

# Pipelining Your Next Generation Sequencing Workflows



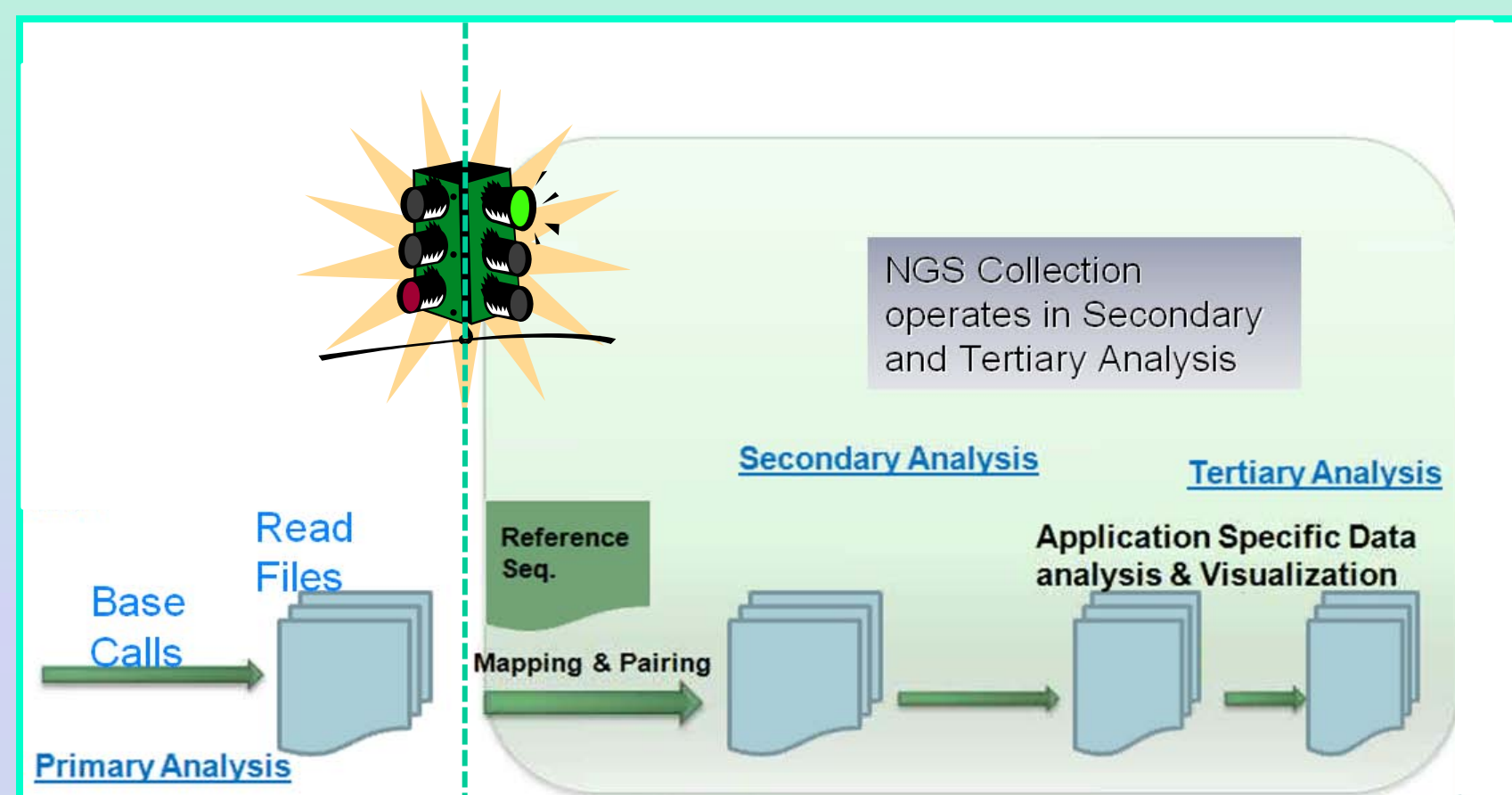
Nancy Miller Latimer, Kristine Briedis, Yi-Shiou Chen, Scott Markel, and David Waner, Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA.

**Abstract:** As the cost of DNA sequencing drops, we are approaching a fundamental shift in how we do science. It may soon become financially feasible to completely sequence the genomes and targeted chromosomal regions of patients in a clinical trial. Successful users of deep sequencing platforms often have in-house bioinformatics expertise and locally customizable software. This is typically not included with the platform vendor's hardware. The informatics challenge is making sense of all the data, especially as not all organizations have in-house expertise or the budget to create customized tools. However, many workflows are routine and predictable and, as such, are amenable to automated analyses. Using a data-pipelining approach we demonstrate several common workflows that start after base calling and continue through variation analysis. The examples are implemented using Pipeline Pilot's Next Generation Sequencing Collection (NGS) soon in beta.

## Problem Statement

For my genome sequencing experiment, I roughly know what analyses and reports I would like at the end, but I want to

- Simplify the end-to-end process from read files to analysis results
- Perform the mapping using several programs and compare results
- Take advantage of frequently changing algorithms without breaking my workflow
- Simplify the organization of mapped reads, reference sequences, features of interest
- Deploy using the processing and analysis workflows using the web or SharePoint as a workbench to deliver results
- Create tracks that can be used by multiple genome browsers



The Pipeline Pilot NGS Collection begins after read files are created.

## Use Case - End to End Workflow

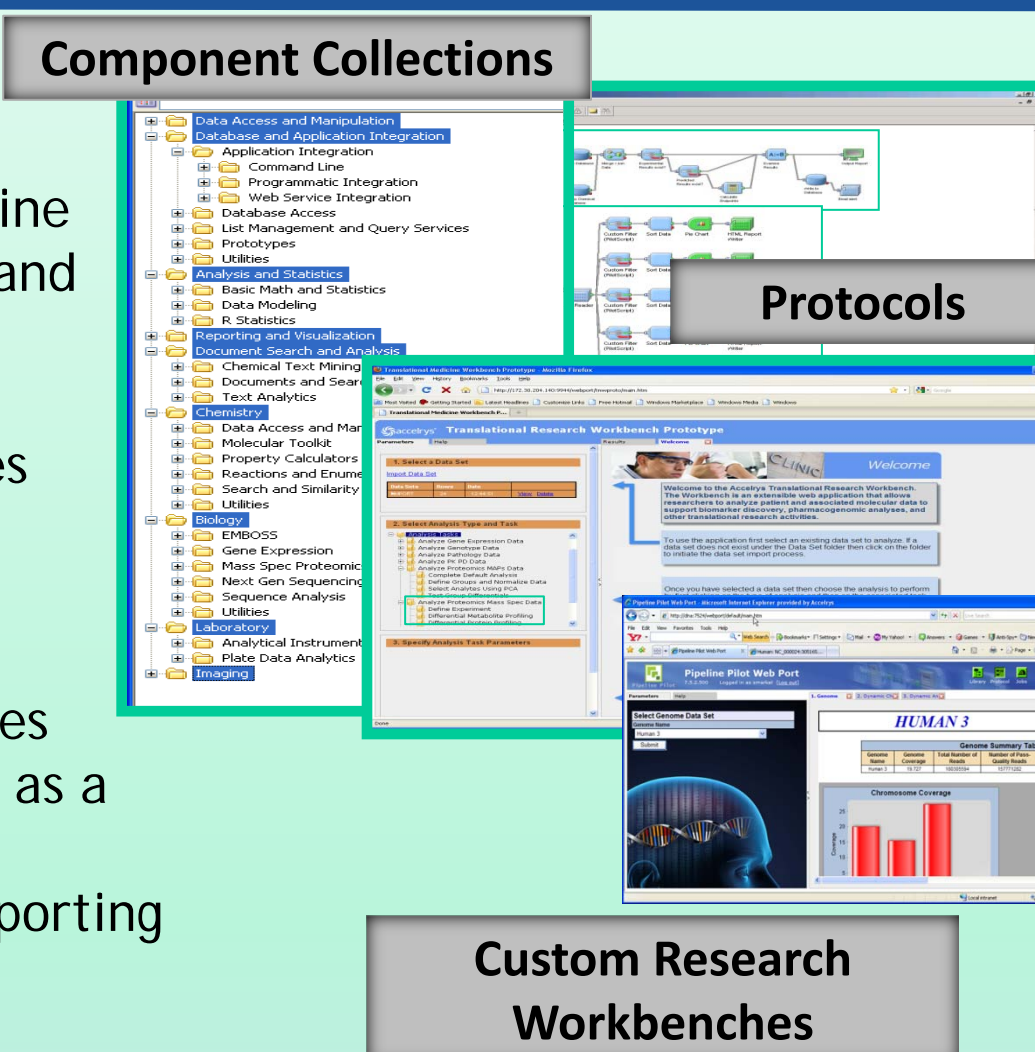
Identify SNPs and indels from an Illumina Human Sequencing Experiment Reads to Variation Report

- Download data from NCBI Short Read Archive
- Download reference sequence(s) and SNPs from NCBI
- Create a NGS data repository
- Map reads to reference(s)
- Create summary of data coverage and quality
- Identify SNPs and indels for the regions of interest
- Create tracks (input files) for GBrowse and interactively explore results
- Create an interactive summary report for SNPs that links to the literature

## General Approach

A pipelining platform, such as Pipeline Pilot, facilitates quick prototyping and automation of analysis. It also can

- Integrate disparate data-sources and data-types
- Integrate or wrap third-party applications and workflows
- Simple encoding of business rules
- Deploy using web or SharePoint as a workbench or to deliver results
- Use interactive charting and reporting components

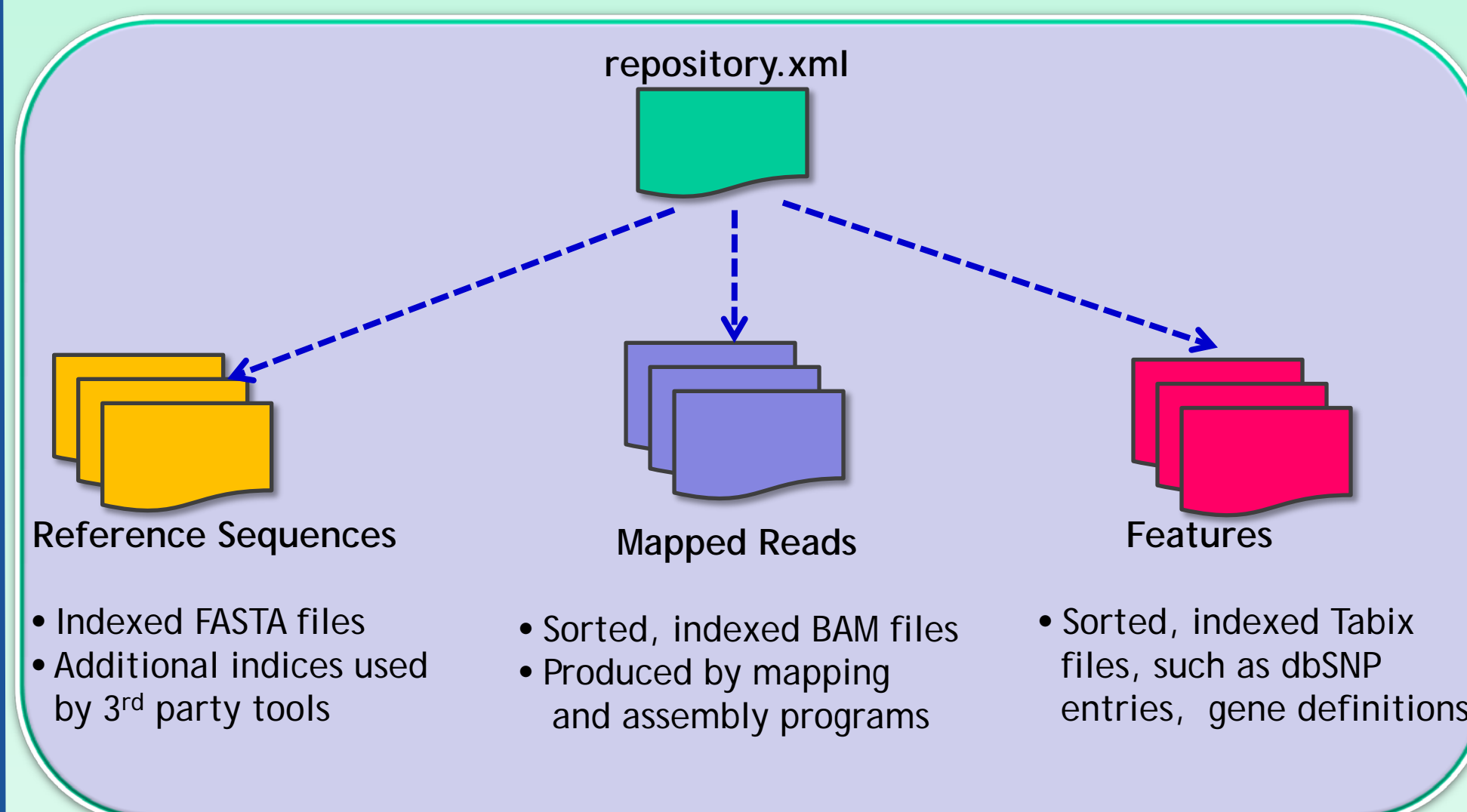


## Top 10 Reasons Why the NGS Collection for Pipeline Pilot is a Good Fit for Deep Sequencing Analysis

1. You go from a deluge of reads generated by Illumina, SOLiD, or 454 deep sequencing platforms to quick data-quality assessments and insightful analyses—whether you are an expert or a novice.
2. You can perform common workflows such as *de novo* assembly, mapping to a reference genome or reference sequences, and variation detection by using NGS components and sample protocols.
3. The NGS Collection supports unpaired read and paired reads in both base space and color space.
4. Through use of a flexible XML-defined repository, Pipeline Pilot components access genomic features and reference genomes and manage your data.
5. The data readers and writers use common formats, such as SAM, BAM, GFF3, or FASTQ, as appropriate, to optimize integration of open source programs in this quickly evolving domain.
6. Since the Collection is built on the flexible Pipeline Pilot platform, third-party applications may be integrated and new algorithms may be implemented, without breaking existing data pipelines.
7. The Sequence Analysis Collection can operate on output from NGS components.
8. Pipeline Pilot manages the data for your experiments.
9. New reports are quickly built with the Pipeline Pilot Reporting Collection.
10. Easily incorporate and encapsulate R functions and scripts with the R-Stats Collection.

## Data Management - Repository

The *Repository* virtually defines your experimental database. The reference sequence(s) provides the location context. A given repository contains mapped reads from your experiment, the reference sequence(s), and features that are located on the reference sequence(s). The *repository.xml* file includes the user-specified location, name, and description of each file contained in the *Repository*.



## With Pipeline Pilot's Next Generation Sequencing Collection You Can:

- Add features to any repository, e.g. genes and dbSNP entries
- Assess run quality and coverage through custom interactive web-based reports
- Filter data based on pairs proximity, quality, and repeated sequences
- Generate track information for viewing in your favorite browser
- Map reads with BWA, Bowtie, or *mapreads*
- Perform *de-novo* assembly with MIRA3
- View mapping stats for your experiment
- Identify single nucleotide polymorphisms
- Identify insertions and deletions
- Detect copy number variation

Pipeline Pilot functions as both a DAS server and a DAS client