Scalable subclonal reconstruction of cancer cells in DNA sequencing data using a penalized likelihood model

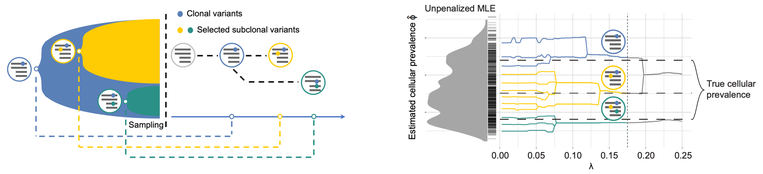
Yujie Jiang[1,2,#], Matthew D Montierth[1,3,#], Yu Ding[1], Kaixian Yu[4], Quang Tran[1], Aaron Wu[1], Ruonan Li[1], Shuangxi Ji[1], Xiaoqian Liu[1,5], Seung Jun Shin[6], Shaolong Cao[1], Yuxin Tang[7], Tom Lesluyes[8], Marek Kimmel[2], Jennifer R. Wang[9], Maxime Tarabichi[10], Hongtu Zhu[11], Peter Van Loo[8,12,13], Wenyi Wang[1,*]

[1]Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030 [2]Department of Statistics, Rice University, Houston, TX 77005, [3]Graduate program in Quantitative Computational Biology, Baylor College of Medicine, Houston, TX 77030, [4]Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, [5]Current address: Department of Statistics, The University of California, Riverside, CA 92507, [6]Department of Statistics, Korea University, Seoul, South Korea, 02841, [7]Department of Computer Science, Rice University, Houston, TX 77005, [8]Cancer Genomics Group, The Francis Crick Institute, London, UK , [9]Department of Head and Neck Surgery, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, [10]IRIBHM-J.E. Dumont, Brussels, Belgium, B1070, [11]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, [12]Department of Genetics, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, [13]Department of Genomic Medicine, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, [#]authors contributed equally, [*]Correspondence: WWang7@mdanderson.org

Cancer
UT MD Anderson

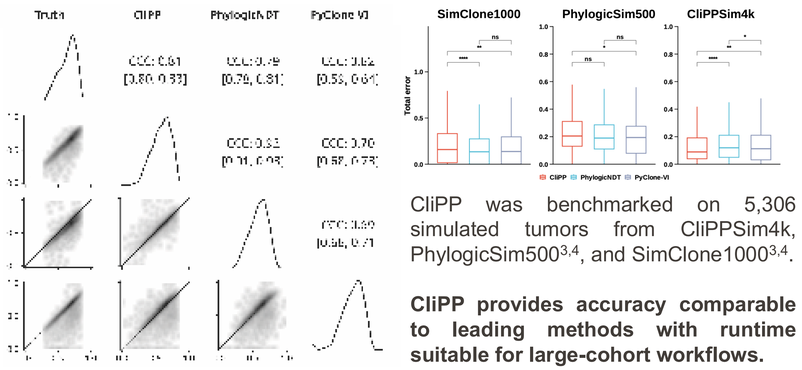
CliPP: Clonal structure identification through Pairwise Penalization

CliPP[1] estimates mutation-specific cellular prevalence from SNV read counts after adjusting for tumor purity and allele-specific copy number. Pairwise SCAD[2] penalization fuses mutations with similar cellular prevalence, producing clonal and subclonal mutation clusters without pre-specifying the number of clones.



Input: SNV read counts + tumor purity + allele-specific copy number
Output: mutation clusters + cellular-prevalence estimates + clonal/subclonal labels

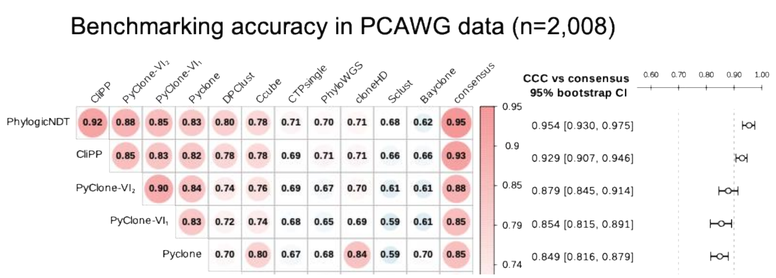
CliPP is fast and accurate in simulated and real data

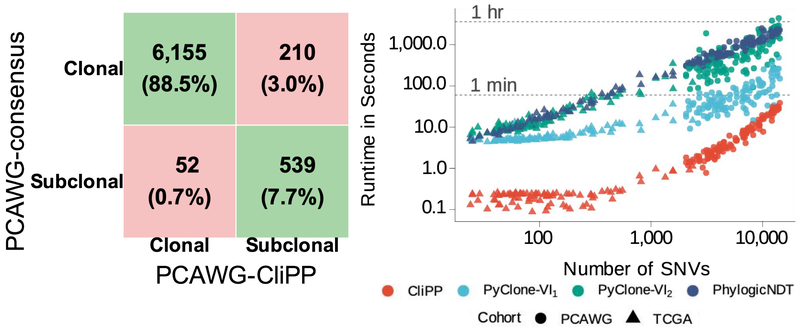


CliPP was benchmarked on 5,306 simulated tumors from CliPPSim4k, PhylogicSim500[3,4], and SimClone1000[3,4].

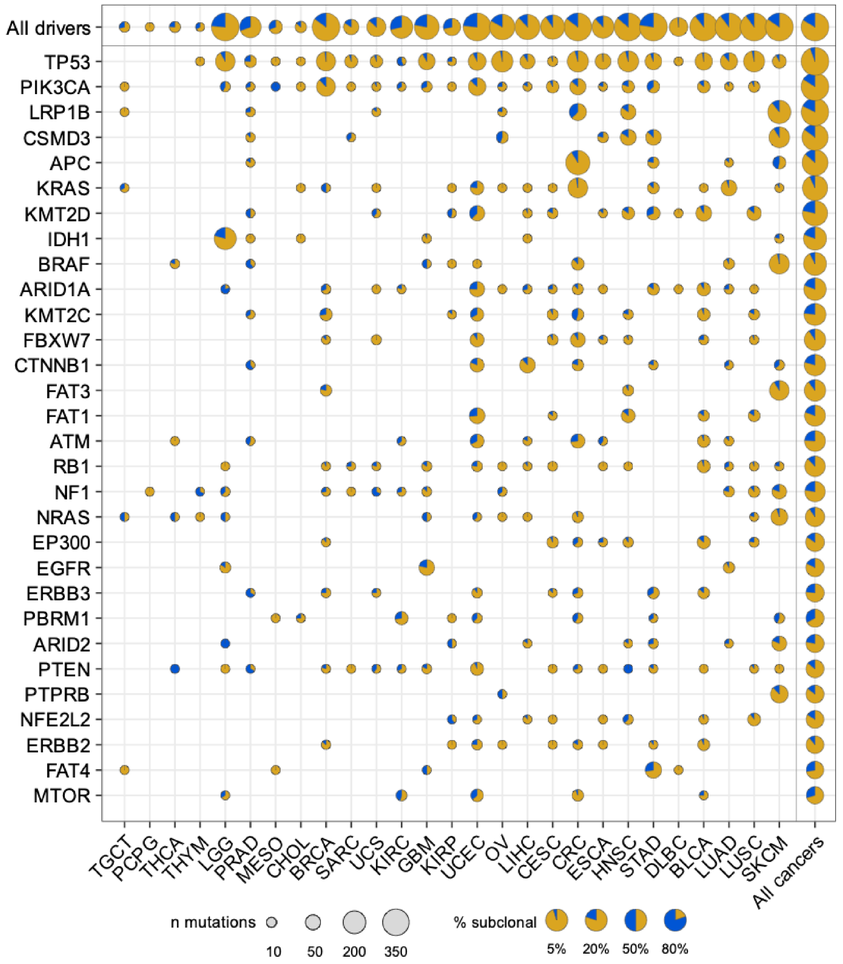
CliPP provides accuracy comparable to leading methods with runtime suitable for large-cohort workflows.

Benchmarking accuracy in PCAWG data (n=2,008)



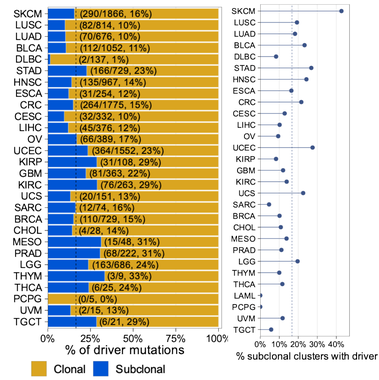


Pan-cancer driver clonality landscape



Each bubble represents one gene × cancer-type pair. Bubble area scales with the number of driver[5] mutations in that cell; pie wedges show the clonal versus subclonal fraction. Cancer types on the x-axis are ordered by median tumor mutational burden, and genes on the y-axis are ordered by total mutation count. The top row summarizes all drivers within each cancer type; the rightmost column summarizes each driver gene across all cancers.

TP53, APC, and KRAS were mostly clonal across cancers, consistent with early founding roles in tumor evolution. In contrast, KMT2C, CTNNB1, and ATM showed higher subclonal fractions, suggesting later acquisition in a subset of tumors.



CliPP annotated 13,666 high-confidence driver mutations in 274 genes across 4,218 tumors; 28 cancer types are shown in the final driver-clonality panels.

2,256 / 13,666 driver mutations were subclonal (16.5%).

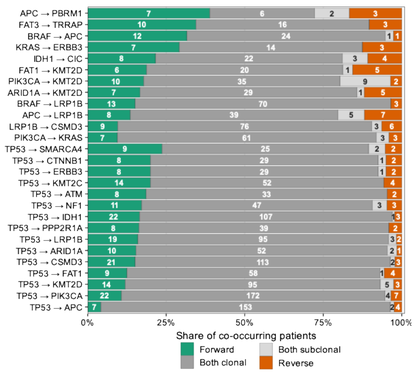
Drivers appeared in 16.7% of subclonal mutation clusters. Gene-level patterns were heterogeneous: TP53, APC, and KRAS were mostly clonal, whereas KMT2C, CTNNB1, and ATM were more often subclonal.

Inferring timing of driver pairs

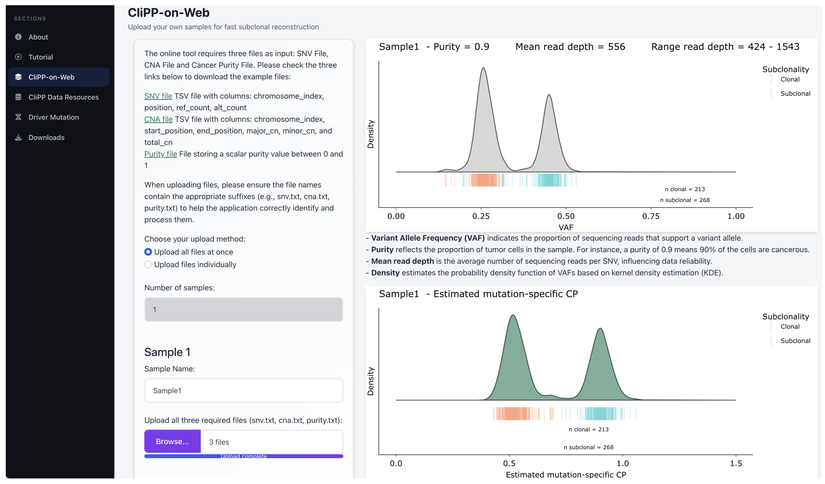
We infer driver order from CliPP clonal/subclonal status: if gene A is clonal and gene B is subclonal in the same tumor, A is interpreted as earlier.

Pairs are classified as Forward, Reverse, Both clonal, or Both subclonal. Plotted pairs have ≥10 resolvable patients, where resolvable = Forward + Reverse.

Most pairs were both clonal, reflecting shared early acquisition and limited timing resolution from single-biopsy bulk sequencing, with forward directions capturing known biology.



CliPP-on-web for coding-free subclonal reconstruction



CliPP-on-web provides a coding-free interface for subclonal reconstruction and visualization. Users can upload SNV, copy-number, and purity files; run CliPP; inspect VAF and cellular-prevalence distributions; query TCGA and PCAWG samples; check clonal/subclonal status of TCGA driver mutations; and download CliPP output files.

CliPP GitHub

CliPP bioRxiv

CliPP-on-web

1. Jiang, YJ, *et al*. Scalable subclonal reconstruction of cancer cells in DNA sequencing data using a penalized likelihood model, *bioRxiv* , (2026)
2. Fan, J. & Li, R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Am. Stat. Assoc.* **96**, 1348–1360 (2001).
3. Dentro, S. C. *et al*. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239-2254.e39 (2021).
4. Dentro, S. PCAWG Intra-Tumor Heterogeneity Simulations. **1**, (2021).
5. Martínez-Jiménez, F. *et al*. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).

CliPP provides a fast, accurate, and scalable framework for bulk-DNA subclonal reconstruction. By combining pairwise penalized likelihood with standard SNV, purity, and allele-specific copy-number inputs, CliPP achieves high concordance with PCAWG consensus reconstructions while enabling cohort-scale TCGA analysis. Applied across TCGA, CliPP identifies cancer-type-, gene-, and driver-pair-specific patterns of driver subclonality, generating a resource for studying tumor evolution and driver mutation timing.