The Veterinary Journal 202 (2014) 543–549

Contents lists available at ScienceDirect

The Veterinary Journal

journal homepage: www.elsevier.com/locate/tvjl

Risk of false positive genetic associations in complex traits with underlying population structure: A case study

Carrie J. Finno [a,*], Monica Aleman [b], Robert J. Higgins [c], John E. Madigan [b], Danika L. Bannasch [a]

[a] Department of Population Health and Reproduction, University of California, Davis, CA 95616, USA
[b] Department of Medicine and Epidemiology, University of California, Davis, CA 95616, USA
[c] Department of Pathology, Microbiology and Immunology, University of California, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Accepted 13 September 2014

Keywords:
Equine degenerative myeloencephalopathy
Horse genome
Single nucleotide polymorphisms
Vitamin E

ABSTRACT

Genome-wide association (GWA) studies are widely used to investigate the genetic etiology of diseases in domestic animals. In the horse, GWA studies using 40–50,000 single nucleotide polymorphisms (SNPs) in sample sizes of 30–40 individuals, consisting of only 6–14 affected horses, have led to the discovery of genetic mutations for simple monogenic traits. Equine neuroaxonal dystrophy is a common inherited neurological disorder characterized by symmetric ataxia. A case-control GWA study was performed using genotypes from 42,819 SNP marker loci distributed across the genome in 99 clinically phenotyped Quarter horses (37 affected, 62 unaffected).

A significant GWA was not achieved although a suggestive association was uncovered when only the most stringently phenotyped NAD-affected horses ($n = 10$) were included (chromosome 8:62130605 and 62134644 [$\log(1/P) = 5.56$]). Candidate genes (*PIK3C3*, *RIT2*, and *SYT4*) within the associated region were excluded through sequencing, association testing of uncovered variants and quantitative RT-PCR. It was concluded that variants in *PIK3C3*, *RIT2*, and *SYT4* are not responsible for equine neuroaxonal dystrophy. This study demonstrates the risk of false positive associations when performing GWA studies on complex traits and underlying population structure when using 40–50,000 SNP markers and small sample size.

Published by Elsevier Ltd.

Introduction

Over the past 25 years, two approaches have been applied to uncover genes contributing to specific diseases in domestic animals. The first approach targets candidate genes selected based on their role in comparative diseases in other species, while the second involves mapping disease traits of interest to a chromosomal location using genetic markers (Andersson and Georges, 2004). The recent availability of large panels of single nucleotide polymorphisms (SNPs) in domestic animal species has led to an expansion in the search for genomic regions associated with genetic diseases through the use of genome-wide association (GWA) studies.

In the horse, the first-generation SNP array (Illumina EquineSNP50 Beadchip) contains 54,602 SNPs (McCue et al., 2012). This array was used to identify chromosomal regions, leading to mutation discovery, for various monogenic traits in the horse using relatively small sample sizes (Brooks et al., 2010; Fox-Clipsham et al., 2011; Andersson et al., 2012). In addition to simple monogenic traits, GWA studies using the equine SNP chips have identified quantitative trait loci for more complex traits (Dupuis et al., 2011; Corbin et al., 2012; Kulbrock et al., 2013). Despite the apparent success in mapping more complex traits and diseases, underlying genetic mutations have not been uncovered for many of these traits. The purpose of the present study was to demonstrate the risks of false positive associations when performing a GWA study using small sample sizes in populations with complex traits and underlying population structure.

Equine neuroaxonal dystrophy (NAD) is characterized by a symmetric ataxia and proprioceptive deficits, developing between 6 and 12 months of age with no sex predilection (Beech and Haskins, 1987; Aleman et al., 2011; Finno et al., 2013). Equine degenerative myeloencephalopathy (EDM) is considered a more pathologically advanced form of NAD, in which evidence of histological damage extends to the white matter of the spinal cord. Because we have previously established that cases of NAD and EDM can occur within families with the same underlying environmental risk factors and that EDM appears to be a pathologically more extensive form of NAD (Aleman et al., 2011), the disease has subsequently been termed NAD/EDM.

Currently there is no means to establish an ante-mortem diagnosis of NAD/EDM. Vitamin E plays an important role in the development of NAD/EDM in genetically predisposed foals (Blythe et al., 1991) but low serum vitamin E is not consistently reported

* Corresponding author. Tel.: +1 530 752 2739.
E-mail address: cjfinno@ucdavis.edu (C.J. Finno).

http://dx.doi.org/10.1016/j.tvjl.2014.09.013
1090-0233/Published by Elsevier Ltd.

in all cases (Dill et al., 1990). There is very strong evidence that, in susceptible families, dietary vitamin E in the susceptible foal modifies the severity of the phenotype (Aleman et al., 2011). Definitive diagnosis, for the purposes of a genetic investigation, requires identification of characteristic lesions in the brainstem and spinal cord at post-mortem.

We have previously demonstrated a complex mode of inheritance for NAD/EDM and excluded putative variants in a strong candidate gene, α-tocopherol transfer protein (*TTPA*), as causative for NAD/EDM (Finno et al., 2013). Based on comparative knowledge of the clinical and histological presentation of NAD/EDM, there are no additional candidate genes to evaluate and either a linkage or association study is required to further investigate the etiology. Once a candidate region is discovered, prioritized genes for further evaluation will include those genes involved in synaptic function because studies have demonstrated accumulation of synaptic proteins in EDM-affected horses (Siso et al., 2003).

A GWA study was performed using two populations: (1) 37 clinically affected NAD/EDM Quarter horses (QHs) and 62 clinically unaffected QHs, and (2) a subset of the previous population that included only the 10 affected NAD/EDM QHs in which the disease was confirmed at post-mortem plus the same 62 clinically phenotyped unaffected QHs. Candidate genes were prioritized within the region(s) of significant association for further investigation. We hypothesized that, although NAD/EDM appears to be a complex disease trait, a major gene involved in synaptic transmission would be identified through the GWA study and a variant within that gene would be significantly associated with NAD/EDM. This study highlights the importance of adequate sample size and evaluation of underlying population structure when performing GWA studies to identify true regions of association using low coverage SNP arrays in complex traits.

Materials and methods

Horses

DNA was collected from 99 clinically phenotyped QHs, including 37 clinically NAD/EDM-affected and 62 unaffected QHs. Horses were defined as clinically affected with a mean ataxia score ≥2 and unaffected with a mean ataxia score of 0 as previously described (Aleman et al., 2011). Ten of the NAD/EDM-affected and four unaffected horses were confirmed at post-mortem examination, with histological findings as previously described (Aleman et al., 2011).

Of the population used for the GWA study, 33/37 affected and 53/62 control horses overlapped with horses used in a previous genetic study (Finno et al., 2013). Unaffected horses were ≥1 year of age and 1-year-old horses were re-examined at 3 years of age to confirm an unaffected phenotype, because cases of NAD/EDM have been reported up to 3 years of age (Beech and Haskins, 1987). Horses under 1 year of age were not included in this group because clinical signs of NAD/EDM may not be apparent during the first year of life. Of the post-mortem confirmed affected cases, five were by the same sire and five were unrelated within three generations.

Unaffected horses were sampled from the same farm as affected cases and 47/62 of these were unrelated to NAD/EDM-affected horses within three generations. Fifteen unaffected horses were related to NAD/EDM-affected horses within two generations but were raised under identical environmental conditions as NAD/EDM-affected horses and demonstrated no evidence of neurological disease. Therefore, these horses served as ideal controls to balance relatedness between NAD/EDM-affected and unaffected cases.

All protocols were approved by the University of California, Davis Institutional Animal Care and Use Committee (Protocol 15963).

Genome wide association (GWA)

Horses were genotyped for 54,602 SNPs using the Equine SNP50 genotyping array (Illumina). SNPs were selected that passed quality control settings (minor allele frequency >1% and genotyping across individuals >90%). A case/control standard allelic GWA study was performed and population stratification assessed by estimating the genomic control inflation factor (λ) using GenABEL (Aulchenko et al., 2007b), implemented in the R program (R Development Core Team, 2013). When $\lambda = 1$, there is no population stratification and association results should not be influenced by population structure (Wu et al., 2011).

To account for the population substructure and relatedness in this population of horses, a linear mixed model was implemented, using two distinct algorithms. The first program utilized an approximation method to perform the linear mixed model. Genome-wide rapid association using mixed model and regression (GRAMMAR), implemented in the R package GenABEL, first estimates the residuals from the linear mixed model under the null model and then treats these residuals as phenotypes for further genome-wide analysis by a standard linear model (Aulchenko et al., 2007a). The second program used an exact method to perform the linear mixed model, thereby avoiding repeatedly estimating variance components when performing each test, and was implemented through GEMMA (Zhou and Stephens, 2012). Both algorithms perform linear mixed models based on clustering that accounts for both population substructure and relatedness through use of a kinship matrix estimated from identical by descent distances.

Significance thresholds

A Bonferroni correction for 42,819 tests (corresponding to the number of usable SNPs), defined by significant $P_{\text{genome-wide}} = 0.05$, was determined, yielding a threshold of 1.17×10^{-6} (significant, $\log [1/P] = 5.93$). As Bonferroni corrections to control type 1 error in genetic association studies are highly conservative and have been shown to 'overcorrect' SNPs that are not truly independent, it has been recommended to apply both significant and suggestive *P*-value thresholds to properly control for type I error (Duggal et al., 2008). There is a wide lack of consistency in suggestive *P*-values applied in GWA studies using 30–50,000 SNP markers, both within and across species (Dupuis et al., 2011; Do et al., 2014; Zhang et al., 2014). False discovery rate (FDR) has lower incidence of type II error (Verhoeven et al., 2005) and setting the FDR at 0.10 and 0.05 has been recommended as criteria for suggestive and significant linkage, respectively (Benjamini and Yekutieli, 2005). Therefore, to define a suggestive association using an FDR set at 0.10, empirical *P*-values (P_{emp}) from the m independent tests were ranked from $P_1 \ldots P_m$ for each locus (i–m) and then tested against the partial inequality $P_i \leq \alpha i/m$, where $\alpha = 0.10$ and i was the rank of that test based on P_{emp} in ascending order. Where $(P_{\text{emp}})_i \leq \alpha i/m$, the null hypothesis and those with lower P_i were rejected (Verhoeven et al., 2005).

Haplotype analysis

For any SNP with a suggestive association, haplotypes were reconstructed on that particular chromosome using Haploview (Barrett et al., 2005). Association testing of both single markers and haplotypes was performed with the number of permutations based on the number of markers on that particular chromosome, with the adjusted haplotype-wide significance threshold then set at $P_{\text{corrected}} = 0.05$.

Candidate gene sequencing

The highest genome-wide suggestive regions, upon correction for genomic inflation, were screened for candidate genes using the equine reference database, available at UCSC.[1] Within 1 Mb of the candidate region surrounding the two most highly associated SNPs on ECA8, there were three genes, all of which demonstrate expression in the central nervous system, *PIK3C3*, *RIT2* and *SYT4*. Two post-mortem-confirmed NAD/EDM affected and one post-mortem-confirmed unaffected horses (from GWA study) were selected for sequencing. *PIKC3* and *RIT2* were sequenced in both genomic and cDNA, while *SYT4* was sequenced in cDNA only.

The equine orthologs to human genes *PIK3C3* (NM_002647) and *RIT2* (NM_002930) were identified by the equine BLAT search[1] and exons identified within the September 2007 2.0 draft assembly of the domestic horse (*Equus caballus*)[2]. In addition to exons and ≥200 base pairs of flanking intronic sequence, both *PIK3C3* and *RIT2* were evaluated in the February 2009 human assembly[3] for promoter-associated sequence and variants in the 5′ untranslated region (UTR) and splice sites using Ensembl[4]. Sequences were compared to the horse assembly by an equine BLAT search and regions were included for sequencing. Primers flanking each region were designed[5] (Rozen and Skaletsky, 2000) and PCR performed using primer-specific melting temperatures (Appendix: Supplementary Table S1). Sequences were scanned for variants and the equine reference sequence was used as an additional unaffected sample.

Fine structure mapping

A custom-designed SNP genotyping platform was created using the 88 variants (84 SNPs, four insertions/deletions) uncovered from sequencing of *PIK3C3* and *RIT2* only as non-synonymous variants were not uncovered through sequencing of

[1] See: http://genome.ucsc.edu.
[2] See: http://www.ncbi.nlm.nih.gov/genome/145.
[3] See: http://www.ncbi.nlm.nih.gov/genome/51.
[4] See: http://useast.ensembl.org.
[5] See: http://hokker.wi.mit.edu/primer3/.