Benchmarking Computational Tools For Calling SNPs And Indels In Complex Microbial Populations

Philip Sweet[1], Natalie Ball[1], Bárbara Müller[2], Sandra Vu[1], Lisa Anderson [1], Sadie Downing[1], Amy Gresser[1], Aditya Hindupur[1], John Hogan[1], Hiromi Kagawa[1], Aphrodite Kostakis[1], Matthew Paddock[1], Kevin Sims [1], Kevin Tyre[3], Fuzhong Zhang[2], Fang Bai[2], Frances Donovan[1], **A. Mark Settles**[1]

[1]NASA Ames Research Center, Bioengineering Branch, Mountain View, CA, [2]University of Florida, Agricultural Sciences, Gainesville, FL, [3]NASA Kennedy Space Center, Space Life Sciences Lab , Merritt Island, FL

Abstract

The NASA BioNutrients missions seek to understand the suitability of microorganisms for bioproduction during space flight. One topic of interest is the stability of microbial genomes during long-term ambient storage and subsequent rehydration and growth. To address these questions, samples from 8 species were flown to ISS for 5 years of desiccated storage at ambient temperature (Stasis Packs) and 2 species were packaged along with powdered media inside a bioreactor system to allow hydration and growth in microgravity (Production Packs). For both systems, Whole Genome Sequencing (WGS) of the DNA extracted from the returned samples and paired ground controls will be conducted to identify changes in genome stability due to time, storage conditions and growth in space. Across the technical replicates, ground controls, 10 timepoints, and multiple experimental conditions, ~300 samples have been selected for initial analysis with WGS sequencing to 100x coverage. A flexible and resource efficient mutation calling pipeline is needed to process this large dataset and allow for comparisons between species.

Many bioinformatics tools for calling Indels and Single Nucleotide Variants (SNVs) are designed for use with pure isolates, where true variations from the reference genome are expected to dominate the reads aligning to the location of mutation. In contrast, DNA from the Stasis Pack (SP) samples was collected directly after recovery from desiccated storage and the Production Pack (PP) samples were collected after fermentation. In this context, reads with mutations are expected to be less frequent than reads that align with the reference genome, as each sample will include multiple lines of cells. Thus, BioNutrients samples are expected to be similar to samples from cancer cell or “pooled” sequencing approaches. In preparation for the analysis of the BioNutrients samples, we have tested three mutation calling tools (GATK for Microbes, BreSeq and DiscoSNP) designed for complex samples.

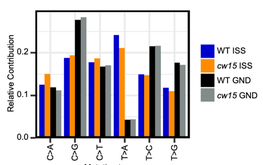
A challenge of validating mutation identification pipelines is a lack of “Ground Truth” datasets, especially for complex samples. To compare these three tools, we sought to identify mutations in pre-existing WGS data collected from populations of *Chlamydomonas reinhardtii* that were exposed to UV mutagenesis and growth in LEO as part of the Space Algae-1 mission. Here we present a summary of these tools against the analysis originally conducted using the CRISP tool. Critical metrics are compared such as runtime, the number of SNPs, the number and size of Indels, and patterns of transversion and transitions identified by each tool are reported. By sharing these benchmarking results collected in support of the BioNutrients mission, we aim to guide others seeking to identify SNVs in similarly complex microbial samples.

Introduction

Space Algae-1

- *Chlamydomonas reinhardtii* (*C. reinhardtii*) is a unicellular microalgae with potential applications in in-flight resource recycling and food generation.
- During the Experiment Verification Test (EVT) phase of the Space Algae-1 Mission, three SNP calling pipelines were tested on short read WGS data collected from populations of *C. reinhardtii* after UV mutagenesis. **CRISP[1] was identified as the most sensitive tool for identifying SNPs from these complex genomic samples[2]**
- During implementation of Space Algae-1, two strains of *C. reinhardtii* were cultured on the International Space Station (ISS) for 20 days (~30 rounds of cell division).
- Comparison of the ISS cultures vs Ground Control cultures using CRISP revealed a **space-specific mutagenic signature in the bias toward A—T mutations in the flown samples**.

Astronaut Alex Gerst holding a Space Algae-1 culture on the ISS. NASA Image



Relative compositions of SNPs identified by CRISP in cultures grown on ISS and Ground Controls for two strains (WT and cw15) of C. reinhardtii

BN-1 Production Packs during fermentation on ISS (left) and on earth (right). NASA Image

BioNutrients 1 & BioNutrients 2

- The BioNutrients-1 (BN-1)[3] and BioNutrients-2 (BN-2)[4] missions seek to demonstrate the feasibility of synthetic biology approached to address long duration crewed mission objectives, such as producing essential nutrients.
- **An ideal space production species would maintain viability, genetic fidelity and high yields, even after several years of desiccated storage in space.**
- For BN-1, in addition to DNA samples for mutation rate analysis, the viability (Stasis Packs) and nutrient yields (Production Packs) were collected at multiple time points across 5 years of storage.
- DNA samples from 8 species, including bacteria and yeast, have been collected as part of BN-1 and BN-2 for short read WGS.
- **Desired features for the variant calling tool to analyze BioNutrients WGS:**
  - **Short-read based**
  - **Species agnostics**
  - **Designed for somatic or pooled samples**
  - **Report SNPs and Indels**
  - **Open source**

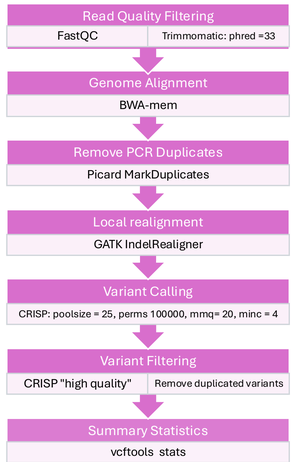
Methods

Workflow	Input	SNP Calling Framework	Publication Year	Last Updated
CRISP[1]	**bam**	**Contingency table**	**2010**	**2024**
GATK for Microbes[5]	**bam**	**Mutect2/Bayesian**	**2021**	**2021**
DiscoSNP[6]	**fastq**	**de Bruijn graphs**	**2017**	**2022**
BreSeq[7]	**fastq**	**Mpileup/Bayesian**	**2014**	**2024**

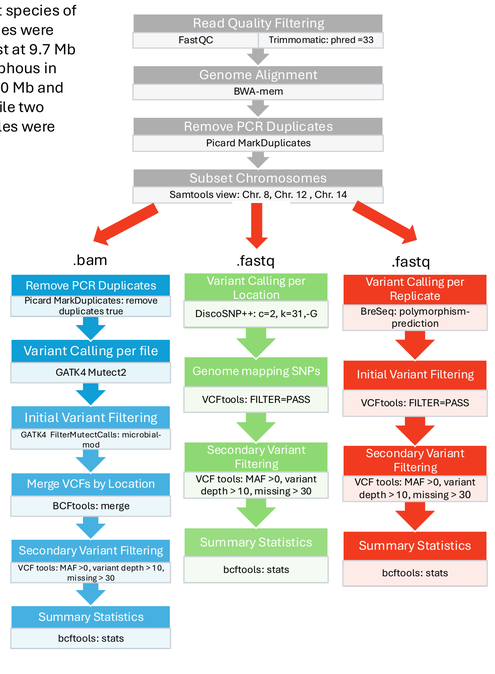
Benchmarking Design

To reflect the genome size of bacteria and yeast species of BN-1 and BN-2, three *C. reinhardtii* chromosomes were selected for benchmarking. Chr. 12 is the largest at 9.7 Mb and was noted as having a high rate of polymorphous in the EVT study. Chr. 8 and Chr. 14 are smaller (5.0 Mb and 4.1Mb) with lower rates of polymorphisms . While two strains were included in EVT, only the WT samples were used for benchmarking.

Initial CRISP Analysis[2]



Reanalysis

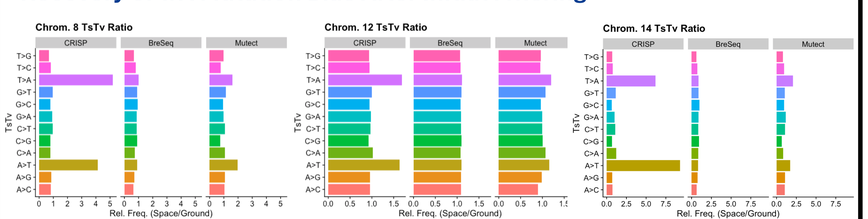


Results

Variant Depth per Sample

	Mean Chrom. Total Depth	Mean Chrom. Depth Post-Dup. Removal	Mean Variant Depth: CRISP	BreSeq	GATK Mutect2	DiscoSNP (All)	DiscoSNP (Mapped)
Chr. 8	**204**	**146**	**125**	**91**	**87**	**225**	**8**
Chr. 12	**180**	**131**	**120**	**104**	**107**	**120**	**7.5**
Chr. 14	**211**	**150**	**150**	**75**	**75**	**366**	**7.25**

Recovery of T/A Mutation Bias After Initial Filtering

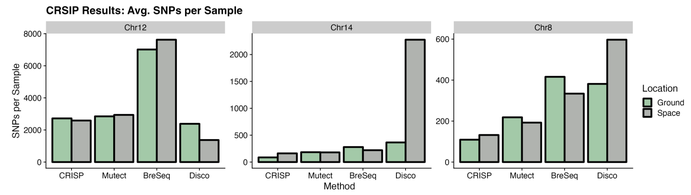


Results

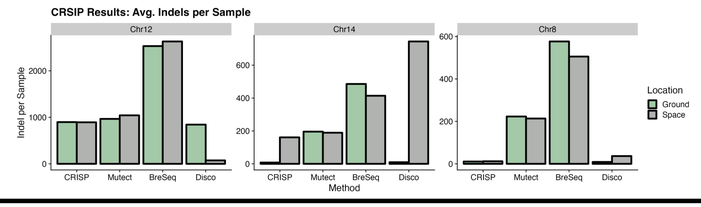
Computational time on EC2 R8g.xlarge

Tool	Avg. Time
DiscoSNP	1 min/Gb of .fastq
BreSeq	45 min/Gb of .fastq
GATK Mutect2	4 min/Gb of .bam

Avg. Number of SNPs per sample with default filters



Avg. Number of Indels per sample with default filters



Conclusions

SNP Calling Tools

- **BreSeq:** Took the most computational time, detected the greatest number of SNPs and Indels, but not the previously observed T/A mutation bias.
- **GATK for Microbes:** Detected a comparable number of SNPs as CRISP as well as slight bias toward A/T mutations in the ratio of TsTv’s.
- **DiscoSNP:** Very fast but the majority of reads accumulated on unmappable sequences and total variant calls are inconsistent with CRISP.

Next Steps

- Implement secondary quality filtering in BreSeq and GATK for Microbes results.
- Compare the percent of overlapping SNP locations between tools.

Funded by

STMD Game Changing Development

Supported by

U.S. NATIONAL LABORATORY

Citations

1. Bansal V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics. 2010 Jun
2. Zhang J, Müller BSF, Tyre KN, Hersh HL, Bai F, Hu Y, Resende MFR Jr, Rathinasabapathi B, Settles AM. Competitive Growth Assay of Mutagenized *Chlamydomonas reinhardtii* Compatible With the International Space Station Veggie Plant Growth Chamber. Front Plant Sci. 2020 May 25
3. BioNutrients-1: Development of an On-Demand Nutrient Production System for Long-Duration Missions Natalie Ball , Hiromi Kagawa , Aditya Hindupur , Kevin Sims. ICES-2020-119
4. BioNutrients-2: Improvements to the BioNutrients-1 Nutrient Production System. Natalie Ball , Aditya Hindupur , Hiromi Kagawa3 , Aphrodite Kostakis, Amy L. Gresser , Kevin Sims , Sean Sharif , Alyssa G. Villanueva, Frances Donovan , A. Mark Settles. ICES-2021-331
5. Introducing GATK for microbes. (GATK Team). https://gatk.broadinstitute.org/hc/en-us/articles/360060004292-Introducing-GATK-for-Microbes
6. DiscoSnp++: de novo detection of small variants from raw unassembled read set(s). Pierre Peterlongo, Chloé Riou, Erwan Drezen, Claire Lemaitre. bioRxiv 209965
7. Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol. 2014