Benchmarking and Quality Control for genomic variant calling

2025

Johannes Köster

Bioinformatics and Computational Oncology
Institute for AI in Medicine (IKIM)
University of Duisburg-Essen

Genomic variant calling and why it is challenging

Seeking for genomic variants

DNA

sequencing ▼

sequencing reads

```
AACCGATTAACCGGAGTCCCTCGGTAGTTATTTACC
AACCGGAGTCCCTCGGTAGTTATTTACCCTCTCCGC
AGTCCCTCGGTAGTTATTTACCCTCTCCGCGTCCTTTC
ATCCGGAGTCGCAACCGATTAACCGGAGTCCCT
GAGTCGCAACCGATTAACCGGAGTCCCTCGGTAGTTAT
```

read alignment ▼

aligned reads

```
                    AACCGATTAACCGGAGTCCCGCGGTAGTTATTTACC
                          AACCGGAGTCCCGCGGTAGTTATTGACCCTCTCCGC
                                AGTCCCTCGGTAGTTATTTACCCTCTCCGCGTCCTTTC
               ATCCGGAGTCCCAACCGATTAACCGGAGTCCCT
                    GAGTCGCAACCGATTAACCGGAGTCCCTCGGTAGTTAT
...GTAATCCGGAGTCGCAACCGATTAACCGGAGTCCCGCGGTAGTTATTTACCCTCTCCGCGTCCTTTCTA...
```

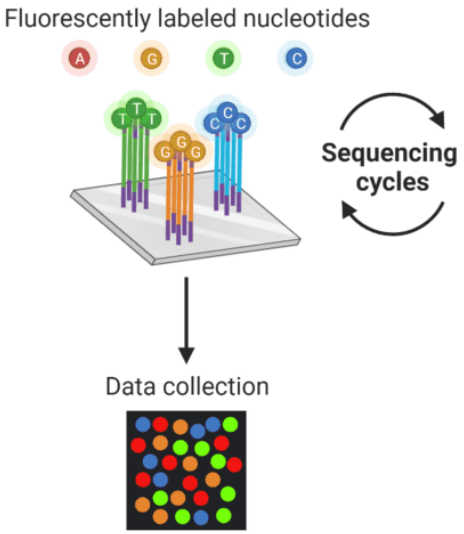
variant calling ▼

genomic variants

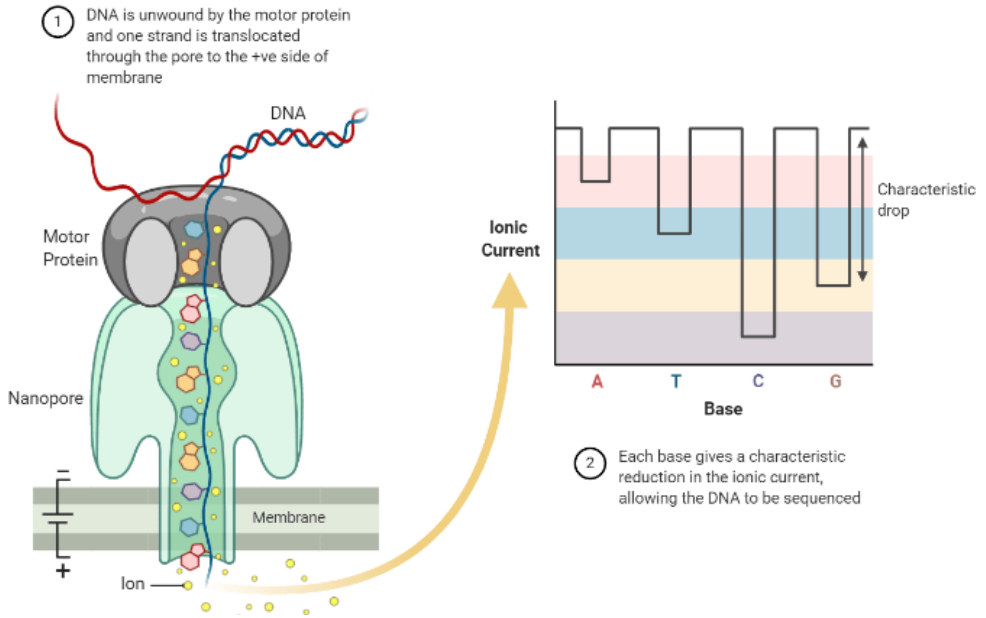
```
                    AACCGATTAACCGGAGTCCCGCGGTAGTTATTTACC
                          AACCGGAGTCCCGCGGTAGTTATTGACCCTCTCCGC
                                AGTCCCTCGGTAGTTATTTACCCTCTCCGCGTCCTTTC
               ATCCGGAGTCCCAACCGATTAACCGGAGTCCCT
                    GAGTCGCAACCGATTAACCGGAGTCCCTCGGTAGTTAT
...GTAATCCGGAGTCGCAACCGATTAACCGGAGTCCCGCGGTAGTTATTTACCCTCTCCGCGTCCTTTCTA...
```

Sequencing

short reads:



long reads:



basecalling uncertainty:

posterior probability of incorrect base (base quality)

Read alignment

```
AACCGATTAACCGGAGTCCCTCGGTAGTTATTTACC
AACCGGAGTCCCTCGGTAGTTATTTACCCTCTCCGC
AGTCCCTCGGTAGTTATTTACCCTCTCCGCGTCCTTTC
ATCCGGAGTCGCAACCGATTAACCGGAGTCCCT
GAGTCGCAACCGATTAACCGGAGTCCCTCGGTAGTTAT
```

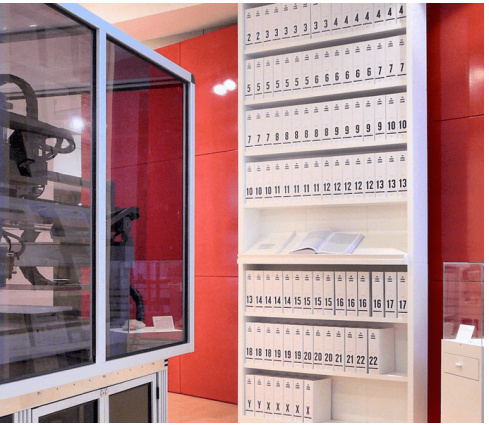
for each read:

find best position of a short text in a very long text (alphabet: A,C,G,T)

▼

challenges:

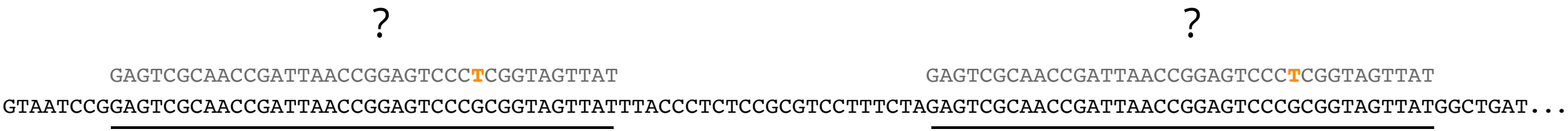
- repetetive regions
- sequencing errors
- variants



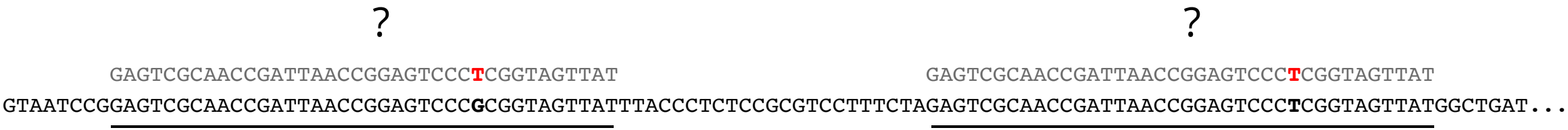
```
                    AACCGATTAACCGGAGTCCCGCGGTAGTTATTTACC
                          AACCGGAGTCCCGCGGTAGTTATTGACCCTCTCCGC
                               AGTCCCTCGGTAGTTATTTACCCTCTCCGCGTCCTTTC
      ATCCGGAGTCCCAACCGATTAACCGGAGTCCCT
           GAGTCGCAACCGATTAACCGGAGTCCCTCGGTAGTTAT
...GTAATCCGGAGTCGCAACCGATTAACCGGAGTCCCGCGGTAGTTATTTACCCTCTCCGCGTCCTTTCTA...
```

Alignment uncertainty

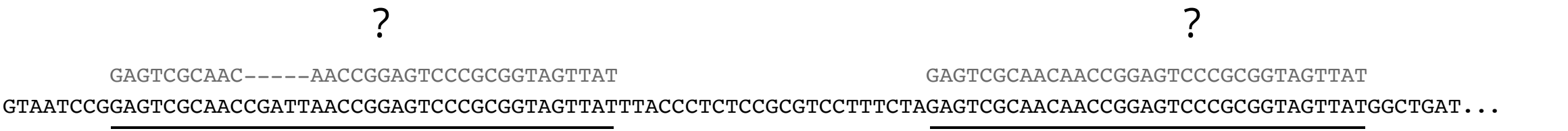
repetitive regions:



sequencing errors:



variants:



Global benchmarking of genomic variant calling

Benchmarking genomic variant calling

Given:

- a plethora of methods to choose from, each with a plethora of parameters
- continuous innovations and method improvements

Thus:

- take (real) datasets with known truth
- fight overfitting by using many benchmark datasets
- continuous benchmarking!!

Exemplary benchmarks

Genome in a Bottle

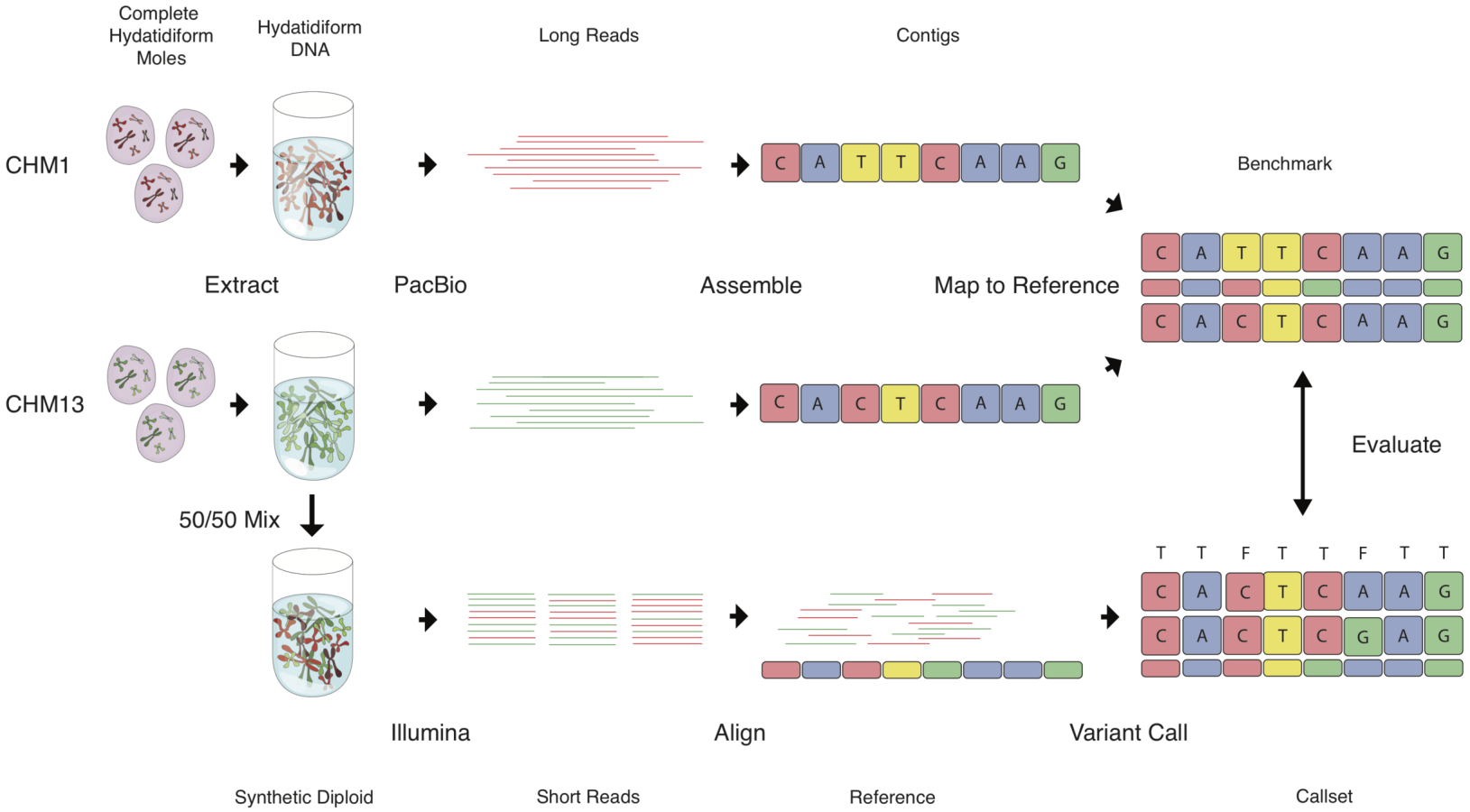
- NIST-lead consortium for generating benchmark datasets for genomic variant calling
- NA12878/HG001 (well-studied sample from the HapMap project)
- three further sets of samples from family pedigrees
- Snakemake-based pipeline for consensus HQ variant calls from multiple technologies and callers

https://www.nist.gov/programs-projects/genome-bottle

https://github.com/usnistgov/defrabb

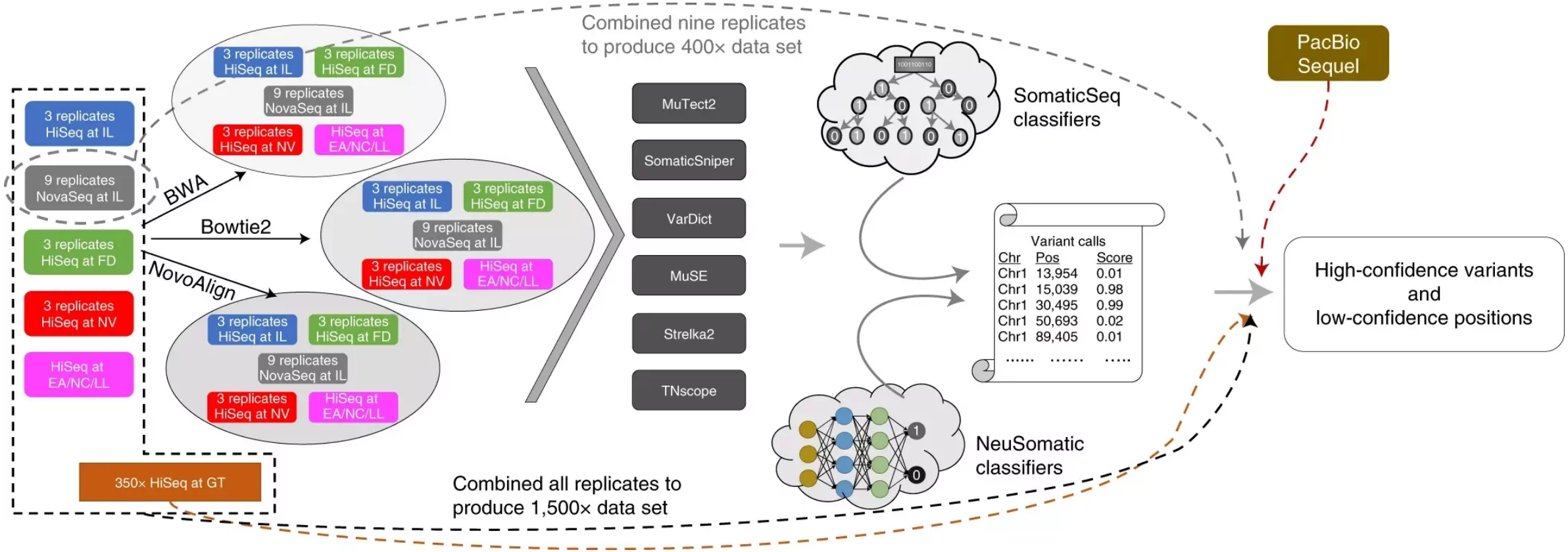
CHM-eval

Synthetic, pseudo-diploid benchmark sample as mixture of two haploid cell lines (CHM13, CHM1)



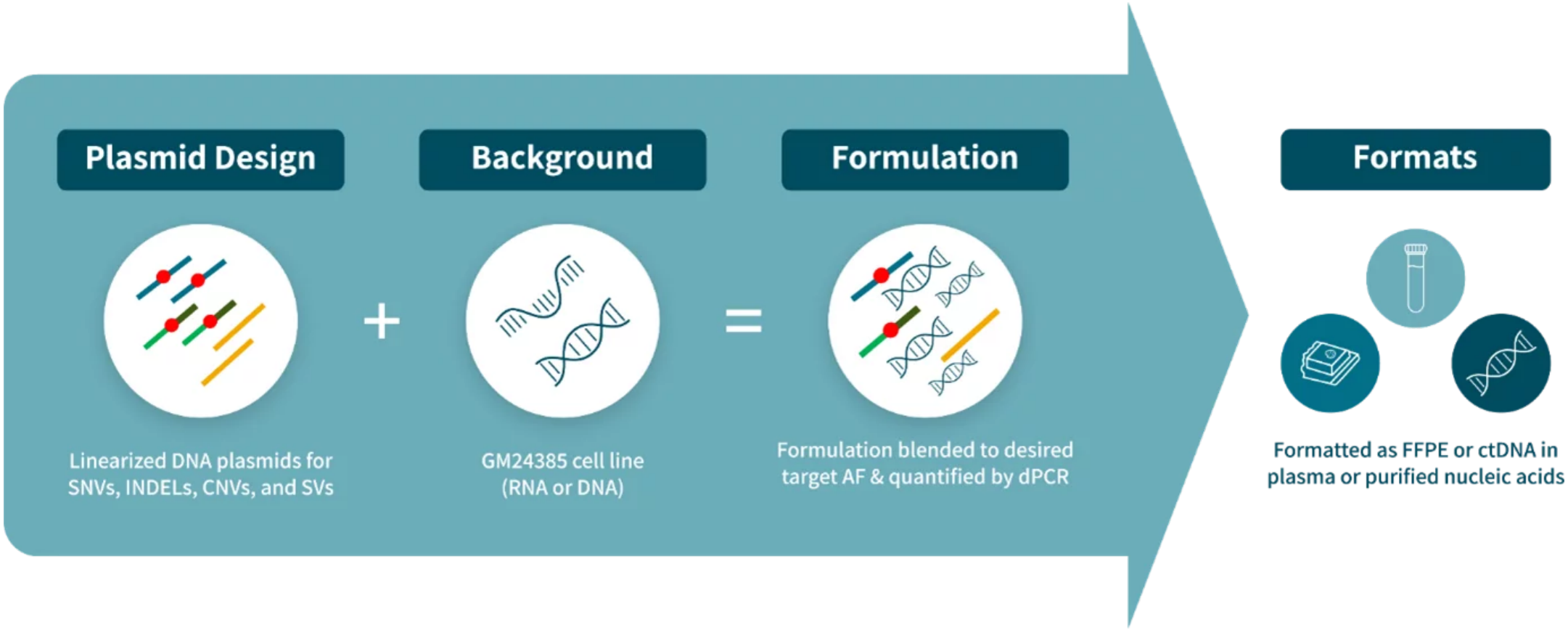
Li et al. Nature Methods 2018

SEQC2



Fang et al. Nature Biotechnology 2021

Seracare "somatic cancer" reference material



https://www.seracare.com/Controls---Reference-Materials-NGS-Somatic-Cancer

NCBench


NGS-CN

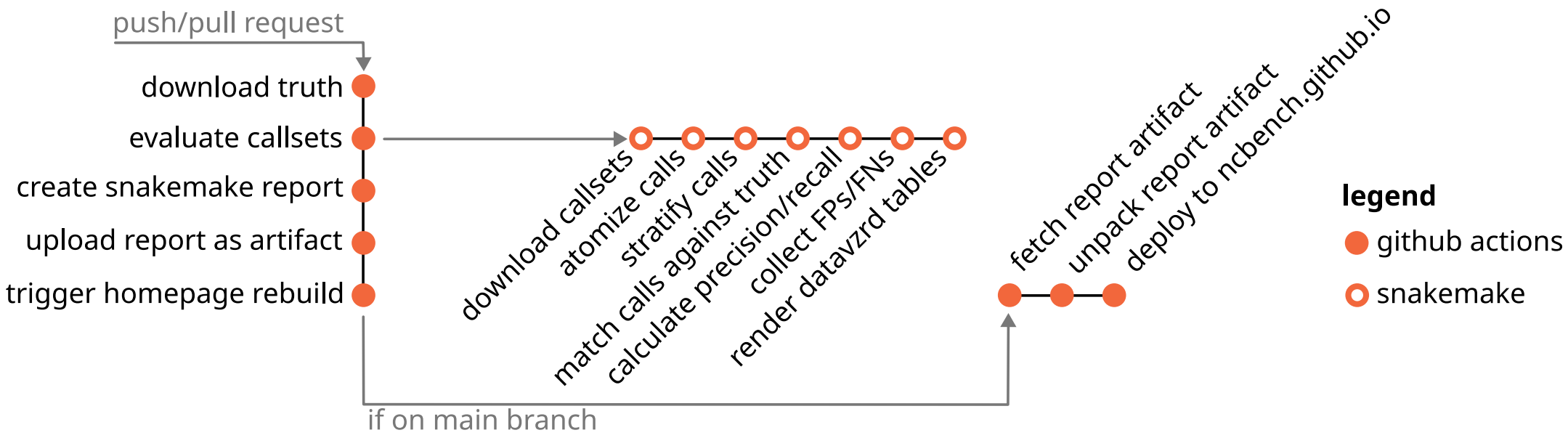

DRESDEN
concept
genome
center
COMPETENCE
CENTRE FOR
GENOMIC
ANALYSIS KIEL


WEST GERMAN
GENOME CENTER
NCCT


GHGA


MEDIZIN
INFORMATIK
INITIATIVE
PM4ONCO

NCbench in a nutshell



F1000Research F1000Research 2024, 12:1125 Last updated: 13 NOV 2024

Check for updates

RESEARCH ARTICLE

REVISED **NCBench: providing an open, reproducible, transparent, adaptable, and continuous benchmark approach for DNA-sequencing-based variant calling**

[version 2; peer review: 2 approved]

Friederike Hanssen[1], Gisela Gabernet[1], Famke Bäuerle[1-4], Bianca Stöcker[5], Felix Wiegand[5], Nicholas H. Smith[6], Christian Mertes[6-8], Avirup Guha Neogi[9], Leon Brandhoff[9,10], Anna Ossowski[9], Janine Altmueller[9,11,12], Kerstin Becker[9], Andreas Petzold[13], Marc Sturm[14], Tyll Stöcker[15], Sugirthan Sivalingam[16], Fabian Brand[17], Axel Schmidt[18], Andreas Buness[19], Alexander J. Probst[20], Susanne Motameny[9,10], Johannes Köster[5,21]

Datasets

- GIAB
- CHM-eval
- SEQC2
- Seracare (upcoming)

Key technologies:

- Snakemake
- Datavzrd
- Github Actions
- Github Pages

https://snakemake.github.io

https://datavzrd.github.io

Representing and comparing variants

```
1       10167   .       CCTAACCCTAACCTA CCTTAACCTTAACCTA
```

```
CCTAACCC TAACCTA
CCTTAACCTTAACCTA
```

▼

```
1    10167    CCTAACCCTAA    C
1    10170    A    T
1    10172    C    A
1    10174    C    CT
```

```
CC TAACCCTAACCTA
CCTTAACCTTAACCTA
```

▼

```
1    10167    CCTAACCCTAA    C
1    10168    C    CT
1    10174    C    T
```

▶

- apply bcftools norm with same version to truth and callset
- use rtg vcfeval for determining TP, FP, FN

https://samtools.github.io/bcftools

https://github.com/RealTimeGenomics/rtg-tools

Calling precision and recall

recall:

should go down with weaker evidence

precision:

should be independent of evidence strength

Hence:

stratification by read depth

- determine read depth with mosdepth
- stratify truth and callsets with bedtools

https://github.com/brentp/mosdepth

https://bedtools.readthedocs.io

Genotyping precision and recall

For genotyping, precision may decrease with lower read depth (because it is MLE), for calling not.

- Genotyping and calling should be analyzed separately.

https://ncbench.github.io

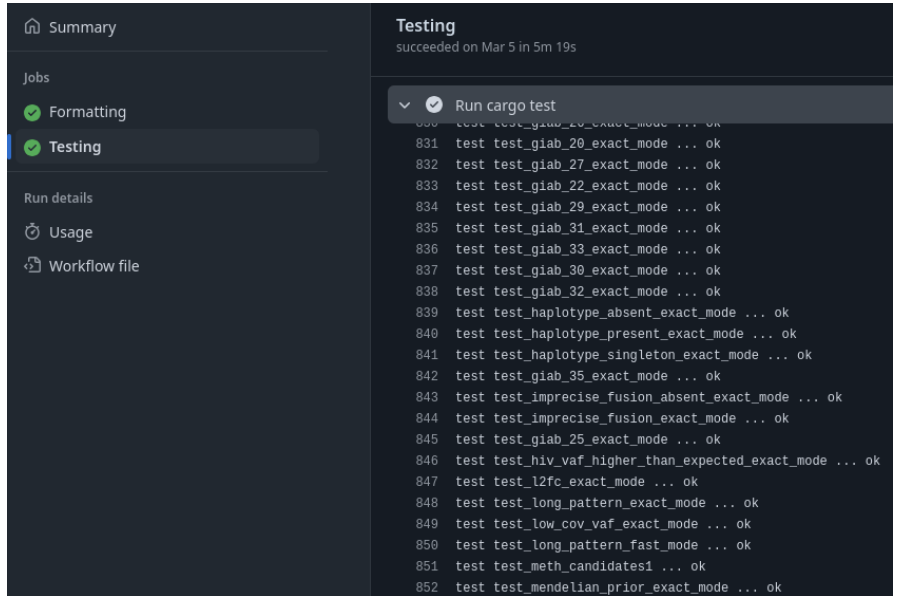
Local benchmarking

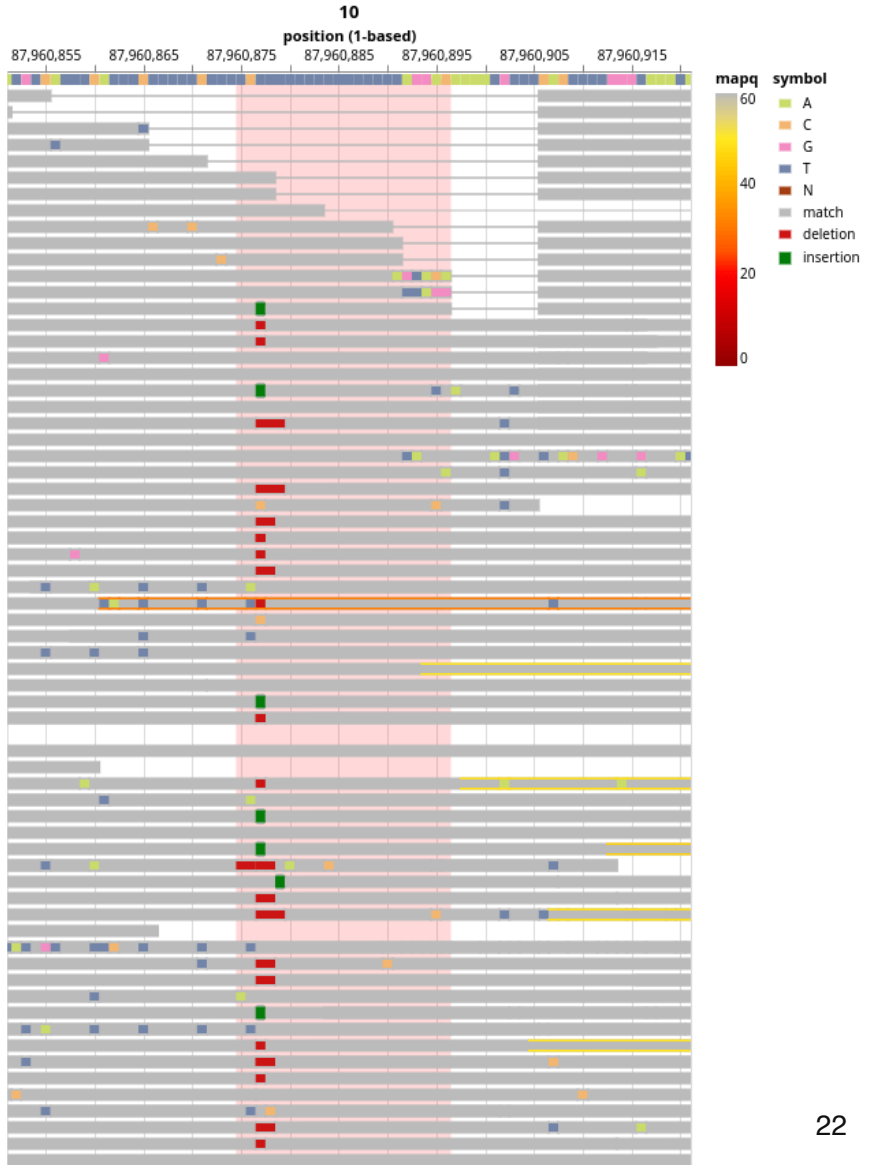
Issues with global benchmarks

- Truth can be biased towards used tools
- Limited set of confidence regions
- At individual loci, performance may differ a lot from a global measure
- Expensive to compute ▸ rarely repeated

Good complement: local benchmarks

- individual variants *or non-variants*
- truth is known by prior knowledge, orthogonal validation, or manual inspection
- fast to test
- can be embedded in continuous integration platforms like github actions





Automatic testcase generation in Varlociraptor

Automatic test case generation:

```
varlociraptor call variants --testcase-prefix testcase --testcase-locus CHROM:POS generic \
  --scenario scenario.yaml --obs tumor=tumor.bcf normal=normal.bcf
```

145

public testcases (simulated + real benchmarks)

66

private testcases (clinical)

Köster et al. Genome Biology 2020

https://varlociraptor.github.io

Lähnemann et al. Nature Communications 2021

Conclusion

- Genomic variant calling is still a hard problem
- Global benchmarks help to understand performance of method and parameter combinations
- As methods evolve fast, it is crucial to apply them continuously
- Combination of Github/Snakemake/Datavzrd makes that easy
- Local benchmarks complement global ones by being faster and specific for problematic regions

Acknowledgements

Susanne Motameny, Gisela Gabernet, Friederike Hansen, Famke Bäuerle, Bianca Stöcker, Hamdiye Uzuner, Adrian Prinz, Felix Mölder, Felix Wiegand, Nicholas H. Smith, Christian Mertes, Avirup Guha Neogi, Leon Brandhoff, Anna Ossowski, Janine Altmueller, Kerstin Becker, Andreas Petzold, Marc Sturm, Tyll Stöcker, Sugirthan Sivalingam, Fabian Brand, Axel Schmidt, Andreas Buness, Alexander J. Probst, Narci Kübra, Anna Bennet-Pages