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

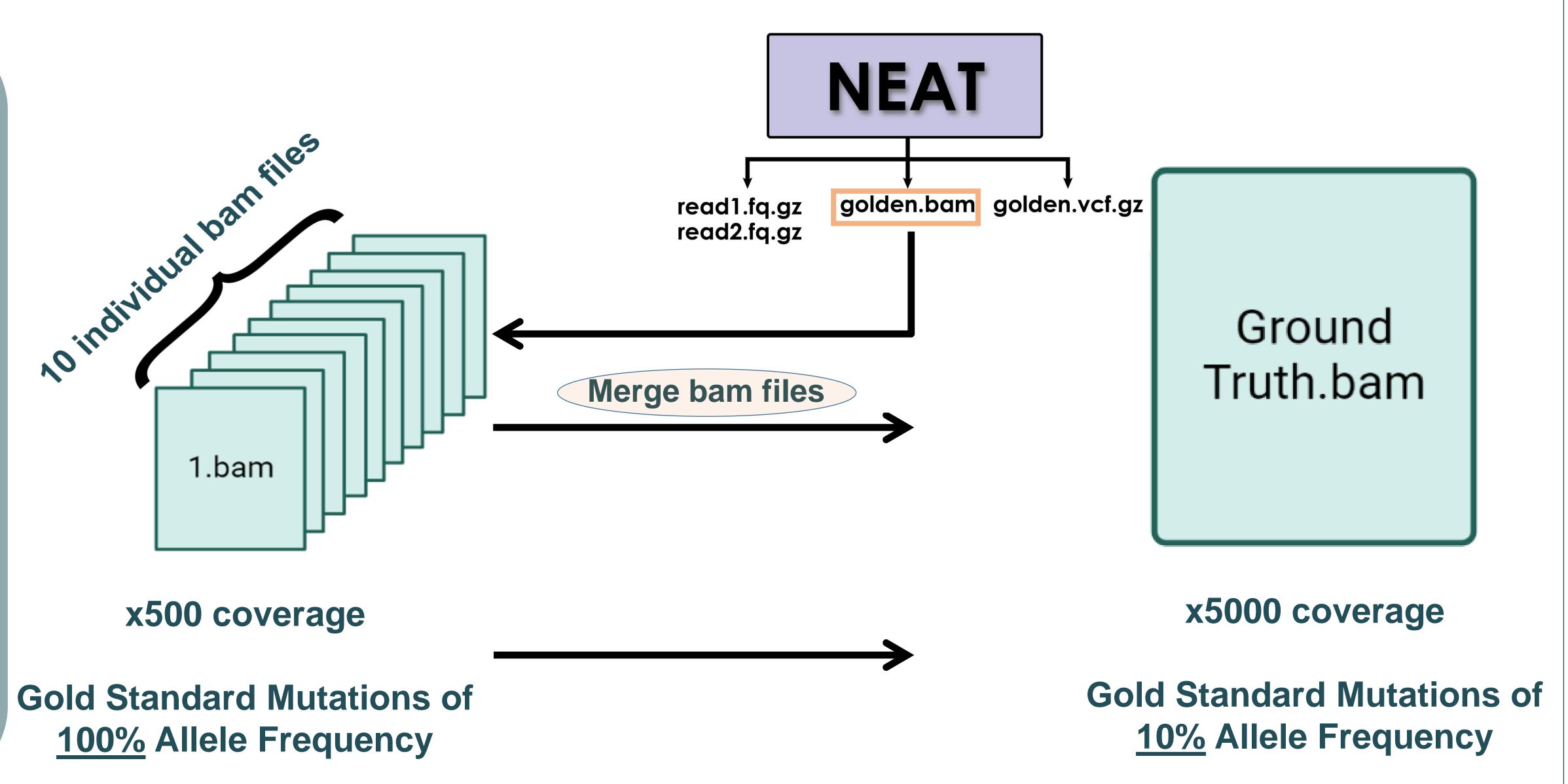
Fragkouli Styliani-Christina^{1, 2}, Pechlivanis Nikos¹, Agathangelidis Andreas², Psomopoulos Fotis¹

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

Synthetic «Gold Standard» Dataset Generation

Highlights

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



2 Benchmarking GATK-Mutect2 Variance in AF Density plots of «Ground Truth» Mutations per DNA Base **Down-sampling of coverage of** Differences in AF of **«Ground Truth» Mutations «Ground Truth» Mutations** Misrepresentation of information 5,000 4,000 -10.0% 3,000 Coverage (No. 6 1.0% 1,000 0.1% **Ground Truth Mutect2 Ground Truth Mutect2** 25% 50% 75% 100% 25% 50% 75% 100% 25% 50% 75% 100% 25% 50% 75% 100% **Allele Frequency** Difficulty to identify mutations at low frequencies **Venn plot of the Overall Mutations Ground Truth GATK**

Alterati 10337 5031 1034 (63.0%) (30.7%)(6.3%)5,000 15,000 10,000 **Chromosomal Position** To learn more Allele Frequency Category about our work

 5%
 10%
 20%
 50%
)

please visit Divergence in the identification of SNVs and their AF of «Ground Truth» Mutations

