**Project Name: Watermelon**

**Project Description:** A modular pipeline to efficiently operationalize the RNA-Seq workflow of the Bioinformatics Core.

**Problem Statement**

*Why are we doing this?*

The Bioinformatics Core needs an efficient, up-to-date and modular pipeline that can handle large RNA-Seq project loads and produce reliable high quality results. The existing Legacy RNA-Seq pipeline is out-of-date, inflexible, crashes frequently and takes too long to run. The lack of documentation combined with design complexity makes it difficult to update, debug and maintain functionality.

We propose developing Watermelon—a modular pipeline to efficiently operationalize the RNA-Seq workflow. The goal of Watermelon is to increase the quality and efficiency of RNA-Seq analysis, cut analysis time significantly, and lead to faster project throughput. It will generate high quality, reproducible results that can meet the basic RNA-Seq analysis needs of our customers.

*We will know that we have succeeded when:*

* The new pipeline replaces essential capabilities in the legacy pipeline
* The new pipeline adds:
  + more robust QC
  + more efficient alignment
  + multiple diffex analysis callers and better support for multi-group comparisons, covariants, etc.
  + better summary reporting
* The new pipeline is deployed and in use by BFXCore analysts; the old pipeline is “removed”

**Key Goals for Pipeline Implementation**

1. Simple, modular standard pipeline framework
   * Enables regular, updates to established, industry-standard protocols and workflows
2. Reproducible
   * Usable by all Bioinformatics core staff; reproducible across analysts (any analyst at the core can run it and get the exact same deliverables)
   * Documented (Quickstart, Operations, Deliverables, and Maintenance docs)
3. Fast analysis:
   * Simple project setup
   * Improved compute time and compute utilization
   * Robust (crash-worthy)
   * Modular and flexible (can be efficiently re-started at any step in the pipeline)
   * Runs on all compute environments (Comp3, Comp5, Flux etc.).
4. Efficient maintenance and development:
   * Designated BFXCore maintainers (more than one person) can make changes to the pipeline.
   * Pipeline and dependencies are version controlled
   * Simple, documented mechanisms to
     + add/update component tools (e.g. Cuffdiff vs. DeSeq2, STAR vs. Tophat)
     + config new compute environments
     + add or update species-specific sequences/annotations
   * Automated tests to confirm the integrity/config of pipeline and compute environments reduce regressions and enable faster development
5. Generates self-explanatory, comprehensive, scientifically sound, customer-facing output:
   * Summary results
   * Detailed QC
   * Detailed results
   * References/Citation
   * Methods
   * Executed commands
   * Logs

Additionally, the pipeline implementation should enable these service goals:

* + Validation of pipeline against standard references (e.g. ERCC RNA spike in)
  + Stringent validation of project results though automated checks and better analyst-enabled QC
  + Regular, scheduled review of pipeline components/approach with scientific advisory committee

**Work Breakdown Structure**

Project will be delivered in staged releases detailed below. The conclusion of each release will include:

* A brief meeting with bioinformatics core to demonstrate progress and review/adjust contents of upcoming sprint
* An update in the github repository
* A deployed RNA-Seq pipeline that is available for general use by the bioinformatics core analysts.

Release 1: Basic implementation: Concludes: 9/30/2016

Get most of the core use cases working with the existing (Legacy) tools. This might entail changes to how we invoke existing tools to make them work correctly.

* SE analysis
* QC (FastQC)
* Quality, end trimming (cutadapt)
* Genome and transcriptome alignment (tophat)
* Read counts (HTSeq)
* Gene and isoform diffex (cuffdiff, cummerbund)
* Gene diffex (DESeq2)
* Cuffdiff report
* Quickstart doc/implementation diagram

Release 2: Extend supported use cases (MVP): Concludes 11/15/2016 ?

Complete functionality that would enable deprecation of legacy pipeline.

* PE analysis
* Stranded analysis
* Support for concat reads
* Adjusted intron length
* Basic support for multiple runs

Release 3: Transition to DESeq2?: Concludes ?

* Extended support for multiple runs
* DESeq2 report/plots
* Module-based genome reference/annotation
* Continued support for cuffdiff
* Preliminary research on alternative isoform diffex approaches
* Spec comprehensive experiment report
* Updated deliverables doc
* Maintenance doc

Release 4: Better tools/reporting: Concludes ?

* Better QC
  + FastQScreen
  + QoRTS
* Better alignment
  + STAR
  + HISAT2
* Better diffex
  + Other (tbd)
* Better reporting
  + MultiQC
  + Comprehensive experiment report