Open-source *guideseq* software for analysis of GUIDE-seq data

To the Editor:

We recently described the genome-wide, unbiased identification of double-strand breaks (DSBs) enabled by sequencing (GUIDE-seq) technology, a sensitive method for detecting global off-target DSBs induced by RNA-guided CRISPR–Cas9 nucleases in living cells[1]. The experimental component of GUIDE-seq is straightforward and encompasses capture of a double-stranded oligodeoxynucleotide (dsODN) into Cas9-induced DSBs in cells, selective amplification of these integration events, and next-generation sequencing of genomic DNA adjacent to the dsODN. However, analysis of the resulting sequencing data is a multistep process that, as described in our original published report[1], required multiple custom-built software components. Here we describe *guideseq*, a streamlined, open-source Python package that enables any user to readily perform analysis of GUIDE-seq experiment data (**Fig. 1a**). The software is simple to use and requires only basic technical knowledge to set up and run.

The *guideseq* software performs analysis based on raw sequencing data and a sample manifest in YAML format (http://yaml.org/). The sample manifest organizes the required information for bioinformatic analysis of GUIDE-seq runs, including the location of raw sequencing read files, the names of the biological samples and control, the sequences of dual-index barcodes, and the intended target site sequence.

In an initial step, our GUIDE-seq analysis pipeline prepares sequencing reads for alignment by demultiplexing a pooled multisample sequencing run into sample-specific read files. PCR duplicates are consolidated based on 8-bp unique molecular indexes (UMIs) in order to improve quantitative interpretation of GUIDE-seq read counts (https://github.com/aryeelab/umi). Next, off-target identification is performed through read alignment, site identification, false positive filtering, and reporting steps (**Supplementary Methods**). Off-target cleavage sites are sorted by GUIDE-seq read count, and figures are produced of the sequence alignment (**Fig. 1b**). The pipeline can either be run end-to-end with a single command or, if preferred, the component steps can be executed individually.

The *guideseq* Python package is provided under an open-source (AGPLv3) license and should broadly enable researchers to analyze GUIDE-seq experiments. Source code, installation, and up-to-date running instructions will be maintained at http://github.com/aryeelab/guideseq (see **Supplementary Note** for version of instructions at the time of this publication).

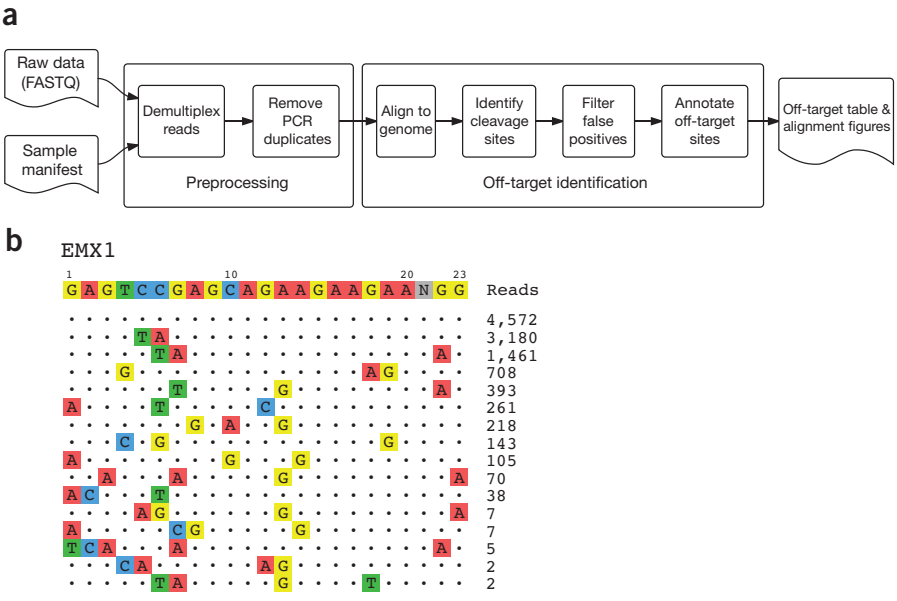

Figure 1 Overview of software analysis pipeline for processing of GUIDE-seq data and example output visualization. (**a**) Representation of data preprocessing and analysis pipeline for GUIDE-seq data by the *guideseq* program. (**b**) Example of an off-target sequence alignment visualization produced by the *guideseq* package using GUIDE-seq data for a gRNA targeted to the *EMX1* gene in human U2OS cells[1]. The top row is the intended on-target sequence, and the subsequent rows illustrate the alignment and GUIDE-seq read count of every detected DSB. Mismatches between a DSB and the on-target site are depicted by a colored box containing the mismatched base; otherwise, a black dot is shown.

The authors declare competing financial interests: details are available in the online version of the paper (doi:10.1038/nbt.3534).

Shengdar Q Tsai[1–4,6], Ved V Topkar[1–3,6], J Keith Joung[1–4] & Martin J Aryee[1,2,4,5]

[1]Molecular Pathology Unit, Massachusetts General Hospital, Charlestown, Massachusetts, USA. [2]Center for Cancer Research, Massachusetts General Hospital, Charlestown, Massachusetts, USA. [3]Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, Massachusetts, USA. [4]Department of Pathology, Harvard Medical School, Boston, Massachusetts, USA. [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [6]These authors contributed equally to this work.
e-mail: M.J.A. (aryee.martin@mgh.harvard.edu) or J.K.J. (jjoung@mgh.harvard.edu)

1. Tsai, S.Q. *et al. Nat. Biotechnol.* **33**, 187–197 (2015).

Editor's note: This article has been peer-reviewed.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nbt.3534).

ACKNOWLEDGMENTS

J.K.J. is supported by an NIH Director's Pioneer Award (DP1 GM105378) and the Jim and Ann Orr MGH Research Scholar Award. M.J.A. is supported by the MGH Department of Pathology.

AUTHORS CONTRIBUTIONS

S.Q.T. conceived and developed the initial GUIDE-seq analysis algorithm. M.J.A. developed UMI processing and PCR deduplication code. V.V.T. developed the software package infrastructure, filtering and visualization modules, and wrote documentation with input from M.J.A. and S.Q.T. J.K.J. and M.J.A. supervised the project. S.Q.T., V.V.T., J.K.J., and M.J.A. wrote the manuscript.

COMPETING FINANCIAL INTERESTS