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

Styliani-Christina Fragkouli^{1, 2}, Nikos Pechlivanis¹, Andreas Agathangelidis², Fotis Psomopoulos¹

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 1 Synthetic «Ground Truth» Dataset Generation NEAT Highlights Sindividual barnfile's Generation of synthetic golden.bam golden.vcf.gz genomics data based on read1.fq.gz read2.fq.gz TP53 gene Define «Ground Truth» Truth.bam Merge bam files SNPs and INDELs in order to benchmark 1.bam somatic variant callers Investigate the impact of x5000 coverage x500 coverage variant callers in variants at low frequencies "Gold Standard" Variants of 10% Allele Frequency 100% Allele Frequency















