MiTCR: software for T-cell receptor sequencing data analysis

To the Editor: High-throughput sequencing technologies have transformed the field of antigen receptor diversity studies, enabling deep and quantitative analysis for deciphering adaptive immunity function in health and disease[1–4]. As more data are produced each year, there is steadily growing demand for standardized analysis software.

Here we report MiTCR, an open-source software for rapid, robust and comprehensive analysis of hundreds of millions of raw high-throughput sequencing reads containing sequences encoding human or mouse α or β T-cell antigen receptor (TCR) chains (**Supplementary Software**). Raw data in FASTQ format generated via Illumina, 454 or Ion Torrent sequencing can be used as input for analysis. The only requirement is that sequence encoding conserved positions flanking the complementarity-determining region 3 (CDR3), Cys104 and Phe118 or Trp118, is covered by a sequencing read.

The approaches MiTCR uses to analyze TCR sequencing data have been shown to be efficient in previous studies[5,6]. The software performs CDR3 extraction, identifies V, D and J segments, assembles clonotypes, filters out or rescues low-quality reads[5] and provides advanced correction of PCR and sequencing errors[1,5] using either a predefined or user-specified strategy (**Fig. 1a** and **Supplementary Notes 1–3**). Simple command-line parameters, human-readable configuration files and a well-documented application programming interface (API) optimized for use in scripts make the software flexible enough for routine data extraction by immunologists as well as for more advanced analysis and customization by bioinformaticians (**Supplementary Note 1** and **Supplementary Data 1**).

We computationally optimized and parallelized the algorithms such that MiTCR can efficiently extract CDR3 information at a speed of more than 50,000 sequencing reads per second (0.3–0.6 gigabases min^{-1}) on standard PC hardware. For example, Illumina MiSeq run of 10 million reads can be analyzed in ~3 min, and a HiSeq lane of 100 million reads can be analyzed in ~20 min (**Supplementary Note 4**).

Output is provided in a tab-delimited text file that contains exhaustive information regarding TCR clonotype composition, abundance and aggregated sequence quality (**Supplementary Data 2**). Additionally, we developed MiTCR Viewer software that works with a custom (*.cls) format produced by MiTCR, enabling convenient visualization, filtering and *in silico* spectratyping of the data (http://mitcr.milaboratory.com/viewer/; **Supplementary Data 3**).

We demonstrated the accuracy and specificity of MiTCR for the analysis of both cDNA-based and genomic DNA–based high-throughput sequencing datasets (**Supplementary Tables 1** and **2**).

To verify software performance with datasets of known clonotype composition, we generated Illumina-like data sets *in silico*, based on real rates of PCR and sequencing errors (**Supplementary Note 5**). We determined the efficiency of human TCR-α and TCR-β CDR3 extraction, clonotype generation and error correction for the model data (**Fig. 1b,c**). Accuracy of V and J segment identification was 97–99%. MiTCR efficiency was superior compared to that of existing CDR3-extraction packages (**Supplementary Note 6**, **Supplementary Table 3** and **Supplementary Fig. 1**).

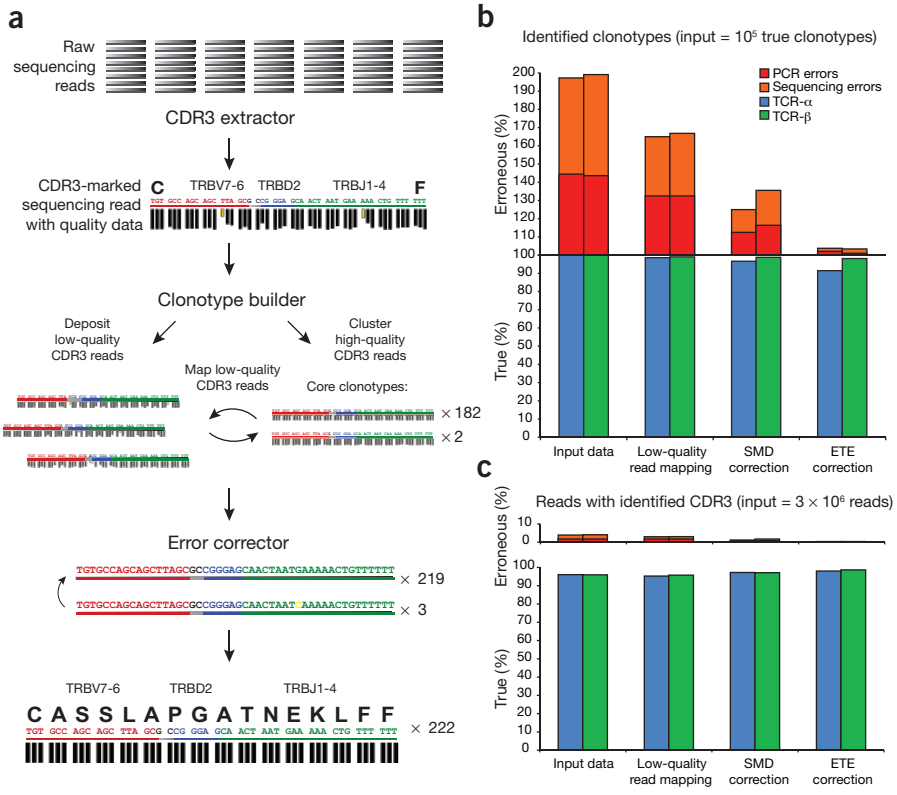

Figure 1 | MiTCR data analysis. (**a**) Analysis pipeline. Vertical bars show sequence quality. Horizontal bars represent raw sequencing reads. (**b**,**c**) Analysis of *in silico*–simulated data of 3 × 10^6 sequencing reads (average Phred quality = 33) containing 10^5 human TCR-α or TCR-β CDR3 clonotypes. Efficiency of CDR3 identification and correction of PCR and sequencing errors is shown for the input clonotypes (**b**) and CDR3-containing reads (**c**). The error-correction strategies (SMD, save my diversity and ETE, eliminate these errors) are described in **Supplementary Notes 1–3**.

MiTCR modules are checked by more than 80 comprehensive unit tests, which improved the reliability and correctness of the code. The MiTCR API package can be used in Java projects through Maven and in Groovy scripts using Groovy Grapes. MiTCR is regularly updated; Windows installer, cross-platform binaries and source code are available from http://mitcr.milaboratory.com/ under the terms of the GNU GPL v3 license.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2555).

ACKNOWLEDGMENTS

We are grateful to T. Schumacher and C. Linnemann for valuable technical discussions and to M. Eisenstein for English editing. This work was supported by the Molecular and Cell Biology program of the Russian Academy of Sciences and Russian Foundation for Basic Research (12-04-33139-mol-a, 12-04-33065-mol-a, 12-04-00229-a, 13-04-00998-a and 13-04-40251).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Dmitriy A Bolotin[1,3], Mikhail Shugay[1,3], Ilgar Z Mamedov[1], Ekaterina V Putintseva[1], Maria A Turchaninova[1], Ivan V Zvyagin[1,2], Olga V Britanova[1] & Dmitriy M Chudakov[1,2]

[1]Shemiakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia. [2]Central European Institute of Technology, Masaryk University, Brno, Czech Republic. [3]These authors contributed equally to this work. e-mail: chudakovdm@mail.ru

PUBLISHED ONLINE 28 JULY 2013; DOI:10.1038/NMETH.2555

1. Nguyen, P. *et al. BMC Genomics* **12**, 106 (2011).
2. Robins, H.S. *et al. Blood* **114**, 4099–4107 (2009).
3. Venturi, V. *et al. J. Immunol.* **186**, 4285–4294 (2011).
4. Warren, R.L. *et al. Genome Res.* **21**, 790–797 (2011).
5. Bolotin, D.A. *et al. Eur. J. Immunol.* **42**, 3073–3083 (2012).
6. Britanova, O.V. *et al. Bone Marrow Transplant.* **47**, 1479–1481 (2012).

ESTOOLS Data@Hand: human stem cell gene expression resource

To the Editor: We developed ESTOOLS Data@Hand, a resource to facilitate exploration of published gene expression array data in stem cell research. The resource, updated four times a year, offers efficient sample identification, preprocessing that enables cross-experiment comparisons and computational analysis.

High-throughput studies provide large amounts of data on human embryonic stem cells (hESCs) and human induced pluripotent stem cells (hiPSCs), their parent cells as well as their differentiated progeny cells[1]. Published genome-wide expression data on stem cells can be exploited to answer questions other than those addressed in the original studies. This is one reason why such data are actively stored in the public repositories ArrayExpress[2] and Gene Expression Omnibus (GEO)[3]. Nevertheless, the means to perform such reanalyses are currently limited. Often, sample information is only available as free text in publications or public databases, hindering identification of appropriate samples. Moreover, the measurement data are often available in heterogeneous formats, and the lack of systematic preprocessing hampers cross-experiment comparisons. Some databases have been developed for stem cell research (**Supplementary Note 1** and **Supplementary Table 1**), but none provide both a wide range of human pluripotent stem cell gene expression data and typical analysis tools.

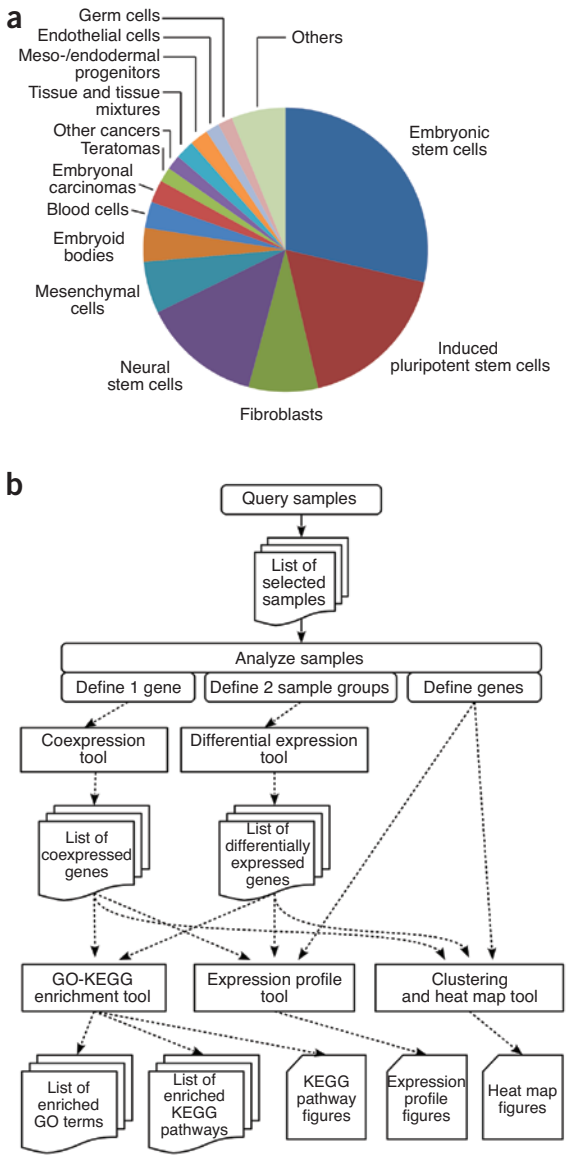

Figure 1 | Gene expression data and analysis workflows in ESTOOLS Data@Hand. (**a**) Cell and tissue types of the samples. (**b**) The data analysis steps available on the user interface.

ESTOOLS Data@Hand regroups gene expression array data and annotations primarily from experiments including hESCs or hiPSCs and thus involves stem cell pluripotency, differentiation and cell dedifferentiation. We selected data from GEO and ArrayExpress manually, preferentially including large experimental series; the selection covers pluripotent cells as well as dozens of other cell and tissue types reported in the same studies (**Fig. 1a**, **Supplementary Methods** and **Supplementary Tables 2–4**). Meta-analysis, a statistical approach to combine results from independent but related studies is a relatively inexpensive option that has the potential to increase both the statistical power and generalizability of single-study analysis[4]. We established two sample meta-datasets of 408 and 245 jointly normalized samples using the two most common array types for this data collection, Affymetrix and Illumina