High-Resolution cfDNA Sequencing Data Simulation for the Development of a Minimal Cancer Residual Disease Monitoring Model

Augustas Markevičius Ramunė Bružaitė Mindaugas Morkūnas

Institute of Data Science and Digital Technologies, Vilnius University

The problem and proposal

Liquid biopsy sequencing data can be used to track cancer without a direct tumor tissue biopsy, but in such a dataset only a small fraction of DNA fragments may come from the tumour. Methods for tracking this signal over time need test data, however real-life ctDNA datasets with known ctDNA fractions, subclonal architectures, etc. are slow to collect, costly and restricted by patient privacy.

We propose a method to generate synthetic ctDNA datasets comprising ground truth extraction from tumor WGS data, determining sampling and fragment feature rules, combining these into a data generation scheme and validation using existing somatic mutation calling tools.

Ground truth extraction

The DREAM SMC-HET challenge is used for ground truth information. 51 simulated tumors are considered, each built by spiking mutations into a phased high-depth HG002 genome. Subclone structure, copy-number events and mutations are preset by design, and all tumours share one germline.

We extract:

1. The single nucleotide variant (SNV) list
2. Each SNV's subclone
3. Each cluster's cancer cell fraction and the clonal tree
4. The design purity
5. Allele-specific copy-number states

Fragment sampling

Each DNA molecule in the synthetic sample is either ctDNA, Clonal hematopoiesis (CH) DNA, or healthy cfDNA. Sampling is based on variant allele frequency (equation in Fig. 1, derived from [1]). The ctDNA fraction f is swept from 10^{-1} (advanced disease) to 10^{-4} or lower (MRD) [2] [3]. Parameters CCF, N_T, m come from the DREAM truth.

For each tracked site having SNVs (DREAM ground truth), the number of mutant molecules is drawn from a Poisson distribution with mean (genome equivalents × VAF).

CH variants are sampled from empirical VAF distributions [3], in CH-dominated genes chosen from [4]. Their VAFs are scaled by $(1 - f)$.

All remaining molecules are healthy cfDNA carrying HG002 germline variants.

Fragment features

The source of the fragment (ctDNA, clonal hematopoeisis DNA or healthy) is assigned first. This determines the fragment's length and end motifs. Clonal hematopoeisis DNA fragments are similar in length to healthy DNA. ctDNA fragments are shorter than healthy whether or not they carry a tracked SNV, though their length varies by cancer type [5].

A fragment is placed randomly across target regions, and only those overlapping a target are kept.

Further fragment features are summarized in Table 1.

Table 1: Fragment features by fragment type

Fragment feature	Healthy, CH DNA	Tumour DNA	Source
Length mode	∼ 166 bp	∼ 143-145 bp	[6], [5], [7]
% fragments <150 bp	∼ 21%	∼ 41%	[5]
10 bp periodicity below 150 bp	Weak	Stronger	[6], [5]
5' end motifs	CC-rich ends	Fewer CC ends	[6], [7]

References

[1] M. Tarabichi *et al.*, "A practical guide to cancer subclonal reconstruction from DNA sequencing," *Nature Methods*, vol. 18, pp. 144–155, 1 2021.

[2] S. Akbarinejad *et al.*, "Tumor-naïve ctDNA detection with deep learning-enhanced error suppression for sensitive mutation calling," *Genome Medicine*, vol. 18, pp. 90–, 6 2026.

[3] P. Razavi *et al.*, "High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants," *Nature Medicine*, vol. 25, pp. 1928–1937, 11 2019.

[4] D. W. Brown *et al.*, "Fragmentomic-based algorithm to computationally predict tumor-somatic, germline, and clonal hematopoiesis variant origin in liquid biopsy," *The Journal of Liquid Biopsy*, vol. 9, p. 100311, 7 2025.

[5] F. Mouliere *et al.*, "Enhanced detection of circulating tumor DNA by fragment size analysis," *Science Translational Medicine*, vol. 10, pp. eaat4921–, 11 2018.

[6] Y. M. D. Lo *et al.*, "Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies," *Science*, vol. 372, pp. eaaw3616–, 4 2021.

[7] W. A. Tsui *et al.*, "Cell-free DNA fragmentomics in cancer," *Cancer Cell*, vol. 43, pp. 1792–1814, 10 2025.

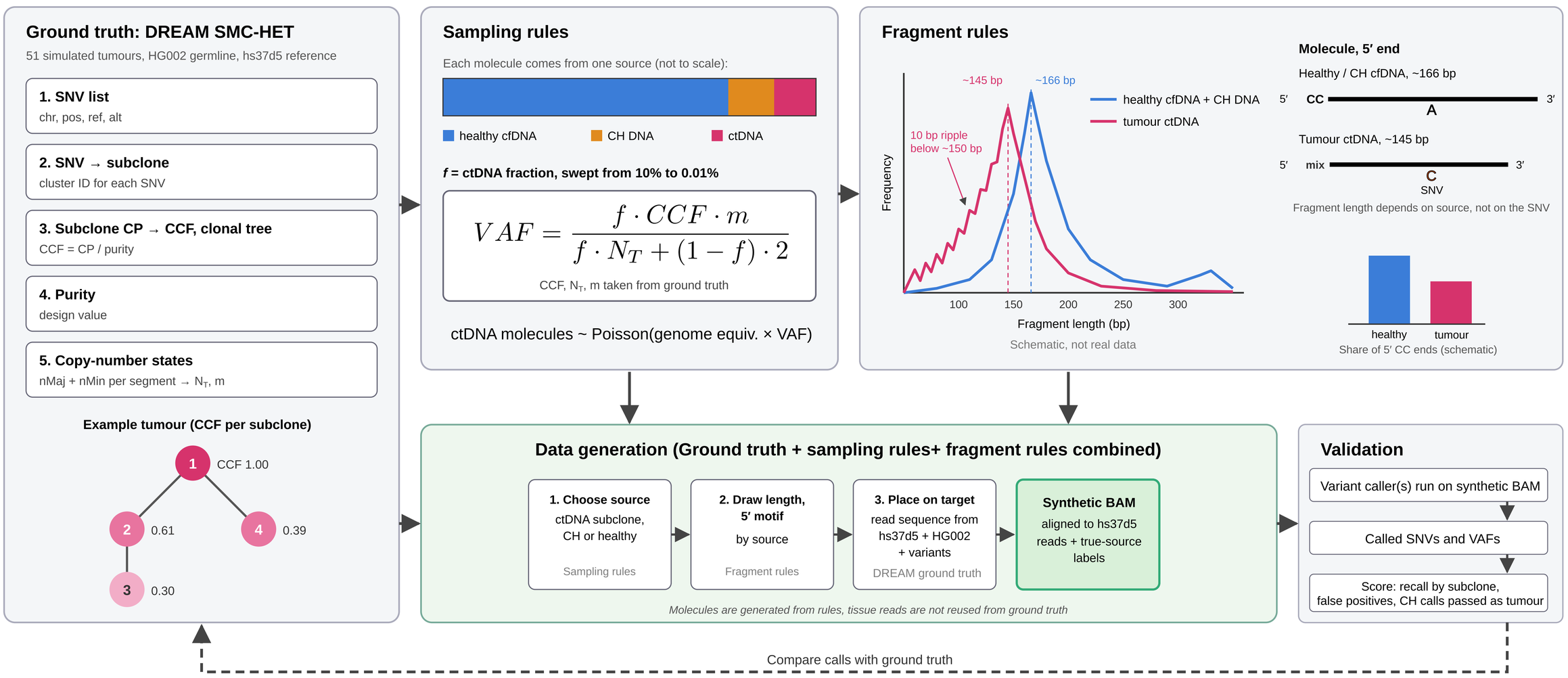

Figure 1: Overcall schema of the proposed methodology