BAP – Bacterial Analysis Pipeline

CoP Bioinformatics

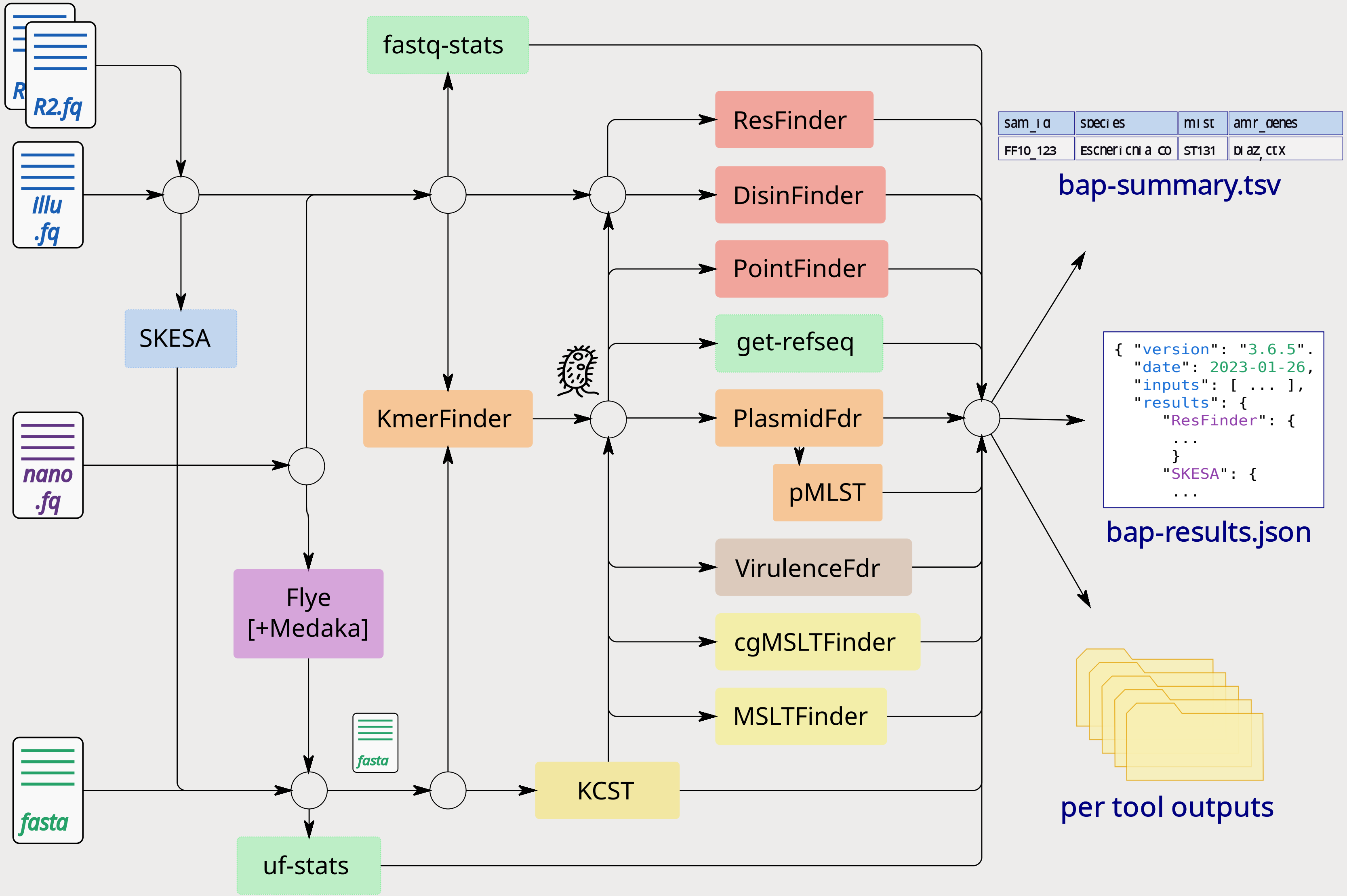
2 May 2024

Marco van Zwetselaar

The “BAP”

- CGE Bacterial Analysis Pipeline
  - https://github.com/zwets/cge-bap
- History
  - 2013/14 KCRI starts WGS: collaboration with DTU
  - Sequencing *and* bioinformatics analysis at KCRI (Moshi, Tanzania)
  - DTU CGE services & databases “ported” from Computerome to Laptop (workstation)
  - Online CGE services: BAP project at DTU
    - Run all services in one go
    - Run multiple samples in batch
  - KCRI CGE BAP started from the “CGE BAP Backend” (Docker)

CGE BAP 3.7.0 (“The BAP”)



sam_id	species	mlst	amr_genes
FF10_123	Escherichia coli	ST131	blaz, ctx

bap-summary.tsv

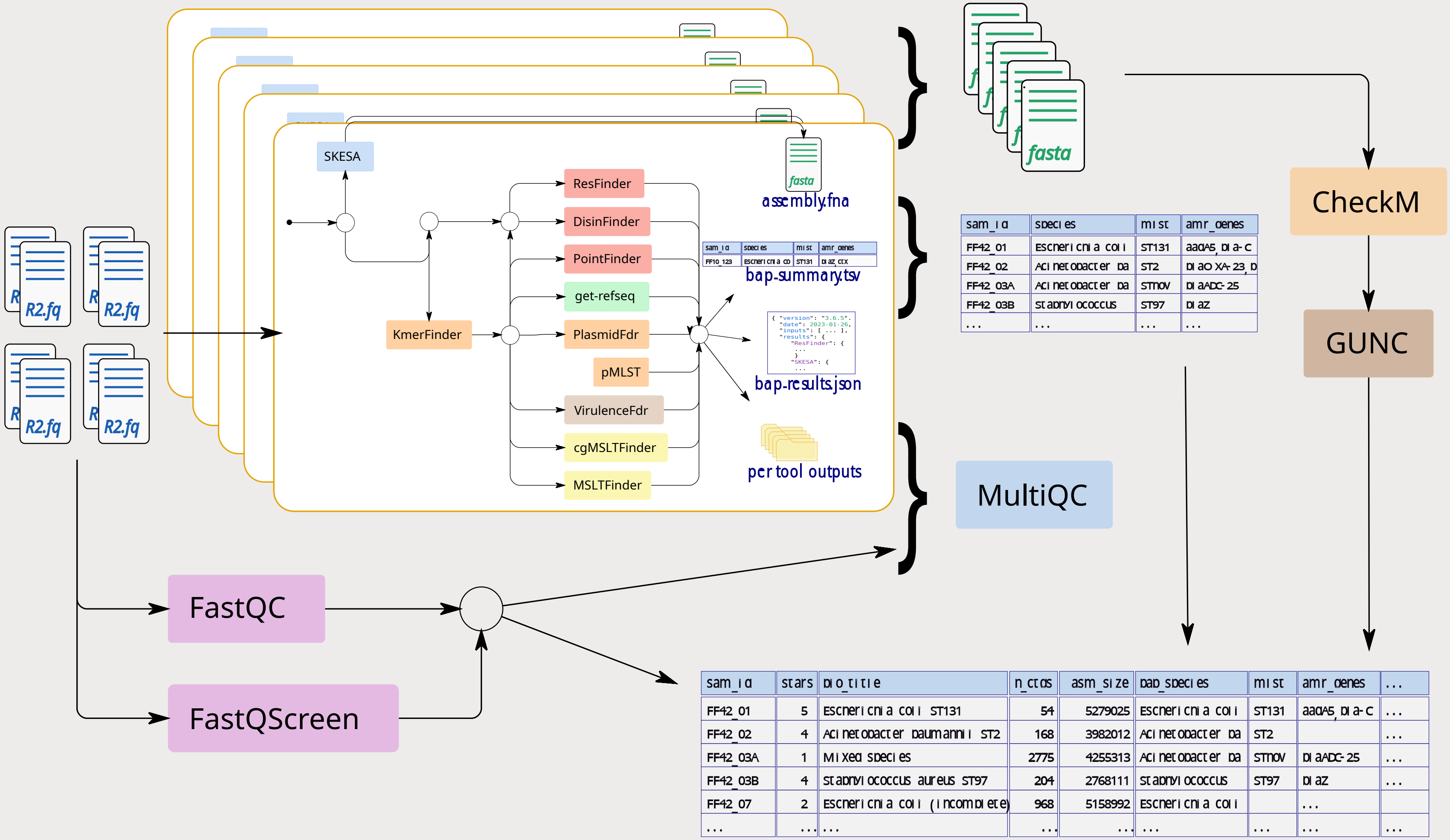
```
{ "version": "3.6.5".
  "date": 2023-01-26,
  "inputs": [ ... ],
  "results": {
     "ResFinder": {
      ...
      }
     "SKESA": {
      ...
```

bap-results.json

Bacterial Analysis Pipeline

- Standard workflow
  - Integrates the CGE Services
  - Default analyses & settings
  - Configurable “a la carte”
- One isolate at a time
  - Illumina or Nanopore reads
  - Genome assembly
- Open Source
  - github.com/zwets/cge-bap
- Moderate hardware
  - 32G laptop (Linux / MacOS)

The “Project BAP” workflow



sam_id	species	mlst	amr_genes
FF42_01	Escherichia coli	ST131	aadA5, bla-C
FF42_02	Acinetobacter ba	ST2	blaOXA-23, b
FF42_03A	Acinetobacter ba	STnov	blaADC-25
FF42_03B	Staphylococcus	ST97	blaZ
...	...	...	...

sam_id	stars	bio_title	n_ctgs	asm_size	bap_species	mlst	amr_genes	...
FF42_01	5	Escherichia coli ST131	54	5279025	Escherichia coli	ST131	aadA5, bla-C	...
FF42_02	4	Acinetobacter baumannii ST2	168	3982012	Acinetobacter ba	ST2		...
FF42_03A	1	Mixed species	2775	4255313	Acinetobacter ba	STnov	blaADC-25	...
FF42_03B	4	Staphylococcus aureus ST97	204	2768111	Staphylococcus	ST97	blaZ	...
FF42_07	2	Escherichia coli (incomplete)	968	5158992	Escherichia coli		...	
...	...	...	...	...	...	...	...	...

Project BAP

- “Batching the BAP”
  - Process all isolates of a project
  - Collate & aggregate
  - Bundle for customer
- Add QC, flag issues
  - Genome completeness
  - Contamination & Mixes
  - Quality rating: 0-5 star
  - “Label” assembly title
- Utilise HPC facilities
  - HTCondor job batching
  - HTCondor workflow control
  - Include heavyweight tools
  - The ‘genomics inventory’ (gi)

Concluding

- Stable workhorse for routine work
  - Minimal hands-on time, fully automated
  - Default settings, standard outputs
  - Keep up to date (centrally) with latest tools & databases
  - Can run on moderate hardware (current-day mid-range laptop)

But

  - World has moved on (esp internet connectivity!)
  - Not well suited for ONT work
  - All command-line, still need “bioinformatician support”

The
Fleming Fund
Regional Grants

Thank you

NOGUCHI MEMORIAL INSTITUTE
FOR MEDICAL RESEARCH
UNIVERSITY OF GHANA , LEGON

NATIONAL INSTITUTE FOR
COMMUNICABLE DISEASES
Division of the National Health Laboratory Service

INSTITUT
PASTEUR
de Dakar

UKaid
from the British people

MOTT
MACDONALD

SEQAFRICA

DTU

This programme is being funded by the UK Department of Health and Social Care.
The views expressed do not necessarily reflect the UK Government’s official policies.