

Synthetic Genomics Data Generation And Evaluation For The Use Case Of Benchmarking Somatic Variant Calling Algorithms

Styliani – Christina Fragkouli

PhD Candidate

✉ sfragkoul@certh.gr

🔄 [sfragkoul](https://github.com/sfragkoul)

🐦 [@scfragkoul](https://twitter.com/scfragkoul)



National and Kapodistrian University of Athens
Faculty of Sciences
Department of Biology
Section of Genetics & Biotechnology



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

MOTIVATION

- Variant calling plays an important role in identifying genetic lesions.
- In the case of variants at low frequency ($\leq 10\%$) identification becomes more challenging.
- The challenge that arises is the absence of a ground truth for reliable and consistent identification and benchmarking.

✉ sfragkoul@certh.gr

🔄 [sfragkoul](https://github.com/sfragkoul)

🐦 [@scfragkoul](https://twitter.com/scfragkoul)

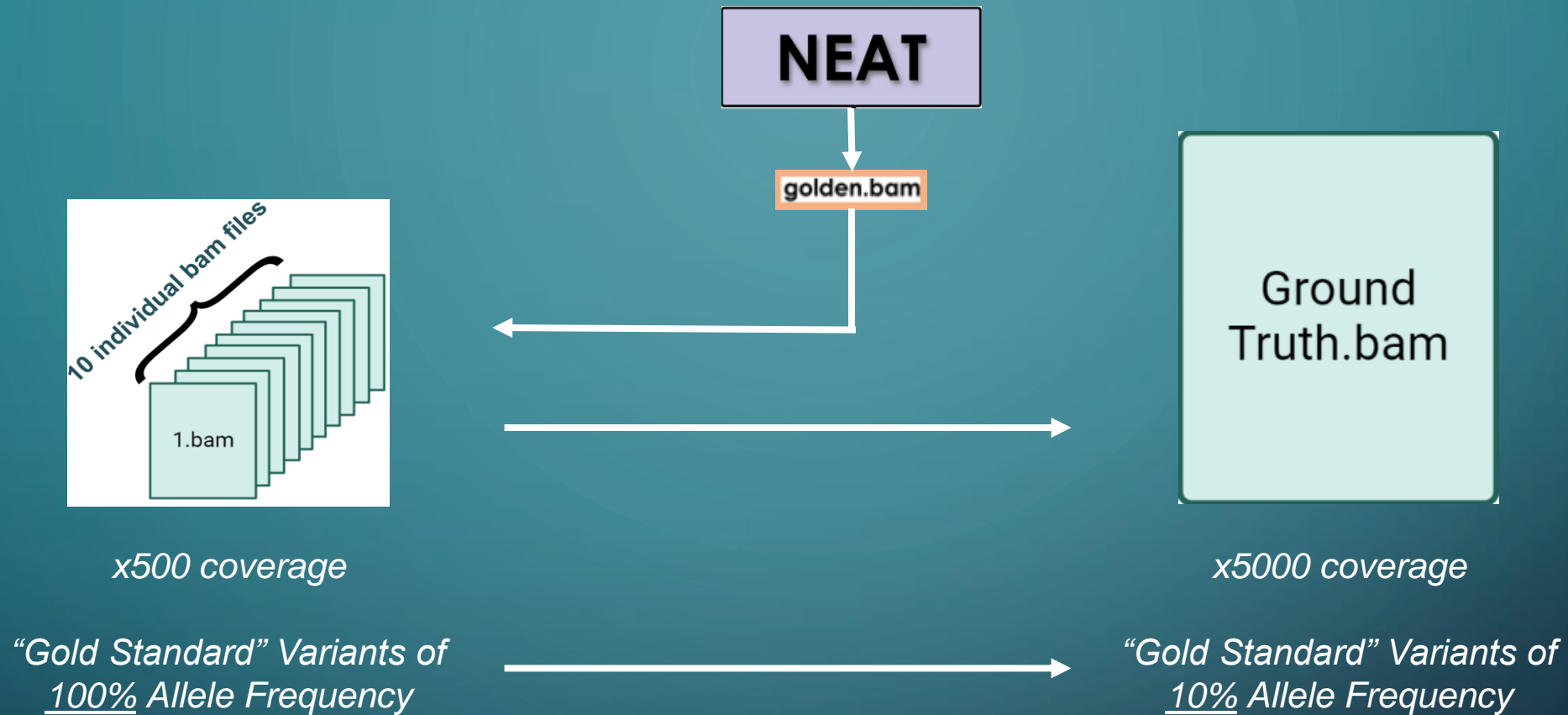


National and Kapodistrian University of Athens
Faculty of Sciences
Department of Biology
Section of Genetics & Biotechnology



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

SYNTHETIC «GOLD STANDARD» DATASET GENERATION



✉ sfragkoul@certh.gr

🔄 [sfragkoul](https://github.com/sfragkoul)

🐦 [@scfragkoul](https://twitter.com/scfragkoul)



National and Kapodistrian University of Athens
Faculty of Sciences
Department of Biology
Section of Genetics & Biotechnology

INAB
INSTITUTE OF APPLIED BIOSCIENCES
INSTITOYTO ΕΦΑΡΜΟΖΜΕΝΩΝ ΒΙΟΓΕΝΕΤΙΚΩΝ
CENTRE for RESEARCH and TECHNOLOGY-HELLAS



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

COMPARING ALGORITHMS

*In-silico generated dataset that contains
«Ground Truth» SNIPs and INDELs*



VS

*Results from GATK somatic
variant calling algorithm*



✉ sfragkoul@certh.gr

🔄 [sfragkoul](https://github.com/sfragkoul)

🐦 [@scfragkoul](https://twitter.com/scfragkoul)

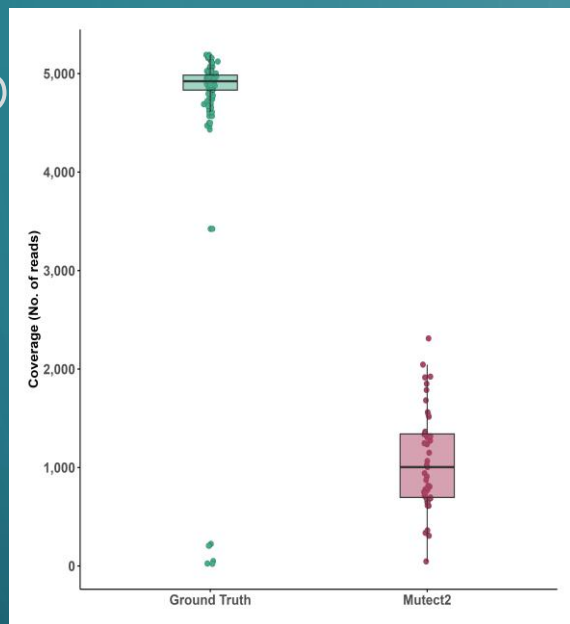


National and Kapodistrian University of Athens
Faculty of Sciences
Department of Biology
Section of Genetics & Biotechnology

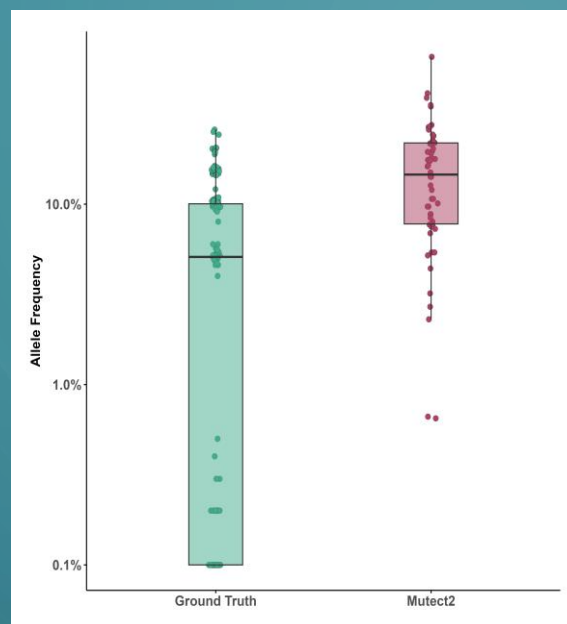


CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

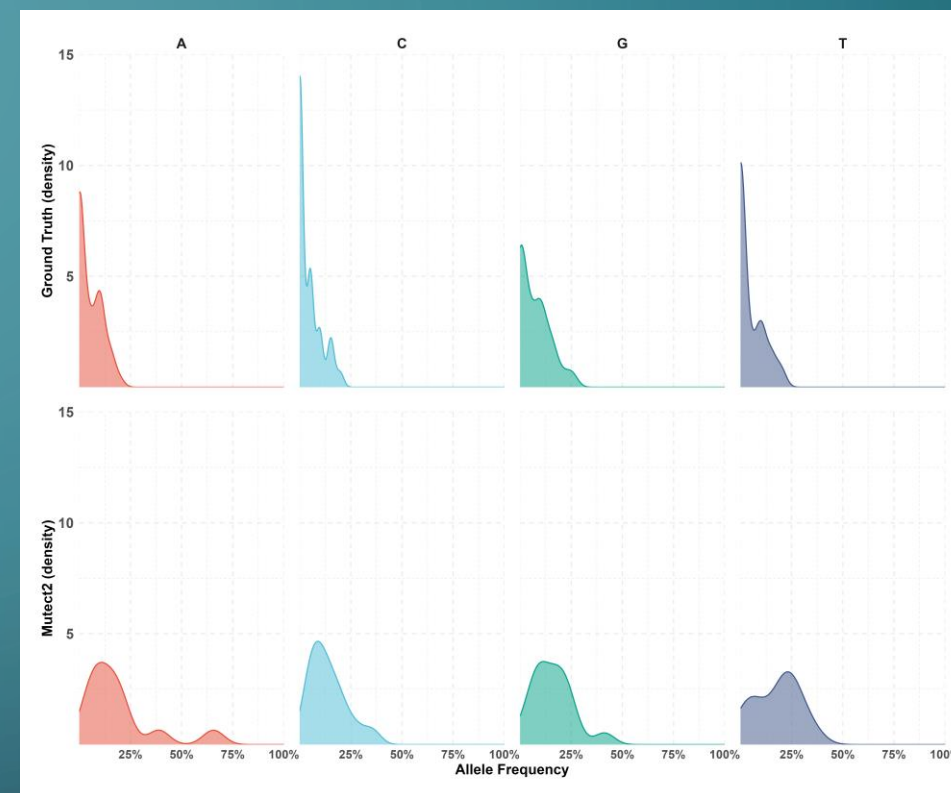
BENCHMARKING GATK-MUTECT2



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

✉ sfragkoul@certh.gr

🔄 [sfragkoul](https://github.com/sfragkoul)

🐦 [@scfragkoul](https://twitter.com/scfragkoul)



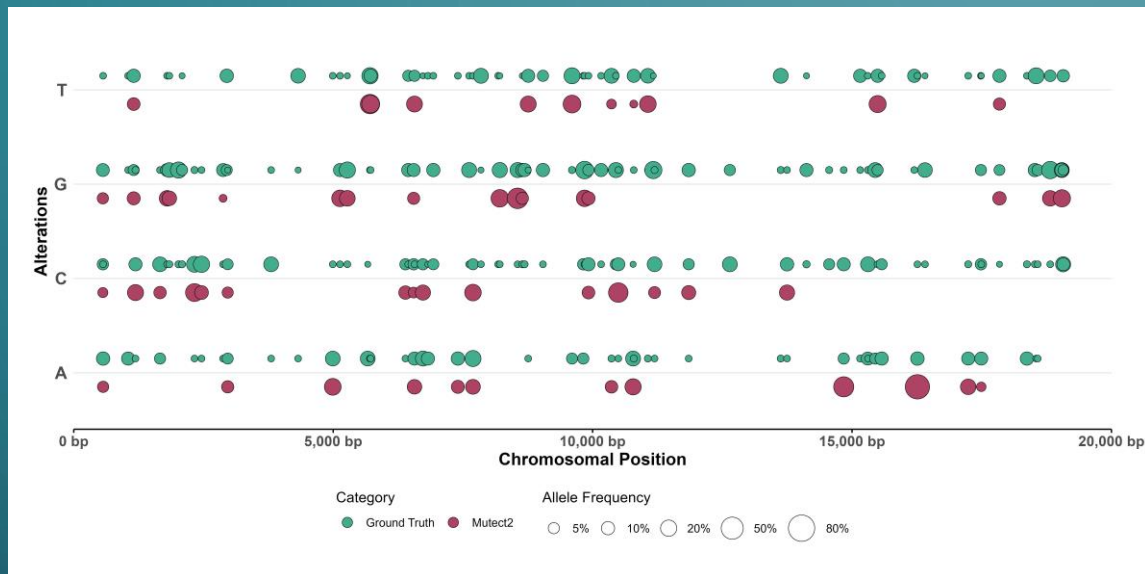
National and Kapodistrian University of Athens
Faculty of Sciences
Department of Biology
Section of Genetics & Biotechnology

INAB
INSTITUTE OF APPLIED BIOSCIENCES
INSTITUTOYTO EPAPMOZMENON BIOENIZHTHMON
CENTRE for RESEARCH and TECHNOLOGY-HELLAS

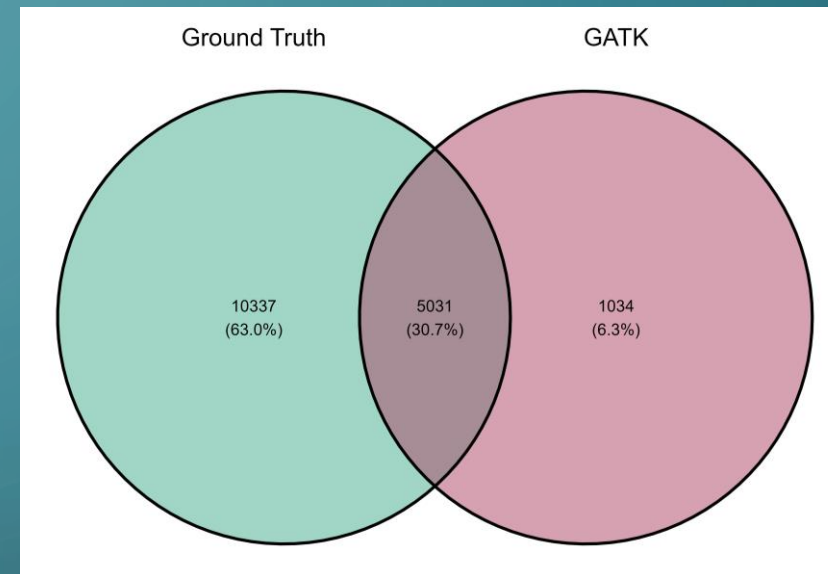


CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

BENCHMARKING GATK-MUTECT2



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



Venn plot of the Overall Variants

✉ sfragkoul@certh.gr

🔄 [sfragkoul](https://github.com/sfragkoul)

🐦 [@scfragkoul](https://twitter.com/scfragkoul)



National and Kapodistrian University of Athens
Faculty of Sciences
Department of Biology
Section of Genetics & Biotechnology

INAB
INSTITUTE OF APPLIED BIOSCIENCES
ΙΝΣΤΙΤΟΥΤΟ ΕΦΑΡΜΟΣΜΕΝΩΝ ΒΙΟΓΕΝΕΤΙΚΩΝ
CENTRE for RESEARCH and TECHNOLOGY-HELLAS



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

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**

To learn more about our work please visit



✉ sfragkoul@certh.gr

🔄 [sfragkoul](https://github.com/sfragkoul)

🐦 [@scfragkoul](https://twitter.com/scfragkoul)



National and Kapodistrian University of Athens
Faculty of Sciences
Department of Biology
Section of Genetics & Biotechnology

INAB
INSTITUTE OF APPLIED BIOSCIENCES
ΙΝΣΤΙΤΟΥΤΟ ΕΦΑΡΜΟΣΜΕΝΩΝ ΒΙΟΓΕΝΕΤΙΚΩΝ
CENTRE for RESEARCH and TECHNOLOGY-HELLAS



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

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

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

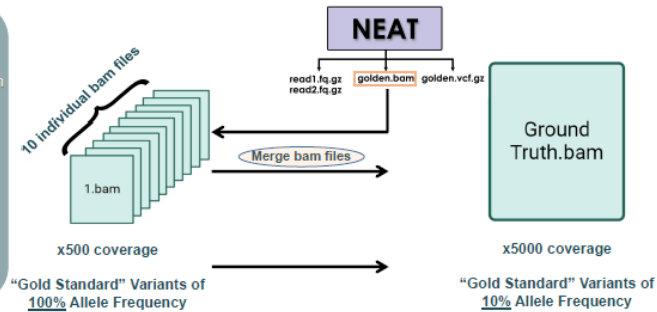
¹Institute of Applied Biosciences, Centre of Research and Technology Hellas, Thessaloniki, Greece

²Department of Biology, National and Kapodistrian University of Athens, Athens 10670, GR

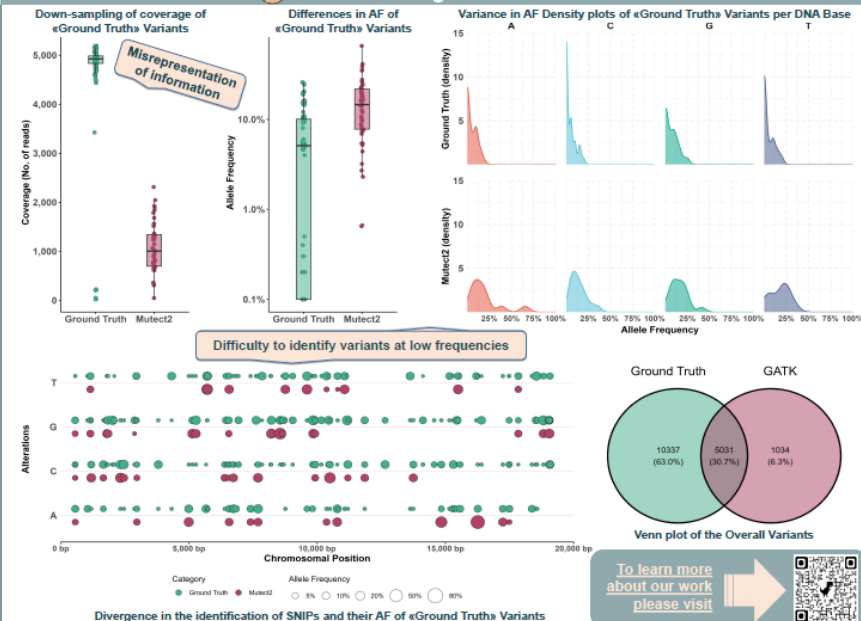
1 Synthetic «Gold Standard» Dataset Generation

Highlights

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



2 Benchmarking GATK-Mutect2



✉ sfragkoul@certh.gr

🌐 [sfragkoul](https://www.sfragkoul.com)

🐦 [@scfragkoul](https://twitter.com/scfragkoul)



National and Kapodistrian University of Athens
Faculty of Sciences
Department of Biology
Section of Genetics & Biotechnology



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

Come say Hi :)

Monday 24th of July

Poster Session A

Track: BOSC

Poster: A-102