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

includes 1.7 Mb of contiguous, non–X-degenerate sequence not examined in the earlier publication. The remaining 1.5 Mb reported by the authors is a superficial sampling of the 'ampliconic' portions of the chimpanzee Y chromosome. The ampliconic regions of primate Y chromosomes are of great biological and medical interest[3–10]. These regions are difficult but not impossible to sequence systematically and comprehensively (see refs. 3 and 5 and our unpublished results), and, in man, they comprise 10.2 Mb, or nearly half of the Y chromosome's male-specific euchromatin[3]. If a similar fraction of the chimpanzee Y chromosome is ampliconic, then large and biologically significant portions of the chromosome have yet to be sequenced and analyzed.

The authors[1] reported more genes within the X-degenerate regions of the chimpanzee and human Y chromosomes than did investigators in earlier studies[2,3], but these additions, we suggest, do not withstand scrutiny. Unlike prior studies of the human and chimpanzee Y chromosomes[2,3], the authors' inferences were based on very limited electronic analyses and were not validated experimentally. This may explain why several pseudogenes or disrupted genes—some explicitly identified as such in earlier studies—were treated as functional genes despite previous experimental evidence to the contrary (**Supplementary Note** and **Supplementary Figure 1** online). These include the *TMSB4Y* and *USP9Y* pseudogenes on the chimpanzee Y chromosome[2], the *GYG2* pseudogene on the human and chimpanzee Y chromosomes[2,3] and the *CD24L4* pseudogene on the human Y chromosome.

Finally, the authors appear to have overestimated the nucleotide divergence between the two chimpanzee Y chromosomes represented by the PTB1 and CHORI-251 libraries. We aligned the PTB1 and CHORI-251 sequences (**Supplementary Tables 1–3** online; sequence alignments can be found at http://jura.wi.mit.edu/page) and found their divergence to be 0.002%, or roughly 20 times lower than the 0.0422% reported by the authors[1]. (The authors similarly overestimated divergence between PTB1 and a third chimpanzee Y chromosome, represented by the RPCI-43 library; our sequence alignments can be found at http://jura.wi.mit.edu/page. Note that all CHOR-251 and RPCI-43 sequences included in our alignments with PTB1 were publicly available, as finished sequence, prior to the study by Kuroki *et al.*) Our calculation of divergence between the Y chromosomes of PTB1 and CHORI-251 is so low (~1 in 50,000 nucleotides) that sequencing errors (estimated at less than 1 in 200,000 nucleotides in each study) could account for about one in every three substitutions that appear to differentiate the chromosomes (**Supplementary Note**). It is unclear how the authors calculated a divergence 20-fold higher than ours when comparing the same sequences.

Jennifer F Hughes[1], Helen Skaletsky[1], Steve Rozen[1], Richard K Wilson[2] & David C Page[1]

[1]Howard Hughes Medical Institute, Whitehead Institute, and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts, 02142, USA. [2]Genome Sequencing Center, Washington University School of Medicine, 4444 Forest Park Boulevard, St. Louis, Missouri 63108, USA.
e-mail: dcpage@wi.mit.edu

1. Kuroki, Y. *et al. Nat. Genet.* **38**, 158–167 (2006).
2. Hughes, J.F. *et al. Nature* **437**, 100–103 (2005).
3. Skaletsky, H. *et al. Nature* **423**, 825–837 (2003).
4. Yen, P. *Nat. Genet.* **29**, 243–244 (2001).
5. Kuroda-Kawaguchi, T. *et al. Nat. Genet.* **29**, 279–286 (2001).
6. Rozen, S. *et al. Nature* **423**, 873–876 (2003).
7. Hurles, M.E. & Jobling, M.A. *Nat. Genet.* **34**, 246–247 (2003).
8. Anonymous. *Nat. Genet.* **35**, 195 (2003).
9. Tyler-Smith, C. & McVean, G. *Nat. Genet.* **35**, 201–202 (2003).
10. Repping, S. *et al. Nat. Genet.* **35**, 247–251 (2003).

Kuroki *et al.* reply:

We very much appreciate that Dr. Page and his colleagues have spared their valuable time to carefully evaluate and re-analyze the data from our chimpanzee Y chromosome comparative analysis[1].

First, we would like to clarify several misunderstandings concerning our paper, particularly about the sequenced region in our paper[1]. We determined complete sequences for 271 kb of the Y-specific pseudoautosomal region 1 and 12.7 Mb of the male-specific region. We produced high-quality sequence data for almost half of the entire chimpanzee Y chromosome and carried out detailed comparative analyses between human and chimpanzee Y chromosomes. We contrasted this with similar analyses that were carried out on the autosomes, namely human chromosome 21 and chimpanzee chromosome 22 (now renumbered to 21), and the non-recombining portions of the Y chromosome[1]. In addition to the interspecies analyses, we examined the diversity in chimpanzees using publicly available sequence data[1–3] and verified that the diversity in the chimpanzee Y chromosome was very low. These analyses include the entire X-degenerate region and parts of the ampliconic region[1]; the remaining parts of the ampliconic region have not yet been fully examined. Our results showed that the structure and sequence identity of the regions were extremely different between human and chimpanzee. We found that both human and chimpanzee Y chromosomes retain the same basic structure: that is, many palindromic structures have accumulated on the Y chromosome, but the regions involved in the palindrome conformation are species specific, such as the chimpanzee-specific palindrome CSP1 (see Supplementary Fig. 6 in ref. 1 and our website, http://stt.gsc.riken.jp/ (RIKEN Genomic Sciences Center)). We agree it is important to completely sequence the chimpanzee Y chromosome ampliconic region and to carry out more comprehensive comparative analyses.

Concerning the differing results of our gene annotation, this is dependent on the method used for the analysis[1,2]. As described in the papers, we used the human Y chromosome gene set manually annotated by the HAVANA group (http://vega.sanger.ac.uk/homo_sapiens/), whereas Hughes *et al.* used their own annotation data for human Y chromosome. Hughes *et al.* emphasized that our annotation results were wrong and that their results, which were verified by RT-PCR, were more reliable. We feel that Hughes *et al.* not only overestimated the reliability of the RT-PCR technology but misunderstood our conclusions: that is, we did not assert whether these four genes, *USP9Y*, *TMSB4Y*, AC002992.5 (*GYG2-like*) and *CD24L4*, were functional, because it could not be determined from our sequence-based analysis. Even if we had RT-PCR data, that it in itself would not be conclusive as to the functionality of these genes. Further support, such as full-length cDNA cloning or RNA blot analysis, would be more convincing. We agree that further careful analysis for gene identification, annotation and verification is necessary to identify the functional and biological meaning not only of the chimpanzee genes but also of their human counterparts, an area of ongoing research.

Hughes *et al.* recalculated the chimpanzee diversity and had some concerns about

the difference in our results. We suggest that the difference is due to a combination of alternate calculation methods and slightly different data sets (at the time of our analysis, the assembly of Hughes *et al.*, DP000054, was not available). We used a local alignment method, while they used a global alignment method. The reason we used a local alignment method was to compare the diversity of the Y chromosome with that of the autosomes—in this case, human chromosome 21 and chimpanzee chromosome 22. We re-analyzed the chimpanzee Y chromosome diversity using different parameters, 90% and 99% identity over a length of 1,000 bp, and obtained the values 0.042% to 0.011%, respectively, between CH251 and PTB1 (**Supplementary Note** online). In either case, the conclusion we first reported has not changed: that the diversity of the chimpanzee Y chromosome is much lower than expected, as pioneered by Stone *et al.*[4] and reconfirmed by the independent analysis of Hughes *et al.*[2] using their different strategy.

Yoko Kuroki[1], Todd D Taylor[1], Hideki Noguchi[1,2], Takehiko Ito[3], Atsushi Toyoda[1], Yoshiyuki Sakaki[1] & Asao Fujiyama[1]

[1]RIKEN Genomic Sciences Center, Yokohama, Kanagawa 230-0045, Japan. [2]Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-0882, Japan. [3]Mitsubishi Research Institute, Tokyo 100-8141, Japan.

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

1. Kuroki, Y. *et al. Nat. Genet.* **38**, 158–167 (2006).
2. Hughes, J.F. *et al. Nature* **437**, 100–103 (2005).
3. The International Chimpanzee Chromosome 22 Consortium. *Nature* **429**, 382–388 (2004).
4. Stone, A.C. *et al. Proc. Natl. Acad. Sci. USA* **99**, 43–48 (2002).

Normalization procedures and detection of linkage signal in genetical-genomics experiments

To the Editor:

Recent microarray-based experiments, designed to measure the influence of genetic variation on gene expression at a near–genome-wide scale, have offered the first evidence of heritability of mRNA levels between individuals[1–6]. The collective approach has been to treat expression values for each transcript across individuals as a molecular phenotype, in a massively parallel linkage analysis. In genetical-genomics experiments, each array measures expression levels from an individual with a different genetic background (*e.g.*, BxD mouse[4]). Unprocessed data from microarrays representing individuals may show quite different distributional characteristics due to experimental vagaries rather than true biological influence. If these are not removed by normalization[7], then the expression levels of a given set of genes will systematically vary across the arrays; this variation can then correlate with genotype distributions at one or more loci, giving rise to spurious 'genetic signal'.

To systematically examine the influence of normalization on reproducibility, or otherwise, of genetic linkages associated with individual or small groups of genes, we reanalyzed recombinant inbred strain data of both our own and external groups[4–6]. We applied a range of normalization procedures (**Supplementary Methods** online) that remove varying degrees of systematic structure from raw expression data, and we examined the concordance of linkage analysis results using the 'correspondence at the top' (CAT) plot format of ref. 8 (**Fig. 1**).

We predicted a gradual decrease in the numbers of genes being identified, commensurate with increasingly sensitive artifact removal, leaving a common 'core' of transcripts under genuine biological influence from well-defined loci. In contrast, we found a startling lack of agreement between results in both the transcripts identified as variant and the loci implicated in their variation (**Supplementary Fig. 1** online). This lack of concordance was present in data collected by both single- and dual-channel experimental platforms (**Fig. 1**). Generally, transcripts with higher variance in expression demonstrated better linkage concordance than those with lower variance (**Supplementary Fig. 2** online).

It can be argued that the observation of limited overlap in sets of genes exhibiting linkage is a result of the statistical fluctuations common to microarray data analyses and is therefore

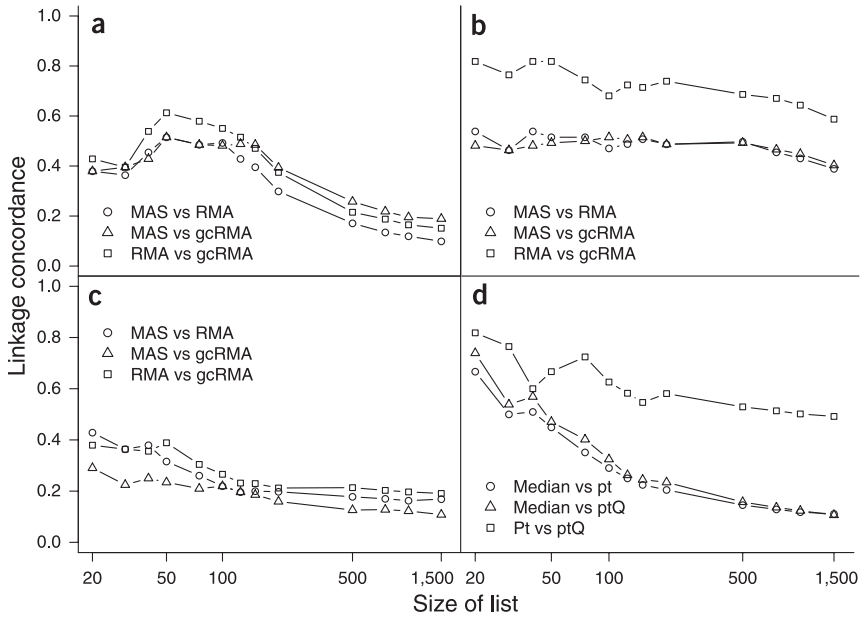

Figure 1 Concordance analysis of linkage statistics based upon three normalization methods. Data taken from (**a**) ref. 4 (**b**) fat tissue from ref. 6 (**c**) HSC samples from ref. 5 and (**d**) 31-BxD mouse liver (C.J.C. and P.F.R.L., unpublished. data). The graphs are in CAT format[8]. The *x* axis represents the top-ranking group containing *N* gene and marker pairs (a group of 20, 30, or more gene-marker pairs). This is plotted against the concordance of the identity of the genes and markers in the *N* group when the appropriate two normalizations are compared (corresponding to symbols defined in figure).