SOFTWARE TOOL ARTICLE

The khmer software package: enabling efficient nucleotide sequence analysis [version 1; peer review: 2 approved, 1 approved with reservations]

Michael R. Crusoe\textsuperscript{1}, Hussien F. Alameldin\textsuperscript{2}, Sherine Awad\textsuperscript{3}, Elmar Boucher\textsuperscript{4}, Adam Caldwell\textsuperscript{5}, Reed Cartwright\textsuperscript{6}, Amanda Charbonneau\textsuperscript{7}, Bede Constantinides\textsuperscript{8}, Greg Edvenson\textsuperscript{9}, Scott Fay\textsuperscript{10}, Jacob Fenton\textsuperscript{11}, Thomas Fenzl\textsuperscript{12}, Jordan Fish\textsuperscript{11}, Leonor Garcia-Gutierrez\textsuperscript{13}, Phillip Garland\textsuperscript{14}, Jonathan Gluck\textsuperscript{15}, Iván González\textsuperscript{16}, Sarah Guermond\textsuperscript{17}, Jiarong Guo\textsuperscript{18}, Aditi Gupta\textsuperscript{1}, Joshua R. Herr\textsuperscript{1}, Adina Howe\textsuperscript{19}, Alex Hyer\textsuperscript{20}, Andreas Härpfer\textsuperscript{21}, Luiz Irber\textsuperscript{11}, Rhys Kidd\textsuperscript{22}, David Lin\textsuperscript{23}, Justin Lippi\textsuperscript{24}, Tamer Mansour\textsuperscript{3,25}, Pamela McA’Nulty\textsuperscript{26}, Eric McDonald\textsuperscript{11}, Jessica Mizzi\textsuperscript{27}, Kevin D. Murray\textsuperscript{28}, Joshua R. Nahum\textsuperscript{29}, Kaben Nanlohy\textsuperscript{30}, Alexander Johan Nederbragt\textsuperscript{31}, Humberto Ortíz-Zuazaga\textsuperscript{32}, Jeramia Ory\textsuperscript{33}, Jason Pell\textsuperscript{11}, Charles Pepe-Ranney\textsuperscript{34}, Zachary N. Russ\textsuperscript{35}, Erich Schwarz\textsuperscript{36}, Camille Scott\textsuperscript{11}, Josiah Seaman\textsuperscript{37}, Scott Sievert\textsuperscript{38}, Jared Simpson\textsuperscript{39,40}, Connor T. Skennerton\textsuperscript{41}, James Spencer\textsuperscript{42}, Ramakrishnan Srinivasan\textsuperscript{43}, Daniel Standage\textsuperscript{44,45}, James A. Stapleton\textsuperscript{46}, Susan R. Steinman\textsuperscript{47}, Joe Stein\textsuperscript{48}, Benjamin Taylor\textsuperscript{11}, Will Trimble\textsuperscript{49}, Heather L. Wiencko\textsuperscript{50}, Michael Wright\textsuperscript{11}, Brian Wyss\textsuperscript{11}, Qingpeng Zhang\textsuperscript{11}, en zyme\textsuperscript{51}, C. Titus Brown\textsuperscript{1,3,11}

\textsuperscript{1}Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA
\textsuperscript{2}Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI, USA
\textsuperscript{3}Population Health and Reproduction, University of California, Davis, Davis, CA, USA
\textsuperscript{4}Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR, USA
\textsuperscript{5}Biology Department, San Jose State University, San Jose, CA, USA
\textsuperscript{6}School of Life Sciences and The Biodesign Institute, Arizona State University, Tempe, AZ, USA
\textsuperscript{7}Genetics, Michigan State University, East Lansing, MI, USA
\textsuperscript{8}Computational and Evolutionary Biology, Faculty of Life Sciences, University of Manchester, Manchester, UK
\textsuperscript{9}Micron Technology, Seattle, WA, USA
\textsuperscript{10}Invitae, San Francisco, CA, USA
\textsuperscript{11}Computer Science and Engineering, Michigan State University, East Lansing, MI, USA
\textsuperscript{12}Independent Researcher, Munich, Germany
\textsuperscript{13}Mathematics Institute, University of Warwick, Warwick, UK
\textsuperscript{14}Eastlake Data, Seattle, WA, USA
\textsuperscript{15}Graduate Program, University of Maryland, College Park, MD, USA
\textsuperscript{16}Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, USA
\textsuperscript{17}Independent Researcher, Seattle, WA, USA
\textsuperscript{18}Center for Microbial Ecology, Michigan State University, East Lansing, MI, USA
Abstract
The khmer package is a freely available software library for working efficiently with fixed length DNA words, or k-mers. khmer provides implementations of a probabilistic k-mer counting data structure, a compressible De Bruijn graph representation, De Bruijn graph partitioning, and digital normalization. khmer is implemented in C++ and Python, and is freely available under the BSD license at https://github.com/dib-lab/khmer/.

Keywords
bioinformatics, dna sequencing analysis, k-mer, kmer, khmer, online, low-memory, streaming
Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** C. Titus Brown ([titus@idyll.org](mailto:titus@idyll.org))

**Competing interests:** No competing interests were disclosed.

**Grant information:** khmer development has largely been supported by AFRI Competitive Grant no. 2010-65205-20361 from the USDA NIFA, and is now funded by the National Human Genome Research Institute of the National Institutes of Health under Award Number R01HG007513, as well as by the the Gordon and Betty Moore Foundation under Award number GBMF4551, all to CTB.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Copyright:** © 2015 Crusoe MR *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Crusoe MR, Alameldin HF, Awad S *et al*. The khmer software package: enabling efficient nucleotide sequence analysis [version 1; peer review: 2 approved, 1 approved with reservations] F1000Research 2015, 4:900 (https://doi.org/10.12688/f1000research.6924.1)

**First published:** 25 Sep 2015, 4:900 (https://doi.org/10.12688/f1000research.6924.1)
**Introduction**

DNA words of a fixed-length k, or “k-mers”, are a common abstraction in DNA sequence analysis that enable alignment-free sequence analysis and comparison. With the advent of second-generation sequencing and the widespread adoption of De Bruijn graph-based assemblers, k-mers have become even more widely used in recent years. However, the dramatically increased rate of sequence data generation from Illumina sequencers continues to challenge the basic data structures and algorithms for k-mer storage and manipulation. This has led to the development of a wide range of data structures and algorithms that explore possible improvements to k-mer-based approaches.

Here we present version 2.0 of the khmer software package, a high-performance library implementing memory- and time-efficient algorithms for the manipulation and analysis of short-read data sets. khmer contains reference implementations of several approaches, including a probabilistic k-mer counter based on the CountMin Sketch1, a compressible De Bruijn graph representation built on top of Bloom filters2, a streaming lossy compression approach for short-read data sets termed “digital normalization”3, and a generalized semi-streaming approach for k-mer spectral analysis of variable-coverage shotgun sequencing data sets4.

khmer is both research software and a software product for users: it has been used in the development of novel data structures and algorithms, and it is also immediately useful for certain kinds of data analysis (discussed below). We continue to develop research extensions while maintaining existing functionality.

The khmer software consists of a core library implemented in C++, a CPython library wrapper implemented in C, and a set of Python “driver” scripts that make use of the library to perform various sequence analysis tasks. The software is currently developed on GitHub under https://github.com/dib-lab/khmer, and it is released under the BSD License. There is greater than 87% statement coverage under automated tests, measured on both C++ and Python code but primarily executed at the Python level.

**Methods**

**Implementation**

The core data k-mer counting data structures and graph traversal code are implemented in C++, and then wrapped for Python in hand-written C code, for a total of 10.5k lines of C/C++ code. The command-line API and all of the tests are written in 13.7k lines of Python code. C++ FASTQ and FASTA parsers came from the SeqAn library5.

Documentation is written in reStructuredText, compiled with Sphinx, and hosted on ReadTheDocs.org.

We develop khmer on github.com as a community open source project focused on sustainable software development6, and encourage contributions of any kind. As an outcome of several community events, we have comprehensive documentation on contributing to khmer at https://khmer.readthedocs.org/en/latest/dev/. Most development decisions are discussed and documented publicly as they happen.

**Operation**

khmer is primarily developed on Linux for Python 2.7 and 64-bit processors, and several core developers use Mac OS X. The project is tested regularly using the Jenkins continuous integration system running on Ubuntu 14.04 LTS and Mac OS X 10.10; the current development branch is also tested under Python 3.3, 3.4, and 3.5. Releases are tested against many Linux distributions, including RedHat Enterprise Linux, Debian, Fedora, and Ubuntu. khmer should work on most UNIX derivatives with little modification. Windows is explicitly not supported.

Memory requirements for using khmer vary with the complexity of data and are user configurable. Several core data structures can trade memory for false positives, and we have explored these details in several papers, most notably Pell et al. 20127 and Zhang et al. 20148. For example, most single organism mRNAseq data sets can be processed in under 16 GB of RAM9,10, while memory requirements for metagenome data sets may vary from dozens of gigabytes to terabytes of RAM.

The user interface for khmer is via the command line. The command-line interface consists of approximately 25 Python scripts; they are documented at http://khmer.readthedocs.org/ under User Documentation. Changes to the interface are managed with semantic versioning9 which guarantees command line compatibility between releases with the same major version.

khmer also has an unstable developer interface via its Python and C++ libraries, on which the command line scripts are built.

**Use cases**

khmer has several complementary feature sets, all centered on short-read manipulation and filtering. The most common use of khmer is for preprocessing short read Illumina data sets prior to de novo sequence assembly, with the goals of decreasing compute requirements for the assembler as well as potentially improving the assembly results.

**Prefiltering sequence data for de novo assembly with digital normalization**

We provide an implementation of a novel streaming “lossy compression” algorithm in khmer that performs abundance normalization of shotgun sequence data. This “digital normalization” algorithm eliminates redundant short reads while retaining sufficient information to generate a contig assembly11. The algorithm takes advantage of the online k-mer counting functionality in khmer to estimate per-read coverage as reads are examined; reads can then be accepted as novel or rejected as redundant. This is a form of error reduction, because the net effect is to decrease not only the total number of reads considered for assembly, but also the total number of errors considered by the assembler. Digital normalization results in a decrease of the amount of memory needed for de novo assembly of high-coverage data sets with little to no change in the assembled contigs.

Digital normalization is implemented in the script normalize-by-median.py. This script takes as input a list of FASTA or FASTQ files, which it then filters by abundance as described above;
see 3 for details. The output of the digital normalization script is a
downsampl ed set of reads, with no modifications to the individual
reads. The three key parameters for the script are the k-mer size, the
desired coverage level, and the amount of memory to be used for
k-mer counting. The interaction between these three parameters and
the filtering process is complex and depends on the data set being
processed, but higher coverage levels and longer k-mer sizes result
in less data being removed. Lower memory allocation increases the
rate at which reads are removed due to erroneous estimates of their
abundance, but this process is very robust in practice'.

The output of normalize-by-median.py can be assembled using a de novo assembler such as Velvet[6], IDBA[11], Trinity[12] or SPAdes[13].

K-mer counting and read trimming
Using a memory-efficient CountMin Sketch data structure, khmer
provides an interface for online counting of k-mers in streams
of reads. The basic functionality includes calculating the k-mer
frequency spectrum in sequence data sets and trimming reads at
low-abundance k-mers. This functionality is explored and bench-
marked in 1.

Basic read trimming is performed by the script filter-abund.py, which takes as arguments a k-mer count graph (created by
khmer's load-into-counting.py script) and one or more
sequence data files. The script examines each sequence to find
k-mers below the given abundance cutoff, and truncates the
sequence at the first such k-mer. This truncates reads at the location
of substitution errors produced by the sequencing process. When
processing sequences from variable coverage data sets, filter-
abund.py can also be configured to ignore reads that have low
estimated abundance.

K-mer abundance distributions can be calculated using the script
abundance-dist.py, which takes as arguments a k-mer count
graph, a sequence data file, and an output filename. This script
determines the abundance of each distinct k-mer in the data file
according to the k-mer count graph, and summarizes the abundances
in a histogram output.

We recently extended digital normalization to provide a general-
ized semi-streaming approach for k-mer spectral analysis'. Here,
we examine read coverage on a per-locus basis in the De Bruijn
graph and, once a particular locus has sufficient coverage, call
errors or trim bases for all following reads belonging to that graph
locus. The approach is "semi-streaming" because some reads must
be examined twice. This semi-streaming approach enables few-pass
analysis of high coverage data sets. More, the approach also makes
it possible to apply k-mer spectral analysis to data sets with uneven
coverage such as metagenomes, transcriptomes, and whole-genome
amplified samples.

Because our core data structure sizes are preallocated based on
estimates of the unique k-mer content of the data, we also provide
fast and low-memory k-mer cardinality estimation via the script
unique-kmers.py. This script uses the HyperLogLog algorithm
to provide a probabilistic estimate of the number of unique k-mers
in a data set with a guaranteed upper bound'. A manuscript on this
implementation is in progress (Irber and Brown, unpublished).

Partitioning reads into disconnected assembly graphs
We have also built a De Bruijn graph representation on top of a
Bloom filter, and implemented this in khmer. The primary use for
this so far has been to enable memory efficient graph partitioning, in
which reads contributing to disconnected subgraphs are placed into
different files. This can lead to an approximately 20-fold decrease
in the amount of memory needed for metagenome assembly', and
may also separate reads into species-specific bins'.

Reformatting collections of short reads
In support of the streaming nature of this project, our preferred
paired-read format is with pairs interleaved in a single file. As an
extension of this, we automatically support a "broken-paired" read
format where orphaned reads and pairs coexist in a single file. This
enables single input/output streaming connections between tools,
while leaving our tools compatible with fully paired read files as
well as files containing only orphaned reads.

For converting to and from this format, we supply the scripts
extract-paired-reads.py, interleave-reads.py, and split-paired-reads.py to respectively extract fully
paired reads from sequence files, interleave two files containing
read pairs, and split an interleaved file into two files containing read
pairs.

In addition, we supply several utility scripts that we use in our
own work. These include sample-reads-randomly.py for
performing reservoir sampling of reads and readstats.py for
summarizing sequence files.

Summary
The khmer project is an increasingly mature open source scientific
software project that provides several efficient data structures and
algorithms for analyzing short-read nucleotide sequencing data.
khmer emphasizes online analysis, low memory data structures
and streaming algorithms. khmer continues to be useful for both
advancing bioinformatics research and analyzing biological data.

Software availability
Software available from
https://khmer.readthedocs.org/en/v2.0/

Link to source code
https://github.com/dib-lab/khmer/releases/tag/v2.0

Link to archived source code as at time of publication
http://dx.doi.org/10.5281/zenodo.312581

Software license
Copyright: 2015, The Regents of the University of California.
Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of the Michigan State University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Author contributions
CTB is the primary investigator for the khmer software package. MRC is the lead software developer from July 2013 onwards. Many significant components of khmer have their own paper describing them (see “Use Cases”, above). The remaining authors each have one or more Git commits in their name.

Competing interests
No competing interests were disclosed.

Grant information
khmer development has largely been supported by AFRI Competitive Grant no. 2010-65205-20361 from the USDA NIFA, and is now funded by the National Human Genome Research Institute of the National Institutes of Health under Award Number R01HG007513, as well as by the the Gordon and Betty Moore Foundation under Award number GBMF4551, all to CTB.

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

This paper describes version 2 of the khmer software suite. The software is developed to provide both a set of directly usable tools (e.g. normalize-by-median for digital normalization) as well as an experimental framework for developers looking to design new algorithms and methods. It has proven very useful on both these fronts. The repository is highly watched and starred on GitHub, the developers are very responsive (see more below), and both the senior author's group and other researches seem to be leveraging this framework to build new tools and algorithms.

The paper itself does a good job of describing the software at a high level, including the overall design and goals. I would have appreciated slightly more detail about the motivation behind the design decisions, and the tradeoffs they entail (e.g. Why have a Python front-end? Why use hand-written binding code rather than a binding generator, like SWIG, that would allow interfaces to other languages as well?). I understand that a comprehensive description is not feasible in a manuscript of this length. It would be very interesting to know, however, the cost paid for using the high-level interface rather than the C++ library directly. When the underlying computation is trivial, simply having to iterate over an enormous number of things in Python could add non-trivial overhead. Despite these desiderata, I find that the paper is generally well written and does a good job of describing what a new user might want to know about khmer, and so I approve of this manuscript.

Like Daniel, I also downloaded and built the software using the instructions provided in the ReadTheDocs documentation. The process was simple, and worked well, with the exception of a minor glitch running the tests. After debugging the cause of the problem, I posted an issue to the GitHub repository, and received a response in less than a day. I bring this up because, while not an aspect of the paper itself, good developer support is crucial to the long-term survival and utility of a software package — khmer seems to have this.

This brings me to my final point, about the (currently) controversial authorship policy on this paper, which is ancillary to the quality of the paper (and software) itself. At this point, I must reserve judgement on whether I think the authorship policy adopted by this paper is “good” or “bad” (for science, the community, etc.). Incidentally, this is a dichotomy that does not capture the subtlety or importance of this issue well. In the manuscript, the authors state "We develop khmer on github.com as a community open source project focused on sustainable software development, and encourage contributions of any kind." Thus, contributions to khmer are of a potentially wide variety in character (and also, I believe, not simply related to improving or maintaining the code). Those who contribute to the design, improve the usability, work on documentation, support new and existing users, and develop and propagate best practices are all
contributing something valuable to the khmer software “ecosystem”. It is unreasonable to expect a piece of software that is ~25k lines of code (and growing) to be actively developed, maintained, and supported by only a small contingent of people, many of whom may be graduate students soon to graduate and move on. Thus, if we are interested in the long-term viability and quality of such software, we must adopt a system of credit that values and recognizes a variety of different types of contribution. On the other hand, I do share the concern that, in the midst of the current authorship system, bestowing that recognition in the form of authorship may have the adverse effect of diminishing the public perception of the very credit one is trying to grant. Perhaps there is a solution along the lines adopted by this paper, or perhaps something drastically different needs to be considered.

**Competing Interests:** No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

---

Daniel S. Katz
Computation Institute, University of Chicago, Chicago, IL, USA

Regarding the paper, it is a fairly straightforward description of a software package, containing all the things that such a paper should have - a description of the goals, the implemented methods, the hardware and software dependencies (systems on which the software has been tested), some guidance on usage, pointers to the software and documentation, and references.

Regarding the software, I did download and build the software, which seemed to work, other than a fair number of warnings. I was not able to successfully test the software, however, due to issues in https://khmer.readthedocs.org/en/v2.0/user/install.html#run-the-tests Does this mean I should not approve the article? Or should I ask the authors for help in understanding the error and hold off on submitting this report?

I would have liked to have chosen "Approved with reservations" for the status of this review, but my reservations are with the F1000 system for this type of paper, not with this specific paper, so in fairness to the authors, given the lack of clarity of what I should be doing as a reviewer for a software paper, I approve this paper based on its quality as a good description of the software, and not on the quality of software (and related documentation) itself.

**Competing Interests:** No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

---

Reviewer Response 06 Oct 2015

Daniel S. Katz, University of Illinois at Urbana-Champaign, USA
In addition to my report, regarding software papers in general under F1000, I believe that much more should be required from their reviewers, and what is required should be made clear. Software journals (e.g., Ubiquity Press's Journal of Open Research Software, Elsevier's Software X) have specific statements of what a reviewer should do, which say a lot about the quality of the review. For JORS, this is defined on a web page (http://openresearchsoftware.metajnl.com/about/editorialPolicies/). For Software X, the criteria are not on the web (as far as I know) but are embedded in the review form/process, and are roughly equivalent.

**Competing Interests:** none

---

Referee Report 05 October 2015

https://doi.org/10.5256/f1000research.7456.r10514

Ewan Birney
European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK

This is an update of a widely used tool, khmer, which is in broad use in the technical community around de Bruijn graphs and short reads, based on Bloom filters. It is a good update, provides link to the code, and is sensibly written with tests. I have no concern about the scientific aspect of this paper.

I do find the author inclusion list taking a concept and going to the extreme, and I don't think it is sensible to have an anonymised author (en zyme) on the list, with in effect no way to attribute to a person this. Science's openness in publication is also about attribution. Although I understand Titus' consistency of having all git committers as authors, I think it is sensible to make a distinction of substantial/scientific changes, of which the vast majority of the authors are. Acknowledgements are precisely there to handle these other cases.

I believe it is uncontroversial to appropriately trim the author list, to use the acknowledgements for anonymous improvement (happens regularly in science) and small details (again, a commonplace practice).

**Competing Interests:** No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
| The peer review process is transparent and collaborative |
| Your article is indexed in PubMed after passing peer review |
| Dedicated customer support at every stage |

For pre-submission enquiries, contact research@f1000.com