BoxPlotR: a web tool for generation of box plots

To the Editor: In biomedical research, it is often necessary to compare multiple data sets with different distributions. The bar plot, or histogram, is typically used to compare data sets on the basis of simple statistical measures, usually the mean with s.d. or s.e.m. However, summary statistics alone may fail to convey underlying differences in the structure of the primary data (**Fig. 1a**), which may in turn lead to erroneous conclusions. The box plot, also known as the box-and-whisker plot, represents both the summary statistics and the distribution of the primary data. The box plot thus enables visualization of the minimum, lower quartile, median, upper quartile and maximum of any data set (**Fig. 1b**). The first documented description of a box plot–like graph by Spear[1] defined a range bar to show the median and interquartile range (IQR, or middle 50%) of a data set, with whiskers extended to minimum and maximum values. The most common implementation of the box plot, as defined by Tukey[2], has a box that represents the IQR, with whiskers that extend 1.5 times the IQR from the box edges; it also allows for identification of outliers in the data set. Whiskers can also be defined to span the 95% central range of the data[3]. Other variations, including bean plots[4] and violin plots, reveal additional details of the data distribution. These latter variants are less statistically informative but allow better visualization of the data distribution, such as bimodality (**Fig. 1b**), that may be hidden in a standard box plot.

Despite the obvious advantages of the box plot for simultaneous representation of data set and statistical parameters, this method is not in common use, in part because few available software tools allow the facile generation of box plots. For example, the standard spreadsheet tool Excel is unable to generate box plots. Here we describe an open-source application, called BoxPlotR, and an associated web portal that allow rapid generation of customized box plots. A user-defined data matrix is uploaded as a file or pasted directly into the application to generate a basic box plot with options for additional features. Sample size may be represented by the width of each box in proportion to the square root of the number of observations[5]. Whiskers may be defined according to the criteria of Spear[1], Tukey[2] or Altman[3]. The underlying data distribution may be visualized as a violin or bean plot or, alternatively, the actual data may be displayed as overlapping or nonoverlapping points. The 95% confidence interval that two medians are different may be illustrated as notches defined as $\pm(1.58 \times \mathrm{IQR}/\sqrt{n})$ (ref. 5). There is also an option to plot the sample means and their confidence intervals. More complex statistical comparisons may be required to ascertain significance according to the specific experimental design[6]. The output plots may be labeled; customized by color, dimensions and orientation; and exported as publication-quality .eps, .pdf or .svg files. To help ensure that generated plots are accurately described in publications, the application generates a description of the plot for incorporation into a figure legend.

The interactive web application is written in R (ref. 7) with the R packages shiny, beanplot[4], vioplot, beeswarm and RColorBrewer, and it is hosted on a shiny server to allow for interactive data analysis. User data are held only temporarily and discarded as soon as the session terminates. BoxPlotR is available at http://boxplot.tyerslab.com/ and may be downloaded to run locally or as a virtual machine for VMware and VirtualBox. .

© 2014 Nature America, Inc. All rights reserved.

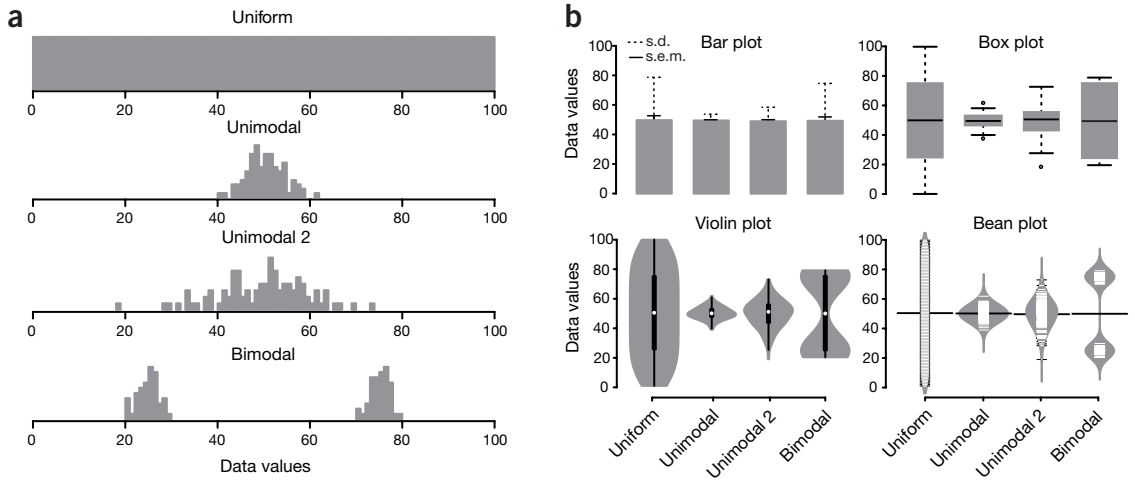

Figure 1 | Data visualization with box plots. (**a**) Hypothetical sample data sets of 100 data points each that are uniform, unimodal with one of two different variances or bimodal. Simple bar plot representations and statistical parameters may obscure such different data distributions. (**b**) Comparison of data visualization methods. Bar plots typically represent only the mean and s.d. or s.e.m. Box plots visualize the five-number summary of a data set (minimum, lower quartile, median, upper quartile and maximum). Violin and bean plots represent the actual distribution of the individual data sets.

ACKNOWLEDGMENTS
This work was supported by the Wellcome Trust through a Senior Research Fellowship to J.R. (084229), a core grant to the Wellcome Trust Centre for Cell Biology (092076), a European Research Council grant (233457) to M.T., a Genome Québec International Recruitment Award to M.T. and a Canada Research Chair in Systems and Synthetic Biology to M.T.

AUTHOR CONTRIBUTIONS
M.S. and J.W. conceived of the box plot tool, and M.S. developed the tool with input from all authors; J.W. implemented the server architecture; M.S., J.W., J.R. and M.T. wrote the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Michaela Spitzer[1,4], Jan Wildenhain[1], Juri Rappsilber[1,2] & Mike Tyers[1,3]

[1]Wellcome Trust Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK. [2]Department of Biotechnology, Technische Universität Berlin, Berlin, Germany. [3]Institute for Research in Immunology and Cancer, Department of Medicine, Université de Montréal, Montréal, Québec, Canada. [4]Present address: Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada.
e-mail: juri.rappsilber@ed.ac.uk or md.tyers@umontreal.ca

1. Spear, M.E. *Charting Statistics* (McGraw-Hill, 1952).
2. Tukey, J.W. *Exploratory Data Analysis* (Addison-Wesley, 1977).
3. Altman, D.G. *Practical Statistics for Medical Research* (Chapman and Hall, 1991).
4. Kampstra, P. *J. Stat. Softw.* **28**, c01 (2008).
5. McGill, R., Tukey, J.W. & Larsen, W.A. *Am. Stat.* **32**, 12–16 (1978).
6. Nieuwenhuis, S., Forstmann, B.U. & Wagenmakers, E.J. *Nat. Neurosci.* **14**, 1105–1107 (2011).
7. Ihaka, R. & Gentleman, R. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).

E-CRISP: fast CRISPR target site identification

To the Editor: Advances in genome sequencing technologies over the past years have led to a large increase in structural genomic data, but functional interpretation often lags behind. Scalable technologies to modulate gene function are necessary to link genotypes to phenotypes. To this end, clustered, regularly interspaced short palindromic repeats (CRISPR) in combination with a CRISPR-associated nuclease 9 (Cas9) were recently described as a new tool for genome engineering[1–6].

CRISPR-Cas was first discovered as a bacterial defense mechanism against foreign (viral) DNA[7,8]. Every CRISPR encodes an RNA (crRNA), consisting of a guide RNA (gRNA) and transactivating CRISPR RNA parts. A processed crRNA fragment is incorporated into the Cas9 protein, guiding it to the target DNA, where the Cas9 nuclease introduces a double-strand break[9,10]. The CRISPR-Cas system has been successfully used in human induced pluripotent stem cells, mice, zebrafish and flies, among other organisms, to disrupt gene function.

Here we describe E-CRISP, a web application to design gRNA sequences (**Fig. 1a**). It provides flexible output and experiment-oriented design parameters, enabling design of multiple libraries and thereby systematic analysis of the influence of different parameters. E-CRISP identifies target sequences complementary to the gRNA ending in a 3′ protospacer-adjacent motif (PAM), N(G or A)G, which is required for the recruited Cas9 nuclease to cut the DNA double strand. E-CRISP uses a fast indexing approach to find binding sites and a binary interval tree for rapid annotation of putative gRNA target sites (**Supplementary Note 1**). Using these algorithms, it is feasible to create genome-scale libraries for several organisms in a few hours. For instance, to design a library covering the *Drosophila melanogaster* genome requires less than 1 h (**Supplementary Fig. 1** and **Supplementary Table 1**).

Off-target effects and target-site homology are evaluated by E-CRISP using the alignment program Bowtie2 (**Supplementary Note 2**). Designs are shown in the output if the number of off-targets does not exceed a user-specified threshold. If more than one design is found targeting a desired locus, designs are ranked according to on-target specificity and number of off-targets. E-CRISP can also be used to reevaluate CRISPR constructs for on- or off-target sites and targeted genomic loci. As an example, we searched for designs to target *let-7* for gene disruption in zebrafish, fly, worm and human (**Fig. 1b**). We found at least one gRNA design per locus. In worm, fly and human, the cuts are located at the site that is transformed to mature microRNA and thus should lead to mutations blocking its proper function. In zebrafish the cut is located in the predicted hairpin structure.

E-CRISP is available for twelve organisms and can be easily extended. E-CRISP will help to further develop and deploy the

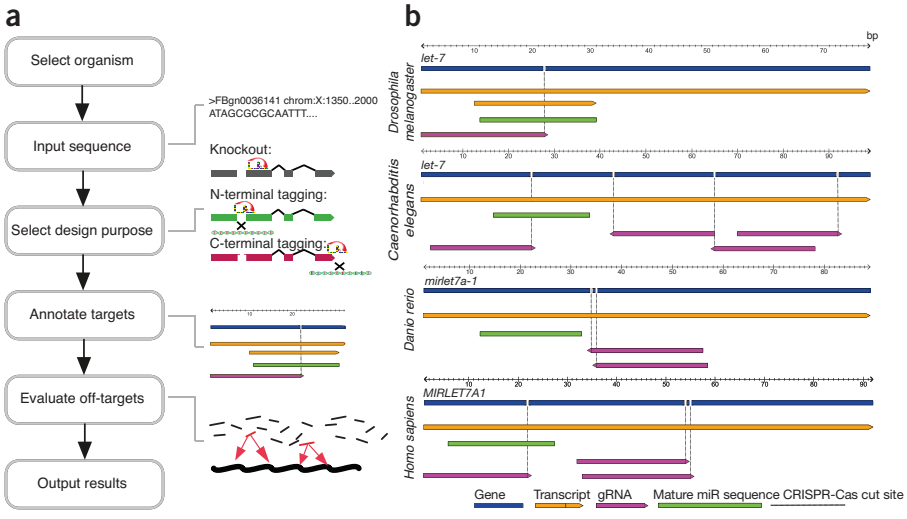

Figure 1 | E-CRISP workflow and example. (**a**) The main steps of the E-CRISP workflow. First, the user chooses the organism and the target sequence. This target can be a gene symbol, an ENSEMBL ID or a FASTA sequence. Second, the user specifies the purpose of the editing experiment. Depending on the purpose, E-CRISP will target different regions of the gene sequence. Third, E-CRISP filters the results according to gene annotation information. Fourth, off-targets are analyzed on the basis of sequence alignment of each design to the reference genome. Finally, E-CRISP produces a user-defined output page. (**b**) Guide RNAs were designed against the *let-7* locus of the indicated species. Sequence and location of the mature gRNAs have been retrieved from miRBase.

© 2014 Nature America, Inc. All rights reserved.