



Regions of high differentiation—Worth a check

Bettina Harr

Genome Res. 2006 16: 1193-1194

Access the most recent version at doi:[10.1101/gr.5787706](https://doi.org/10.1101/gr.5787706)

References

This article cites 1 articles, 1 of which can be accessed free at:
<http://genome.cshlp.org/content/16/10/1193.full.html#ref-list-1>

Article cited in:

<http://genome.cshlp.org/content/16/10/1193.full.html#related-urls>

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#)

To subscribe to *Genome Research* go to:
<http://genome.cshlp.org/subscriptions>

Regions of high differentiation—Worth a check

Bettina Harr

Institute for Genetics, Department of Evolutionary Genetics, 50674 Köln, Germany

B. Boursot and K. Belkhir in this issue of *Genome Research* point to an ascertainment bias in my use of SNPs to identify regions of high differentiation between a pair of house mouse subspecies. Here, I discuss additional points to consider and argue that the ultimate test for such regions should be an independent confirmation using unrelated samples. I provide evidence that regions of high differentiation as identified from laboratory strains and potentially biased SNP markers can be confirmed and therefore are worth a deeper investigation.

[Supplemental material is available online at www.genome.org.]

Boursot and Belkhir (2006) point to an ascertainment bias in my use of SNPs to identify regions of high differentiation between a pair of mouse subspecies (Harr 2006). The argument is that genomic regions in lab strains that contain a mixture of genes from two subspecies are biased toward detection of SNPs that differentiate the two subspecies. To see the cause of the bias, consider two lines that are completely fixed for different alleles at every base. If a lab strain is composed entirely of one line, no SNPs will be detected. However, SNPs are detectable in regions where DNA material from both lines is segregating in the lab strain, and these SNPs will indicate strong differences between the two lines. In reality, many SNPs are detected based on within-subspecies polymorphisms, but the effect of introgression still results in a bias toward detecting regions of introgression as regions of differentiation. In the predominantly *Mus musculus domesticus* lab strains used to isolate the SNPs, ~20% of the genome contains introgressed material from *Mus musculus musculus*, and regions of introgression will be biased toward SNPs that identify differences between subspecies. I am grateful to Boursot and Belkhir for pointing out this bias, which I had overlooked in my previous publication (Harr 2006).

As noted by Boursot and Belkhir (2006), it is difficult to determine how much this bias makes the peaks of regional differentiation I identified artifacts. Boursot and Belkhir reanalyzed the data and identified nine peaks of regional differentiation along chromosome stretches where they were also able to assess the presence of introgression (of these nine, I considered only six to be significant) (Harr 2006). Boursot and Belkhir suggested that six of the nine peaks (three of the six from my publication) (Harr 2006) were “obviously” associated with introgression of material from *musculus*. Although this needs to be more rigorously tested, for the purpose of discussion, I accept their conclusion. Assuming that *musculus* introgression occurs across 20% of the genome (following Boursot and Belkhir), an expected 1.8 peaks (9×0.2) should fall in regions of introgression by chance alone. The difference between 6 and 1.8 is 4.2, implying that ~50% of the peaks Boursot and Belkhir identified are artifacts (i.e., 4.2 of 9 peaks). The number may be less than this, however.

The SNPs in the Wellcome Trust database were largely selected on the basis of polymorphism in at least some of eight strains, but >15,000 were provided from other sources. SNPs closer than 50 kb with identical strain distribution patterns were thinned out, and gaps >500 kb were filled with SNPs from other data sets (see <http://www.well.ox.ac.uk/mouse/INBREDS/>). Thus,

regions with expected high SNP density (because of introgression of *musculus* DNA) were pruned of SNPs more than regions that show no introgression. In regions with no introgression, the SNPs must have been identified as a result of polymorphism within the subspecies *domesticus*; such within-subspecies polymorphism must also contribute to the SNPs within the introgressed regions. Thus, it is possible that relatively few of the SNPs in the regions of introgression were originally identified as a result of the introgression. Such SNPs may inflate the height of the peaks of differentiation, but perhaps not the peaks themselves.

Given that SNP pruning is unlikely to be complete, I assessed the extent of introgression in the Wellcome Trust data set by estimating SNP density along each chromosome. If SNPs were identified entirely without regard to chromosomal position, and between-subspecies differences contributed importantly to SNPs, introgression should correspond to regions of high SNP density. I compared these estimates with regions of introgression for the chromosomes where this measure is available in Boursot and Belkhir (2006) and also with measures of differentiation between subspecies (this latter approach differs from the analysis in Harr [2006], where measures of differentiation were presented for 60 SNP windows, rather than a fixed length along the chromosome). The results show a general correspondence between regions of high SNP density and introgression, as identified by Boursot and Belkhir. They also highlight some other regions of high SNP density that Boursot and Belkhir were unable to assess directly for introgression. Of a total of 11 peaks of differentiation between subspecies, six coincide with peaks of high SNP density. The association is less obvious for the other five.

I also assessed *musculus* introgression from this data set, by calculating the average G_{st} between the eight lab strains that were used to identify or retain the SNPs (a subset of these lab strains [strains A/J, 129S1/SvImJ, AKR, BALB/c, C57BL/6], and DBA2/J] was used by Boursot and Belkhir [2006] to assess *musculus* introgression but using the Perlegen data) and the eight pure *domesticus* individuals for which data are available from the Wellcome Trust. High values of G_{st} imply that some lab strains are *musculus*-like in a particular region. While my identification of *musculus*-like fragments is less powerful than the method of Boursot and Belkhir because fewer SNPs are in the Wellcome Trust database, it has the advantage that the assignment of *musculus*-likeness is based on multiple strains derived from each pure subspecies and not a single *musculus* and *domesticus* strain as in Boursot and Belkhir (2006) (there is a rather large fraction of SNPs that segregate in both subspecies, i.e., 22% of all polymorphic SNPs, and these might not be correctly assigned based on just a single in-

E-mail harrb@uni-koeln.de; fax 49-221-470-5975.

Article is online at <http://www.genome.org/cgi/doi/10.1101/gr.5787706>.

dividual per subspecies). Not all peaks of high SNP density correspond to regions of *musculus* introgression as estimated by this method, suggesting that some of the high-SNP-density regions might not be contaminated. More importantly, there are several peaks of high *musculus* introgression that do not colocalize with regions of elevated differentiation between subspecies. Such difficulties of interpretation imply that not all peaks that lie in either regions of introgression or of high SNP density should be immediately dismissed as artifacts.

In addition to the bias noted by Boursot and Belkhir (2006), several other factors may make the presence of any particular peak doubtful whether or not it is in a region of introgression. These include the lines chosen for original establishment and Type 1 errors. I suggest that the best approach is to check the data independently through analysis of completely different samples, without dismissing any peaks as artifacts prior to the check. I reported some results in my earlier publication (Harr 2006) and present some additional data here (Supplemental Table 1). A good example is the region of high differentiation around position 30 Mb on chromosome 8. As pointed out by Boursot and Belkhir (2006), the central portion of this peak corresponds to introgressed material. Independent sequencing from natural populations shows that the two fragments that lie within this region of introgression show relatively low G_{st} values (although

G_{st} values are still $\sim 2\times$ higher than the average of the reference regions). These low values may support Boursot and Belkhir's suggestion that introgression is responsible for the high levels of differentiation observed in the lab strains. However, one neighboring fragment located outside of the introgressed (but still highly differentiated) region shows 12 fixed differences between the subspecies in the analysis of natural populations (Harr 2006). I have now analyzed one fragment from the natural populations located on the other side of the introgressed region on chromosome 8 (but also still located in the region of high differentiation). This too shows an extremely high G_{st} value (between *domesticus* and *musculus* there are 11 fixed differences, $G_{st} = 1.0$) (Supplemental Table 1). This suggests that the chromosome 8 peak should not be dismissed as an artifact. We are presently investigating the possibility that the chromosome 8 segment, which shows extremely high differentiation at its borders, represents an inversion that is fixed between the subspecies.

References

- Boursot, B. and Belkhir, K. 2006. Mouse SNPs for evolutionary biology: Beware of ascertainment biases. *Genome Res.* (this issue).
Harr, B. 2006. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16**: 730–737.