Synthetic Genomics Data Generation and Evaluation for the Use Case of Benchmarking Somatic Variant Calling Algorithms
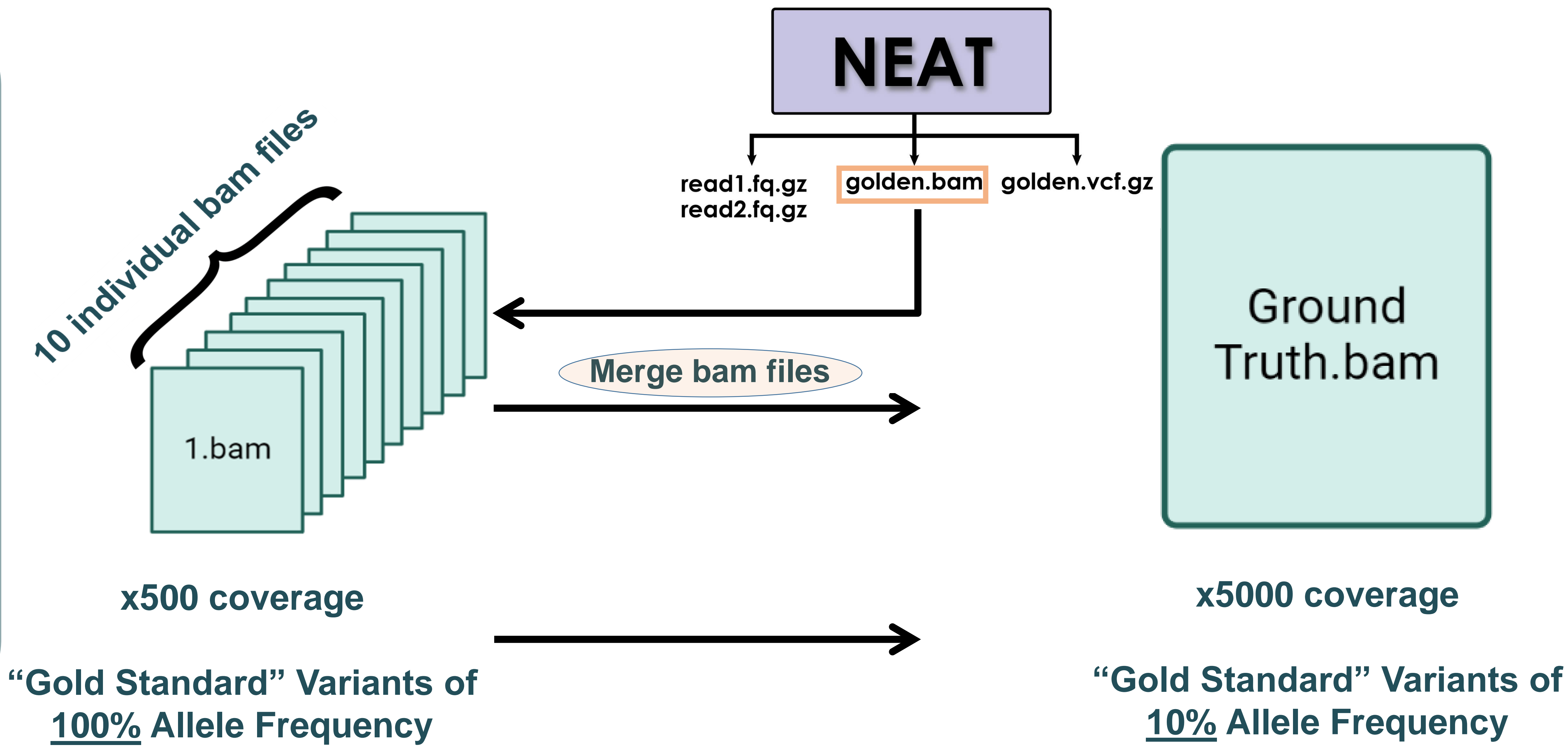
Styliani-Christina Fragkouli[1, 2], Nikos Pechlivanis[1], Andreas Agathangelidis[2], Fotis Psomopoulos[1]
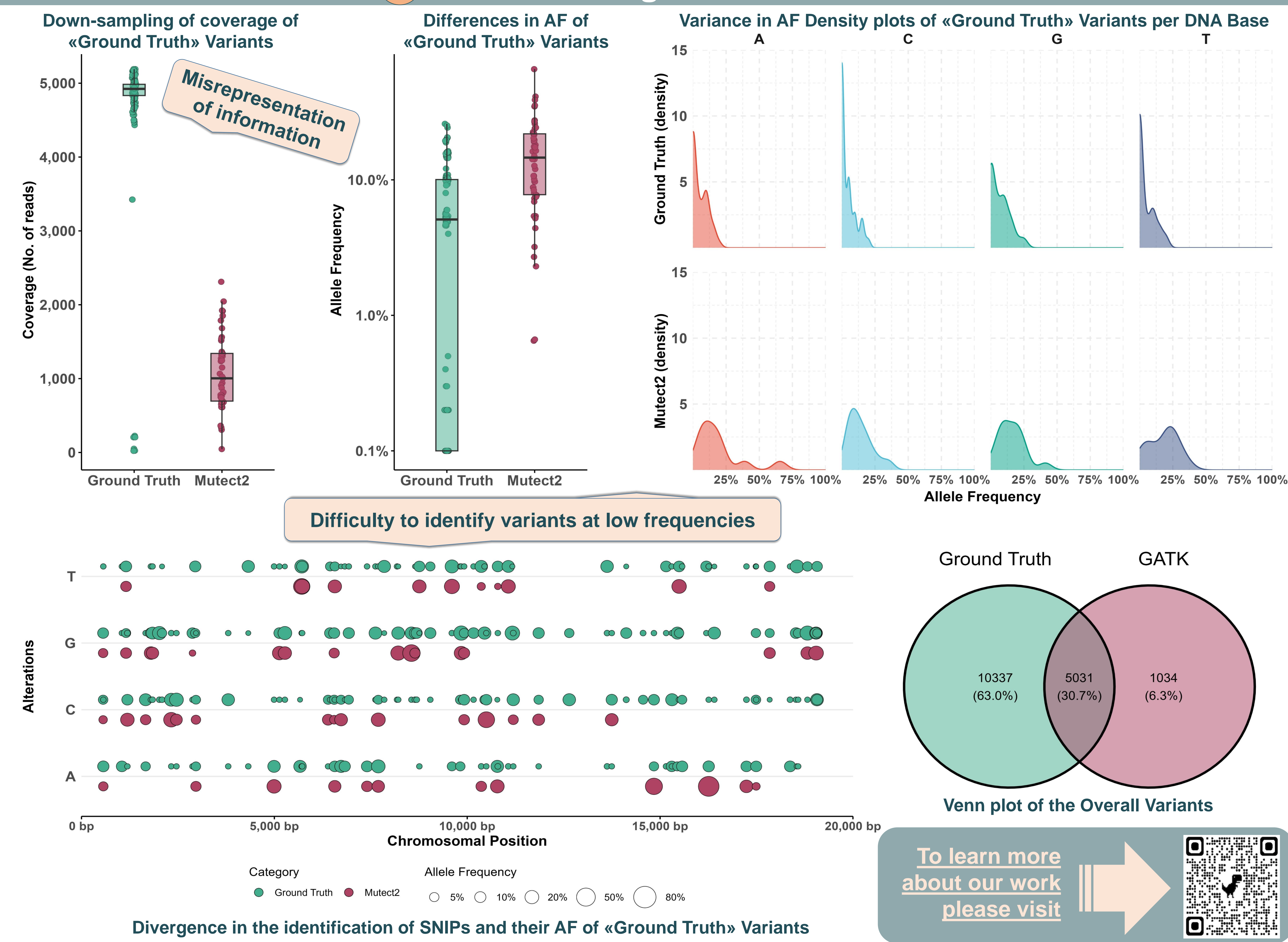
[1]Institute of Applied Biosciences, Centre of Research and Technology Hellas, Thermi, 57001, Thessaloniki, Greece
[2]Department of Biology, National and Kapodistrian University of Athens, Athens 10679, GR

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**

NEAT

read1.fq.gz   golden.bam   golden.vcf.gz
read2.fq.gz

10 individual bam files

1.bam

Merge bam files

Ground Truth.bam

x500 coverage

"Gold Standard" Variants of <u>100%</u> Allele Frequency

x5000 coverage

"Gold Standard" Variants of <u>10%</u> Allele Frequency

2 Benchmarking GATK-Mutect2

Down-sampling of coverage of «Ground Truth» Variants

Misrepresentation of information

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

Category
- Ground Truth
- Mutect2

Allele Frequency
5% 10% 20% 50% 80%

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

Ground Truth   **GATK**

10337 (63.0%)   5031 (30.7%)   1034 (6.3%)

Venn plot of the Overall Variants

To learn more about our work please visit