The Discovery of a Pseudo SNP and its Usage in Gender Test

Jie Huang, MD, MPH

Abstract

Single Nucleotide Polymorphism (SNP) technology has been widely used in genome-wide association studies (GWAS). Quality control procedures usually includes testing minor allele frequency, hardy–weinberg equilibrium, call frequency, and concordance rate. However, little effort has been reported to validate a SNP itself. Here we report how a SNP is found not to be actual polymorphic at one locus, instead from two sequences on two different chromosomes. *[N A J Med Sci. 2009;2(2):41.]*

Introduction

The dbSNP database shows the SNP rs12743401 only has two genotypes - C/T and T/T, while the other combination C/C does not exist in all four HapMap populations. This irrational distribution leads to a search for the explanation of how this SNP was detected by genotyping technology.

Methods

Sequence match across genome

The FASTA sequence surrounding this SNP was obtained from RefSNP Cluster Report.[1] Subsequently, this FASTA sequence was fed into BLAST (Basic Local Alignment Search Tool) for a genome-wide nucleotide sequence comparison.[2]

Distribution of genotype across gender

About 3,000 samples of publicly available GWAS data was downloaded from illumina iControlDB.[3] The genotype of SNP rs12743401 was tabulated with gender to identify potential sex chromosomes related bias.

Results

Figure 1 shows that the genotype of this SNP for about 3,000 samples of GWAS data revealed that all females have the genotype of T/T while all the males have the genotype of C/T, with exception of 1 sample.

Figure 2 shows actually there are two hits for the FASTA sequence across the whole genome, one on chromosome 1 and another one on chromosome Y.

Based on the above two findings, it is not difficult to conclude that this SNP rs12743401 is not really polymorphic. Instead, heterozygosity in male is actually a detection of two regions, one on chromosome 1 and the other on chromosome Y. This explains well the existence of only two genotypes for this SNP. Interestingly, this non-polymorphic pseudo SNP turns out to be a great marker for gender test.

Discussion

While we found a genetic marker that is suitable for gender verification, this "SNP" can generates spurious association when gender is not well matched in a case-control GWAS study. Therefore, it is important for thorough data quality control. This discovery also invites further investigation on the sequence similarity between autosomal and sex chromosomes.

Jie Huang, MD, MPH
Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114

References

1. Illumina iControlDB. http://www.illumina.com/. 1/10/2009.
2. RefSNP. http://www.ncbi.nlm.nih.gov/SNP. 1/10/2009.
3. BLAST. *http://blast.ncbi.nlm.nih.gov/Blast.cgi*. 1/10/2009.

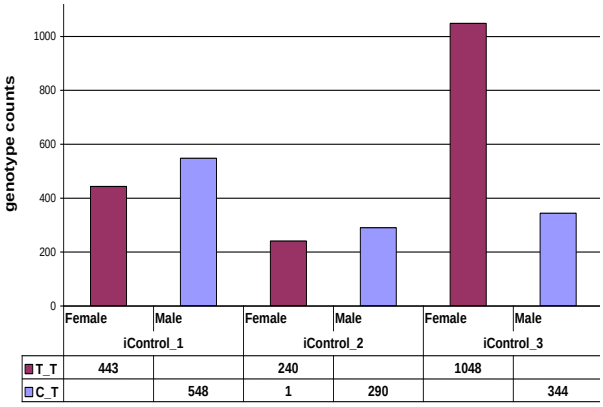

Figure 1. Tabulation of alleles of rs12743401 by gender in three.

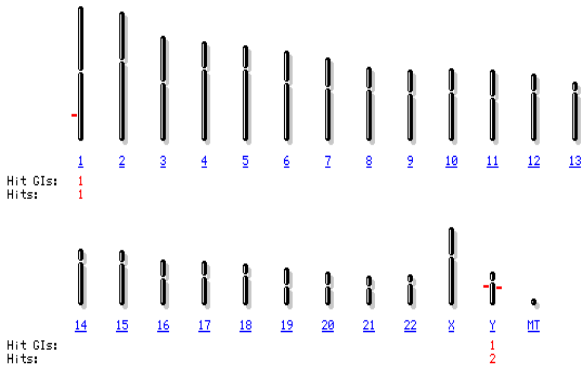

Figure 2. BLAST of SNP rs12743401 shows 2 matches.