AACR 2021

A Comprehensive Sample Tracking and Data Processing Workflow for Next Generation Sequencing

Abstract No. 2280

Chandra Sekhar Pedamallu[1], Joon Sang Lee[1], Shu Yan[1], Adalis Maisonet[1], Aleksandr Sidoruk[2], Tenghui Chen[1], Yulia Kamyshova[2], Mariia Zueva[2], Mark Magid[1], Quan Wan[1], Jeffrey Thompson[3], Valerie Zebrouck[3], Immanuel Gadaczek[3], Mikhail Alperovich[2], Brian McNatt[3], Alexei Protopopov[1], Donald Jackson[1], Jack Pollard[1]

[1]Sanofi Precision Oncology Research, Cambridge, MA, [2]EPAM, Boston, MA, [3]Sanofi Digital R&D, Cambridge, MA

Introduction

- Systems to manage and document sequencing experiments from sample receipt through data generation and data processing play a vital role in application of Next Generation Sequencing (NGS) to drug discovery.
- To track NGS lab processes, we adopted Benchling,™ which combines a digital notebook and a laboratory information management system (LIMS).
- We developed FONDA (Framework Of Next generation sequencing Data Analysis) to process NGS data and implemented it on a dockerized Amazon Web Services cloud platform.
- The current FONDA version (as of Nov 2020) has pipelines for single cell multi-omics (CITE-seq and scTCR/BCR-seq) and bulk RNA-seq.
- Integration of FONDA with Benchling enables users to access sequencing data, sample metadata and wet lab covariates (e.g., library construction kit information, sample QC metrics) easily without manual intervention.
- Availability and implementation: FONDA is implemented in Java and released under the Apache License 2.0. FONDA can be downloaded from GitHub at https://github.com/epam/fonda.

Benchling LIMS

- Benchling is a SaaS (software as a service) solution accessible from any lab computer with no installation required.
- Benchling documents and automates steps in the NGS process including sample registration, nucleic acid extraction, library construction, flow cell construction, sequencer sample sheet generation and BCL2FASTQ conversion (as shown in Figure 1).
- The flexible design of Benchling enables wet lab scientists to
  - Design and modify wet lab workflows.
  - Ease retrieve of appropriate protocol for each sample and sequencing library type.
  - Share wet lab workflows / protocols with other wet lab scientists across the groups
- Benchling captures meta data such as project name, study name, project type (clinical / pre-clinical), study type (target ID, MOA, biomarker), sample origin (tissue, PBMC, Blood, nucleic acid library), patient ID (applicable for clinical samples), sample type (tumor, normal), treatment information and reagents used in the lab with batch information.
- Benchling generates an “analysis ready sample sheet” with project and study information, location of FASTQ, sample species and library type. Then, this sample sheet is provided to FONDA for further analysis.
- Connectivity from Benchling to QC instruments and NGS sequencers enables us to
  - Track sample and library quality easily
  - Monitor sequencing runs and track NGS data quality.

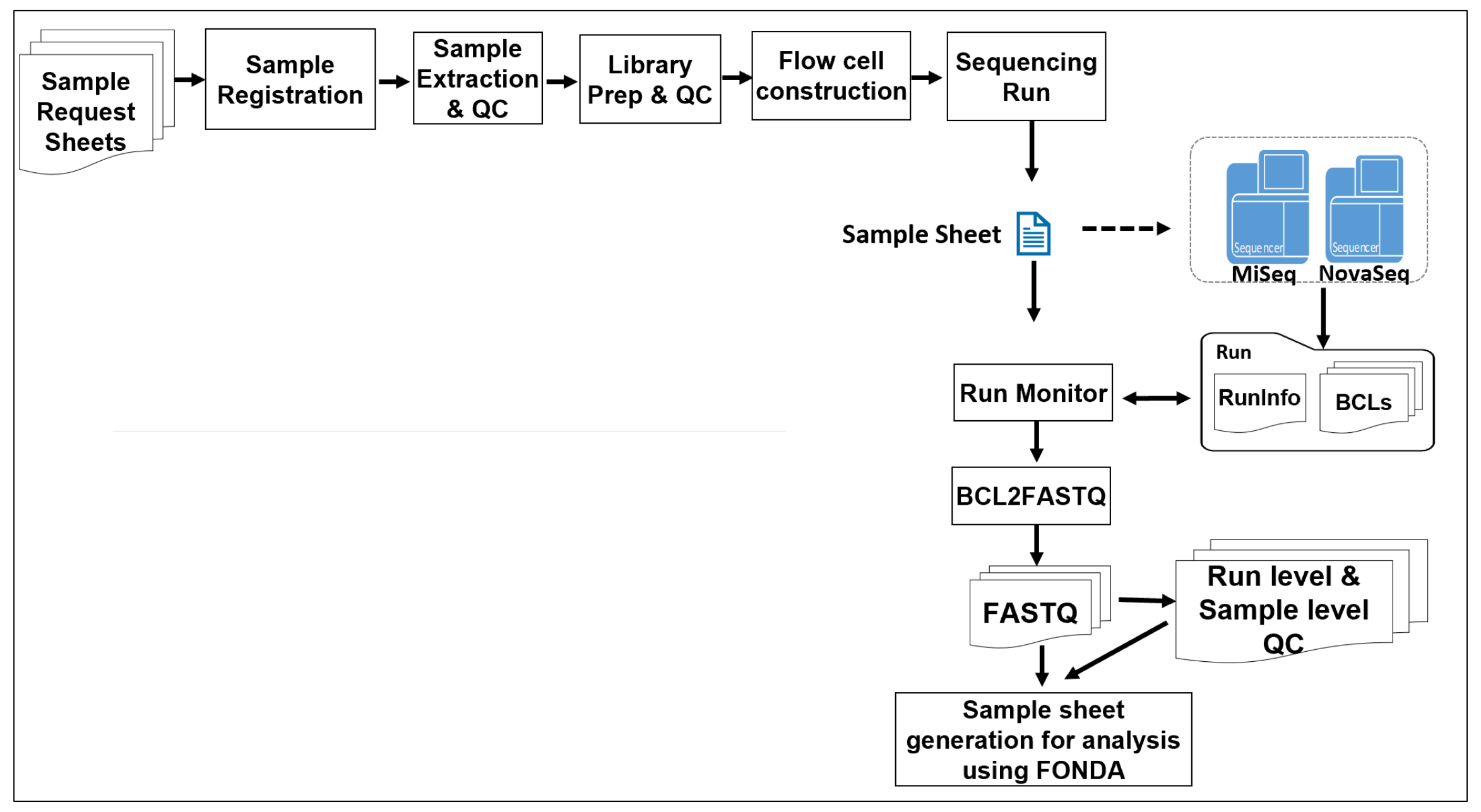


Figure 1. High level Benchling NGS process

Workflows in Benchling

- We developed sequencing protocol specific wet lab workflows using the reagent manufacturers protocols :
  - Bulk RNA sequencing (mRNA, Target Enrichment)
  - Bulk DNA sequencing (WGS, WES, Targeted sequencing)
  - Single Cell sequencing (scRNA, CITE-seq, scTCR-seq)
- We developed sequencing library type specific QC workflows to capture sequencing run level and sample level QC
- We developed workflows to transfer sequencing data and metadata to FONDA.

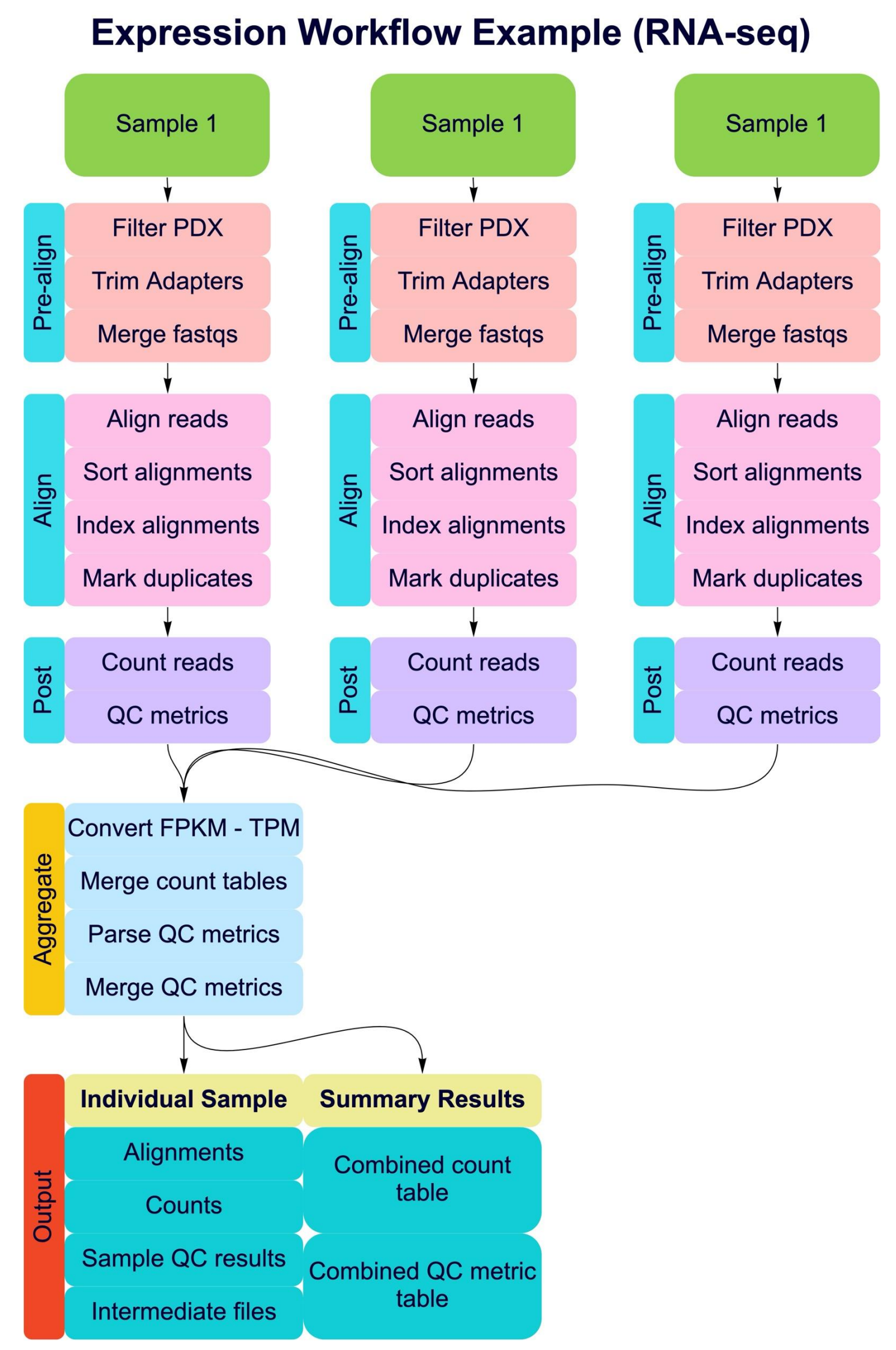
FONDA – Overview & Design

- FONDA compared to publicly available analytical frameworks such as snakemake, nextflow, invoke, gkno, WDL, Galaxy, and Toil has
  - Better module sharing between the pipelines
  - Ease to build and maintain complex workflows for data analysis
  - Portability of pipelines between SUNGRID engine and AWS
  - Good debugging ability
- To process data generated from workflows in a highly managed environment such as Benchling, FONDA was designed with a focus on automation, scalability and extensibility.
- FONDA is based on the organization of individual pipelines that are highly specific to wet lab workflows, and each pipeline's logic is hard coded into the framework as workflow objects.

Pipeline Details

- Each sample is processed individually and then sample level outputs are aggregated.
- For a given experiment incoming from Benchling, FONDA uses a predefined logic to generate shell scripts that process individual samples and their aggregated summary results into an output ready for secondary analysis. (Figure 2)

A)



B)

	RNA-seq	scRNA-seq
Input	FASTQ	FASTQ
Tools	Star, RNA-SeQC, Cufflinks, featureCount,	cell ranger, Doubletdetection, Scrublet
Outputs	FPKM, TPM, raw counts, qc metric	UMI counts (or mtx), qc metric

Figure 2. A) An example FONDA workflow (bulk RNA-seq). Individual samples are processed first, followed by aggregation of individual results and QC metrics. Blocks in figure correspond to individual scripts except for pre-align and align which are combined into a single script per sample, as well as output being actual files. B) Table indicating individual components of pipelines in FONDA.

FONDA - Implementation

- FONDA is written in Java and is executed as a portable JAR. It is implemented on SUNGRID engine enabled AWS.
- An internal template engine (Thymeleaf) collects pre-defined commands for individual tools within a pipeline (e.g. bwa, cufflinks) and combines them into independent shell scripts.
- Scripts are then executed as local processes or submitted to job schedulers such as Grid Engine. These scripts control pipeline processes directly.
- FONDA requires 3 configuration inputs:
  1) Sample manifest - table outlining sample information and input file path
  2) Global configuration - config file detailing dependency information and reference files
  3) Study configuration - config file detailing project specific information
- FONDA results include:
  - Result files produced in each sample directory
  - Aggregate and QC analysis results collected in combined project directory
  - Logs : All process outputs are captured by log files.

FONDA - Characteristics

- Ease
  - Pipelines are organized into pre-defined workflows for different types of sequencing data. Also, pipeline configuration provides the flexibility to set pipeline parameters.
- Scalability
  - Due to the “scatter and gather" structure of FONDA pipelines, each sample is processed parallelly and then aggregated once all samples finished processing. Total number of process initiated is always number of samples + 1.
- Extensibility
  - New workflows can also be accommodated through definition of new pipeline logic.
  - Pipelines can adapt existing tool templates using Thymeleaf (www.thymeleaf.org) templating engine to avoid conflicts between pipeline versions.

Summary

- Benchling interconnects the sample information including sample identifiers, cDNA/DNA and library identifiers, and all QC results.
- FONDA provides reproducible and scalable pipelines for NGS data analysis.
- Integrating Benchling and FONDA enables better traceability and reproducibility.