dbGuide: a database of functionally validated guide RNAs for genome editing in human and mouse cells

Alexander A. Gooden†, Christine N. Evans†, Timothy P. Sheets†, Michelle E. Clapp and Raj Chari ‡

Laboratory Animal Sciences Program, Frederick National Lab for Cancer Research, Frederick, MD 21702, USA

Received August 14, 2020; Revised September 03, 2020; Editorial Decision September 14, 2020; Accepted September 22, 2020

ABSTRACT
With the technology’s accessibility and ease of use, CRISPR has been employed widely in many different organisms and experimental settings. As a result, thousands of publications have used CRISPR to make specific genetic perturbations, establishing in itself a resource of validated guide RNA sequences. While numerous computational tools to assist in the design and identification of candidate guide RNAs exist, these are still just at best predictions and generally, researchers inevitably will test multiple sequences for functional activity. Here, we present dbGuide (https://sgrnascorer.cancer.gov/dbguide), a database of functionally validated guide RNA sequences for CRISPR/Cas9-based knockout in human and mouse. Our database not only contains computationally determined candidate guide RNA sequences, but of even greater value, over 4000 sequences which have been functionally validated either through direct amplicon sequencing or manual curation of literature from over 1000 publications. Finally, our established framework will allow for continual addition of newly published and experimentally validated guide RNA sequences for CRISPR/Cas9-based knockout as well as incorporation of sequences from different gene editing systems, additional species and other types of site-specific functionalities such as base editing, gene activation, repression and epigenetic modification.

INTRODUCTION
CRISPR/Cas9-based genome editing has been an indispensable technology for understanding the biology of living organisms (1,2). As a result, a tremendous effort has been invested in aiding researchers in designing and executing CRISPR-based experiments by providing critical resources such as reagents and protocols as well as software tools that allow a user to rapidly identify guide RNA sequences with specific predicted on-target and off-target characteristics (3–18) (Table 1). However, these sequences are essentially predictions, often requiring assessment of multiple candidate guide RNAs.

Over the last 7 years, there has been an explosion of literature utilizing CRISPR/Cas9 to knockout specific genes or genomic regions, of which the majority is in human or mouse model systems. The culmination of all of this work has provided an opportunity to mine this information for guide RNA sequences shown to be functional at either a genotypic and/or phenotypic level. Variation in reporting standards, typographical errors and relegation of sequences into supplementary materials are among a few reasons why curation of this information has been non-trivial.

To this end, we present dbGuide, a database of functionally validated guide RNA sequences for CRISPR/Cas9 mediated knockout experiments in human or mouse cells (http://sgrnascorer.cancer.gov/dbguide). We have manually curated guide RNA sequences from over 1000 peer-reviewed articles with each sequence having direct reference to the original publication. We are also making results of targeted amplicon sequencing for ~2000 unique sgRNA sequences tested individually in human (293T) or mouse (NIH-3T3 or P19) cultured cells publicly searchable in our database. In total, these efforts encompass nearly 6000 unique guide RNA sequences for which some level of validated activity exists. To our knowledge, this represents the largest database of functionally validated guide RNA sequences. We expect this to be a continually growing resource through multiple mechanisms. We have provided a downloadable template to encourage researchers to submit their newly published/validated sequences and our computational framework will also allow for the inclusion of sequences for CRISPR systems other than for Streptococcus pyogenes, modalities such as base editing, gene activation/repression and epigenetic modifications, and sequences used in other species.

†To whom correspondence should be addressed. Tel: +1 301 846 711; Fax: +1 301 846 6165; Email: raj.chari@nih.gov
‡The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

© The Author(s) 2020. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Table 1. List of sgRNA design tools

| Name                        | URL                                      | Species         | Design? | Validated guides |
|-----------------------------|------------------------------------------|-----------------|---------|------------------|
| CCTop (12)                  | https://crispr.cos.uni-heidelberg.de/   | Many            | Yes     | No               |
| CHOPCHOP (6)                | http://chopchop.cbu.uib.no/             | Many            | Yes     | No               |
| CRISPOR (3)                 | http://crispor.tefor.net/               | Many            | Yes     | No               |
| CRISPRseek (16)             | http://bioconductor.org/packages/release/bioc/html/CRISPRseek.html | Many            | Yes     | No               |
| CRISPRz (13)                | https://research.nhgri.nih.gov/CRISPRz/ | Many            | Yes     | No               |
| CRISPR-ERA (7)              | http://crispr-era.stanford.edu/index.jsp| Many            | Yes     | No               |
| CRISPR MultiTargeter (11)   | http://www.multicrispr.net/             | Many            | Yes     | No               |
| Cas-Designer (8)             | http://www.rgenome.net/cas-designer/    | Many            | Yes     | No               |
| DeepHF (14)                 | http://www.deeplf.com/index/#/Predict   | Human           | No      | No               |
| E-CRISP (4)                 | http://www.e-crisp.org/E-CRISP/         | Many            | Yes     | No               |
| EuPaGDT (9)                 | http://grna.ctegd.uga.edu/              | Many            | Yes     | No               |
| inDelphi (18)               | https://indelphi.giffordlab.mit.edu/    | Human, Mouse    | Yes     | No               |
| sgRNAcas9 (15)              | http://www.biootools.com/software.html   | Many            | Yes     | No               |
| sgRNA Designer (17)         | https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design | Many            | Yes     | No               |
| sgRNA Scorer 2.0 (32)       | https://sgrnascorer.cancer.gov/          | Many            | Yes     | No               |
| WU-CRISPR (5)               | http://crispr.wustl.edu/                | Many            | Yes     | No               |
| COMMERCIAL                  |                                          |                 |         |                  |
| Addgene                     | https://www.addgene.org/crispr/reference/grna-sequence/ | Many            | No      | Yes              |
| CRISPR gRNA Design Tool     | https://www.atum.bio/eCommerce/cas9/input| Many            | Yes     | No               |
| Benchling                   | www.benchling.com                       | Many            | Yes     | No               |
| CRISPR Design Tool (Horizon Discovery) | https://horizondiscovery.com/en/products/tools/CRISPR-Design-Tool | Many            | No      | No               |
| CRISPR sgRNA Design Tool    | https://www.genscript.com/gRNA-design-tool.html | Many            | No      | No               |
| CRISPR Cas9 Guide RNA Checker (IDT DNA) | https://www.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE | Many            | Yes     | No               |
| True Designer Genome Editor (ThermoFisher Scientific) | https://www.thermofisher.com/us/en/home/life-science/genome-editing/geneart-crispr/geneart-crispr-search-and-design-tool.html | Many            | Yes     | No               |

MATERIALS AND METHODS

Master list of human and mouse sgRNA sequences

In addition to functionally validated sgRNA sequences, in order to provide further utility to the database, computationally designed sequences were also obtained from a variety of sources, primarily focused on protein-coding genes (17,19–26). These sources are listed in Supplementary Table S1. CSV-delimited files containing all sgRNA sequences present in our database for human (https://sgrnascorer.cancer.gov/downloads/hg_guide_info.csv.gz) and mouse (https://sgrnascorer.cancer.gov/downloads/mm_guide_info.csv.gz) can be downloaded directly from the dbGuide home page.

Published sgRNA sequences

A broad search of the PubMed database for ‘CRISPR OR Cas9’ was performed and yielded over 15 000 indexed citations. Subsequently, review articles, publications not using human or mouse cells, publications not using S. pyogenes Cas9 and publications not performing knock-out experiments were excluded. In total, guide RNA sequences were sourced from a total of 1322 peer-reviewed articles (Supplementary Table S2) from which the guide RNA sequences used could be ascertained. Although it is likely some published sequences have been missed, a data template is provided for which these missed sequences can be submitted and incorporated into subsequent database updates.

Targeted amplicon sequencing data

Quantitative editing data for nearly 2000 sgRNA sequences were generated internally from either transfection of Cas9/sgRNA plasmids (1 μg) or Cas9 protein (4 μg)/in vitro transcribed (IVT) (2.25 μg) RNA into mouse (NIH-3T3 or P19) or human (HEK293T) cells in a 24-well format. Cas9 protein was produced using plasmid Addgene-62731, a gift from Niels Geijsen (Addgene plasmid #62731; http://n2t.net/addgene:62731; RRID:Addgene62731) (27). IVT RNA was produced using a similar protocol as previously published (28). Cas9/RNP complexing and transfection using Lipofectamine 2000 were also performed similarly to as previously described (29). Polymerase chain reaction from genomic DNA was performed and amplicons were sequenced using the Illumina MiSeq V2 300 cycle kit using the PE (2 × 150) format. List of all sgRNAs sequences tested are listed in Supplementary Table S3.

Mapping information and on/off-target scoring metrics

For all sgRNA sequences obtained, genomic locations of the corresponding target sites were obtained/verified using UCSC BLAT against either the hg38 or mm10 reference genome sequences downloaded from the UCSC...
Figure 1. Structure of the dbGuide database. (A) Components of the dbGuide database. The user interface was developed using html and javascript and the application is managed on an Apache web server using django. All underlying data is stored in a MariaDB (MySQL) database. (B) Schema of the MySQL database. Each species, currently limited to human and mouse, has a table of sgRNA sequences with pre-calculated metrics and a table with gene annotation information. sgRNA sequences obtained from publications are stored in a single table and the ‘species’ field is used to determine which species the guide RNA was used. Similarly, sequences from targeted amplicon sequencing data also have the ‘species’ field for this purpose.

RESULTS AND DISCUSSION

dbGuide uses a simple HTML interface which utilizes the datatables and highcharts javascript libraries for displaying data in tabular and graphical formats, respectively (Figure 1A). The application is built in python using django with a MySQL database used for data storage and retrieval. The database can be accessed without e-mail registration or login.

Within the MySQL relational database, for both human and mouse, there was central sgRNA table with genome target position and all metrics pre-calculated and a gene annotation table which has the location of all protein coding genes based on Gencode annotation. Finally, there are two separate tables for summarizing the amplicon sequencing data and the publication-validated sequences. A depiction of the schema is shown in Figure 1B.

Opening user interface

The introductory user interface is very simple (Figure 2A). The user first specifies whether to search in the human or mouse genome and then can provide either a chromoso-
Figure 2. Screen shots of the key user interfaces. (A) Opening window of the dbGuide database. The user must first select whether to search within the human/mouse genome and then can specify a HGNC gene symbol, chromosomal coordinate in BED format, or ENSEMBL gene/transcript ID. (B) Window depicting results of a search by gene symbol. Selectable rows of sgRNA sequences are returned with various pre-computed metrics and PubMed identifiers, if the sequence had been used in a peer-reviewed publication. (C) Graphical representation of sgRNA sequences for which targeted amplicon sequencing data were generated. Stacked bar plots display the percentage of NHEJ mutation events that results in an ‘in-frame’ (black) or ‘out of frame’ (blue) amino acid change.
mal position (in BED format), gene symbol, *Ensembl* gene or transcript ID, or an sgRNA spacer sequence (without the PAM). In addition, a link to an excel spreadsheet template is provided for researchers wishing to contribute to the database.

Search results when searching by gene/coordinate/accession

Upon entering a valid search term, the user will be shown a table of results detailing all of the guide RNA sequences identified based on the search criteria (Figure 2B). The table displays the following columns:

- **Gene-Symbol**: most recent official gene symbol for a gene,
- **guide_rna**: sgRNA spacer sequence (without PAM),
- **crispr_system**: Cas9 ortholog, currently restricted to Sp-Cas9,
- **position**: genomic location of the sgRNA target site in either human (hg38) or mouse (mm10) assembly,
- **in_protein_coding_exon**: whether the guide RNA targets a protein coding exon,
- **num_transcripts**: the number of GENCODE transcripts targeted,
- **transcript_id_list**: list of GENCODE transcripts (by ENSEMBL ID),
- **sgrnascorer**: predicted activity using the sgRNA Scorer 2.0 algorithm (−3 to 3). Higher the value, greater the predicted activity,
- **rule_set_2**: predicted activity using the Rule Set 2 algorithm (0–1). Higher the value, greater the predicted activity,
- **guide_scan_off**: guidescan off-target score (0–1). Higher the value, higher the specificity,
- **forecast**: Favorable Outcomes of Repair Events at Cas9 targets (FORECasT) score for predicting mutational outcomes (in frame indel %, lower values preferred for knockouts),
- **total_nhej_rate**: if amplicon sequencing data exists, value between 0 and 100, or ‘None’ otherwise,
- **oof_rate**: out of frame mutation rate which is the proportion of mutated reads which would lead to an out of frame mutation,
- **num_publications**: number of published articles using the selected guide RNA sequence,
- **ind_pmid**: list of articles by PubMed ID which used selected guide RNA sequence,
- **num_screens**: number of pooled CRISPR screens in which the guide RNA sequences was enriched or depleted,
- **screen_pmid**: list of articles by PubMed ID of CRISPR-based pooled screens in which this guide was enriched or depleted,
- **sources**: list of sources from which the guide RNA sequence has been curated from,
- **status**: if the sgRNA sequence was functionally validated in a publication or targeted amplicon sequencing data exists, it is termed ‘validated’. Otherwise, it is termed ‘design’. This field allows the user to rapidly filter the list of returned sequences by typing ‘validate’ in the upper-right search box.

Viewing sequence editing data

In addition to sorting by total NHEJ rate, one can view the data in a graphical format. By selecting sgRNA sequences for which data exists, the user can click the ‘View Data’ which will show a stacked bar plot breaking down the total NHEJ rate between ‘in frame’ and ‘out of frame’ percentage (Figure 2C). This can be specifically important for knockout experiments where a higher out of frame (OOF) mutation percentage gives higher probability of protein loss.

Downloading sequences for use in experiments

Upon identifying guide RNA sequences of interest, the user can export the sequences in a ‘ready to order’ format by clicking on the ‘Design’. A tab-delimited text file is generated listing oligonucleotides needed for either ligation into a plasmid or *in vitro* transcription as well as a PDF protocol describing what compatible vectors can be used and how they can be obtained. It is highly recommended to select and download multiple guide RNA sequences when performing knockout experiments to ensure any functional consequences observed may not be due to spurious off-target activity from a single guide RNA.

**DATA AVAILABILITY**

Sorted BAM files for amplicon sequencing data are available from NCBI under Project Identifier PRJNA664634.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**ACKNOWLEDGEMENTS**

We would like to acknowledge Troy Taylor, William Gillette, Jane Jones and Dominic Esposito for kindly providing Cas9 protein. We would also like to acknowledge the Biomedical Informatics and Data Science (BIDS) and Frederick Research Computing Environment (FRCE) for providing assistance in web hosting and computational support.

**FUNDING**

Frederick National Laboratory for Cancer Research’s Technology and Training Fellowship (to T.P.S.). Funding for open access charge: Frederick National Laboratory for Cancer Research.

**Conflict of interest statement.** None declared.

**REFERENCES**

1. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, 339, 823–826.
2. Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339, 819–823.
3. Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J. *et al.* (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.*, 17, 148.
4. Heigwer, F., Kerr, G. and Boutros, M. (2014) E-CRISP: fast CRISPR target site identification. *Nat. Methods*, 11, 122–123.
5. Hiranniramol,K., Chen,Y., Liu,W. and Wang,X. (2020) Generalizable sgRNA design for improved CRISPR/Cas9 editing efficiency. *Bioinformatics*, 36, 2684–2689.

6. Labun,K., Montague,T.G., Krause,M., Torres Cleuren,Y.N., Tjeldnes,H. and Valen,E. (2019) CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.*, 47, W171–W174.

7. Liu,H., Wei,Z., Dominguez,A., Li,Y., Wang,X. and Qi,L.S. (2015) CRISPR-ERA: a comprehensive design tool for CRISPR-mediated gene editing, repression and activation. *Bioinformatics*, 31, 3676–3678.

8. Park,J., Bae,S. and Kim,J.-S. (2015) Cas-Designer: a web-based tool for choice of CRISPR-Cas9 target sites. *Bioinformatics*, 31, 4014–4016.

9. Peng,D. and Tarleton,R. (2015) EuPaGDT: a web tool tailored to design CRISPR guide RNAs for eukaryotic pathogens. *Microb. Genomics*, 1, e000033.

10. Pliatsika,V. and Rigoutsos,I. (2015) 'Off-Spotter': very fast and exhaustive enumeration of genomic lookalikes for designing CRISPR guide RNAs. *Biol. Direct*, 10, 4.

11. Przychozhij,S.V., Rajan,V., Gaston,D. and Berman,J.N. (2015) CRISPR multityrgeter: a web tool to find common and unique CRISPR single guide RNA targets in a set of similar sequences. *PLoS One*, 10, e0119372.

12. Stemmer,M., Thumberger,T., Del Sol Keyer,M., Wittbrodt,J. and Mateo,J.L. (2015) CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS One*, 10, e0124633.

13. Varshney,G.K., Zhang,S., Pei,W., Adomako-Ankomah,A., Fohntung,J., Schaffer,K., Carrington,B., Maskeri,A., Slevin,C., Wolfsberg,T. et al. (2016) CRISPRZ: a database of zebrafish validated sgRNAs. *Nucleic Acids Res.*, 44, D822–D826.

14. Wang,D., Zhang,C., Wang,B., Li,B., Wang,Q., Liu,D., Wang,H., Zhou,Y., Shi,L., Lan,F. et al. (2019) Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.*, 10, 4284.

15. Xie,S., Shen,B., Zhang,C., Huang,X. and Zhang,Y. (2014) sgRNAasCas9: a software package for designing CRISPR sgRNA and evaluating potential off-target cleavage sites. *PLoS One*, 9, e100448.

16. Zhu,L.J., Holmes,B.R., Aronin,N. and Brodsky,J.H. (2014) CRISPRseek: a bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. *PLoS One*, 9, e108424.

17. Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,J., Tothova,Z. and Wolf,Orchard,R. et al. (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, 34, 184–191.

18. Shen,M.W., Arbab,M., Hsu,J.Y., Worstell,D., Culbertson,S.J., Krabbe,O., Cress,A.C., Liu,D.R., Gifford,D.K. and Sherwood,R.I. (2018) Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*, 563, 646–651.

19. Doench,J.G., Hartenian,E., Graham,D.B., Tothova,Z., Hegde,M., Smith,J., Sullender,M., Ebert,B.L., Xavier,R.J. and Rost,D.E. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, 32, 1262–1267.

20. Sanjana,N.E., Shalem,O. and Zhang,F. (2014) Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods*, 11, 783–784.

21. Hart,T., Chandrashekar,M., Aregger,M., Steinhardt,Z., Brown,K.R., MacLeod,G., Mis,M., Zimmermann,M., Fradet-Turcotte,A., Sun,S. et al. (2015) High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, 163, 1515–1526.

22. Chari,R., Mali,P., Moosburner,M. and Church,G.M. (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods*, 12, 823–826.

23. Wang,T., Yu,H., Hughes,N.W., Liu,B., Kundir,A., Klein,K., Chen,W.W., Lander,E.S. and Sabatini,D.M. (2017) Gene essentiality profiling reveals genome networks and synthetic lethal interactions with oncogenic ras. *Cell*, 168, 890–903.

24. Wang,T., Birsoy,K., Hughes,N.W., Krupezk,M.K., Post,Y., Wei,J.J., Lander,E.S. and Sabatini,D.M. (2015) Identification and characterization of essential genes in the human genome. *Science*, 350, 1096–1101.

25. Tzelepis,K., Koike-Yusa,H., De Braekeleer,E., Li,Y., Metzakopan,E., Dovey,O.M., Mupo,A., Grinkevich,V., Li,M., Mazan,M. et al. (2019) A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukemia. *Cell Rep.*, 17, 1193–1205.

26. Henriksson,J., Chen,X., Gomes,T., Ullah,U., Meyer,K.B., Miraghaia,R., Duddy,G., Pramanik,J., Yusa,K., Lahesmaa,R. et al. (2019) Genome-wide CRISPR screens in T helper cells reveal pervasive crosstalk between activation and differentiation. *Cell*, 176, 882–896.

27. D’Astolfo,D.S., Pagliero,R.J., Pras,A., Karthaus,W.R., Cleverson,H., Prasad,V., Lebbink,R.J., Rehmann,H. and Geijsen,N. (2015) Efficient intracellular delivery of native proteins. *Cell*, 161, 674–690.

28. Richardson,C.D., Ray,G.J., Bray,N.L. and Corn,J.E. (2016) Non-homologous DNA increases gene disruption efficiency by altering DNA repair outcomes. *Nat. Commun.*, 7, 12463.

29. Kim,K., Park,S.W., Kim,J.H., Lee,S.H., Kim,D., Koo,T., Kim,K.-E., Kim,J.H. and Kim,J.-S. (2017) Genome surgery using Cas9 ribonucleoproteins for the treatment of age-related macular degeneration. *Genome Res.*, 27, 419–426.

30. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, 12, 656–664.

31. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, 12, 996–1006.

32. Chari,R., Yeo,N.C., Chavez,A. and Church,G.M. (2017) sgRNA Scorer 2.0: a species-independent model to predict CRISPR/Cas9 activity. *ACS Synth. Biol.*, 6, 902–904.

33. Allen,F., Crepaldi,L., Alsinet,C., Strong,A.J., Kleshchevnikov,V., De Angel,P., Palenikopov,P., Khodik,A., Kiselev,V., Kosicki,M. et al. (2018) Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.*, 37, 64–72.

34. Perez,A.R., Pritykin,Y., Vidigal,I.A., Chhangawala,S., Zamparo,L., Leslie,C.S. and Ventura,A. (2017) GuideScan software for improved single and paired CRISPR guide RNA design. *Nat. Biotechnol.*, 35, 347–349.

35. Magoč,T. and Salzberg,S.L. (2011) FLASH: fast length adjustment of random reads to improve genome assemblies. *Bioinformatics*, 27, 2957–2963.

36. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data. (2009) The sequence alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

37. Koster,J. and Balenine,S. (2018) SnakeMake—a scalable bioinformatics workflow engine. *Bioinformatics*, 34, 3600.