

Development and Beta-Testing of a Bioinformatics Pipeline for Metagenomic Analysis

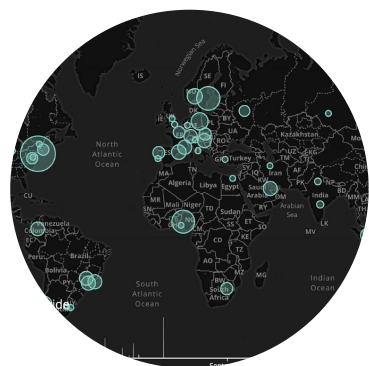
MetaSUB

MetaSub's goal is to map the urban microbiome to build a molecular profile of cities around the globe.
<http://metasub.org/>

The collected samples are then sequenced, using Illumina HiSeq, and analyzed to develop a genetic and epigenetic map detailing the microbiome in each participating city.

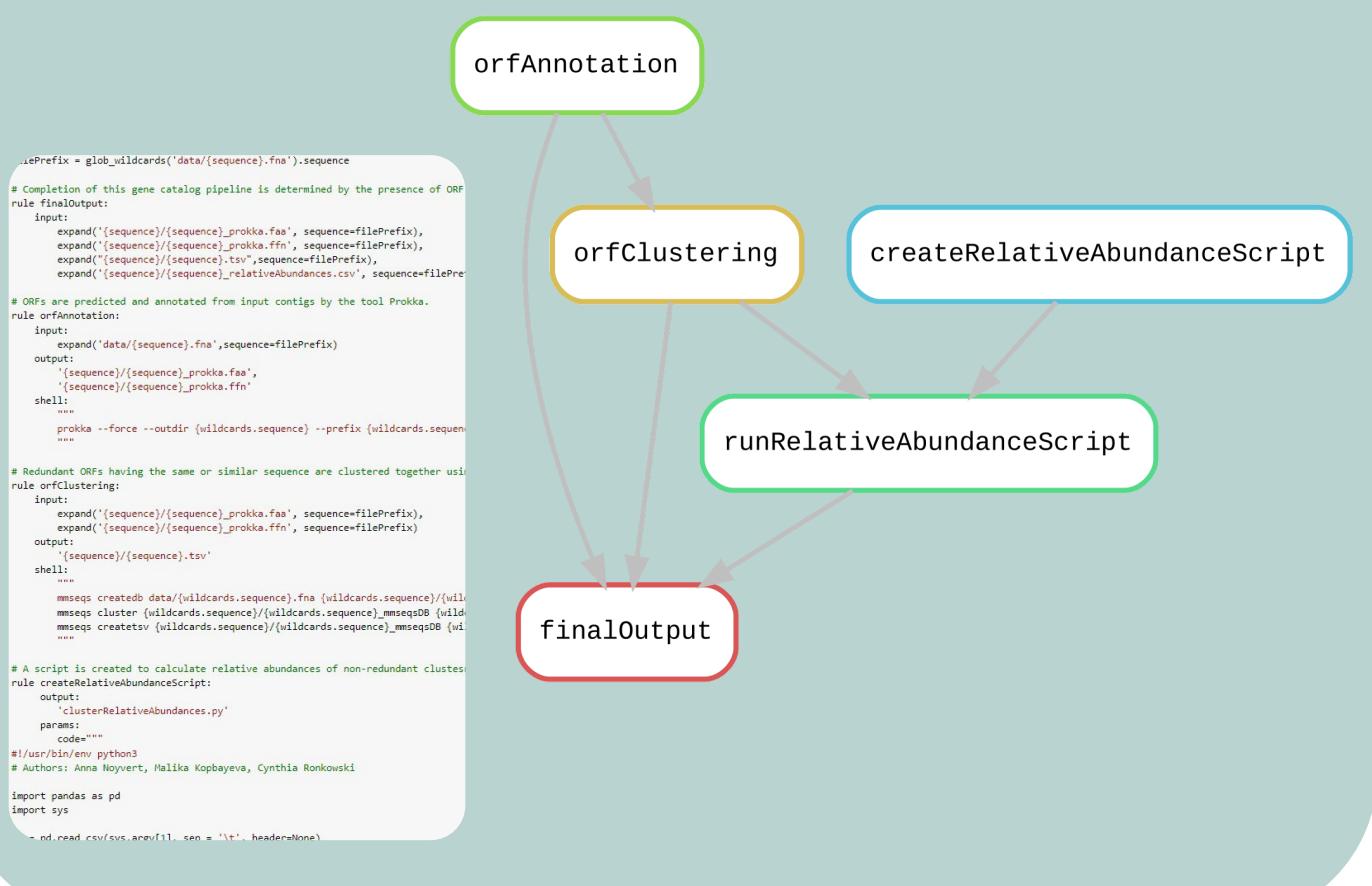
Global City Sampling Day (gCSD) is an event that takes place annually on June 21st. During gCSD, samples from surfaces in subways, buses, airports and other well-traveled public spaces are collected across the world's cities.

The MetaSUB bioinformatics working group is developing a core set of tools that can be run on metagenomes collected during gCSD and is highly reproducible and extendable.



Gene catalog module

- ORF prediction/annotation using Prokka
- Non-redundant ORF clustering using MMseqs2
- Relative ORF abundance quantification with our own script

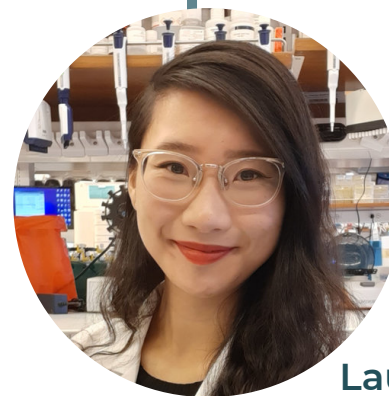


Students: Anna Noyvert, Malika Kopbayeva
 Mentors: Cynthia Ronkowski, Anton Katsuba, Serghei Mangul PhD

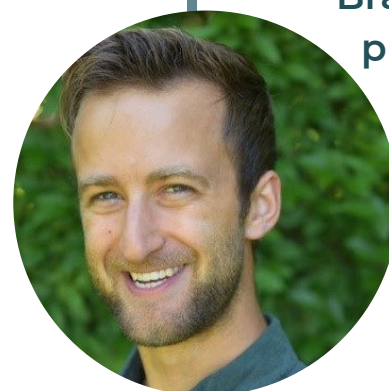
CAMP

Core Analysis Modular Pipeline

- Standardized quality control and abundance outputs
- Alternative to traditional pipeline
- Easily updated by any software developer
- Distributed, long-term development approach
- Integrates cutting-edge bioinformatics tools



Lauren Mak - project leader



Braden Tierney - project leader

- Modules under development:
- Short read quality control
 - Short read assembly
 - Gene catalog
 - Consensus binning
 - Short read taxonomy
 - Assembled virus characterization

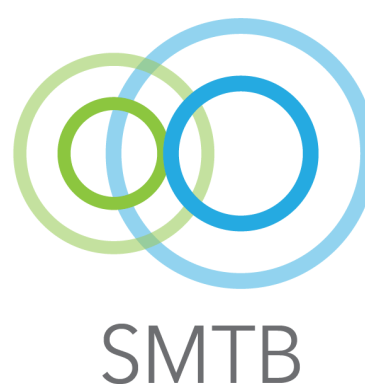


QR code for CAMP

Beta-testing

- Include rulegraph in the main README.
- Document rules in Snakefile with comments.
- References to {{Cookiecutter.module_slug}} are unclear, especially if the meaning of slug is unknown.
- Usage of configuration .yaml files is unclear.
- Link for fixing cases of dependency conflicts is in question.

Acknowledgements



Разработка и бета-тестирование биоинформатического пайплайна для метагеномного анализа

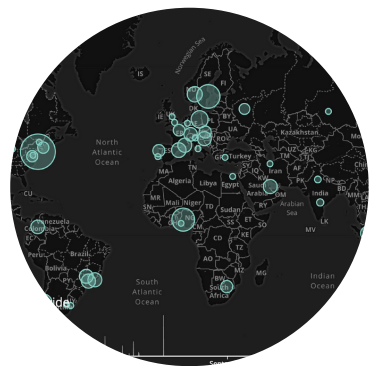
MetaSUB

Цель MetaSub - составить карту городского биома для создания молекулярного профиля городов по всему миру, чтобы улучшить их дизайн, функциональность и влияние на здоровье.

Глобальный день сборки проб в городах - это мероприятие, которое проводится ежегодно 21 июня. Во время гСП в разных городах мира собираются образцы с поверхностей в метро, автобусах, аэропортах и других общественных местах с интенсивным трафиком.

Рабочая группа MetaSUB разрабатывает основной набор инструментов, который может быть использован на каждом метагеноме, собранном во время гCSD, и обладает высокой воспроизводимостью и расширяемостью.

Затем собранные образцы секвенируются с использованием Illumina HiSeq и анализируются для составления генетической и эпигенетической карты микробиома каждого города-участника.



Студенты: Анна Нойверт, Малика Копбаева
Наставники: Синтия Ронковски, Сергей Мангул, Антон Кацубо

CAMP

Core Analysis Modular Pipeline

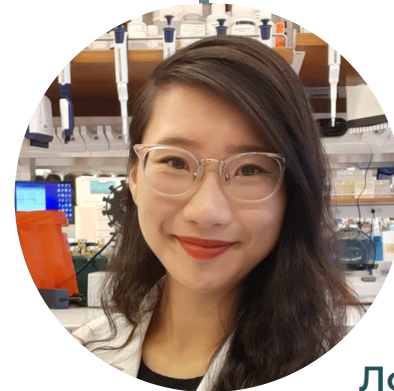
Стандартизированный контроль качества и выходные данные по изобилию

Альтернатива традиционному пайплайну

Легкое для обновления любым разработчиком программного обеспечения

Распределенный, долгосрочный подход к разработке

Интеграция передовых инструментов биоинформатики



Лорен Мак,
ведущий проекта



Брэйден Тирни,
ведущий проекта

Модули в стадии разработки:

- Контроль качества коротких ридов
- Сборка коротких ридов
- Каталог генов
- Консенсусный биннинг
- Таксономия коротких ридов
- Характеризация собранных вирусов



QR код для CAMP

Модуль генного каталога

- Прогнозирование & аннотация OPC (открытых рамок считывания) с помощью Prokka
- Кластеризация избыточных OPC с помощью MMseqs2
- Количественная оценка относительного изобилия OPC с помощью своего кода



orfAnnotation

orfClustering

createRelativeAbundanceScript

runRelativeAbundanceScript

finalOutput

```
!PREFIX = glob_wildcards('data/{sequence}.fna').sequence
# Completion of this gene catalog pipeline is determined by the presence of ORF
rule finalOutput:
  input:
    expand('{sequence}/{sequence}_prokka.faa', sequence=filepathPrefix),
    expand('{sequence}/{sequence}_prokka.ffn', sequence=filepathPrefix),
    expand('{sequence}/{sequence}.tsv', sequence=filepathPrefix),
    expand('{sequence}/{sequence}_relativeAbundances.csv', sequence=filepathPrefix)
  output:
    'finalOutput'
  shell:
    ""

# ORFs are predicted and annotated from input contigs by the tool Prokka.
rule orfAnnotation:
  input:
    expand('data/{sequence}.fna', sequence=filepathPrefix)
  output:
    '{sequence}/{sequence}_prokka.faa',
    '{sequence}/{sequence}_prokka.ffn'
  shell:
    ""
    prokka --force --outdir {wildcards.sequence} --prefix {wildcards.sequence}

# Redundant ORFs having the same or similar sequence are clustered together using MMseqs2.
rule orfClustering:
  input:
    expand('{sequence}/{sequence}_prokka.faa', sequence=filepathPrefix),
    expand('{sequence}/{sequence}_prokka.ffn', sequence=filepathPrefix)
  output:
    '{sequence}/{sequence}.tsv'
  shell:
    ""
    mmseqs createdb {wildcards.sequence}.faa {wildcards.sequence}/{wildcards.sequence}_mmseqsDB
    mmseqs cluster {wildcards.sequence}/{wildcards.sequence}_mmseqsDB {wildcards.sequence}
    mmseqs createtsv {wildcards.sequence}/{wildcards.sequence}_mmseqsDB {wildcards.sequence}

# A script is created to calculate relative abundances of non-redundant clusters.
rule createRelativeAbundanceScript:
  output:
    'clusterRelativeAbundances.py'
  params:
    code=""
  shell:
    ""
    cat <<CODE >>{output}
    #!/usr/bin/env python3
    # Authors: Anna Noyvert, Malika Kopyayeva, Cynthia Ronkouski
    import pandas as pd
    import sys

    # Read contig coordinates from file.
    # Read contig coordinates from file.
```

Beta-testing

- Включение графика правил в основной README.
- Документация правил в Snakefile с комментариями.
- Дополнение документации {{Cookiecutter.module_slug}} с объяснением значения slug.
- Использование конфигурационных файлов .yaml неясно.
- Исправление вспомогательной ссылки для случаев конфликтов зависимостей.

Благодарность



Howard Hughes Medical Institute