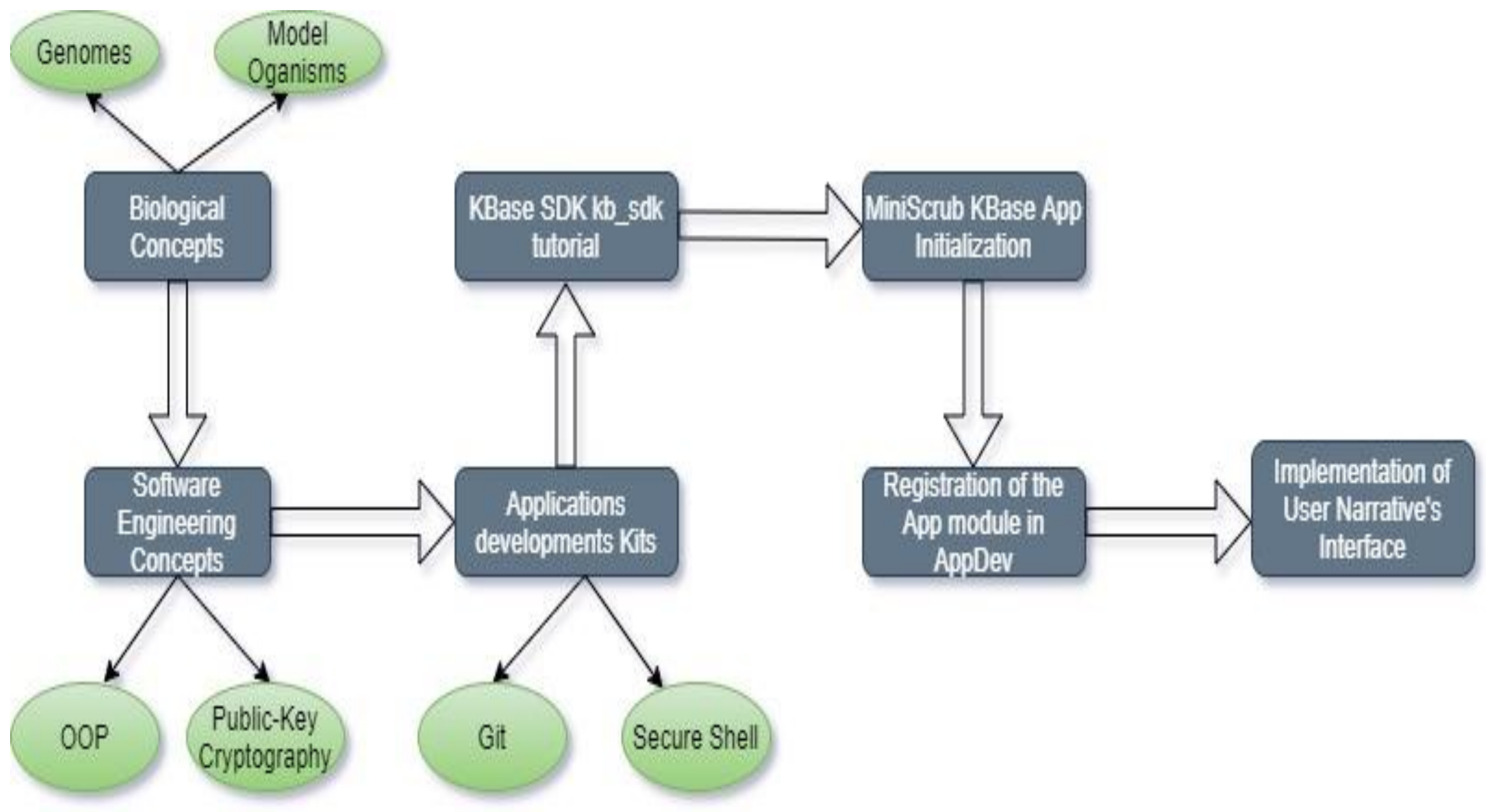
Design and implementation of novel applications for KBase

B. J. Roxane Ouango, Hostos Community College
David Dakota Blair, Computational Science Initiative

Abstract

The Department of Energy Systems Biology Knowledgebase (KBase) is an open-source project enabling collaborative bioinformatics data processing and supporting data provenance, and reproducible science. KBase is designed to be an extensible community resource whose growth is supported by its Software Development Kit (SDK), which enables any developer to build, and deploy new or existing software as KBase apps. Our project at Brookhaven National Laboratory aims to wrap an existing tool called MiniScrub in a KBase app. MiniScrub is a de novo long sequencing read preprocessing method that improves read quality by predicting and removing read segments that have a high concentration of errors. The technologies required include Docker, POSIX shell, python as well as internal KBase APIs. Our approach consists of learning biological topics and software engineering concepts and gaining familiarity with Git, secure shell, and the KBase SDK tool kb_sdk. By implementing MiniScrub in an KBase app, we can promote user-friendly interfaces and contribute new features to KBase to support new avenues of research.

Methods


Genomes
Model Oganisms
Biological Concepts
KBase SDK kb_sdk tutorial
MiniScrub KBase App Initialization
Software Engineering Concepts
Applications developments Kits
Registration of the App module in AppDev
Implementation of User Narrative's Interface
OOP
Public-Key Cryptography
Git
Secure Shell

MiniScrub KBase App Design Flowchart

Results

MiniScrub was wrapped in a KBase app called kb_miniscrub. A demonstration narrative explaining the app's features was created. From the KBase's primary user interface, the Narrative Interface built on the Jupyter10,11 platform, users can (**Figure 2**):

- Upload their private data.
- Search and retrieve extensive public reference data,
- Select and run the kb_miniscrub app on their data,
- View and analyze the results from the app, and
- Record their interpretations along with the analysis steps.

Kb_miniscrub takes a FASTQ file as input. Then, MiniScrub first generates read-to-read alignments by MiniMap, then encodes the alignments into images, and finally builds CNN models to predict low-quality segments that could be scrubbed based on a customized quality cutoff. The app outputs a new FASTQ file with the scrubbed reads.

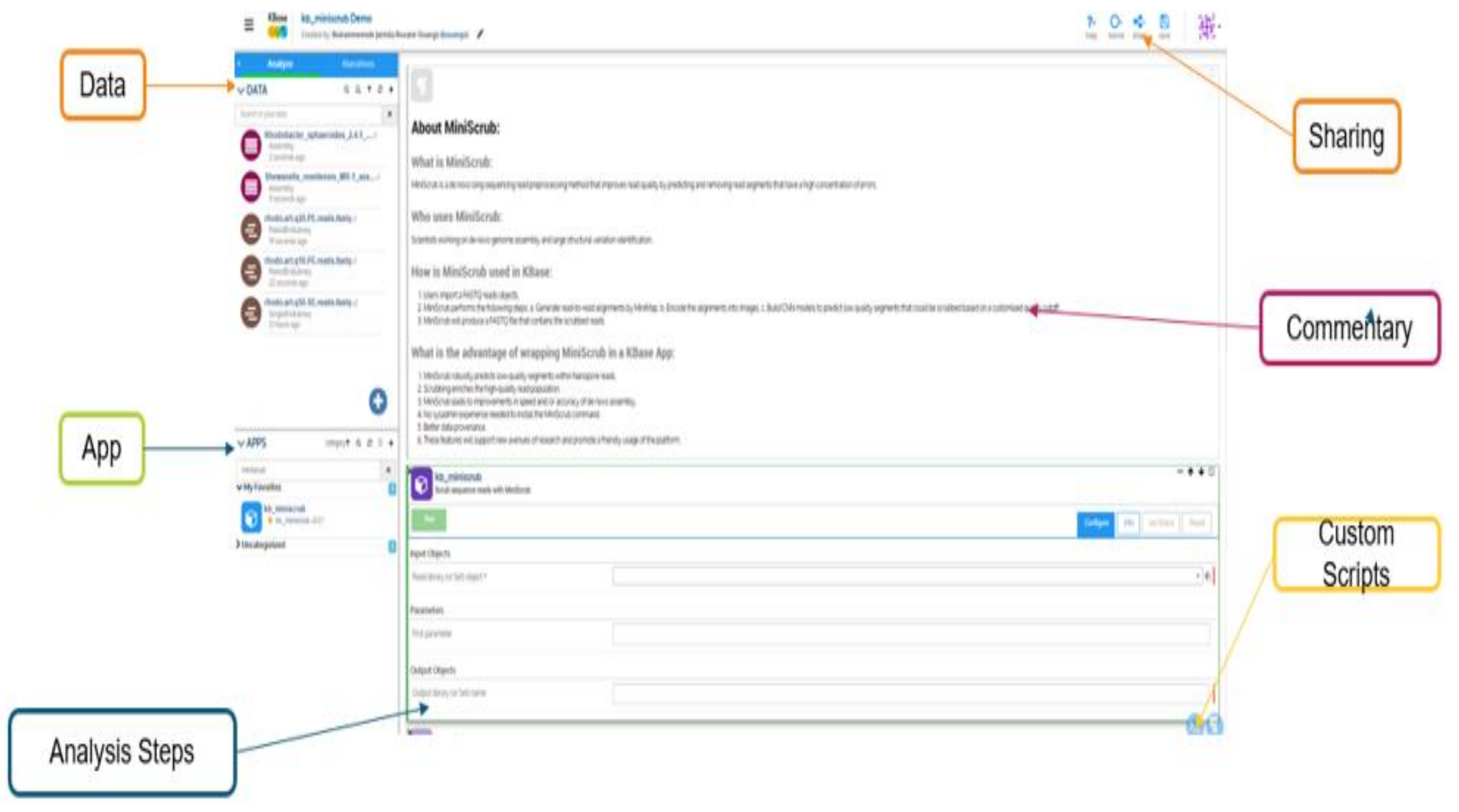


Figure 2: kb_miniscrub user narrative's interface

Introduction

Long read sequencing has become increasingly important in recent years, with sequencing technologies from companies such as Pacific Biosciences and Oxford Nanopore seeing wide use in a variety of applications including genome assembly, detection of antimicrobial resistance genes, sequencing personal transcriptomes, and improving draft genomes. Current single molecule, long sequencing reads also have very high error rates, ranging from 5% to 40% per read and often average about 10% to 20% up to as high as 30-40% depending on variables such as the type and version of the sequencing technology and the experiment being performed. These high error rates make assembly and other applications inefficient or error prone. It is thus critical that methods be developed towards addressing this issue so that the potential of long read sequencing can be fully realized. MiniScrub is a method that performs de novo long read scrubbing using the combined power of fast approximate read-to-read alignments, deep Convolutional Neural Networks, and a novel method for pileup image generation.[1]

Our project at Brookhaven National Laboratory aims to contribute novel applications, modules, and methods to the Department of Energy Systems Biology Knowledgebase (KBase) by wrapping MiniScrub in a KBase app.

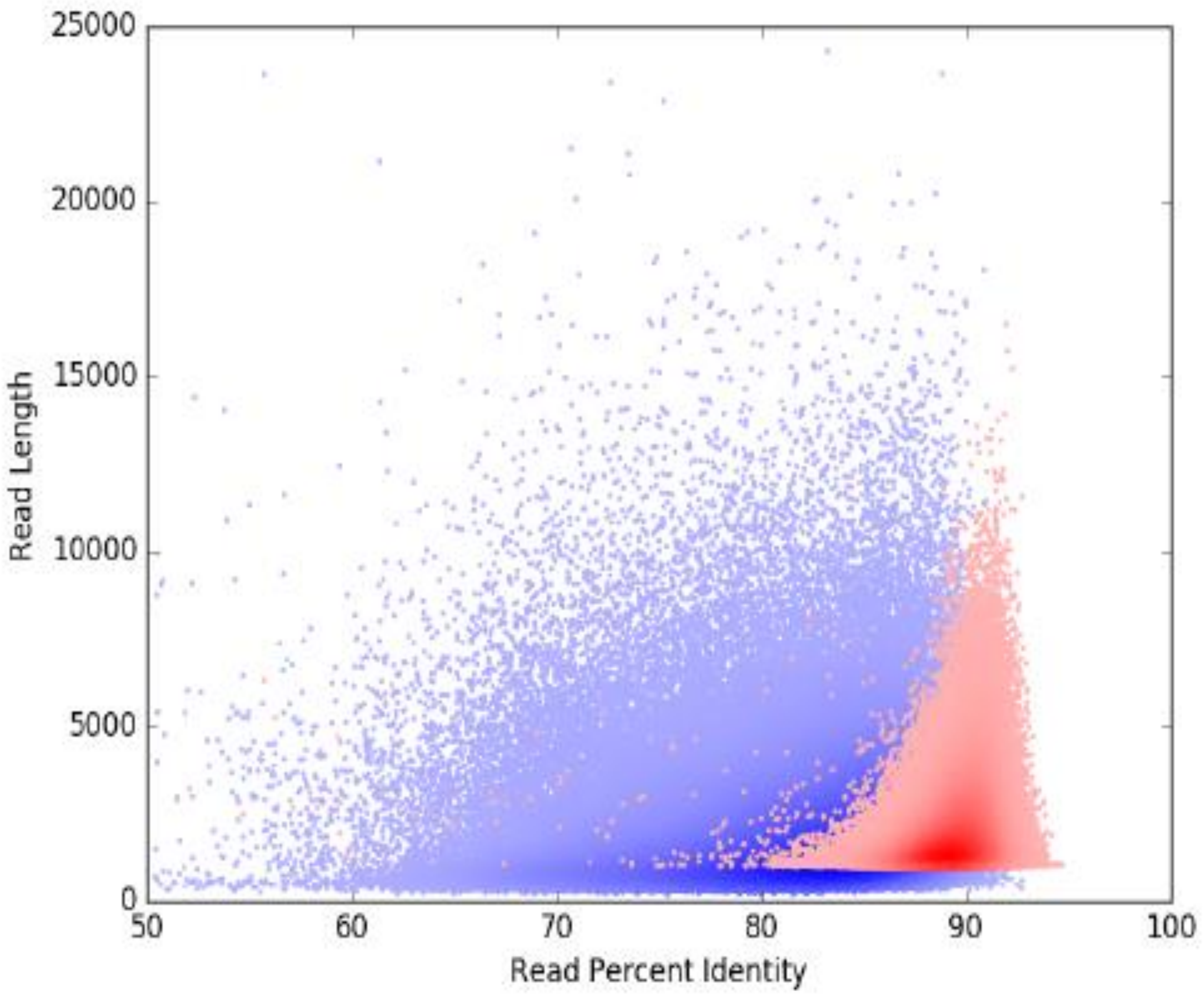


Figure 1: Density scatter plot showing average read quality improvement by MiniScrub versus raw reads. The X-axis shows read percent identity to the reference while the Y-axis shows read length. Raw reads are in blue while scrubbed reads are in red. The darkness of the color indicates increased \density" { more reads fall into a darker region of the graph than the lighter areas. MiniScrub scrubs out most of the low-quality segments in low quality reads while leaving high quality reads intact, increasing average read percent identity by over 3%, from 83.1% to 86.2%. Average read length decreased from 2.6kb to 1.1kb due to splitting reads where low-quality segments were removed..[1]

Conclusions

- MiniScrub robustly predicts low-quality segments within Nanopore reads.
- Scrubbing enriches the high-quality read population.
- MiniScrub leads to improvements in speed and/or accuracy of de novo assembly.
- No sysadmin experience needed to install the MiniScrub command.
- Data provenance is managed.

These features will support new avenues of research and promote a friendly usage of the platform.

Reference

1. MiniScrub: de novo long read scrubbing using approximate alignment and deep learning, Nathan LaPierre, Rob Egan, Wei Wang, Zhong Wang bioRxiv 433573; doi: https://doi.org/10.1101/433573

2. App catalog: https://narrative.kbase.us/#appcatalog

3. MiniScrub's Bitbucket repository: https://bitbucket.org/berkeleylab/jgi-miniscrub

4. MiniScrub's Bitbucket repository: https://bitbucket.org/bouango/jgi-miniscrub/src/master

Acknowledgements

- I would like to thank my mentor Dr. David Dakota Blair for his generous support and dedication to my learning experience.

- Ziming Yang

- This project was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Community College Internships Program (CCI).

www.bnl.gov