BIRN Pipeline Workflow Service-Based Approach to Genomics Computing

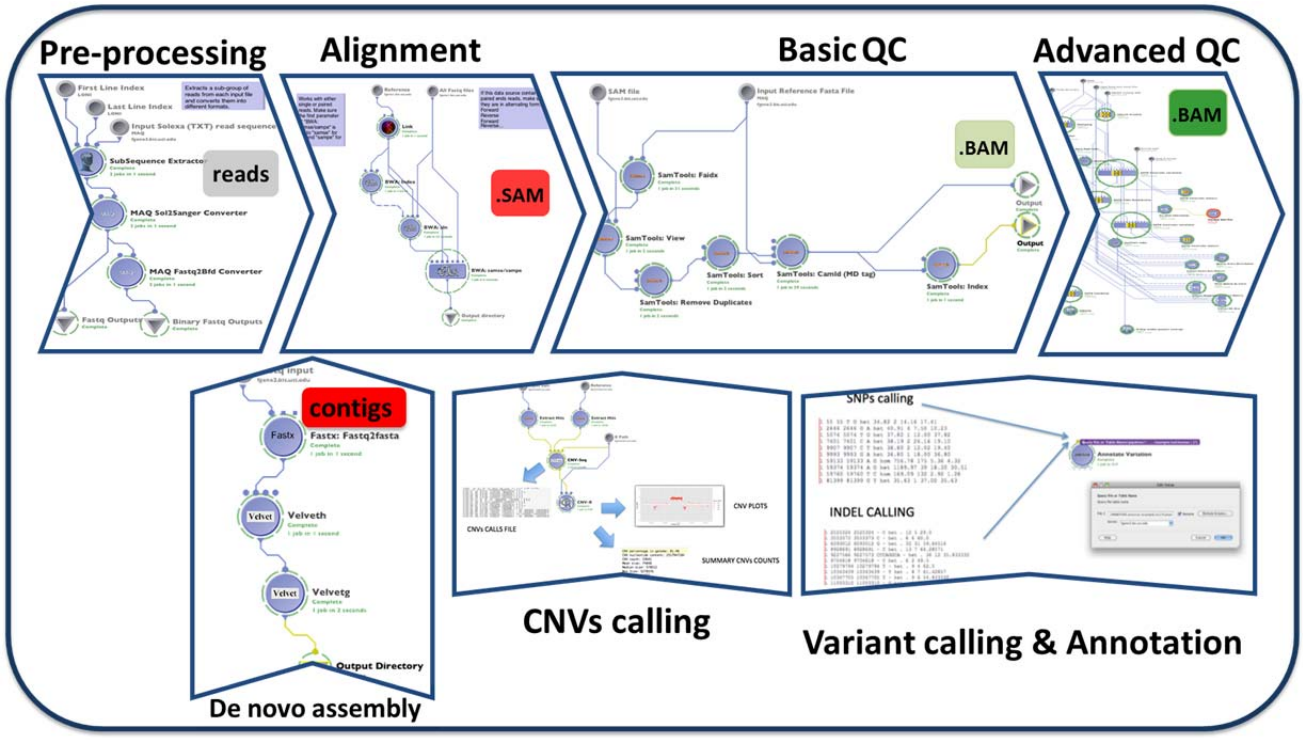
V.2, December 2012

I. Background

By 2020, the size of the entire volume of data stored in the world data is expected to balloon to 40 Zetta Bytes (1ZB=10^6PB). The volume of the genomics data collected in the same period would exceed 20PB. This enormous increase is directly tied to the exponential growth of storage (Kryder's law) and computational power (Moore's law). There are available genomics **data-storage solutions** (e.g., 1000 Genomes, Amazon Data Cloud Service, Broad Institute) and mechanisms for efficient data transfer between sites (e.g., BIRN GlobusOnline, GridFTP). One critical challenge that remains is the problem of efficient, reliable, distributed, user-friendly and reproducible **data processing, analysis and visualization** of large, heterogeneous and dynamic biomedical data.

II. Distributed Graphical Pipeline Genomics Computing

BIRN investigators designed a Graphical Pipeline for Computational Genomics (GPCG) a distributed client-server infrastructure for computational analysis of next generation sequence (NGS) data. The GPCG implements flexible workflows for basic sequence alignment, sequence data quality control, single nucleotide polymorphism analysis, copy number variant identification, annotation, and visualization of results. These workflows cover all the analytical steps required for NGS data, from processing the raw reads to variant calling and annotation. Various software tools for each step of a NGS project are available and researchers can build their own flexible and updatable process. These NGS analysis solutions have direct applications testing of translational hypotheses about the functional role of variants present in the human genome associated with genetic risk factors in complex traits (metabolic, neurologic, psychiatric and other complex disorders).



Hierarchical Pipeline workflow solutions for analyzing NGS data. BIRN Genomics and Pipeline working groups have designed, implemented and validated a number of computational protocols that are openly shared and can be run independently on different Pipeline servers and logically inter-connected on different Pipeline clients. Once the reads have been pre-processed, they can be aligned, undergo to a basic and advanced QC, SNP/Indel and CNVs calling and annotation.

III. Features and Documentation

- Distributed Pipeline Server (**DPS**) software (http://ucla.in/SPDXpO) and web-based **test-service** (http://ucla.in/SPDS5E)
- **Tools** (http://tinyurl.com/ad5vbsb), **Workflows** (http://ucla.in/pbMgUm), **Support** (http://tinyurl.com/b7nxbaf)
- **Videos** (http://tinyurl.com/bhkf9qb), **Screencasts** (http://ucla.in/QJIsXo), **Papers** (http://ucla.in/QJHQBd)