TranscriptSNPView: a genome-wide catalog of mouse coding variation

To the Editor:

With the recent release of the genome-wide sequence for multiple inbred mouse strains[1], and with resequencing data for a large number of additional strains entering the public domain (http://www.niehs.nih.gov/crg/cprc.htm), we are one step closer to being able to identify the underlying genetic variants responsible for the trait characteristics that define each strain. Here, we describe a genome-wide catalog of coding variation in the mouse genome that was developed using an extensive collection of mouse DNA sequence reads, including those recently released by Celera, data from dbSNP[2] and resequencing data generated by Perlegen Sciences for the US National Institute of Environmental Health Sciences (NIEHS). To display these data, we developed a new software tool, TranscriptSNPView, which has been integrated into the Ensembl Genome Browser to take advantage of the evolving mouse genome assembly and the latest Ensembl[3] and Vega gene predictions[4]. TranscriptSNPView can be accessed via the Ensembl Genome Browser (http://www.ensembl.org/Mus_musculus/transcriptsnpview).

TranscriptSNPView displays coding SNP data from 48 mouse strains (**Supplementary Table 1** online). Using the SNP calling algorithm ssahaSNP[5], we computed over 50 million SNPs from the common laboratory *Mus musculus* strains A/J, DBA/2J, 129X1/SvJ and 129S1/SvImJ from whole-genome shotgun sequence reads generated by Celera, and from C3HeB/FeJ and NOD BAC-end sequence reads generated by the Wellcome Trust Sanger Institute. We also generated SNP calls from the *Mus musculus molossinus* strain MSM/Ms using sequence reads generated by RIKEN[6] (**Supplementary Table 1**). Collectively, these SNP calls have been designated 'Sanger SNPs'. The 25 million DNA sequence reads used to generate the Sanger SNP collection represent 7.32-fold coverage of the NCBI mouse build 35 genome assembly and are available via the Ensembl trace repository (http://trace.ensembl.org).

The Sanger SNP calls were distilled to 6.87 million nonredundant genome-wide SNP features and were combined with an additional 6.4 million dbSNP entries (version 126), providing data for an additional 41 mouse strains. By merging these data sets and mapping them against the Ensembl 38.35 mouse gene build, we collated 726,462 coding SNP variants across all strains and computed their amino acid consequences to identify 249,996 nonsynonymous coding changes and 2,667 stop codons. Coding SNP figures for each strain are provided in **Supplementary Table 1**. We also identified instances where stop codons had been lost, and we predicted mutations in introns, invariant intronic splice sites and in untranslated and regulatory regions. These predictions, which can be used as a basis for identifying functional SNP variants, are displayed in TranscriptSNPView. A detailed description of all of the features of TranscriptSNPView is provided in the **Supplementary Note** online.

A data collection of this quality and depth is unprecedented and will provide the means to obtain a high-resolution picture of coding variation in the mouse genome. TrancriptSNPView represents a powerful new tool for functional analysis of the mouse genome and will become a central repository for mouse coding variation data.

Fiona Cunningham[1], Daniel Rios[2], Mark Griffiths[1], James Smith[1], Zemin Ning[1], Tony Cox[1], Paul Flicek[2], Pablo Marin-Garcin[1], Javier Herrero[2], Jane Rogers[1], Louise van der Weyden[1], Allan Bradley[1], Ewan Birney[2] & David J Adams[1]

[1]The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK.
[2]The European Bioinformatics Institute, Hinxton, Cambridgeshire, CB10 1SD, UK.
e-mail: birney@ebi.ac.uk or da1@sanger.ac.uk

Note: Supplementary information is available on the Nature Genetics website.

1. Marris, E. *Nature* **435**, 6 (2005).
2. Sherry, S.T. *et al. Nucleic Acids Res.* **29**, 308–311 (2001).
3. Birney, E. *et al. Nucleic Acids Res.* **34**, D556–D561 (2006).
4. Ashurst, J.L. *et al. Nucleic Acids Res.* **33**, D459–D465 (2005).
5. Ning, Z. *et al. Genome Res.* **11**, 1725–1729 (2001).
6. Abe, K. *et al. Genome Res.* **14**, 2439–2447 (2004).

Has the chimpanzee Y chromosome been sequenced?

To the Editor:

Kuroki *et al.* recently reported "the finished sequence of the chimpanzee Y chromosome"[1]. Their analyses included comparisons with previously reported DNA sequences from the human and chimpanzee Y chromosomes[2,3]. The article[1] was based on the authors' sequencing of 12.7 Mb from the PTB1 library, which represents the genome of one male chimpanzee. We previously sequenced the 9.5-Mb 'X-degenerate' portion of the Y chromosome from a different male chimpanzee, whose genome is represented in the CHORI-251 library[2]. We write to express concerns regarding the conclusions of Kuroki *et al.*, including the gene content of the chimpanzee and human Y chromosomes, and the level of sequence divergence between the two chimpanzee Y chromosomes whose sequences have been explored.

First, the authors' claim of "the finished sequence of the chimpanzee Y chromosome" merits attention[1]. The 12.7 Mb reported in the study overlaps fully the 9.5-Mb X-degenerate region analyzed in the prior study[2]; it also