How data processing infrastructure simplifies biotech computing

Severiano Villarruel August 3, 2023

Personal Background

Personal Intro

Education

- Lewis and Clark College - Biochemistry
- University of Oregon - Bioinformatics and Genomics

Career Experience

- Reverse Translation Oncology Analyst - Genentech
- Web App Developer - Soteria
- Infrastructure Architect - UCSF
- Workflow Developer - Genentech

Expertise and Interests

- Infrastructure Development
- Software Development
- Web App Development

Case study: Extending infrastructure through development of cfDNAme pipeline

Goals and Overview

Goals

- Creating vs extending vs using infrastructure
- Provide a concrete example of extending infrastructure through pipeline development
- Demonstrate how infrastructure can accommodate web apps

Overview

- Pipeline background
- Data import and storage
- Pipeline development
    - Benchmark
    - Optimization
    - Best practices
- Infrastructure extension (Pipeline Integration)
- Web Apps

Pipeline Development

Biology

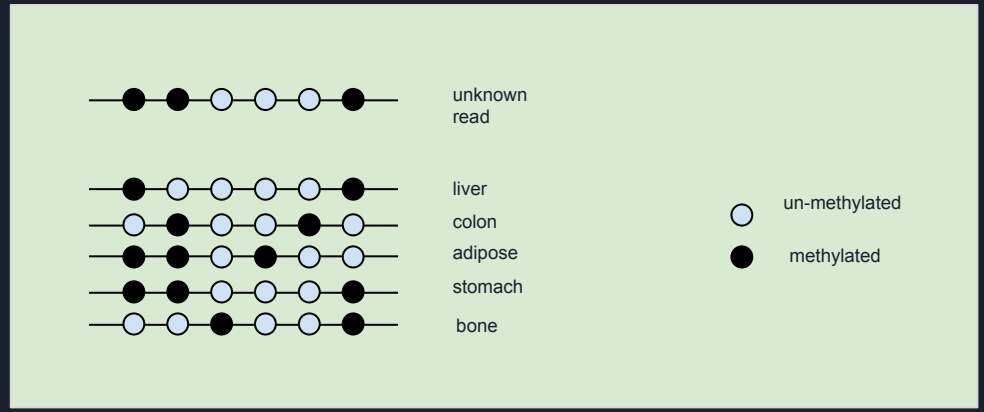
- Cell free DNA is indicative of tissue damage (shedding)
- Methylation signature can be used to identify tissues that harbor cell free DNA (cfDNAme)
- DNA methylation references can be used to detect the tissue signals (tissue damage) in blood plasma

Pipeline Methodology

- Create a tissue methylation reference using publicly available datasets

Pipeline Tasks

- QC
- Mapping
- Signature capture
- Output object formatting



Infrastructure Extension

Input Storage

- Import public datasets into storage system (e.g. AWS, GCP, HPC, etc.)
- Catalogued data on storage system (SQL)
- Accessible via unique ascension id

Pipeline Structure

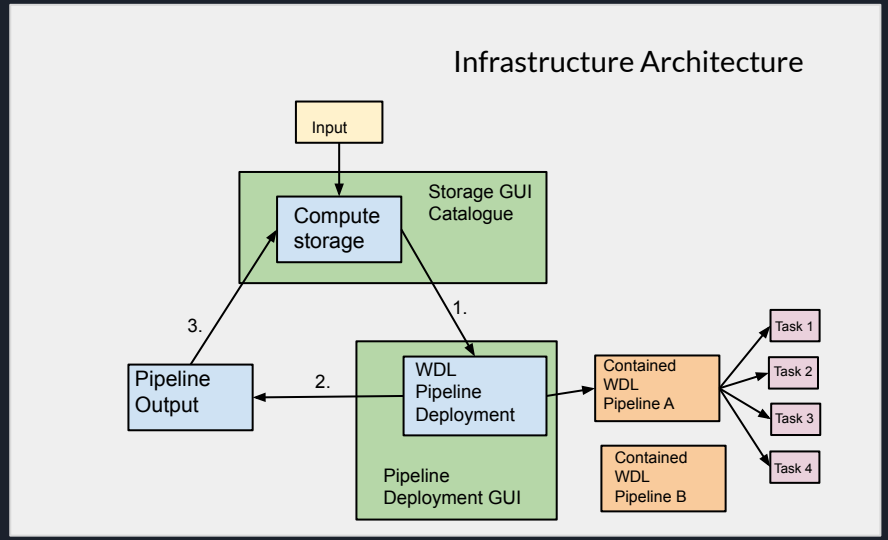
- Tool benchmark and resource estimation
- Pipeline task development within container
- Optimization (parallel-ization, resource requests, etc.)

Pipeline Deployment

- Resource allocation via pipeline management system (e.g. cromwell for WDL)

Output Storage

- Route output back to storage system
- Assign output with ascension, accessible via GUI



Pipeline Results

Metrics

- Processed over 100 different methylated DNA tissue datasets
- Datasets averaged 30GB
- Pipeline took ~18H to run
- Created tissue methylation reference (15 tissue types) in 1 week

Deliverables

- Pipeline is still in use by department
- Apps are still being built to handle output data

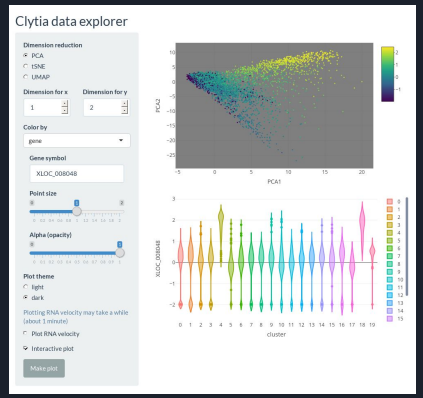
Interactive Web Apps

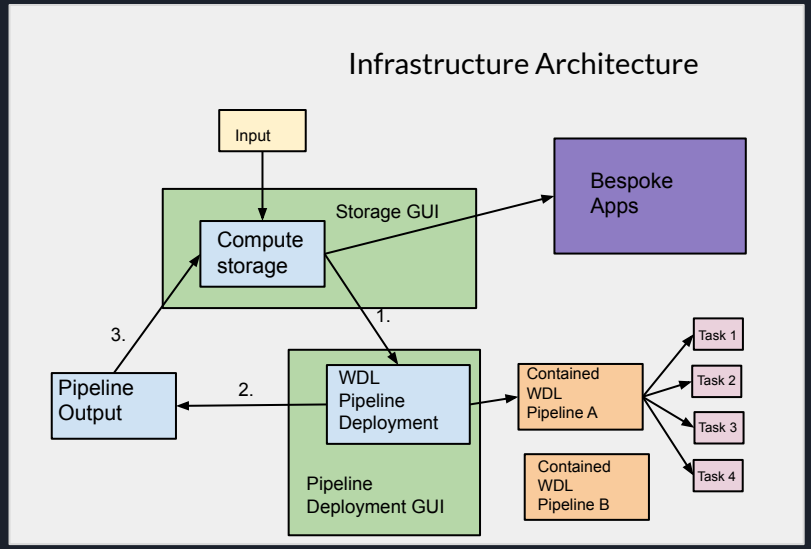
RShiny (Plotly)

- User friendly for data scientists
- Less customization
- Less hosting control
- Slow for big data

Javascript (D3, WebGL, React, Node.js)

- New paradigm
- Customizable and controllable
- Fast regardless of data's size





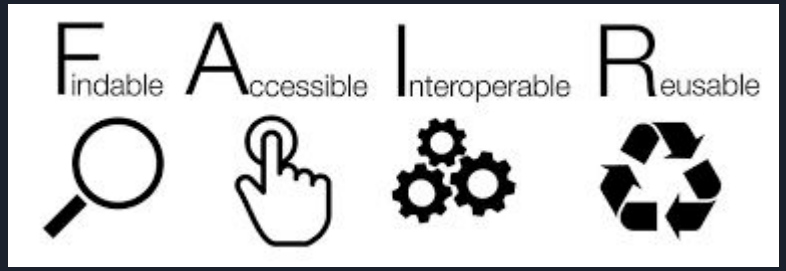
Conclusion

cfDNAme Pipeline takeaways

- Containerized development of individual tasks
- Input/ output handling
- Infrastructure extension (integration)

Infrastructure takeaways

- Extendability
- Organization and Accessibility
- Speed and Automation
- Reproducibility
- User friendly



Thank you!

Conclusion

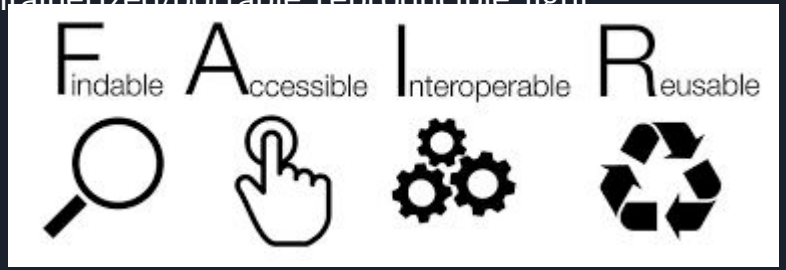
I have extended infrastructure, and also built infrastructure (UCSF)

Organize and make accessible data, data processing and apps

FAIR principles

- *Findability, Accessibility, Interoperability, and Reuse of digital assets*

Tackle the buzzwords (user friendly, automated, containerized/portable, reproducible, light weight)



Import and store input data + Pipeline dev + Integration into architecture (plus input output handling)

Data imported into the storage system (e.g. S3, storage dir) and catalogued in database

Tools in pipeline are benchmarked and the best tool for each step is chosen

Pipeline was optimized using parallelization

automated, portable and reproducible data processing pipeline (WDL, Singularity)

2015 — Graduated Biochemistry/Molecular Biology - Lewis and Clark College

2018 — Smithsonian Intern - Insect Taxonomy

2019 — Worked in Roderick/ Guillespie Lab - UC Berkeley, Biodiversity

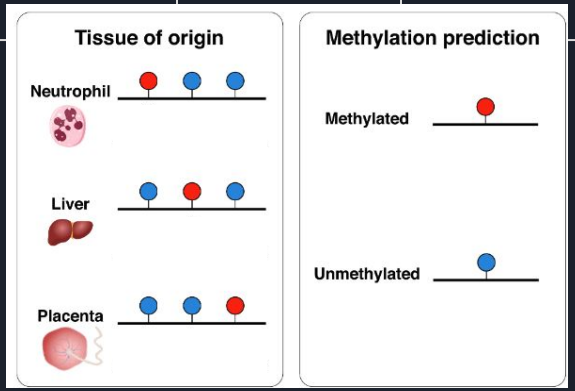
U Oregon Bioinformatics and Genomics MS

Genentech: Pipeline Dev for cfDNAmt data

2022 — UCSF: LIMS Dev

2023 — Genentech: RT Oncology

2024



Tina Moser, Stefan Kühberger, Isaac Lazzeri, Georgios Vlachos, Ellen Heitzer, Bridging biological cfDNA features and machine learning approaches, Trends in Genetics, 39, 4, (285-307), (2023).