplotypes observed in European populations in regions of increased homozygosity. *Hum Hered* 2010; **69**: 184–192.

4 Curtis D: Approaches to the detection of recessive effects using next generation sequencing data from outbred populations. *Adv Appl Bioinform Chem* 2013; **6**: 29–35.

5 Abecasis GR, Auton A, Brooks LD *et al*: An integrated map of genetic variation from 1092 human genomes. *Nature* 2012; **491**: 56–65.

6 He Z, O'Roak BJ, Smith JD *et al*: Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am J Hum Genet* 2014; **94**: 33–46.