A vision for Bioinformatics within the Global FAANG project

Peter Harrison (EMBL-EBI)

peter@ebi.ac.uk

@peterwharrison

Mick Watson (Roslin Institute)

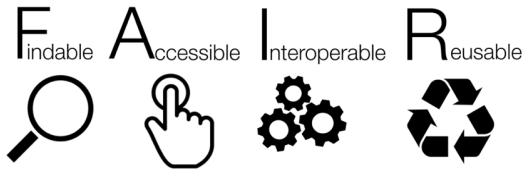
mick.watson@roslin.ed.ac.uk

@BioMickWatson

EMBL

The challenge

- The global FAANG initiative consists of hundreds of researchers, across multiple funded projects and from many institutions spread all over the world.
- Loose coordination through FAANG committees, mailing lists and conference meetings, have achieved a lot scientifically thus far.
- However, from a Bioinformatics perspective how do we with this reality ensure:
  - reproducible and comparable research across multiple projects and species
  - we don't duplicate code developmental effort
  - rapid and accurate research
  - FAANGs principles of open science and FAIR data



Images: CODATA, faang.org

The vision

- Shared development of a complete set of open pipelines across the entire set of FAANG (and associated) projects for all FAANG sequencing technologies and compatible with all species.
- Development based on the principles of open science, open source code and reproducible workflows and environments.
- Researchers improve reproducibility and prevent duplicated development effort by reusing or improving a common set of FAANG pipelines.
- An active and collaborative FAANG Bioinformatics software community.

The technology to achieve this vision already exists

FAANG
Functional Annotation of Animal Genomes

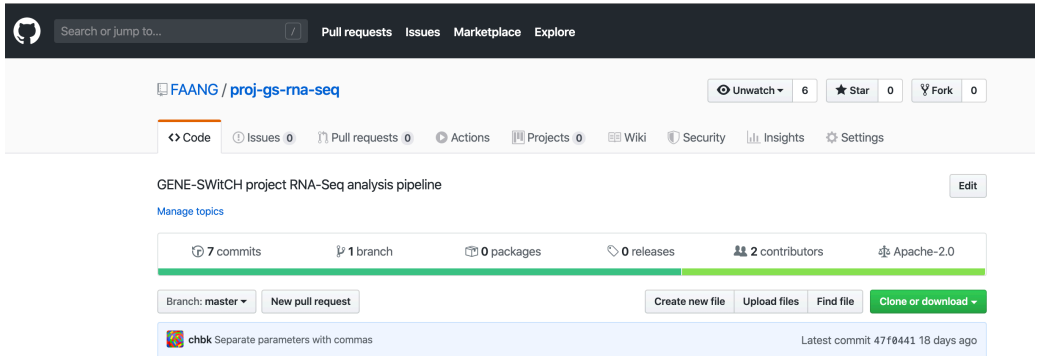
- Ensuring usage of identical pipelines and maximum reproducibility.
- Workflow managers, such as Nextflow, enable easy pipeline construction from existing components and manage data flow scalability.
- Containerisation technologies such as Docker, ensure an identical software environment can be distributed for the pipeline wherever it is installed.
- Cloud technologies that support the above ensure reproducibility as multiple projects use the same platform, accelerate research speed and lower cost.

nextflow

EMBASSY cloud

The first steps

- Ideally go straight to developing a single set of FAANG pipelines, but realistically this will take time.
- For now consortiums developing and maintaining their own pipelines should document and share them more effectively to help achieve our goal.
  - Publicly available permissively licensed pipelines (e.g. Apache or GNU).
  - In data submissions link to the pipeline that made it (there is a field for this).
  - Join Bioinformatics and Data Analysis working group to develop global pipelines.



The final goal

- All consortiums use a single set of approved FAANG pipelines, hosted on FAANG GitHub.
- Pipelines support branch points, e.g. different aligners, and workflows can have multiple end points e.g. an RNA-Seq workflow can have
  - (i) reference based quantification (e.g. StringTie),
  - (ii) reference-free quantification (e.g. Kallisto)
  - (iii) transcriptome assembly (e.g. Trinity)
- Simpler protocols that just record deviations from the default usage.

- All analysis data records in FAANG data portal will have the sample metadata, raw data, detailed protocols, publications, and a GitHub link to pipeline that made it. Full FAIR reproducibility, and ready for comparative analysis.

Key considerations

- Workflows that work on Desktop, HPC and cloud platforms.
- Pipelines should be deigned to exclusively pull data from public repos. Completely reproducible open science.
- Benchmarking is useful, but time consuming. There is no perfect pipeline – each has merits and problems – making things reproducible and documenting is more important that pursuing the perfect pipeline.
- Prevent duplicated effort wherever possible, share your existing pipelines, merge and improve upon a single set.
- Initial effort now will have huge payoff and within the lifetime of current projects.
- For reproducibility, version your software, and importantly also your reference genomes in your analysis.

Pipeline and data aware cloud analysis platforms

- Containerised FAANG analysis pipelines can be easily deployed yourself or be ‘pre-installed’ on cloud analysis platforms.
- Ensures same pipelines are run, lowest cost and maximum reproducibility.
- Choose your own cloud. EMBL-EBI Embassy cloud already has FAANG data colocalised in same data centre. CYVERSE investigating mirroring and preinstallation of approved FAANG pipelines.
- FAANG H2020 GENE-SWitCH project already utilising Embassy cloud for all analysis. Using Nextflow workflow management and will also investigate wrappers for FAANG APIs for data acquisition and automated submission.

EMBASSY cloud

CYVERSE

aws

Google Cloud

GENE-SWitCH: Open science with modern cloud analysis

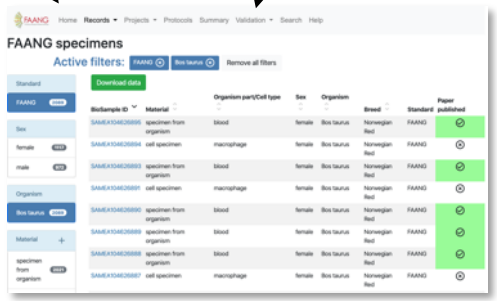
Sample metadata

Raw data immediately submitted

BioSamples

ENA

European Nucleotide Archive



- Data to archives as rapidly as possible.

Images: EMBL-EBI, Illumina

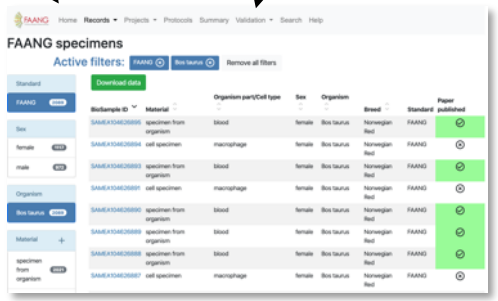
GENE-SWitCH: Open science with modern cloud analysis

Sample metadata

Raw data immediately submitted

BioSamples

ENA
European Nucleotide Archive



Colocalised/ synced raw data for analysis

docker
nextflow
EMBASSY cloud

GitHub

Latest GENE-SWitCH open pipelines (official FAANG pipelines later)

- Data to archives as rapidly as possible.
- Cloud analysis gets data from FAANG.
- Workflow could be automated on new data or updated pipelines.

Images: EMBL-EBI, Illumina

GENE-SWitCH: Open science with modern cloud analysis

Sample metadata

Raw data immediately submitted

BioSamples

ENA
European Nucleotide Archive

Colocalised/ synced raw data for analysis

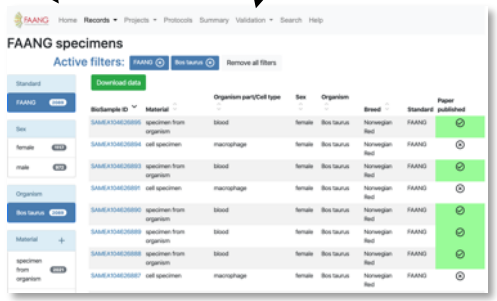
docker
nextflow
EMBASSY cloud

GitHub

Latest GENE-SWitCH open pipelines (official FAANG pipelines later)

ENA
European Nucleotide Archive

Analysis data



- Data to archives as rapidly as possible.
- Cloud analysis gets data from FAANG.
- Workflow could be automated on new data or updated pipelines.

Images: EMBL-EBI, Illumina

How do we get there: how can you contribute to the global effort

- Now is the time for renewed effort, lots of recently funded projects.
- Join the FAANG Bioinformatics and Data Analysis working groups at faang.org.
- Follow FAANG coding guidelines once established later this year.
- Contribute to FAANG pipelines in https://github.com/orgs/FAANG/
- Add your existing pipelines with permissive licenses.
- Don’t start now from scratch, improve and adapt existing pipelines from others.
- Want to be more involved? Request to manage development (issues and pull requests) of a particular global FAANG pipeline in GitHub.
- Always include a link to your pipeline in your FAANG analysis data submissions.

Peter Harrison

peter@ebi.ac.uk

@peterwharrison

- Alexey Sokolov (FAANG Project Manager)
- Jun Fan (FAANG Bioinformatician)
- Guy Cochrane (PI GENE-SWitCH)
- Paul Flicek (PI AQUA-FAANG)
- Daniel Zerbino (PI BovReg)

Mick Watson

mick.watson@roslin.ed.ac.uk

@BioMickWatson

- Richard Kuo (Bioinformatician)
- Andy Law (GENE-SWitCH)
- Alan Archibald (GENE-SWitCH)

Horizon 2020 European Union Funding for Research & Innovation

BBSRC bioscience for the future

Horizon 2020 European Union Funding for Research & Innovation