Synthetic Genomics Data Generation and Evaluation for the Use Case of Benchmarking Somatic Variant Calling Algorithms

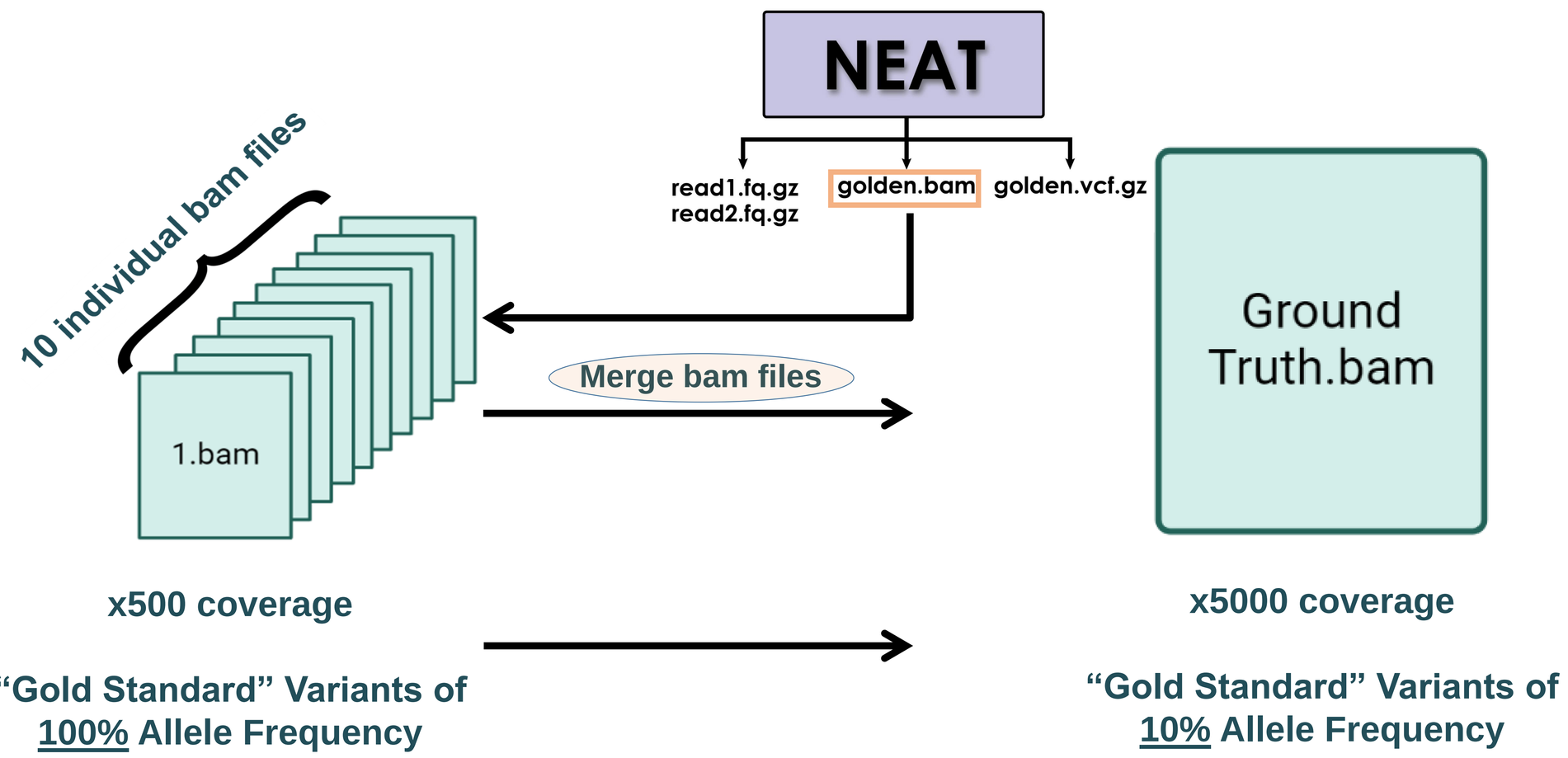
Fragkouli Styliani-Christina[1, 2], Pechlivanis Nikos[1], Agathangelidis Andreas[2], Psomopoulos Fotis[1]

[1]Institute of Applied Biosciences, Centre of Research and Technology Hellas, Thermi, 57001, Thessaloniki, Greece
[2]Department of Biology, National and Kapodistrian University of Athens, Athens 10679, GR

1 Synthetic «Gold Standard» Dataset Generation

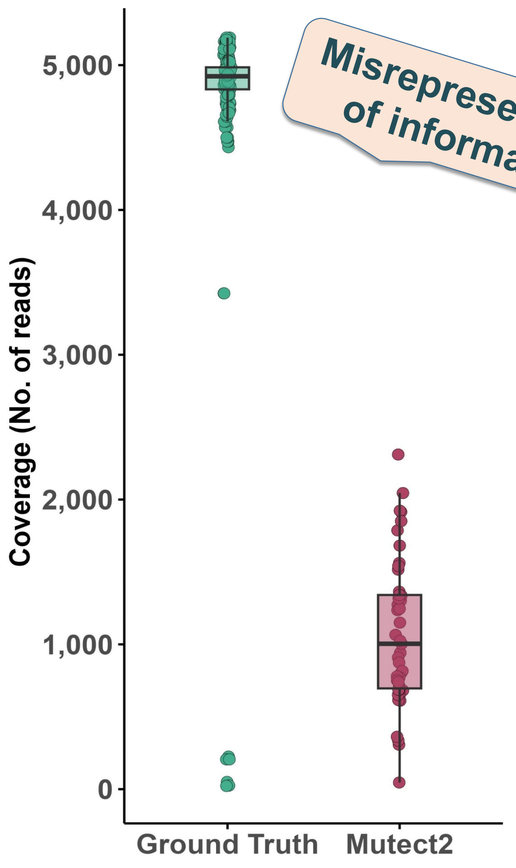
Highlights

- Generation of **synthetic genomics** data based on *TP53* gene
- Define **«Ground Truth»** SNIPs and INDELs in order to **benchmark** somatic variant callers
- Investigate the impact of variant callers in variants at **low frequencies**

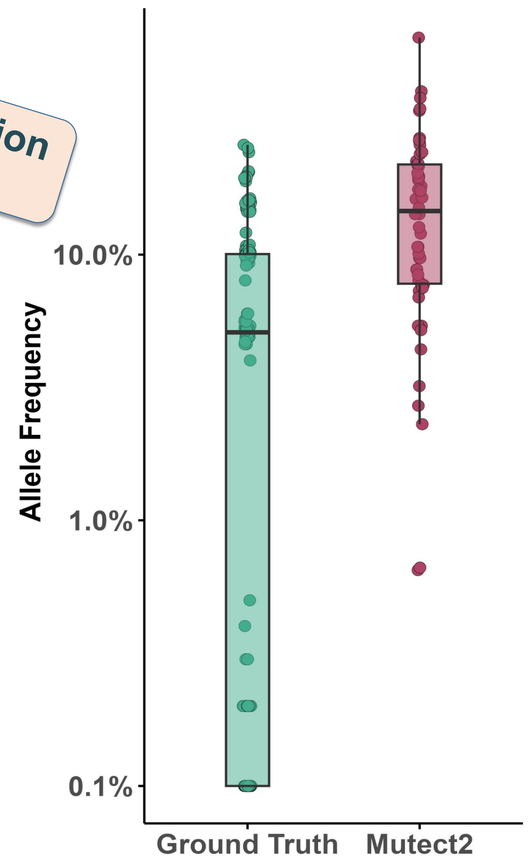


2 Benchmarking GATK-Mutect2

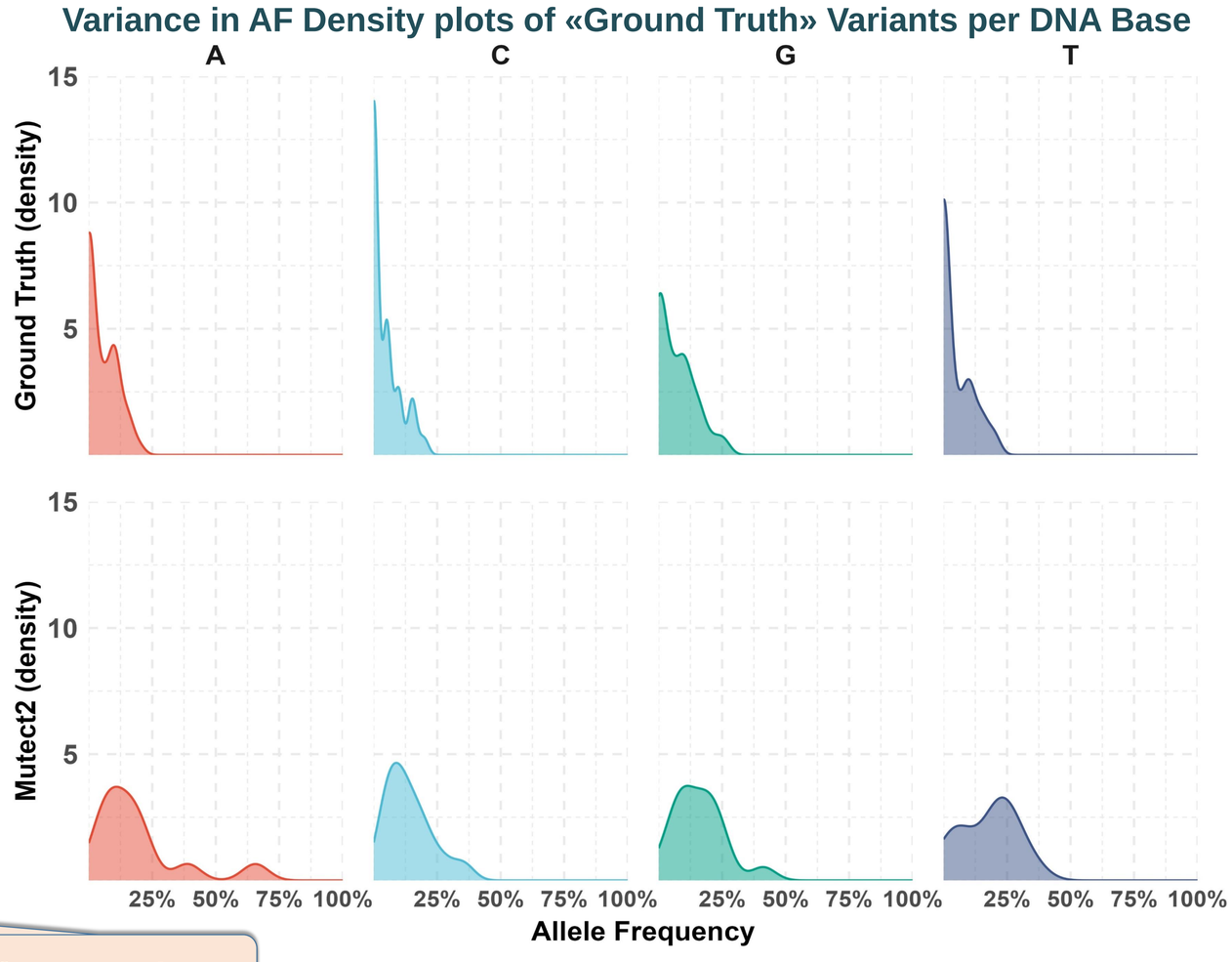
Down-sampling of coverage of «Ground Truth» Variants



Differences in AF of «Ground Truth» Variants



Variance in AF Density plots of «Ground Truth» Variants per DNA Base



Difficulty to identify variants at low frequencies

Divergence in the identification of SNIPs and their AF of «Ground Truth» Variants

Venn plot of the Overall Variants

To learn more about our work please visit

INSTITUTE OF APPLIED BIOSCIENCES
INSTITOYTO EΦAPMOΣMENΩN BIOEΠIΣTHMΩN
CENTRE for RECEARCH and TECHNOLOGY-HELLAS
CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS
sfragkoul@certh.gr
sfragkoul
@scfragkoul
National and Kapodistrian University of Athens
Faculty of Sciences
Department of Biology
Section of Genetics & Biotechnology

This research has been co-financed by the HORIZON-INFRA-2021-EOSC-01-04:
SciLake