
UNCALLED

A Utility for Nanopore Current Alignment to Large Expanses of DNA



A read mapper which rapidly aligns raw nanopore signal to DNA references

Enables software-based targeted sequenceing on Oxford Nanopore (ONT) MinION or GridION via adaptive sampling

Note that **UNCALLED can only be applied to legacy r9.4.1 data**. For r10.4.1 data try ReadFish or ONT's builtin adaptive sampling option.

Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED

Sam Kovaka, Yunfan Fan, Bohan Ni, Winston Timp, Michael C. Schatz

Nature Biotechnology (2020)

For accurate end-to-end nanopore signal alignment, visualization, and analysis see Uncalled4

Installation

```
1 > pip3 install git+https://github.com/skovaka/UNCALLED.git --user
```

OR

```
1 > git clone --recursive https://github.com/skovaka/UNCALLED.git
2 > cd UNCALLED
3 > pip3 install .
```

Requires python >= 3.6, read-until == 3.0.0, pybind11 >= 2.5.0, and GCC >= 4.8.1 (all except GCC are automatically downloaded and installed)

Other dependencies are included via submodules, so be sure to clone with `git --recursive`

We recommend running on a Linux machine. UNCALLED has been successfully installed and run on Mac computers, but real-time ReadUntil has not been tested on a Mac. Installing UNCALLED has not been attempted on Windows.

Indexing

Example:

```
1 > uncalled index -o E.coli E.coli.fasta
```

Positional arguments:

- `fasta-file` reference genome(s) or other target sequences in the FASTA format

Optional arguments:

- `-o/--bwa_prefix` output index prefix (default: same as input fasta)

Note that UNCALLED uses the BWA FM Index to encode the reference, and this command will use a previously built BWA index if all the required files exist with the specified prefix. Otherwise, a new BWA index will be automatically built.

We recommend applying repeat masking your reference if it contains eukaryotic sequences. See masking for more details.

Fast5 Mapping

Example:

```
1 > uncalled map -t 16 E.coli fast5_list.txt > uncalled_out.paf
2 Loading fast5s
3 Mapping
4
5 > head -n 4 uncalled_out.paf
```

```

6 b84a48f0-9e86-47ef-9d20-38a0bde478e 3735 77 328 +
  Escherichia_coli_chromosome 4765434 2024611 2024838 66 228 255 ch:i
  :427 st:i:50085 mt:f:53.662560
7 77fe7f8c-32d6-4789-9d62-41ff482cf890 5500 94 130 +
  Escherichia_coli_chromosome 4765434 2333754 2333792 38 39 255 ch:i
  :131 st:i:238518 mt:f:19.497091
8 eee4b762-25dd-4d4a-8a59-be47065029be 2905 * * * *
  * * * * * 255 ch:i:44 st:i:302369
  mt:f:542.985229
9 e175c87b-a426-4a3f-8dc1-8e7ab5fdd30d 8052 84 154 +
  Escherichia_coli_chromosome 4765434 1064550 1064614 41 65 255 ch:i
  :182 st:i:452368 mt:f:38.611683

```

Positional arguments:

- `bwa-prefix` the BWA reference index prefix generated by `uncalled map`
- `fast5-files` Reads to be mapped. Can be a directory which will be recursively searched for all files with the “.fast5” extension, a text file containing one fast5 filename per line, or a comma-separated list of fast5 file names.

Optional arguments:

- `-l/--read-list` text file containing a list of read IDs. Only these reads will be mapped if specified
- `-n/--read-count` maximum number of reads to map
- `-t/--threads` number of threads to use for mapping (default: 1)
- `-e/--max-events-proc` number of events to attempt mapping before giving up on a read (default 30,000). Note that there are approximately two events per nucleotide on average.

See `example/` for a simple read and reference example.

Real-Time ReadUntil

Warning: in the latest MinKNOW version, an API bug may prevent UNCALLED from properly ejecting reads. You can identify this bug if you do not see a peak of small “adaptive sampling” reads in read length histogram. If this occurs you should stop your sequencing run, briefly start a new sequencing run with MinKNOW’s builtin version of adaptive sampling enabled, then stop that run and restart your UNCALLED run. We have found that this may initialize something in MinKNOW which allows UNCALLED to function properly.

Example:

```

1 > uncalled realtime E.coli --port 8000 -t 16 --enrich -c 3 >
  uncalled_out.paf
2 Starting client
3 Starting mappers
4 Mapping
5
6 > head -n 4 uncalled_out.paf
7 81ba344d-60df-4688-b37f-9064e76a3eb8 1352 * * * * *
   * * * * * 255 ch:i:68 st:i:29101 mt:f:375.93841
   wt:f:1440.934 mx:f:0.152565
8 404113c1-6ace-4690-885c-9c4a47da6476 450 * * * * *
   * * * * * 255 ch:i:106 st:i:29268 mt:f:63.272270
   wt:f:1591.070 en:f:0.010086
9 d9acafe3-23dd-4a0f-83db-efe299ee59a4 1355 * * * * *
   * * * * * 255 ch:i:118 st:i:29378 mt:f:239.50201
   wt:f:1403.641 ej:f:0.120715
10 8a6ec472-a289-4c50-9a75-589d7c21ef99 451 98 369 + Escherichia_coli
    4765434 3421845 3422097 56 253 255 ch:i:490 st:i:29456 mt:f
    :79.419411 wt:f:8.551202 kp:f:0.097424

```

We recommend that you try mapping fast5s via `uncalled map` before real-time enrichment, as run-time issues could occur if UNCALLED is not installed properly.

The command can generally be run at any time before or during a sequencing run, although an error may occur if UNCALLED is run before any sequencing run has been started in the current MinKNOW session. If this happens you should start UNCALLED after the run begins, ideally during the first mux scan. If you want to change the chunk size you must run the command *before* starting the run (see below).

Positional arguments: - `bwa-prefix` the BWA reference index prefix generated by `uncalled map`

Required arguments: - `--enrich` will *keep* reads that map to the reference if included OR - `--deplete` will *eject* reads that map to the reference if included Exactly one of `--deplete` or `--enrich` must be specified

Optional Arguments:

- `-c/--max-chunks` number of chunks to attempt mapping before giving up on a read (default: 10).
- `--chunk-size` size of chunks in seconds (default: 1). Note: this is a new feature and may not work as intended (see below)
- `-t/--threads` number of threads to use for mapping (default: 1)
- `--port` MinION device port.
- `--even` will only eject reads from even channels if included
- `--odd` will only eject reads from odd channels if included

-
- `--duration` expected duration of sequencing run in hours (default: 72)

Altering Chunk Size

The ReadUntil API receives signal is “chunks”, which by default are one second’s worth of signal. This can be changed using the `--chunk-size` parameter. Note that `--max-chunks-proc` should also be changed to compensate for changes to chunk sizes. *If the chunk size is changed, you must start running UNCALLED before sequencing begins.* UNCALLED is unable to change the chunk size mid-sequencing-run. In general reducing the chunk size should improve enrichment, although previous work has found that the API becomes unreliable with chunks sizes less than 0.4 seconds. We have not thoroughly tested this feature, and recommend using the default 1 second chunk size for most cases. In the future this default size may be reduced.

Simulator

Example:

```
1 > uncalled sim E.coli.fasta /path/to/control/fast5s --ctl-seqsum /path/
   to/control/sequencing_summary.txt --unc-seqsum /path/to/uncalled/
   sequencing_summary.txt --unc-paf /path/to/uncalled/uncalled_out.paf
   -t 16 --enrich -c 3 --sim-speed 0.25 > uncalled_out.paf 2>
   uncalled_err.txt
2
3 > sim_scripts/est_genome_yield.py -u uncalled_out.paf --enrich -x E.
   coli -m mm2.paf -s sequencing_summary.txt --sim-speed 0.25
4
5 unc_on_bp      150.678033
6 unc_total_bp   6094.559395
7 cnt_on_bp      33.145022
8 cnt_total_bp   8271.651331
```

The simulator simulates a real-time run using data from two real runs: one control run and one UN-CALLED run. Reads are simulated from the control run, and the pattern of channel activity of modeled after the control run. The simulator outputs a PAF file similar to the real-time mode, which can be interpreted using scripts found in `sim_scripts/`.

Example files which can be used as template UNCALLED sequencing summary and PAF files for the simulator can be found [here](#). The control reads/sequencing summary can be from any sequencing run of your sample of interest, and it does not have to match the sample used in the provided examples.

The simulator can take up a large amount of memory (> 100Gb), and loading the fast5 reads can take quite a long time. To reduce the time/memory requirements you could truncate your control sequencing summary and only the loads present in the summary will be loaded, although this may reduce the

accuracy of the simulation. Also, unfortunately the fast5 loading portion of the simulator cannot be exited via a keyboard interrupt and must be hard-killed. I will work on fixing this in future versions.

Arguments:

- `bwa-prefix` the prefix of the index to align to. Should be a BWA index that `uncalled index` was run on
- `control-fast5-files` path to the directory where control run fast5 files are stored, or a text file containing the path to one control fast5 per line
- `--ctl-seqsum` sequencing summary of the control run. Read IDs must match the control fast5 files
- `--unc-seqsum` sequencing summary of the UNCALLED run
- `--unc-paf` PAF file output by UNCALLED from the UNCALLED run
- `--sim-speed` scaling factor of simulation duration in the range (0.0, 1.0], where smaller values are faster. Setting below 0.125 may decrease accuracy.
- `-t/--threads` number of threads to use for mapping (default: 1)
- `-c/--max-chunks-proc` number of chunks to attempt mapping before giving up on a read (default: 10). Note that for the simulator, altering this changes how many chunks is loaded from each each, changing the memory requirements.
- `--enrich` will *keep* reads that map to the reference if included
- `--deplete` will *eject* reads that map to the reference if included
- `--even` will only eject reads from even channels if included
- `--odd` will only eject reads from odd channels if included

Exactly one of `--deplete` or `--enrich` must be specified

Output Format

UNCALLED outputs to stdout in a format similar to PAF. Unmapped reads are output with reference-location-dependent fields replaced with *s. Lines that begin with “#” are comments that useful for debugging.

Query coordinates, residue matches, and block lengths are estimated assuming 450bp sequenced per second. This estimate can be significantly off depending on the sequencing run. UNCALLED attempts to map a read as early as possible, so the “query sequence length” and “query end” fields correspond to the leftmost position where UNCALLED was able to confidently map the read. In many cases this may only be 450bp or 900bp into the read, even if the read is many times longer than this. This differs from aligners such as minimap2, which attempt to map the full length of the read.

The “query sequence length” field currently does not correspond to the actual read length, rather an

estimate of the number of bases that UNCALLED attempted to align. In most cases this will be equal to “query end”. This may be changed to better reflect the full read length in future versions.

Both modes include the following custom attributes in each PAF entry:

- **mt: map time.** Time in milliseconds it took to map the read.
- **ch: channel.** MinION channel that the read came from.
- **st: start sample.** Global *sequencing* start time of the read (in signal samples, 4000 samples/sec).

`uncalled realtime` also includes the following attributes:

- **ej: ejected.** Time that the eject signal was sent, in milliseconds since last chunk was received.
- **kp: kept.** Time that UNCALLED decided to keep the read, in milliseconds since last chunk was received.
- **en: ended.** Time that UNCALLED determined the read ended, in milliseconds since last chunk was received.
- **mx: mux scan.** Time that the read *would have* been ejected, had it not have occurred within a mux scan.
- **wt: wait time.** Time in milliseconds that the read was queued but was not actively being mapped, either due to thread delays or waiting for new chunks.

pafstats

We have included a functionality called `uncalled pafstats` which computes speed statistics from a PAF file output by UNCALLED. Accuracy statistics can also be included if provided a ground truth PAF file, for example based on [minimap2](https://github.com/lh3/minimap2) alignments of basecalled reads. There is also an option to output the original UNCALLED PAF annotated with comparisons to the ground truth.

Example:

```
1 > uncalled pafstats -r minimap2_alns.paf -n 5000 uncalled_out.paf
2 Summary: 5000 reads, 4373 mapped (89.46%)
3
4 Comparing to reference PAF
5   P      N
6 T  88.74  6.80
7 F   0.60  3.74
8 NA: 0.12
9
10 Speed           Mean      Median
11 BP per sec:    4878.24   4540.50
12 BP mapped:     636.29    362.00
```

13	MS to map:	140.99	89.96
----	------------	--------	-------

Positional arguments - `infile` PAF file output by UNCALLED

Optional arguments - `-n/--max-reads` maximum number of reads to parse - `-r/--ref-paf` ground-truth alignments (from minimap2) to compute TP/TN/FP/FN rates - `-a/--annotate` if used with `-r`, will output PAF with “rf:” tag indicating TP, TN, FP, or FN

Accuracy statistics: - TP: true positive - percent infile reads that overlap reference read locations - FP: false positive - percent infile reads that do not overlap reference read locations - TN: true negative - percent of reads which were not aligned in reference or infile - FN: false negative - percent of reads which were aligned in the reference but not in the infile - NA: not aligned/not applicable - percent of reads aligned in infile but not in reference. Could be considered a false positive, but the truth is unknown.

Practical Considerations

For ReadUntil sequencing, the first decision to make is whether to perform **enrichment** or **depletion** (`--enrich` or `--deplete`). In enrichment mode, UNCALLED will eject a read if it *does not* map to the reference, meaning your target should be the reference. In depletion mode, UNCALLED will eject a read if it *does* map to the reference, meaning your target should be everything except your reference.

Note that enrichment necessitates a quick decision as to whether or not a read maps, since you want to eject a read as fast as possible. Usually ~95% of reads can be mapped within three seconds for highly non-repetitive references, so setting `-c/--max-chunks-proc` to 3 generally works well for enrichment. The default value of 10 works well for depletion. Note these values assume `--chunk-size` is set to the default 1 second.

UNCALLED currently does not support large (> ~1Gbp) or highly repetitive references. The speed and mapping rate both progressively drop as references become larger and more repetitive. Bacterial genomes or small collections of divergent bacterial genomes typically work well. Small segments of eukaryotic genomes can also be used, however the presence of any repetitive elements will harm the performance. Collections of highly similar genomes will not work well, as conserved sequences introduce repeats. See masking for repeat masking scripts and guidelines.

ReadUntil works best with longer reads. Maximize your read lengths for best results. You may also need to perform a nuclease flush and reloading to achieve the highest yield of on-target bases.

UNCALLED currently only supports reads sequenced with r9.4.1/r9.4 chemistry.

Release notes

- v2.2: added event profiler which masks out pore stalls, and added compile-time debug options
- v2.1: updated ReadUntil client for latest MinKNOW version, made `uncalled index` automatically build the BWA index, added hdf5 submodule, further automated installation by auto-building hdf5, switched to using setuptools, moved submodules to submods/
- v2.0: released the ReadUntil simulator `uncalled sim`, which can predict how much enrichment UNCALLED could provide on a given reference, using a control and UNCALLED run as a template. *Also changed the format of certain arguments:* index prefix and fast5 list are now positional, and some flags have changed names.
- v1.2: fixed indexing for particularly large or small reference
- v1.1: added support for altering chunk size
- v1.0: pre-print release