MMseqs2: sensitive protein sequence searching for analysis of massive data sets

Martin Steinegger[1,2] **& Johannes Söding**[1]

[1]Quantitative and Computational Biology group, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany; [2]Department for Bioinformatics and Computational Biology, Technische Universität München, 85748 Garching, Germany

e-mail: johannes.soeding@mpibpc.mpg.de; martin.steinegger@mpibpc.mpg.de

Supplementary Methods

Overview. MMseqs2 (Many-against-Many sequence searching) is a software suite to search and cluster huge protein sequence sets. MMseqs2 is open source GPL-licensed software implemented in C++ for Linux and Mac OS. At the core of MMseqs2 is its sequence search module. It searches and aligns a set of query sequences against a set of target sequences. Queries are processed in three consecutive stages of increasing sensitivity and decreasing speed (**Fig. 1a**): (1) the fast k-mer match stage filters out 99.9 % of sequences, (2) the ungapped alignment stage filters out a further 99 %, and (3) the accurate, vectorized Smith-Waterman alignment thus only needs to align $\sim 10^{-5}$ of the target sequences.

MMseqs2 builds on our MMseqs software suite[8], designed for fast sequence clustering and searching of globally alignable sequences. The k-mer match stage of MMseqs2, which is crucial for its improved sensitivity-speed trade-off, has been developed from scratch, profile-to-sequence and sequence-to-profile searching capabilities have been developed, and many other powerful features and utilities have been added (see **Supplemental Table S2** for an overview of differences).

The software is designed to run on multiple cores and servers and exhibits nearly linear scalability. It makes extensive use of single instruction multiple data (SIMD) vector units which are part of modern Intel and AMD CPUs. For older CPUs without AVX2 support, MMseqs2 falls back to SSE4.1 instructions throughout with minimal speed loss.

k-mer match stage. Most fast methods follow a seed-and-extend approach: a fast seed stage searches for short-word ("k-mer") matches which are then extended to a full, gapped alignment. Since the k-mer match stage processes all sequences, it needs to be much faster than the subsequent stages. Its sensitivity is therefore crucial for the overall search sensitivity. In contrast to BLAST and SWORD[21], most fast methods index the database k-mers instead of the query sequences, using hashes or suffix arrays, and a few index both to streamline random memory access during the identification of k-mer matches[13;9;2]. To increase the seeds' sensitivity, some methods allow for one or two mismatched positions[12;9], others employ reduced alphabets[20;23;9;2]. Many use spaced k-mer seeds to reduce spatial clustering of chance matches[9;2].

Whereas most other tools use only single, exact k-mer matches as seeds, the k-mer match stage of MMseqs2 detects double, consecutive, *similar*-k-mer matches occurring on the *same diagonal* $i-j$. i is position of the k-mer in the query and j is the position of the matching k-mer in the target sequence. This criterion very effectively suppresses chance k-mer matches between nonhomologous sequences as these have a probability of only $\sim 1/(L_{\text{query}} + L_{\text{target}})$ to have coinciding diagonals. In contrast, consecutive k-mer matches between homologous sequences lie on the same diagonal if no alignment insertion or deletion occurred between them. A similar criterion is used in the earlier, two-hit 3-mer seed strategy of BLAST[1]. (The new version reverts to a single-hit strategy but uses 6-mers on a reduced size-15 alphabet instead of 3-mers.[17].)

The double k-mer match criterion effectively bases our decision to trigger the next search stage on 11 to 14 amino acids (depending on the number of common residues between seeds) instead of just 2×3 in BLAST or 12 letters on a size-11 reduced alphabet in DIAMOND. (Even though DIAMOND uses 4 seed patterns with 12 letters, it takes independent decisions for each of the patterns.) The information content in each of the reduced-alphabet letters is 3.0 in contrast to 4.2 bits for the full 20-letter amino acid alphabet.

Query sequences are searched one by one against the target set (**Fig. 1b**, loop 1). For each k-mer starting position in the query (loop 2) we generate a list of all similar k-mers (orange frame) with a Blosum62 similarity above a threshold score. Lower threshold scores (option --k-score <int>) result in higher average numbers of similar k-mers and thereby higher sensitivity and lower speed. The similar k-mers are generated with a linear-time branch-and-bound algorithm[7].

For each k-mer in the list of similar k-mers (loop 3), we obtain from the index table (blue frame) the list of target sequence identifiers `target_ID` and the positions j of the k-mer (green frame). In the innermost loop 4 we go through this list to detect double k-mer matches by comparing the current diagonal $i-j$ with the previously matched diagonal for `target_ID`. If the previous and current

diagonals agree, we store the diagonal $i-j$ and `target_ID` as a double match. Below, we describe how we carry out this computation within low-level, fast CPU cache without random memory access.

Minimizing random memory access. Due to the increase in the number of cores per CPU and the stagnation in main memory speeds in the last decade, main memory access is now often the chief bottleneck for compute-intensive applications. Since it is shared between cores, it also severely impairs scalability with the number of cores. It is therefore paramount to minimize random memory accesses.

We want to avoid the random main memory access to read and update the value of `diagonal_prev[target_ID]` in the innermost loop. We therefore merely write `target_ID` and the diagonal $i-j$ serially into an array `matches` for later processing. Because we write linearly into memory and not at random locations, these writes are automatically buffered in low-level cache by the CPU and written to main memory in batches with minimal latency. When all k-mers from the current query have been processed in loop 2, the `matches` array is processed in two steps to find double k-mer matches. In the first step, the entries (`target_ID`, $i-j$) of `matches` are sorted into 2^B arrays (bins) according to the lowest B bits of `target_ID`, just as in radix sort. Reading from `matches` is linear in memory, and writing to the 2^B bins is again automatically buffered by the CPU. In the second step, the 2^B bins are processed one by one. For each k-mer match (`target_ID`, $i-j$), we run the code in the magenta frame of Fig. 1b. But now, the `diagonal_prev` array fits into L1/L2 CPU cache, because it only needs $\sim N/2^B$ entries, where N is the number of target database sequences. To minimize the memory footprint, we store only the lowest 8 bits of each diagonal value in `diagonal_prev`, reducing the amount of memory to $\sim N/2^B$ bytes. For example, in the 256 KB L2 cache of Intel Haswell CPUs we can process a database of up to 256K $\times\, 2^B$ sequences. To match L2 cache size to the database size, MMseqs2 sets $B = \texttt{ceil}(\texttt{log2}(N/\texttt{L2\_size}))$.

Index table generation. For the k-mer match stage we preprocess the target database into an index table. It consists of a pointer array (black frame within blue frame in Fig. 1b) and k-mer entries (green frame in Fig. 1b). For each of the 21^k k-mers (the 21st letter X codes for "any amino acid") a pointer points to the list with target sequences and positions (`target_ID`, j) where this k-mer occurs. Prior to index generation, regions of low amino acid compositional complexity are masked out using TANTAN (see Masking low-complexity regions)[4].

Building the index table can be done in multithreaded fashion and does not require any additional memory. To that end, we proceed in two steps: counting of k-mers and filling entries. In the first step each thread counts the k-mers, one sequence at a time. All threads add up their counts using the atomic function `__sync_fetch_and_add`. In the second step, we allocate the appropriately sized array for the k-mer entries. We transform the k-mer count array into an offset table. Each thread picks a new sequence from the database, parallelized by OpenMP using `#pragma omp parallel for`, and processes all k-mers in the sequence. Each k-mer is now written into the entry array. We can prevent two threads writing into the same position pointed to by the offset pointer by fetching and incrementing it at the same time using the atomic `__sync_fetch_and_add` instruction. Building the index table file for 3×10^7 sequences takes 7.1 minutes on 2×8 cores.

Memory requirements. The index table needs $4 + 2$ bytes for each entry (`target_ID`, j), and one byte per residue is needed to store the target sequences. For a database of NL residues, we therefore require $NL \times 7$ B. The pointer array needs another $21^k \times 8$ B. The target database set can be split into arbitrary chunk sizes to fit them into memory (see Parallelization).

Ungapped alignment stage. A fast, vectorized algorithm computes the scores of optimal ungapped alignments on the diagonals with double k-mer matches. Since it has a linear time complexity, it is much faster than the Smith-Waterman alignment stage with its quadratic time complexity. The algorithm aligns 32 target sequences in parallel, using the AVX2 vector units of the CPU. To only access memory linearly we precompute for each query sequence a matrix with 32 scores per query residue, containing the 20 amino acid substitution scores for the query residue, a score of -1 for the letter X (any residue), and 11 zero scores for padding. We gather bundles of 32 target sequences with matches on the same diagonal and also preprocess them for fast access: We write the amino acids of position j of the 32 sequences consecutively into block j of 32 bytes, the longest sequence defining the number of blocks. The algorithm moves along the diagonals and iteratively computes the 32 scores of the best alignment ending at query position i in AVX2 register S using $S = \texttt{max}(0, S_{\text{match}} + S_{\text{prev}})$. The substitution scores of the 32 sequences at the current query position i in AVX2 register S_{match} are obtained using the AVX2 (V)PSHUFB instruction, which extracts from the query profile at position i the entries specified by the 32 bytes in block j of the target sequences. The maximum scores along the 32 diagonals are updated using $S_{\max} = \texttt{max}(S_{\max}, S)$. Alignments above 15 bits are passed on to the next stage.

Vectorized Smith-Waterman alignment stage. We extended the alignment library of Mengyao et al.[22], which is based on Michael Farrar's stripe-vectorized alignment algorithm[3], by adding support for AVX2 instructions and for sequence profiles. To save time

when filtering matches, we only need to compute the score and not the full alignment. We therefore implemented versions that compute only the score and the end position of the alignment, or only start and end position and score.

Amino acid local compositional bias correction. Many regions in proteins, in particular those not forming a stable structure, have a biased amino acid composition that differs considerably from the database average. These regions can produce many spurious k-mer matches and high-scoring alignments with non-homologous sequences of similarly biased amino acid distribution. Therefore, in all three search stages we apply a correction to substitution matrix scores developed for MMseqs[8], assigning lower scores to the matches of amino acids that are overrepresented in the local sequence neighborhood. Query sequence profile scores are corrected in a similar way: The score $S(i, \mathrm{aa})$ for amino acid aa at position i is corrected to $S_{\mathrm{corr}}(i, \mathrm{aa}) = S(i, \mathrm{aa}) - \frac{1}{40} \sum_{j=i-20, j\neq i}^{i+20} S(j, \mathrm{aa}) + \frac{1}{L_{\mathrm{query}}} \sum_{j=1}^{L_{\mathrm{query}}} S(j, \mathrm{aa})$.

Masking low-complexity regions. The query-based amino acid local compositional bias correction proved effective, particularly for sequence-to-sequence searches. However, for iterative profile sequence searches a very low level of false discovery rate is required, as false positive sequences can recruit more false positives in subsequent iterations leading to massively corrupted profiles and search results in these instances. We observed that these cases were mainly caused by biased and low-complexity regions in the target sequences. We therefore mask out low-complexity regions in the target sequences during the k-mer matching and the ungapped alignment stage. We use TANTAN[4] with a threshold of 0.9% probability for low complexity.

Iterative profile search mode. The first iteration of the profile-to-sequence search is a straightforward MMseqs2 sequence-to-sequence search. We then realign all matched sequences with E-values below the specified inclusion threshold (option `--e-profile <value>`, default value 0.1). At this stage, we add a score offset of -0.1 bits per matched residue pair to the scores of the substitution matrix to avoid *homologous overextensions* of the alignments, a serious problem causing many false positives in iterative profile searches[5;6]. In all further iterations, we remove from the prefilter results sequences that were previously included in the profile and align only the newly found sequence matches. From the search results we construct a simple star multiple sequence alignment (MSA) with the query as the master sequence. We filter the multiple sequence alignment with 90% maximum pairwise sequence identity and pick the 100 most diverse sequences using C++ code adapted from our HH-suite software package[15]. As in HH-suite, we compute position-specific sequence weights[1], which ensure that MSAs with many matched segments that stretch only part of the query sequence, as occurs often for multidomain proteins, are treated appropriately. We add pseudocounts to the amino acid counts of each profile column as described for HHsearch[18]. All matches included in the profile or achieving an E-value in the last iteration below the value given by `-e <value>` are displayed.

Sequence-to-profile search mode. To enable searching a target profile database, we made four changes to the search workflow (**Supplementary Fig. S11**): (1) We generate a k-mer index table for the target database by looping over all profiles and all k-mer positions and adding all k-mers to the index that obtain a profile similarity score above the threshold. Lower score thresholds lead to more k-mers and higher sensitivity. (2) We only look for the exact query k-mers in the index table. The former loop 3 is omitted. (3) The ungapped alignment for each matched diagonal is computed between the query sequence and the target profile's consensus sequence, which contains at each column the most frequent amino acid. (4) The previous step produced for each query sequence s a list of matched profiles p with score $S_{sp} > 15$bit. However, the gapped alignment stage can only align profiles with sequences and not vice versa. We therefore transpose the scores S_{sp} in memory and obtain for each profile p all matched sequences, $\{s : S_{ps} > 15\mathrm{bit}\}$, which we pass to the gapped alignment stage. Finally, the results are transposed again to obtain for each query sequence a list of matched profiles.

Algorithmic novelty in MMseqs2. MMseqs2 builds upon many powerful previous ideas in the sequence search field, such as inexact k-mer matching[1], finding two k-mers on the same diagonal[1], or spaced k-mers[13]. In addition to carefully engineering every relevant piece of code for maximum speed, we introduce with MMseqs2 several novel ideas that were crucial to the improved performance: (1) the algorithm to find two consecutive, inexact k-mer matches (**Fig. 1b**); (2) the avoidance of random memory accesses in the innermost loop of the k-mer match stage (**Supplementary Fig. S1**); (3) the use of 7-mers, which is only enabled by the fast generation of similar k-mers ($\sim 60\,000$ per k-mer in sensitive mode); (4) iterative profile-sequence search mode including profile-to-sequence vectorized Smith-Waterman alignment; (5) sequence-to-profile search mode; (6) the introduction of a fast, vectorized ungapped-alignment step (**Fig. 1a**); (7) a fast amino acid compositional bias score correction on the query side that suppresses high-scoring false positives.

Parallelization. Due to the stagnation in CPU clock rates and the increase in the number of cores per CPU, vectorization and parallelisation

across multiple cores and servers is of growing importance for highly compute-intensive applications. Besides careful vectorization of time-critical loops, MMseqs2 is efficiently parallelized to run on multiple cores and servers using OpenMP and message passing interface (MPI).

OpenMP threads search query sequences independently against the target database and write their result into separate files. After all queries are processed, the master thread merges all results together.

To parallelize the time-consuming k-mer matching and gapless alignment stages among multiple servers, two different modes are available. In the first, MMseqs2 can split the target sequence set into approximately equal-sized chunks, and each server searches all queries against its chunk. The results from each server are automatically merged. Alternatively, the query sequence set is split into equal-sized chunks and each server searches its query chunk against the entire target set. Splitting the target database is less time-efficient due to the slow, IO-limited merging of results. But it reduces the memory required on each server to $NL \times 7\text{B}/\#\text{chunks} + 21^k \times 8\,\text{B}$ and allows users to search through huge databases on servers with moderate memory sizes. If the number of chunks is larger than the number of servers, chunks will be distributed among servers and processed sequentially. By default, MMseqs2 automatically decides which mode to pick based on the available memory on the master server.

MMseqs2 software suite. The MMseqs2 suite consists of four simple-to-use main tools for standard searching and clustering tasks, 37 utility tools, and four core tools ("expert tools"). The core tool `mmseqs prefilter` runs the first two search stages in Fig. 1a, `mmseqs align` runs the Smith-Waterman alignment stage, and `mmseqs clust` offers various clustering algorithms. The utilities comprise format conversion tools, multiple sequence alignment, sequence profile calculation, open reading frame (ORF) extraction, 6-frame translation, set operations on sequence sets and results, regex-based filters, and statistics tools to analyse results. The main tools are implemented as `bash`-scripted workflows that chain together core tools and utilities, to facilitate their modification and extension and the creation of new workflows.

Design of sensitivity benchmark. Some recent new sequence search tools were only benchmarked against short sequences, using BLAST results as the gold standard [10;9;2;23]. Short matches require fairly high sequence identities to become statistically significant, making BLAST matches of length 50 relatively easy to detect. (For a sequence match to achieve an $E-$value < 0.01 in a search through UniProt requires a raw score of ~ 40 bits, which on 50 aligned residues translates to a sequence identity $\gtrsim 40\%$). Because long-read, third-generation sequencing technologies are becoming widespread, short-read technologies are improving read lengths, and ORFs and putative genes in metagenomics are commonly predicted from assembled contigs, we constructed a benchmark set using full-length queries and database sequences, not just sequences of structured domains as usually done. Including disordered regions, membrane helices, and other low-complexity regions is important since they often give rise to false-positive sequence matches, particularly in iterative sequence searches.

Because we cannot use BLAST or SWIPE [16] as gold standard if we want to compare other tools with them, we use evolutionary relationships that have been determined on the basis of structures as gold standard. SCOP [14] is a database of protein domains of known structure organised by evolutionary relationships.

We defined true positive matches to have annotated SCOP domains from the same SCOP family, false positives match a reversed sequence. In the first benchmark version matches to a sequence with a SCOP domain from a different fold except the beta propellers (which are known to be homologous [19]) are also conside--red false positives. Other cases are ignored. -- The false discovery rate (FDR) For a single search is computed as FDR = FP/(FP + TP), where TP and FP are the numbers of true and false positive matches below a cutoff score in that search. To prevent a few searches with many false positives from dominating the FDR, we computed the FDR for all searches as arithmetic mean over the single-search FDRs.

We measure the sensitivity of search tools using a receiver operating characteristic (ROC) analysis [19]. We search with a large set of query sequences through a database set (see next paragraph) and record for each query the fraction of true positive sequences detected up to the first false positive. This sensitivity is also called area under the curve 1 (AUC1). We then plot the cumulative distribution of AUC1 values, that is, the fraction of query sequences with an AUC1 value larger than the value on the x-axis. The more sensitive a search tools is the higher will its cumulative distribution trace lie. We do not analyse only the best match for every search in order to increase the number of matches and to thereby reduce statistical noise.

Benchmark set. The SCOP/ASTRAL (v. 1.75) database was filtered to 25% maximum pairwise sequence identity (7616 sequences), and we searched with each SCOP sequence through the UniRef50 (06/2015) database, using SWIPE [16] and, for maximum sensitivity, also three iterations of HHblits. To construct the query set, we chose for each of the 7616 SCOP sequences the best matching UniRef50 sequence for the query set if its SWIPE E-value was below 10^{-5}, resulting in 6370 query sequences with 7598 SCOP-annotated domains. In the

first version of the benchmark (**Fig. 2**), query sequences were shuffled outside of annotated regions within overlapping windows of size 10. This preserves the local amino acid composition while precluding true positive matches in the shuffled regions. In the second version of the benchmark, query sequences were left unchanged (**Supplementary Fig. S3, S4, S5, S6**).

To construct the target database, we selected all UniRef50 sequences with SWIPE or HHblits $E-\text{value} < 10^{-5}$ and annotated them with the corresponding SCOP family, resulting in 3 374 007 annotations and a median and average number of sequences per SCOP family of 353 and 2150, respectively. Since the speed measurements are only relevant and quantitative on a database of realistic size, we added the 27 056 274 reversed sequences from a 2012 UniProt release. Again, the reversion preserves the local amino acid composition while ruling out true positive matches[11]. We removed the query sequences from the target database and removed queries with no correct matches in the target database from the query set.

Benchmarking. We evaluated results up to the 4000th match per query (ranked by E-value) and, for tools with an upper limit on the number of reported matches, set this limit via command line option to 4000. The maximum E-value was set to 10 000 to detect at least one false positive and to avoid biases due to slightly different E-value calculations. Because the MMseqs2 prefilter is already very sensitive and returns proper E-values, the Smith-Waterman alignment stage is not needed in the "fast" and "faster" versions. Program versions and calls are found in the **Supplemental Table S3**.

All benchmarks were run on a single server with two Intel Xeon E5-2640v3 CPUs (2×8 cores, 2.6 GHz) and 128GB memory. Run times were measured using the Linux `time` command, with the target database (70 GB, 30.4 M sequences) on local solid state drives. Since some search tools are speed-optimized not only for large target databases but also for large query sets, we duplicated the query set 100 times for the runtime measurements, resulting in 637 000 query sequences. For the slowest tools, SWIPE, BLAST and RAPsearch2, we scaled up the runtime for the original query dataset 100-fold.

Data availability. Parameters and scripts for benchmarking are deposited at `https://bitbucket.org/martin_steinegger/mmseqs-benchmark`.

Code availability. The source code and binaries of the MMseqs2 software suite can be download at `https://mmseqs.org`.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25(17):3389–3402, Sept. 1997.

[2] B. Buchfink, C. Xie, and D. H. Huson. Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1):59–60, 2015.

[3] M. Farrar. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics, 23(2):156–161, Jan. 2007.

[4] M. C. Frith. A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res., 39(4):e23–e23, 03 2011.

[5] M. C. Frith, Y. Park, S. L. Sheetlin, and J. L. Spouge. The whole alignment and nothing but the alignment: the problem of spurious alignment flanks. Nucleic Acids Res., 36(18):5863–5871, 2008.

[6] M. W. Gonzalez and W. R. Pearson. Homologous over-extension: a challenge for iterative similarity searches. Nucleic Acids Res., 38(7):2177–2189, 2010.

[7] M. Hauser, C. E. Mayer, and J. Söding. kclust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics, 14(1):1–12, 2013.

[8] M. Hauser, M. Steinegger, and J. Söding. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics, 32(9):1323–1330, 2016.

[9] H. Hauswedell, J. Singer, and K. Reinert. Lambda: the local aligner for massive biological data. Bioinformatics, 30(17):i349–i355, 2014.

[10] D. H. Huson and C. Xie. A poor man's BLASTX-high-throughput metagenomic protein database search using PAUDA. Bioinformatics, 30(1):38–39, 2014.

[11] K. Karplus, C. Barrett, and R. Hughey. Hidden markov models for detecting remote protein homologies. Bioinformatics, 14(10):846–856, 1998.

[12] J. J. Kent. BLAT–the BLAST-like alignment tool. Genome Res., 12(4):656–664, Apr. 2002.

[13] B. Ma, J. Tromp, and M. Li. PatternHunter: faster and more sensitive homology search. Bioinformatics, 18(3):440–445, 2002.

[14] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., 247(4):536 – 540, 1995.

[15] M. Remmert, A. Biegert, A. Hauser, and J. Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods, 9(2):173–175, Feb. 2012.

[16] T. Rognes. Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. BMC Bioinformatics, 12(1):221+, June 2011.

[17] S. A. Shiryev, J. S. Papadopoulos, A. A. Schäffer, and R. Agarwala. Improved blast searches using longer words for protein seeding. Bioinformatics, 23(21):2949–2951, 2007.

[18] J. Söding. Protein homology detection by HMM–HMM comparison. Bioinformatics, 21(7):951–960, 2005.

[19] J. Söding and M. Remmert. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Curr. Opin. Struct. Biol., 21(3):404–411, 2011.

[20] J. Tan, D. Kuchibhatla, F. L. Sirota, W. A. Sherman, T. Gattermayer, C. Y. Kwoh, F. Eisenhaber, G. Schneider, and S. M. Stroh. Tachyon search speeds up retrieval of similar sequences by several orders of magnitude. Bioinformatics, 28(12):1645–1646, June 2012.

[21] R. Vaser, D. Pavlović, M. Korpar, and M. Šikić. Sword-a highly efficient protein database search. Bioinformatics, 32(17):i680–i684, 2016.

[22] M. Zhao, W.-P. Lee, E. P. Garrison, and G. T. Marth. Ssw library: An simd smith-waterman c/c++ library for use in genomic applications. PLoS One, 8(12), 12 2013.

[23] Y. Zhao, H. Tang, and Y. Ye. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics, 28(1):125–126, Jan. 2012.

Supplementary Figures and Tables

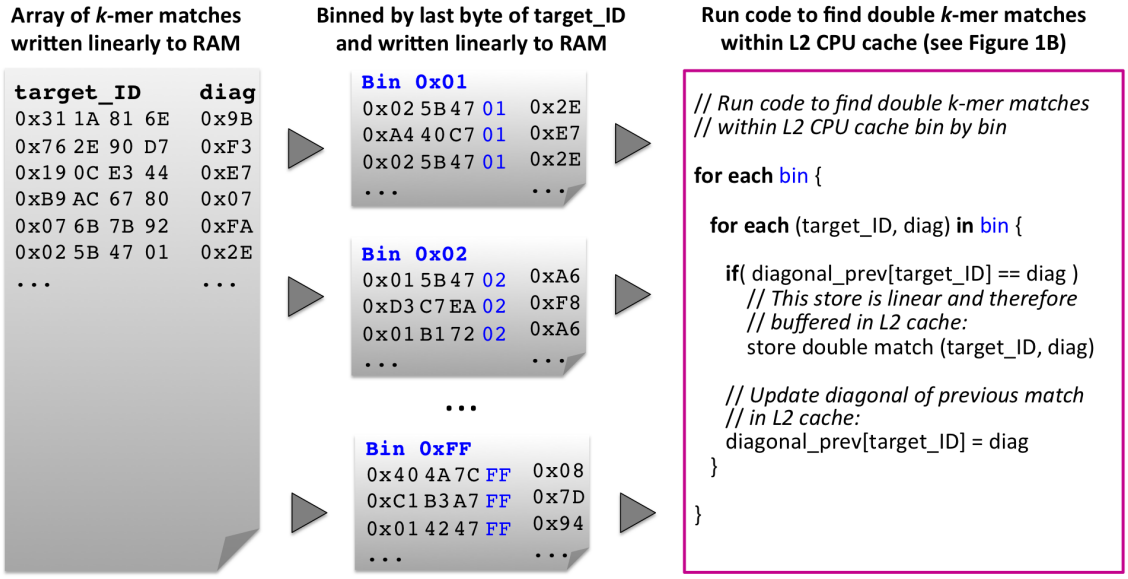

Figure S 1: **Eliminating random memory access during k-mer match stage in MMseqs2** Numbers in this figure are represented in hexadecimal notation (e.g. 0xFF is equal to 255 in decimal). After the end of loop 2 (**Fig. 1B**), the `matches` array on the left, containing single k-mer matches between the query sequence and various target sequences, is processed in two steps to find double k-mer matches. In the first step, the entries (`target_ID`, $i-j$) of `matches` are sorted into 2^B arrays (bins) according to the lowest B bits of `target_ID`. Here, for illustration purposes, we set $B = 8$. In the second step, the 2^B bins are processed one by one. For each k-mer match (`target_ID`, $i-j$), we run the code in the magenta frame of Fig. 1B. But now, the `diagonal_prev` array fits into L1/L2 CPU cache, because it only contains `ceil`$(N/2^B)$ entries, where N is the number of sequences in the target database.

Nature Biotechnology: doi:10.1038/nbt.3988

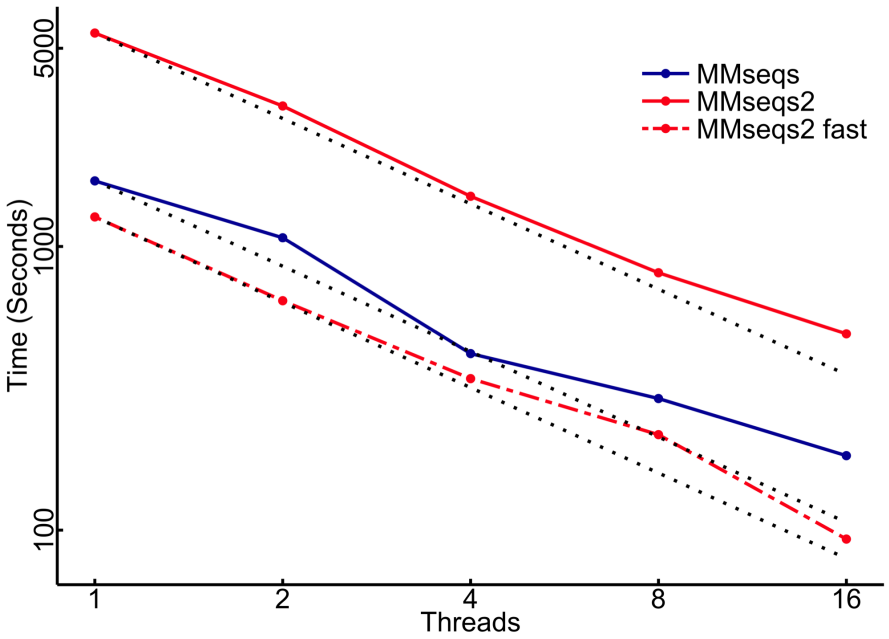

Figure S 2: **Multi-core scaling of MMseqs2** Runtimes of MMseqs and MMseqs2 searches in fast and default sensitivity using 1, 2, 4, 8 and 16 threads on a 2×8 core server with 128 GB main memory. Theoretically optimal scaling is indicated as a dashed black line for each method. We searched with 6370 full length protein queries against 30 Mio. UniProt sequences. On 16 cores, MMseqs achieves 58% and MMseqs2 85% of their theoretical maximum performance interpolated from the single core measurement. The improvement in scaling behaviour from MMseqs to MMseqs2 is owed to minimizing random main memory accesses, as explained in **Fig. S1**.

Nature Biotechnology: doi:10.1038/nbt.3988

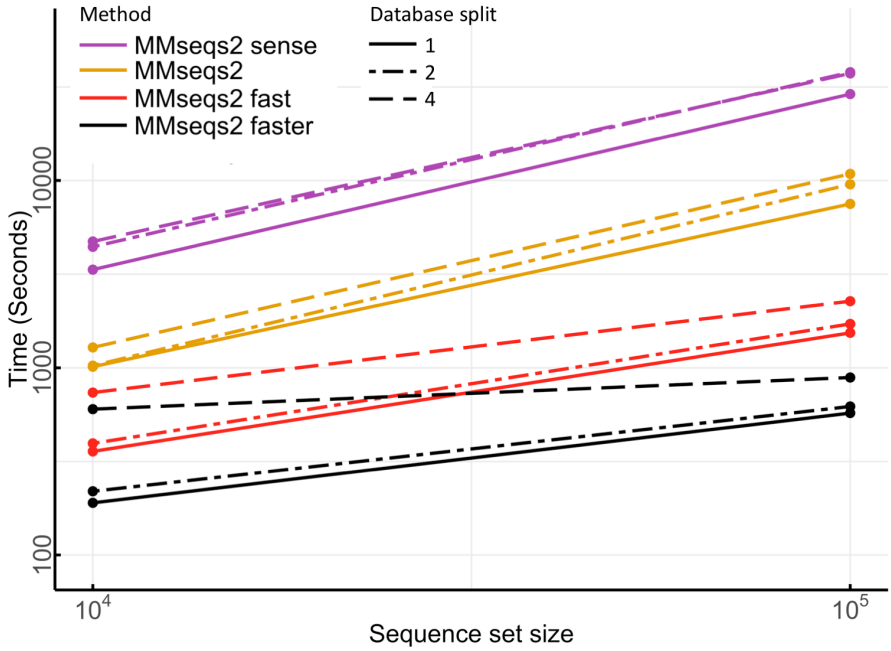

Figure S 3: Runtime of MMseqs2 against the UniProt at different sensitivity and database split settings. We measured the search time with query sets of 10 000 and 100 000 sequences through the UniProt database (Release 2017_03 with 80 204 488 sequences) using four sensitivity settings (faster, fast, default, and sensitive) and splitting the database into 1, 2, and 4 chunks. Runtimes for Refseq/Genbank (Release March 3, 2017 with 81 027 309 sequences) are very similar. The memory consumption of the index table for the split levels of 1, 2, and 4 was 190GB, 101GB, and 57GB respectively. All searches ran on a 2×14-core server with 768GB main memory.

Nature Biotechnology: doi:10.1038/nbt.3988

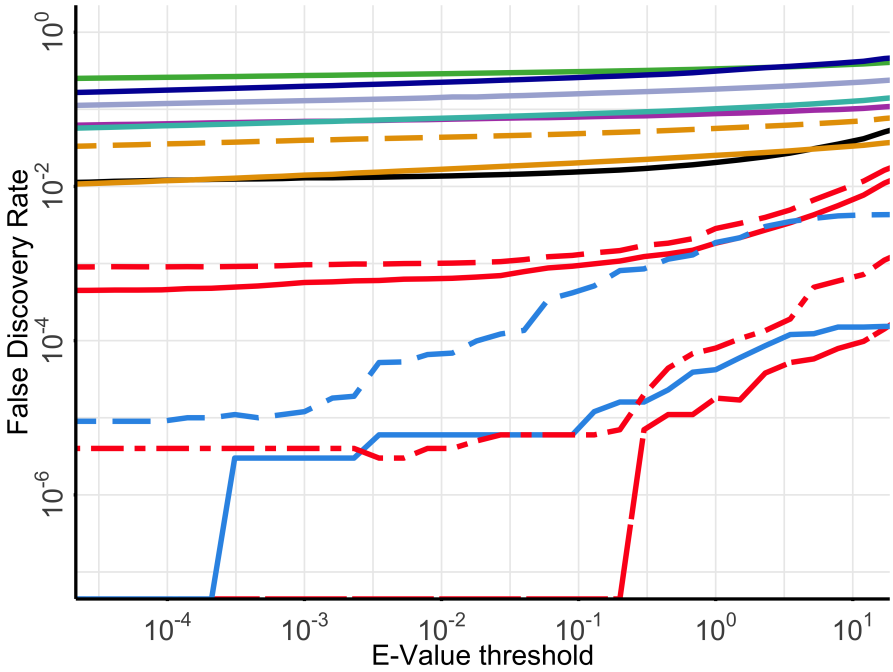

Figure S 4: False discovery rate versus *E*-value threshold in version 2 of the sequence search sensitivity benchmark using unshuffled query sequences. Colors are the same as in **Fig. 2a**

Nature Biotechnology: doi:10.1038/nbt.3988

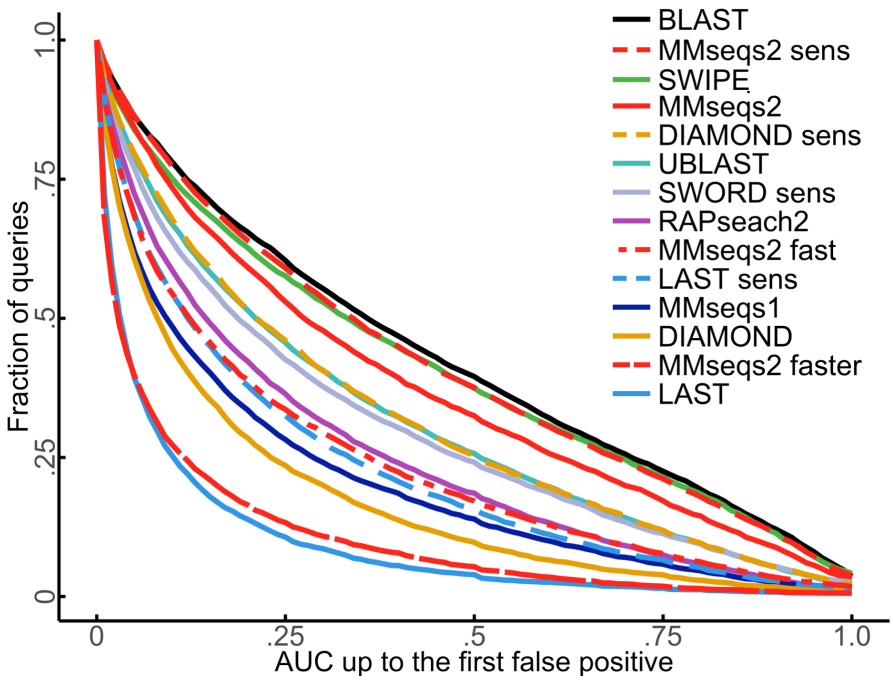

Figure S 5: **Sequence searching sensitivity assessment with unshuffled query sequences.** Cumulative distribution of area under the curve (AUC) sensitivity for all 6324 queries in version 2 of the sequence search sensitivity benchmark using unshuffled query sequences. Higher curves signify higher sensitivity.

Nature Biotechnology: doi:10.1038/nbt.3988

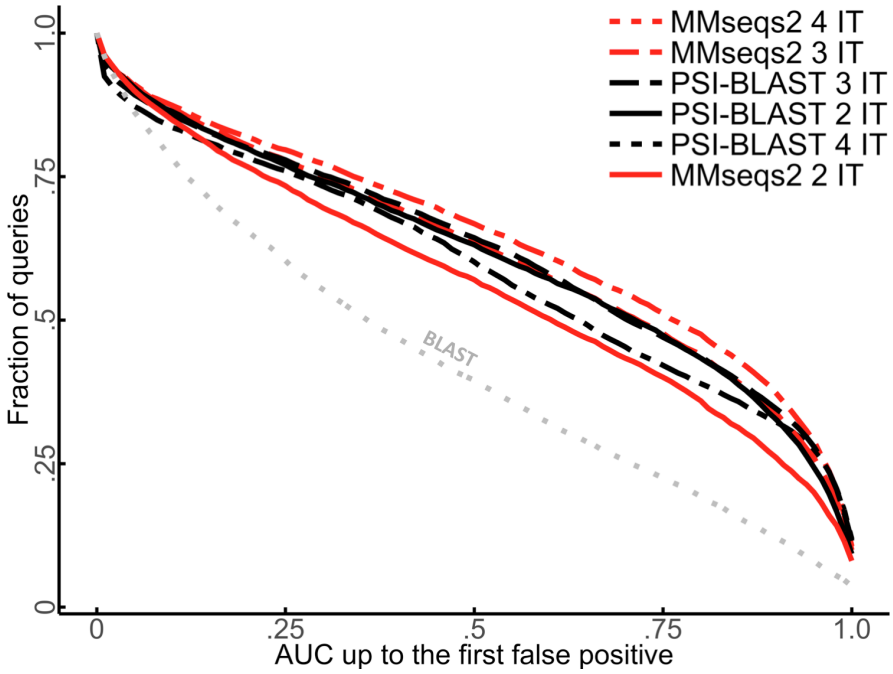

Figure S 6: **Sequence profile searching sensitivity assessment with unshuffled query sequence profiles.** Cumulative distribution of area under the curve (AUC) sensitivity for all 6324 unshuffled query sequences in version 2 of the sequence search sensitivity benchmark using unshuffled query sequences. Higher curves signify higher sensitivity. Higher curves signify higher sensitivity. 2 IT: 2 search iterations etc.

Nature Biotechnology: doi:10.1038/nbt.3988

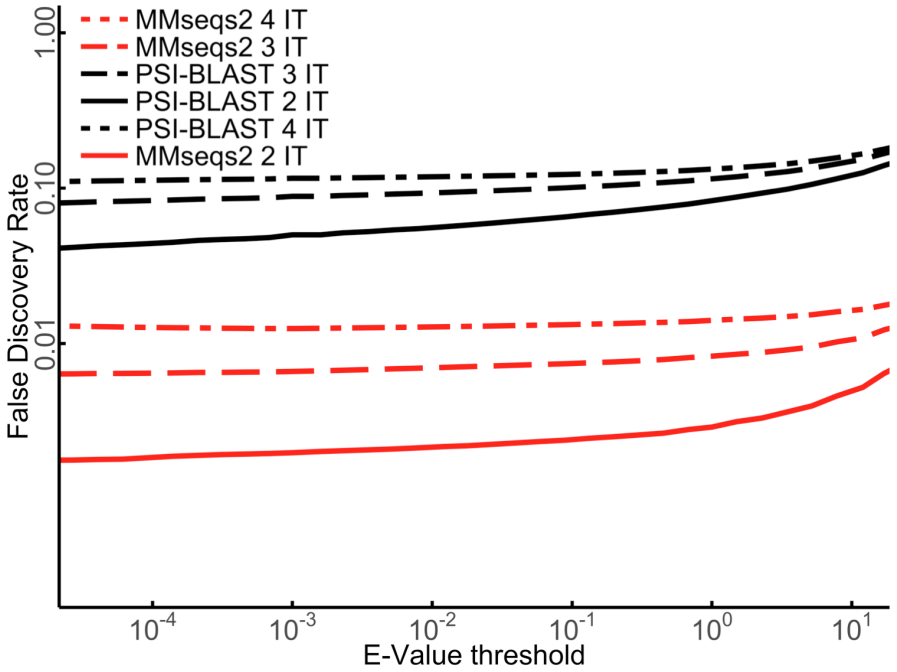

Figure S 7: False discovery rate versus E-value threshold in version 2 of the sequence profile search sensitivity benchmark using unshuffled query sequences.

Nature Biotechnology: doi:10.1038/nbt.3988

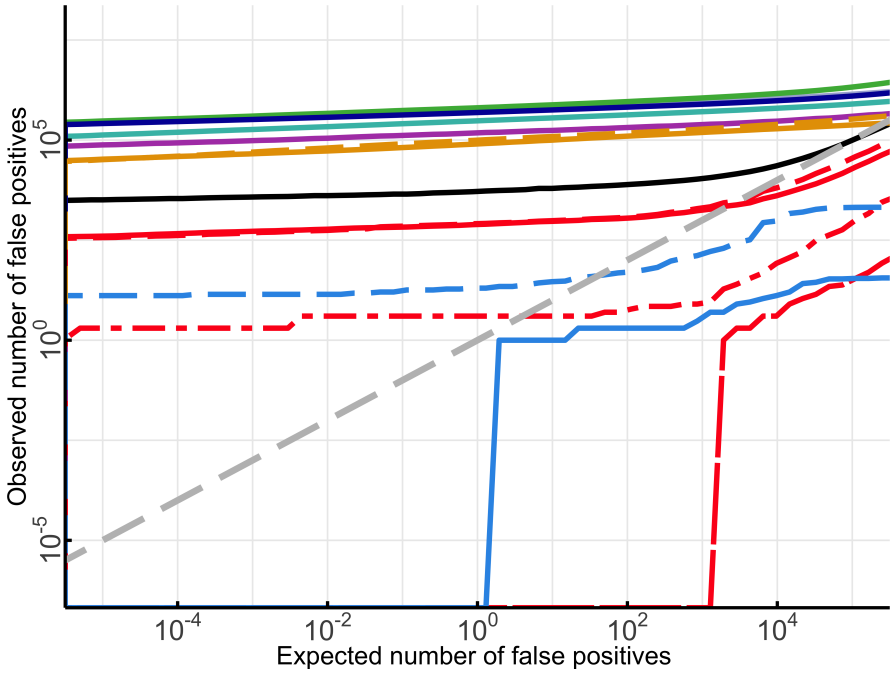

Figure S 8: **Accuracy of reported** *E*-**values.** The expected number of false positives is the *E*-value threshold times the number of searches, $E \times 6324$. The observed number of false positives is the total number of false positives below the E-value threshold in all 6324 searches. If *E*-values were accurate, observed and expected numbers of false positives would coincide (diagonal grey line). LAST and MMseqs2 report the most accurate *E*-values. The false positives shown were obtained with version 2 of the sequence search sensitivity benchmark. Colors are the same as in **Fig. 2a**.

Nature Biotechnology: doi:10.1038/nbt.3988

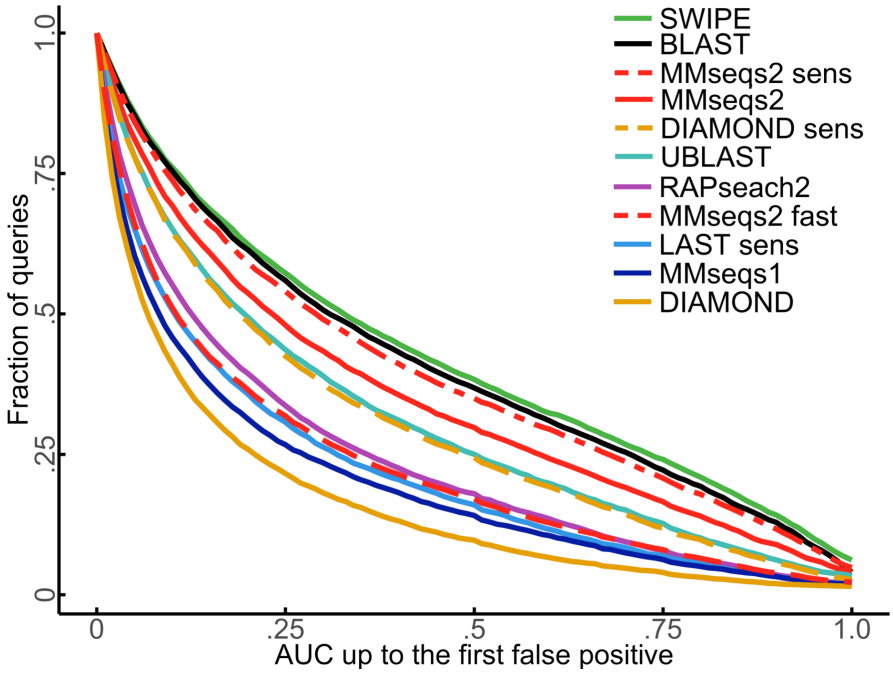

Figure S 9: **Sequence searching sensitivity assessment with single-domain SCOP sequences.** Cumulative distribution of area under the curve (AUC) sensitivity for all 7616 single domain SCOP sequences. Higher curves signify higher sensitivity. AUC up to the first false positive is the fraction of true positive matches found with better E-value than the first false positive match.

Nature Biotechnology: doi:10.1038/nbt.3988

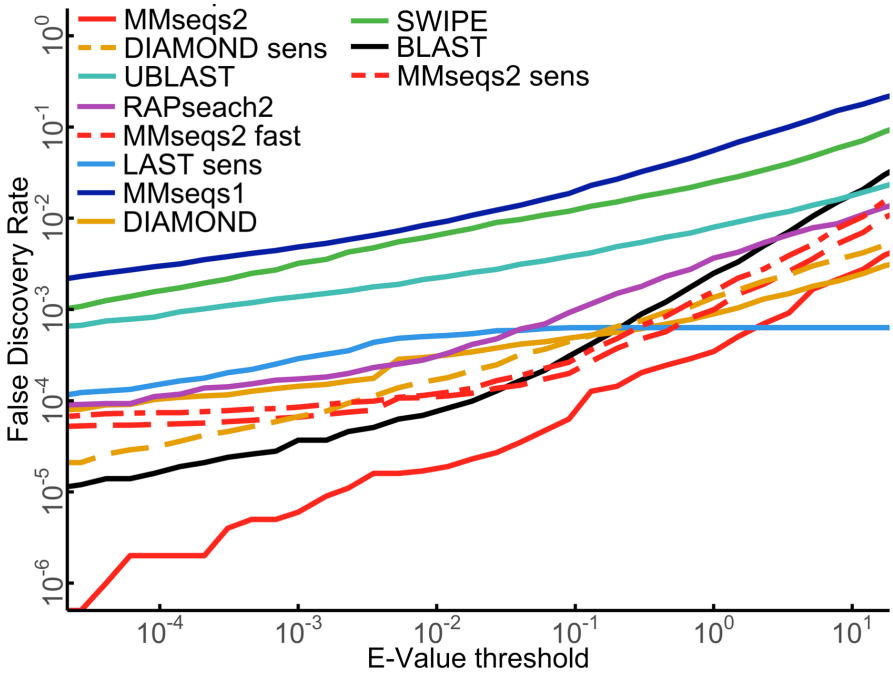

Figure S 10: False discovery rate versus E-value threshold for the single-domain SCOP sequence search sensitivity benchmark.

Nature Biotechnology: doi:10.1038/nbt.3988

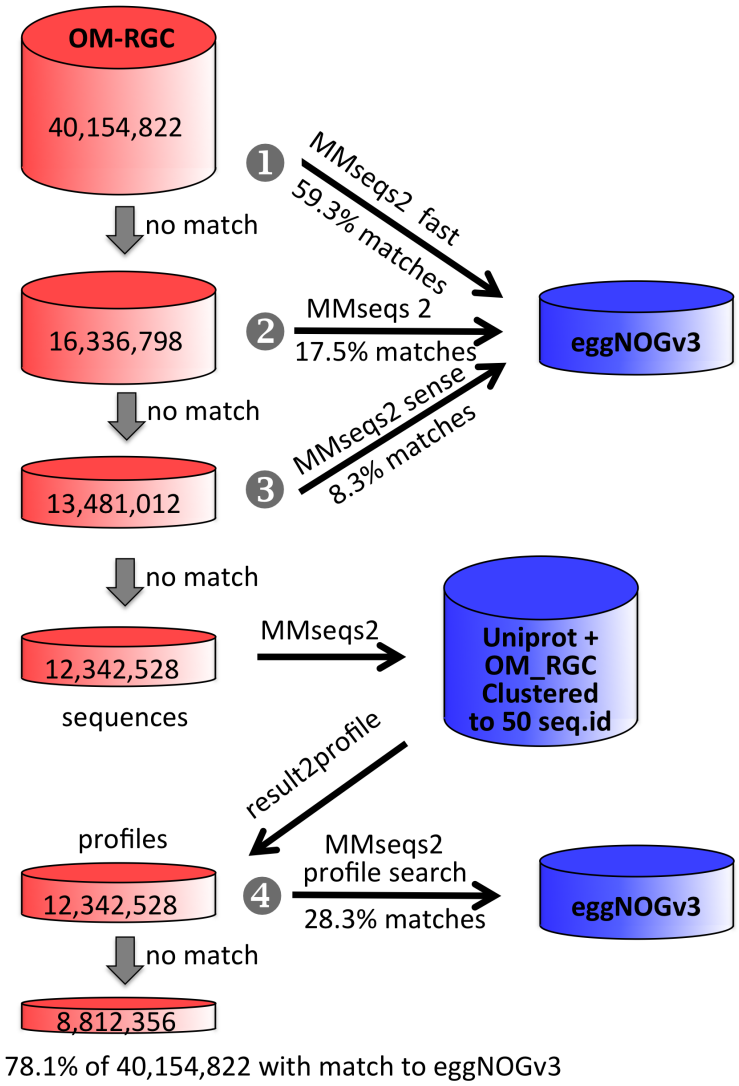

Figure S 11: Workflow for fast and deep annotations of the Ocean Microbiome Reference Gene Catalog (OM-RGC) using MMseqs2.

Nature Biotechnology: doi:10.1038/nbt.3988

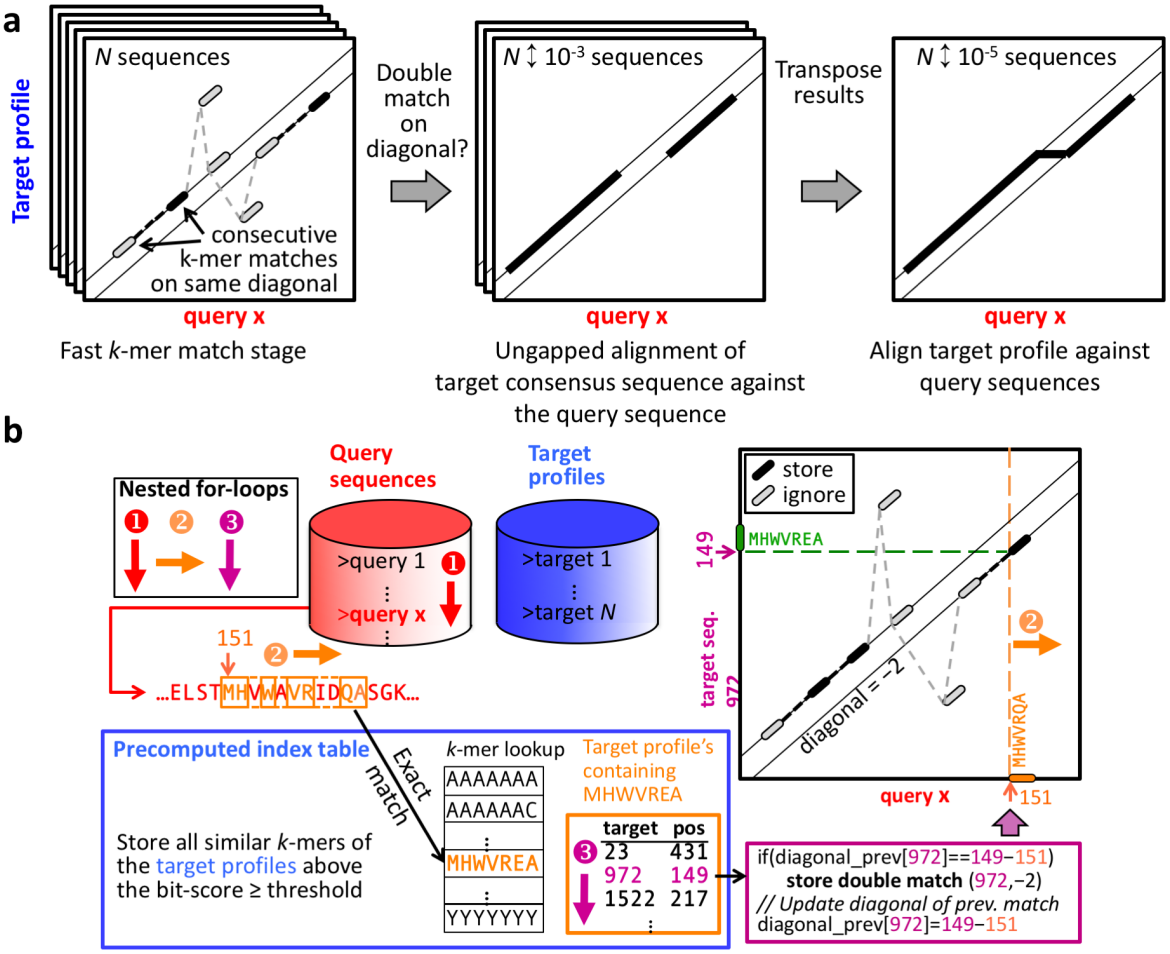

Figure S 12: **Algorithmic changes to perform fast sequence profile searches using MMseqs2.** We precompute all similar k-mers above a similarity threshold for each target profile and store them into the index table. For each query sequence we run over its overlapping, spaced k-mers (loop 2) and look up in the index table (blue frame) only the exact same k-mer. At the ungapped alignment stage we use the target profile consensus sequence. We transpose the results, i.e., we exchange the role of query and target in the results and then, as the last step, align the profiles against all query sequences and transpose back.

Nature Biotechnology: doi:10.1038/nbt.3988

Feature	MMseqs	MMseqs2	Remark
Iterative profile searches	no	yes	Iterative profile searches increase sensitivity far beyond the sensitivity of BLAST
Sequence-to-profile searches	no	yes	Protein sequences can be annotated very fast by searching through databases of profiles, e.g. for Pfam, eggNOG, or PDB
k-mer match stage	Sums up similarity scores of similar 6-mers between pairs of sequences	Finds consecutive double 7-mer matches on the same diagonal	MMseqs aggregates scores of spurious matches across all possible $L_{\text{query}} \times L_{\text{target}}$ start positions. OK for global alignment, but suboptimal for local similarities. MMseqs2's consecutive double-diagonal k-mer match criterion suppresses most spurious matches and also works well for local similarities.
Fast gapless alignment stage	no	yes (AVX2 / SSE2)	Increases sensitivity-versus-speed trade-off by allowing MMseqs2 to evaluate more matches from the k-mer matching stage while still reducing the number of Smith-Waterman alignments
Multicore scalability	Speed-up for 16 cores is 9.3-fold	Speed-up for 16 cores is 13.7-fold	MMseqs2 minimizes random memory access by better utilizing low-level CPU caches (Supplementary Figures S1, S2)
Suppression of false positive matches	Compositional bias score correction on query side in k-mer match stage	Compositional bias score correction on query and target side in all three stages	MMseqs2 eliminates high-scoring false positives much more effectively than MMseqs
Clustering methods	simple greedy strategy	Simple greedy set-cover, single-linkage with depth cut-off	MMseqs2 has an option to reassign cluster members to the best representative
Utility scripts	3	37 (see MMseqs2 userguide on GitHub)	MMseqs2 has added utility tools for format conversion, multiple sequence alignment, sequence profile calculation, ORF extraction, 6-frame translation, operations on sequence sets and results, regex-based filters, and statistics tools to analyse results
Distribution of jobs on computer cluster	no	yes	MMseqs2 uses Message Passing Interface
Split target database among servers	no	yes	Allows MMseqs2 to search or cluster arbitrarily large databases with limited memory
SIMD parallelization	SSE2	AVX2 (SSE4.1 if no support for AVX2)	AVX2 has two-fold higher parallelism and is therefore faster
Lines of code	10 000	30 000	A large proportion of the MMseqs code has been rewritten from scratch and considerably modified for better performance.

Table S 1: **Comparison between MMseqs and MMseqs2.**

Method	Fractions of queries containing a false positive with E-value $< 10^{-3}$	Number of reported false positive matches with E-value$< 10^{-3}$ (Expected number $= 6370 \times 10^{-3} = 0.637$)
MMseqs2 sens	0.001	39
MMseqs2	0.001	54
MMseqs2 fast	0.002	126
MMseqs2 faster	0.001	9
UBLAST	0.107	380807
DIAMOND	0.050	69211
DIAMOND sens	0.067	124906
LAST sense	0.001	7
LAST	0	0
BLAST	0.022	1313
RAPsearch2	0.044	141665
SWORD sens	0.131	676566
MMseqs1	0.159	537403
SWIPE	0.124	788356
MMseqs2 2 IT	0.013	6202
MMseqs2 3 IT	0.024	44270
MMseqs2 4 IT	0.031	94747
PSI-BLAST 2 IT	0.253	
PSI-BLAST 3 IT	0.294	1.01498e+06
PSI-BLAST 4 IT	0.299	1.13397e+06

Table S 2: Analysis of high-scoring false positive matches in 6370 searches with UniProt sequences through a database of 30 million UniProt-derived sequences. For most tools, a seemingly significant E-value (e.g. smaller than 0.001) is not a strong indication of a homologous relationship. In automatic functional annotation pipelines, such unreliable E-values will lead to an increased fraction of false annotations. But note that even unreliable E-values can work very well for ranking matches *within a single search*, and the quality of ranking matches within a single search is what is measured by the AUC1 sensitivity.

Method	Version	Database	Command
MMseqs2 (normal | sense)	2.0	createindex -k 7	search --k-score (95 | 85) -e 10000.0 --max-seqs 4000
MMseqs2 (very fast | fast)	2.0	createindex -k 7	prefilter --k-score (140 | 115) --max-seqs 4000
MMseqs	1.0	fasta2ffindex	--z-score-thr 10.0 -s 4 --max-seqs 4000 -c 0.0 -e 10000.0
SWIPE	2.0.11	makeblastdb -dbtype prot	-e 10000.0 -a 16 -v 4000 -b 4000
RAPsearch2	2.23	makeblastdb -dbtype prot	-v 4000 -z 16 -e 4 -t a -b 0
UBLAST	7.0.1090	-makeudb_ublast	-threads 16 -evalue 10000.0 -ublast
SWORD sens	commit fcb2117		-t 16 -a 4000 --evalue 10000
LAST	last-712	lastdb -cR01 -p -v	-P 16 -u3 -D100
LAST sens	last-712	lastdb -cR01 -p -v	-P 16 -m 4000 -u3 -D100
DIAMOND sens	0.7.9.58	diamond makedb	--max-target-seqs 4000 --evalue 10000.0 -t /dev/shm --threads 16 (--sensitive)
BLAST	2.2.31+	makeblastdb -dbtype prot	-num_descriptions 4000 -num_alignments 4000 -num_threads 16 -evalue 10000.0
PSI-BLAST	2.2.31+	makeblastdb -dbtype prot	-num_descriptions 4000 -num_alignments 4000 -num_threads 16 -num_iterations (2,3,4)
MMseqs2 profile	2.0	createindex -k 7	--num-iterations (2,3,4) -k 7 -s 5.7 -e 10000.0 --max-seqs 4000 --use-index

Table S 3: Program versions and command line parameters of tools used in the benchmark.