Mouse SNPs for evolutionary biology: Beware of ascertainment biases

Pierre Boursot, Khalid Belkhir

Mouse SNPs for evolutionary biology: beware of ascertainment biases

Pierre Boursot[1] and Khalid Belkhir

Laboratoire Génome, Populations, Interactions, Adaptation (UMR5171), Université Montpellier II, 34095 Montpellier Cedex 5, France

[1]Corresponding author.
E-mail boursot@univ-montp2.fr; fax 33-4-67144554.

Recently, Harr (2006) used publicly available Single Nucleotide Polymorphism (SNP) data (http://www.well.ox.ac.uk/mouse/ INBREDS/) to search the mouse genome for regions of elevated differentiation between the subspecies of the house mouse. If they existed, such regions would be of great interest because they would be likely to harbor the genetic factors causing the partial incompatibility between the subspecies, a phenomenon particularly documented in the case of the European subspecies *Mus musculus domesticus* and *Mus musculus musculus* (Storchova et al. 2004; Britton Davidian et al. 2005; Raufaste et al. 2005). Using the 10,265 SNPs spanning the whole genome and that have been typed in seven *domesticus* and eight *musculus* wild-derived strains of various geographical origins, Harr reports that 10 regions located on seven different chromosomes show a significantly higher than average regional proportion of SNPs with alleles alternatively fixed in the two subspecies.

Using SNP data in population genetics and evolution, especially for quantifying differentiation or divergence, requires a careful and unbiased choice of SNPs. Ascertainment biases in the SNP discovery and choice processes can have various origins and can sometimes lead to erroneous conclusions (Morin et al. 2004; Clark et al. 2005). Most mouse SNPs have been discovered by comparing the genomes of classical inbred laboratory mouse strains that have long been known (Bishop et al. 1985) to be hybrids between *M. m. domesticus* (the Western European subspecies) and mice of Asian origin (*M. m. musculus* and/or *Mus musculus castaneus*). Recent surveys have shown that the genomes of these strains are mosaics of fragments of various sizes with either of these two origins, the non-*domesticus* contribution being mostly of *musculus* origin and representing 20%–30% of the laboratory mouse genomes (Wade et al. 2002; Zhang et al. 2005). When using a panel of such strains to discover SNPs, two situations can occur. In regions of the genome where none of the SNP discovery strains is of *musculus* origin (or where they all are, but this is unlikely), one should discover only SNPs that are polymorphic in

domesticus (respectively, *musculus*), but should find none of the SNPs that are alternatively fixed between *domesticus* and *musculus*. In contrast, all SNPs in the latter category should be discovered in regions where some of the test strains are of *musculus* origin. We were therefore concerned that the results reported by Harr could be affected by this bias, the regions with high inter-subspecific differentiation they pinpoint being mostly those for which the SNP discovery strains were of different taxonomic origins.

Most of the SNPs in the data set that Harr used were chosen by comparing the full sequences of five strains, from Celera (strains C57/BL/6J, DBA/2J, A/J, 129S1/SvImJ, and 129X1/SvJ) (Pletcher et al. 2004). Among them, those polymorphic between strains A/J, AKR, BALB/cJ, DBA2/J, C57BL/6J, LP/J, I, and RIIIS/J were retained (http://www.well.ox.ac.uk/mouse/INBREDS/). Fortunately, six of the strains used to define or retain these SNPs have been largely resequenced (http://mouse.perlegen.com/mouse/). Because this data set also includes one *domesticus* wildderived strain (WSB/EiJ), one *musculus* (PWD/PhJ), and one *castaneus* (CAST/EiJ), we could use it to test our hypothesis that the regions of high intersubspecific differentiation detected by Harr are regions where the SNP discovery strains are of mixed origins. We sliced the genome in fragments of 100 kb, and for each fragment we calculated the resemblance (proportion of identical SNPs) of each of the six SNP discovery strains to PWD/PhJ, from which we subtracted the greater of its resemblances to WSB/EiJ and CAST/EiJ. Positive values of the resulting index thus indicate a *musculus* origin; negative values, a non-*musculus* origin. Figure 1 (thin curves) presents the results for the chromosomes on which Harr found significant differentiation peaks. Note that in some regions (such as the distal chromosome 1 and three segments of chromosome 14), the index of the SNP discovery strains oscillates around zero, which prevents classification. This is because in these regions the reference *musculus* strain (PWD/EiJ) is of *domesticus* origin (analysis not shown, but a plausible result since the strain comes from the Czech Republic, close to the natural hybrid zone between *domesticus* and *musculus*). Barring such regions of the genome, it can be seen in the figure that classification of the genomic regions as *musculus* or non-*musculus* is essentially unambiguous based on our index. The figure also shows for comparison the variations of the regional proportion of SNPs alternatively fixed between wild *domesticus* and *musculus* using the same data as Harr (thick curve), and points with arrows to the significant peaks of differentiation found by this investigator. Among 10 such peaks, seven lie in regions where some of the SNP discovery strains are obviously of *musculus* origin. These peaks thus appear to result from the SNP ascertainment bias. However, for the proximal peak on chromosome 2 and that on chromosome 10, no mixed origin of the SNP discovery strains was detected. The results on the proximal peak of the X

chromosome are ambiguous because of a gap in the resequencing data, but suggest some *musculus* contribution in one of the strains. The two former peaks, on chromosomes 2 and 10, and potentially the latter on the X chromosome, could thus be retained as true differentiation peaks, not suffering from the ascertainment bias. It may seem surprising that a high proportion of SNPs were found fixed between the wild *domesticus* and *musculus* when the SNPs were discovered among laboratory strains of apparently pure *domesticus* origin. However, if these genomic regions have undergone recent selective sweeps, most polymorphisms should be rare, which could explain this apparent paradox.

Harr (2006) attempted to verify her findings by sequencing fragments from regions she detected as high and low differentiation in wild-derived *domesticus* and *musculus* mice, and found a significantly higher differentiation (as measured by the G_{st} statistics) in the former category than in the latter, in apparent agreement with her SNP analyses. We believe the agreement is real, but not for the reason given by the author. Although, as we have seen, the ascertainment bias produces many false positives among the high differentiation peaks, it should not contribute to the detection of false low differentiation peaks. Therefore, we suggest that Harr's apparent confirmation follows from her correctly identifying low differentiation regions, but not high differentiation regions as could be claimed without accounting for the ascertainment bias (with the two or three exceptions mentioned above).

References

Bishop, C.E., Boursot, P., Bonhomme, F., and Hatat, D. 1985. Most classical *Mus musculus domesticus* laboratory mouse strains carry a *Mus musculus musculus* Y chromosome. *Nature* **315:** 70–72.

Britton Davidian, J., Fel-Clair, F., Lopez, J., Alibert, P., and Boursot, P. 2005. Postzygotic isolation between the two European subspecies of the house mouse: Estimates from fertility patterns in wild and laboratory-bred hybrids. *Biol. J. Linn. Soc. Lond.* **84:** 379–393.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., and Nielsen, R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15:** 1496–1502.

Harr, B. 2006. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16:** 730–737.

Morin, P.A., Luikart, G., Wayne, R.K., and The SNP Workshop Group. 2004. SNPs in ecology, evolution and conservation. *Trends Ecol. Evolution* **19:** 208–216.

Pletcher, M.T., McClurg, P., Batalov, S., Su, A.I., Barnes, S.W., Lagler, E., Korstanje, R., Wang, X., Nusskern, D., Bogue, M.A., et al. 2004. Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. *PLoS Biol.* **2:** e393.

Raufaste, N., Orth, A., Belkhir, K., Senet, D., Smadja, C., Baird, S.J.E., Bonhomme, F., Dod, B., and Boursot, P. 2005. Inferences of selection and migration in the Danish house mouse hybrid zone. *Biol. J. Linn. Soc. Lond.* **84:** 593–616.

Storchova, R., Gregorova, S., Buckiova, D., Kyselova, V., Divina, P., and Forejt, J. 2004. Genetic analysis of X-linked hybrid sterility in the house mouse. *Mamm. Genome* **15:** 515–524.

Wade, C.M., Kulbokas III, E.J., Kirby, A.W., Zody, M.C., Mullikin, J.C., Lander, E.S., Lindblad-Toh, K., and Daly, M.J. 2002. The mosaic structure of variation in the laboratory mouse genome. *Nature* **420:** 574–578.

Zhang, J., Hunter, K.W., Gandolph, M., Rowe, W.L., Finney, R.P., Kelley, J.M., Edmonson, M., and Buetow, K.H. 2005. A high-resolution multistrain haplotype analysis of laboratory mouse genome reveals three distinctive genetic variation patterns. *Genome Res.* **15:** 241–249.

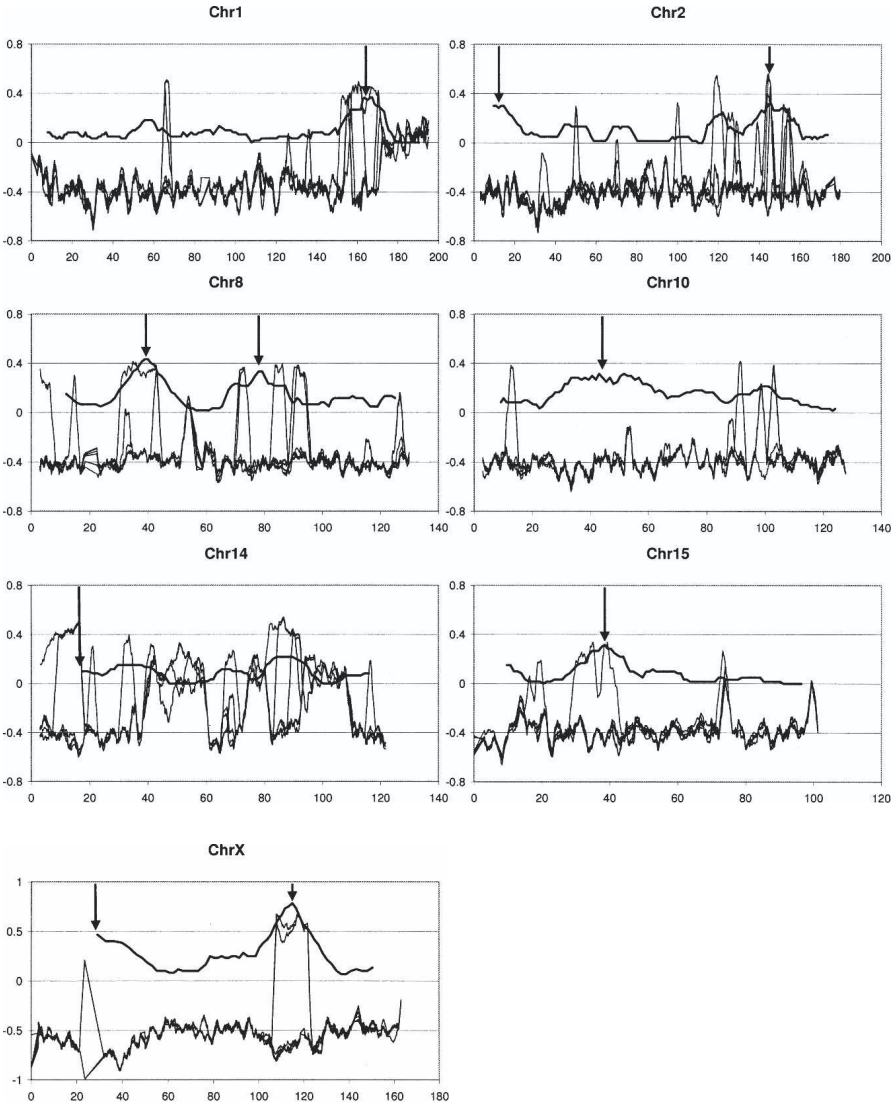

Figure 1. Regional variation along the chromosomes of the resemblance of six resequenced laboratory strains (C57/BL/6J, DBA/2J, A/J, 129S1/SvImJ, AKR/J, and BALB/cByJ) (thin curves) to *Mus musculus musculus* (strain PWD/PhJ) minus their best resemblance to either *Mus musculus domesticus* or *Mus musculus castaneus* (strains WSB/EiJ or CAST/EiJ), compared to the regional variation of the proportion of fixed SNPs between wild-derived *domesticus* (seven strains) and *musculus* (eight strains) (thick curve). The former data set is represented in sliding windows of 2 Mb shifted by 100 kb. In the latter data set, windows of 60 SNPs are shifted by 5 SNPs (as in Harr 2006), and arrows indicate the significant differentiation peaks according to Harr's analyses. Chromosome coordinates for this data set were transformed from build 34 to build 36 in order to match the resequencing data set, using the utility provided at the UCSC Bioinformatics Web site (http://genome.ucsc.edu/).