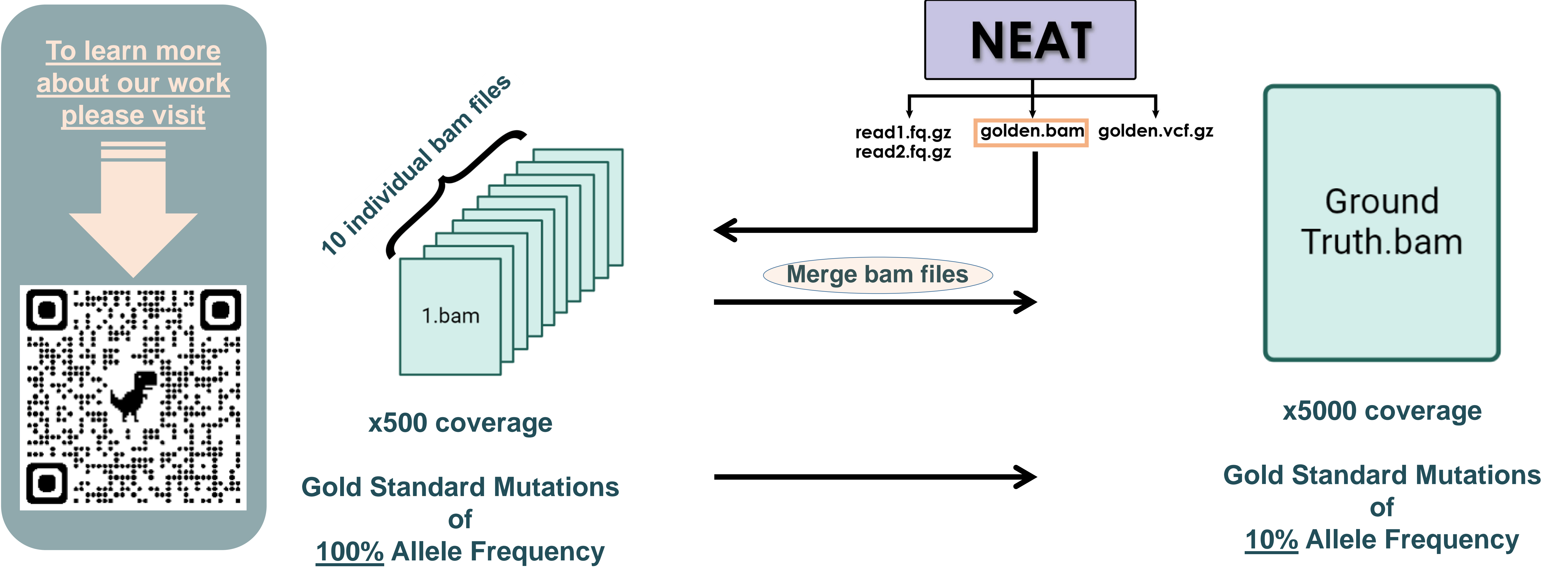# Synthetic Genomics Data Generation and Evaluation for the Use Case of Benchmarking Somatic Variant Calling Algorithms

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

To learn more about our work please visit

NEAT

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

10 individual bam files

1.bam

Merge bam files

Ground Truth.bam

x500 coverage

Gold Standard Mutations of 100% Allele Frequency

x5000 coverage

Gold Standard Mutations of 10% Allele Frequency

2 Benchmarking GATK-Mutect2

Down-sampling effect

Difficulty to identify mutations of low frequencies

Ground Truth   GATK

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

Overlap for Entire Dataset

Category: Ground Truth, Mutect2

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