camlhmp - Classification through yAML Heuristic Mapping Protocol

Robert A. Petit III[1,2], Chayse Rowley[1], Taylor Fearing[1], Jim Mildenberger[1], Rob Christensen[1], Timothy D. Read[2], Joseph M. Reed[1]

1. Wyoming Public Health Laboratory, Wyoming Department of Health, Cheyenne, Wyoming, USA
2. Center of Applied Pathogen Epidemiology and Outbreak Response, Georgia PGCoE, Atlanta, Georgia, USA

camlhmp *("camel hump")*

Typing organisms is a common first-step for understanding their genotypes. Traditionally reliant on PCR-based methods, the field is now moving towards whole genome sequencing, necessitating sequence-based typing tools. Yet, developing these tools often demands significant expertise and occurs in isolation, limiting standardization and community involvement.

Recognizing these challenges, we developed camlhmp (***C**lassification through y**AML H**euristic **M**apping **P**rotocol*). This bioinformatics framework uses YAML to streamline the creation and management of sequence-based typing tools, no matter the experience level.

To demonstrate camlhmp's utility, we *(detailed below)* developed typing tools for *Pseudomonas aeruginosa*, *Streptococcus pneumoniae*, and *Staphylococcus aureus*. Learn more about camlhmp at rpetit3.github.io/camlhmp/

camlhmp powered tools are easy to install and run

camlhmp powered tools prioritize availability from Bioconda as well as user-friendly inputs and outputs. Below is an example from the sccmec tool.

```
(sccmec) rpetit3@wsl2:$ sccmec --input completed-genomes/GCF_022691365.1_ASM2269136v1_genomic.fna.gz
Running sccmec (via camlhmp) with following parameters:
    --input completed-genomes/GCF_022691365.1_ASM2269136v1_genomic.fna.gz
    --yaml-targets /home/rpetit3/miniforge3/envs/sccmec/bin/../share/sccmec/sccmec-targets.yaml
    --yaml-regions /home/rpetit3/miniforge3/envs/sccmec/bin/../share/sccmec/sccmec-regions.yaml
    --targets /home/rpetit3/miniforge3/envs/sccmec/bin/../share/sccmec/sccmec-targets.fasta
    --regions /home/rpetit3/miniforge3/envs/sccmec/bin/../share/sccmec/sccmec-regions.fasta
    --outdir ./
    --prefix sccmec
    --min-targets-pident 90
    --min-targets-coverage 80
    --min-regions-pident 85
    --min-regions-coverage 83
Starting camlhmp for SCCmec Typing (targets)...
Running blastn...
Processing target hits...
Starting camlhmp for SCCmec Typing (regions)...
Running blastn...
Processing region hits...
Final Results...
                                        SCCmec Typing
| sample | type | subtype | mecA | target_schema | target_schema_version | region_schema | region_schema_version | camlhmp_version |
| sccmec | IV | IVj | + | sccmec_targets | 1.2.0 | sccmec_regions | 1.2.0 | 1.1.0 |
Final predicted type written to ./sccmec.tsv
Target-based results against each type written to ./sccmec.targets.details.tsv
Target-based blastn results written to ./sccmec.targets.blastn.tsv
Region-based results against each type written to ./sccmec.regions.details.tsv
Region-based blastn results written to ./sccmec.regions.blastn.tsv
```

It's simple to define a schema

camlhmp schemas are easily defined in YAML format due to its simplicity and readability. Below is an example of a partial schema for typing SCCmec cassettes.

```yaml
%YAML 1.2
---
# metadata: general information about the schema
metadata:
  id: "sccmec_demo"                                  # unique identifier for the schema
  name: "SCCmec Typing"                              # name of the schema
  description: "A demo schema for SCCmec typing"     # description of the schema
  version: "0.0.1"                                   # version of the schema
  curators:                                          # A list of curators of the schema
    - "Robert A. Petit III"

# engine: specifies the computational tools and additional parameters used for sequence
#         analysis.
engine:
  type: blast   # The type of tool used to generate the data
  tool: blastn  # The tool used to generate the data
  params:       # Additional parameters for the tool
    min_pident: 80     # Minimum percent identity for the tool
    min_coverage: 80   # Minimum percent coverage for the tool

# targets: Lists the specific sequence targets such as genes, proteins, or markers that
#          the schema will analyze. These should be included in the associated sequence
#          query data
targets:
  - "ccrA1"
  - "ccrA2"
  - "ccrB1"
  - "ccrB2"
  - "IS431"
  - "IS1272"
  - "mecA"
  - "mecI"
  - "mecR1"

# aliases: groups multiple targets under a common name for easier reference
aliases:
  - name: "ccr Type 1"           # name of the alias
    targets: ["ccrA1", "ccrB1"]  # list of targets that are part of the alias
  - name: "ccr Type 2"
    targets: ["ccrA2", "ccrB2"]
  - name: "mec Class A"
    targets: ["IS431", "mecA", "mecR1", "mecI"]
  - name: "mec Class B"
    targets: ["IS431", "mecA", "mecR1", "IS1272"]

# types: define specific combinations of targets and aliases to form distinct types
types:
  - name: "I"          # name of the profile
    targets:           # list of targets or aliases that are part of the profile
      - "ccr Type 1"
      - "mec Class B"
  - name: "II"
    targets:
      - "ccr Type 2"
      - "mec Class A"
  - name: "IV"
    targets:
      - "ccr Type 2"
      - "mec Class B"
```

Highlights

Avaialble from PyPi and Bioconda
Easily installed through standard Python packaging tools as well as Conda.

Command Line Interface (CLI) Tools
Pre-built CLI commands that can act as wrappers for simple schemas

Application Programming Interface (API)
Available API end points allow developer customization to meet advanced schema needs

Detailed Documentation
Extensive documention for developing with camlhmp is available, including CLI and API

Bactopia Support
All typing tools developed using camlhmp will be supported by Bactopia for seamless analysis

AllTheBacteria Assemblies
Invesitigate nearly 2,000,000 assemblies from the AllTheBacteria project with camlhmp-powered tools

Future Plans

Typing Directly from Reads
Add support for *k*-mer based typing to allow users to directly type sequence reads without assembly

More camlhmp powered Tools
Additional tools will be developed utilizing camlhmp *(hopefully from the community!)*

Examples Tools

With camlhmp you can rapidly start building typing tools for your organism of interest.

Here are a few pre-built examples

pasty

pasty provides a camlhmp schema for serogrouping *Pseudomonas aeruginosa* samples. The serogroup is determined based on the analysis of the O-specific antigen (OSA) gene cluster

https://github.com/rpetit3/pasty

pbptyper

pbptyper provides a camlhmp schema for typing the penicillin binding protein (PBP) in *Streptococcus pneumoniae* samples. PBP typing is based on the alleles present in three PBPs: PBP1a, PBP2b, and PBP2x.

https://github.com/rpetit3/pbptyper

sccmec

sccmec provides a camlhmp schema for typing 15 known SCCmec cassettes, which confer broad-spectrum beta-lactam resistance in Methicillin Resistant *Staphylococcus aureus* (MRSA) samples.

https://github.com/rpetit3/sccmec

Scan to learn more!



Ask for a sticker!