Resources for proteomics in mouse embryonic stem cells

To the Editor: A recent publication in *Nature Methods* described recombinase-mediated cassette exchange (RMCE) for re-engineering gene targeted alleles in mouse embryonic stem cells (ESCs) derived from the International Knock Out Mouse Consortium (IKMC) repositories[1]. We wish to point out that *FlipRosaβgeo* gene–trapped ESC lines in the same repositories[2] can be engineered to encode proteins with N-terminal protein tags using an RMCE-based approach. As do the IKMC's gene-targeted alleles, the *FlipRosaβgeo* gene-trap alleles include site-specific recombinase target sequences that enable RMCE[3] (**Fig. 1a**).

IKMC resources currently contain 25,130 tagging-compatible ESC lines representing 3,695 individual genes (**Supplementary Table 1**). These lines have *FlipRosaβgeo* insertions in the first intron of genes, either downstream of the first noncoding exons or of exons that encode relatively short peptides without apparent functional domains. To knock a sequence encoding a protein tag into these loci, we designed an exchange cassette consisting of a hygromycin resistance gene (hygromycin phosphotransferase; *hygro*) fused via a 2A virus polyprotein cleavage sequence (P2A) to a modified N-terminal localization and affinity purification (*nLAP*) tag encoding enhanced GFP (EGFP)[4]. This *hygro-P2A-nLAP* cassette includes splice acceptor (SA) and splice donor (SD) sites upstream and downstream of the cassette, respectively (**Fig. 1a**); therefore it is a portable exon. For RMCE, the tagging exon was flanked by heterotypic FLP recombinase target sequences identical to those inserted by the gene trap (**Fig. 1a**). RMCE induces the expression of a fusion transcript in which the tagging exon is spliced to the endogenous exons of the trapped gene. Translation yields a protein that is cleaved at the P2A site, so that the tagged endogenous protein is expressed independently (**Fig. 1a**).

We tagged the trapped *Myh9*, *Cdk4*, *Jup*, *Fgd4*, *Trp53*, *Prdx1*, *Sesn2*, *Chm*, *Ctnnd1* and *Fkbp5* ESC lines from the IKMC repository (**Supplementary Table 2**). After electroporating the tagging exon together with a codon-optimized FLP (FLPo) recombinase expres-

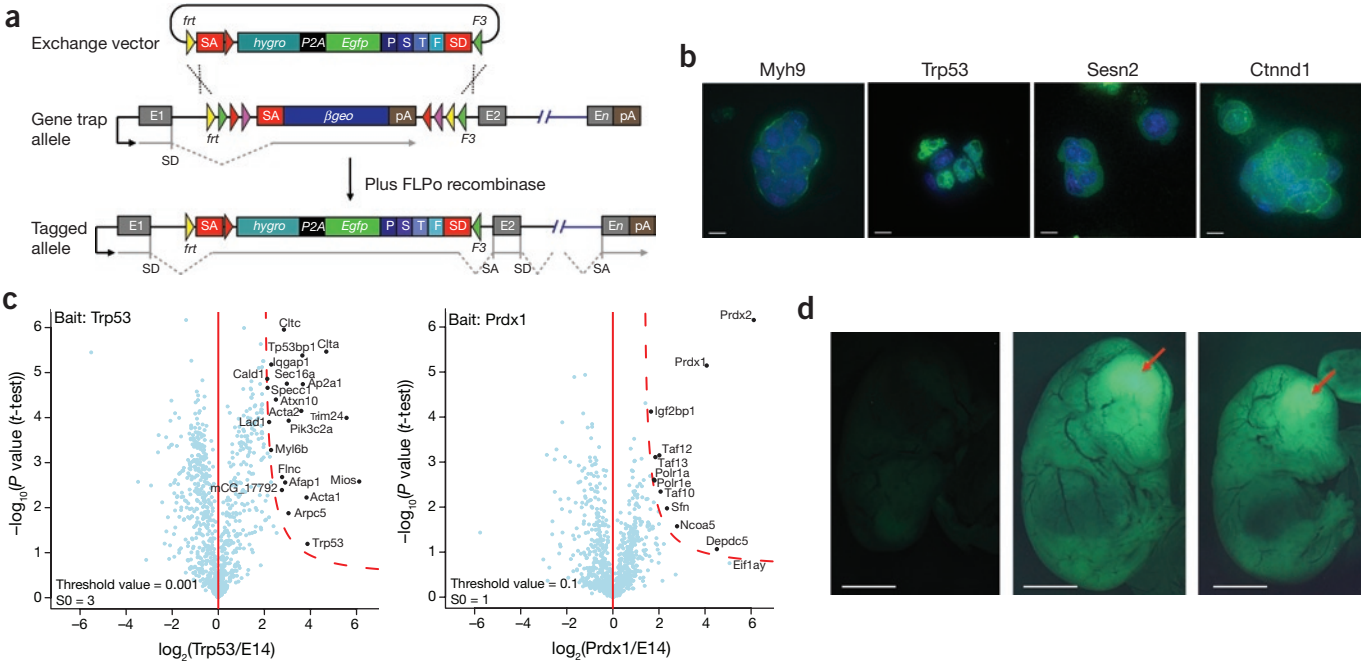

Figure 1 | Proteome analysis in trapped ESC lines. (**a**) Schematic of the *in situ* protein tagging strategy. A protein-tagging cassette is introduced as a portable exon into a *FlipRosaβgeo* gene-trap locus by RMCE via FLPo-mediated recombination. E1–*En*, exons; Frt and F3, heterotypic target sequences for the FLPo recombinase; *loxP* (red triangles) and *lox5171* (purple triangles), heterotypic target sequences for the Cre recombinase; SA, adenovirus type II splice acceptor; *βgeo*, β-galactosidase–neomycinphosphotransferase fusion gene; pA, bovine growth hormone polyadenylation sequence; P, PreScission cleavage site; S, S-peptide; T, TEV protease cleavage site; F, Flag tag; and SD, adenovirus type II splice donor. (**b**) Live-cell imaging of cells expressing the indicated nLAP-tagged proteins (green) to determine localization. Blue, DAP1 stain. Scale bar, 10 μm. (**c**) Volcano plots showing Trp53 and Prdx1 interactors. Tagged proteins were pulled down with antibody to EGFP from ESC extracts; wild-type E14Tg2a (E14) ESCs served as negative controls. Each dot represents an identified protein, and significant interaction partners are represented by black dots; *x* axis, $\log_2$ of the ratio of relative protein intensity in the pulldown and control; *y* axis, $-\log_{10}P$ values of the *t*-test from triplicate experiments. The red line represents the plot-specific false positive rate[5] with its threshold values indicated. S0, curve bend. (**d**) EGFP fluorescence in isolated embryonic day 14.5 embryos from the same litter derived from *nLAP-Trp53* transgenic ESCs. Arrows highlight increased regional fluorescence in forebrains of two highly chimeric embryos. A low chimeric embryo is shown on the left. Scale bar, 2.2 mm.

sion plasmid into the cell lines, and selecting them in the presence of hygromycin (**Supplementary Methods**), an average of 75% of the hygromycin resistant subclones exhibited unique and correctly inserted exchange cassettes (**Supplementary Fig. 1**). Each of these clones expressed correctly spliced fusion transcripts whose translation products were of the expected size (**Supplementary Figs. 2** and **3**). The tagged proteins reflected the known localization patterns of their native counterparts (**Fig. 1b**, **Supplementary Figure 4** and **Supplementary Table 3**).

The physiological amounts of the tagged proteins were sufficient to enable protein-protein interaction studies by a label-free, quantitative affinity purification–mass spectrometry approach[5]. Purification of nLAP-tagged Trp53 together with Prdx1 and their endogenous interaction partners by single-step affinity purification coupled to high-resolution LTQ-Orbitrap mass spectrometry (liquid chromatography–tandem mass spectrometry) recovered the baits plus several known interaction partners, such as TRIM24, Tp53BP1 and CLTC for Trp53 or Prdx2 for Prdx1 (**Fig. 1c**).

The modified ESCs expressed high levels of Oct4, Nanog and Sox2 proteins, suggesting that they are pluripotent (**Supplementary Fig. 5**). *nLAP-Trp53* ESCs efficiently contributed to all cell lineages of a transgenic embryo and replicated the enhanced Trp53 expression seen in the forebrain of embryonic day 14.5 mouse embryos (**Fig. 1d**) (http://www.eurexpress.org/).

In situ protein tagging in *FlipRosaβgeo* ESC lines enables systematic protein localization and protein-protein interaction studies under physiological conditions. It will be useful for applications ranging from proteome analysis in ESC differentiation cultures to the definition of tissue-specific proteomes in mice. The strategy is relevant for over 25,000 characterized and validated gene-trap lines currently available from the German Genetrap Consortium (http://www.genetrap.de/) and European Conditional Mouse Mutagenesis Program (http://www.eucomm.org/) resources.

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

ACKNOWLEDGMENTS

We thank L. von Melchner for useful comments and for help in editing the manuscript, and J. Mühl, A.-T. Tieu, R. Eitz, D. Herold and H. Grunert for technical assistance. This work was supported by grants from the Bundesministerium für Bildung und Forschung to the NGFNplus-DiGtoP consortium (01GS0858), the Deutsche Forschungsgemeinschaft to H.v.M. (ME 82/5-1) and the European Union (EUCOMM project LSHG-CT-2005-018931 to H.v.M. and W.W.).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Frank Schnütgen[1], Franziska Ehrmann[1], Ina Poser[2], Nina C Hubner[3], Jens Hansen[4], Thomas Floss[4], Ingrid deVries[2], Wolfgang Wurst[4,5], Anthony Hyman[2], Matthias Mann[3] & Harald von Melchner[1]

[1]Department for Molecular Hematology, University of Frankfurt Medical School, Frankfurt am Main, Germany. [2]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [3]Max Planck Institute of Biochemistry, Martinsried, Germany. [4]Institute for Developmental Genetics Helmholtz Zentrum München, Germany. [5]German Center for Neurodegenerative Diseases, Technische Universität München, Neuherberg, Germany.
e-mail: melchner@em.uni-frankfurt.de

1. Osterwalder, M. *et al. Nat Methods* **7**, 893–895 (2010).
2. Schnutgen, F. *et al. Proc. Natl. Acad. Sci. USA* **102**, 7221–7226 (2005).
3. Schebelle, L. *et al. Nucleic Acids Res.* **38**, e106 (2010).
4. Poser, I. *et al. Nat. Methods* **5**, 409–415 (2008).
5. Hubner, N.C. *et al. J. Cell Biol.* **189**, 739–754 (2010).

Data transformation practices in biomedical sciences

To the Editor: In over a century since it was first introduced by William Sealy Gosset (under the pseudonym Student), the *t*-test has become one of the most common tests in many fields of research[1] and is now a basic element in a biologist's toolkit for statistical hypothesis testing. Our screen of the first 2010 issue of medical and biological science journals with an impact factor higher than 15 revealed that in 88 of the 213 research articles, the authors had used *t*-tests to analyze their data (**Supplementary Methods**).

Applying a *t*-test is now so routine that many biologists may have forgotten that data should meet certain assumptions, and reminders of its correct use have been published[2,3]. For example, for a valid two-sample *t*-test, the assumptions are that the samples are independent and drawn from populations with equal variances, and that the variable is normally distributed in each group. Although the robustness of Student's *t*-test to the violation of these assumptions is a matter of debate[4], they are seldom verified and data are sometimes transformed in ways that guarantee that the assumptions are no longer met.

A frequent practice sometimes imposed by the nature of the data (such as interexperiment variability) is to normalize data (**Fig. 1**) before applying a *t*-test (not to be confused with the transformation applied to data to approach a Gaussian distribution). Of the 88 articles presenting *t*-test results, in 24 articles the data had been normalized to control samples, which had been given arbitrary values of 1 or 100. Such normalization can be performed in two ways (**Fig. 1a**). In the first approach, all values of control and treatment samples are divided by the mean of the control sample, which thus becomes the arbitrary reference value ('normalization i'). Such normalization conserves the distribution and the relative variance of the samples, allowing the subsequent use of a *t*-test. A second way to perform the normalization is to divide the control and treatment values from each experimental run by the control value ('normalization ii'). Such normalization converts the distribution of the control sample into a uniform distribution with zero variance and renders the normalized data unsuitable for a *t*-test. We encountered the latter normalization in 15 articles.

We investigated the consequences of using normalization ii in terms of type I and type II error rates (percentage of false positives and false negatives, respectively). We used Monte-Carlo simulations (**Supplementary Methods**) to compare error rates when we normalized or did not normalize the same data to control values before applying a two-sample *t*-test.

Applying normalization ii resulted in increased type I error rates as compared to those obtained for unnormalized data (**Fig. 1c**). The use of 'robust' versions of the *t*-test (Welch *t*-test) that do not require the assumption of equal variance[2] only marginally compensated for this increase in type I error rates (data not shown). For sample sizes with $n > 15$, the incidence of false positives increased with data variability (**Fig. 1c** and **Supplementary Fig. 1**). The departure of type II error rates from expected values after normalization ii was minimal for low-variability data but strong for more variable data.

The direction of the change depended on whether the control sample had the smaller or the greater of the two compared means (**Fig. 1d–f**): type II error rates decreased when the control sample had the lower mean (for example, in gene expression upregulation) and increased when the opposite was true (for example, in gene expression