Pan-cancer subclonal mutation analysis of 7,827 tumors predicts clinical outcome

Yujie Jiang[1,2,#], Matthew D Montierth[1,3,#], Kaixian Yu[4,#], Shuangxi Ji[1], Shuai Guo[1], Quang Tran[1], Xiaoqian Liu[1], Seung Jun Shin[5], Shaolong Cao[1], Ruonan Li[1], Yuxin Tang[6], Tom Lesluyes[7], Scott Kopetz[8], Jaffer Ajani[8], Pavlos Msaouel[9], Sumit K Subudhi[9], Ana Aparicio[9], Padmanee Sharma[9], John Paul Shen[8], Anil K. Sood[10], Maxime Tarabichi[11], Jennifer R. Wang[12], Marek Kimmel[2], Peter Van Loo[7,13,14], Hongtu Zhu[15], 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]*Department of Statistics, Korea University, Seoul, South Korea, 02841*
[6]*Department of Computer Science, Rice University, Houston, TX 77005*
[7]*Cancer Genomics Group, The Francis Crick Institute, London, UK*
[8]*Department of Gastrointestinal Medical Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030*
[9]*Department of Genitourinary Medical Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030*
[10]*Department of Gynecologic Oncology and Reproductive Medicine, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030*
[11]*Institut de Recherche Interdisciplinaire en Biologie Humaine et moléculaire, Université Libre de Bruxelles, Bruxelles, Belgium, B1070*
[12]*Department of Head and Neck Surgery, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030*
[13]*Department of Genetics, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030*
[14]*Department of Genomic Medicine, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030*
[15]*Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599*

[#]*authors contributed equally*

[*]*Correspondence: WWang7@mdanderson.org*

Abstract

Intra-tumor heterogeneity is an important driver of tumor evolution and therapy response. Advances in precision cancer treatment will require understanding of mutation clonality and subclonal architecture. Currently the slow computational speed of subclonal reconstruction hinders large cohort studies. To overcome this bottleneck, we developed Clonal structure identification through Pairwise Penalization, or CliPP, which clusters subclonal mutations using a regularized likelihood model. CliPP reliably processed whole-genome and whole-exome sequencing data from over 12,000 tumor samples within 24 hours, thus enabling large-scale downstream association analyses between subclonal structures and clinical outcomes. Through a pan-cancer investigation of 7,827 tumors from 32 cancer types, we found that high subclonal mutational load (sML), a measure of latency time in tumor evolution, was significantly associated with better patient outcomes in 16 cancer types with low to moderate tumor mutation burden (TMB). In a cohort of prostate cancer patients participating in an immunotherapy clinical trial, high sML was indicative of favorable response to immune checkpoint blockade. This comprehensive study using CliPP underscores sML as a key feature of cancer. sML may be essential for linking mutation dynamics with immunotherapy response in the large population of non-high TMB cancers.

Introduction

Mutations in tumor cells accumulate over time, from the mutations present in a single cancer-initiating cell to the subsequent myriad of mutations present across a mature tumor[1,2]. As tumors keep accumulating mutations, they are comprised of heterogeneous subpopulations of cells, i.e., (sub)clones, with additional and distinct mutations from those that were inherited from the ancestral cancer cell[3–5] (**Fig. 1a**). Over the past decade, understanding of the temporal and spatial hallmarks of mutation events across tumor clones has substantially improved[6–10]. Many pan-cancer investigations[11–15] have contributed to these findings, which in turn present opportunities for the development of new cancer therapies.

Early pan-cancer studies identified a strong association of tumor total mutation load with patient outcomes[14,16], and later with response to immunotherapy[17,18]. In June 2020, the US Food and Drug Administration (FDA) approved the first histology-agnostic programmed cell death (PD-1) therapy for patients with solid tumors exhibiting a tumor mutation burden (TMB, calculated as the number of mutations per million bases) of at least 10 (high-TMB)[19]. Among these high-TMB cancers treated with immune checkpoint inhibitors (CPIs) across seven cancer types, clonal (but not subclonal) TMB was found to be the strongest predictor of CPI response[20]. These studies have paved a promising path to assess simultaneously how selection shapes mutation patterns as well as patient survival[14].

Recent pan-cancer studies have further demonstrated distinct evolutionary paths and immune environments between cancers with high versus low mutational load. They find that selection is much stronger in low mutational load cancers[21], while only high mutational load cancers show a tendency towards focal loss of heterozygosity of HLA-I as a mechanism of immune evasion[22]. As most cancers present less than 10 mutations per million bases, including 88% (6,766/7,708) of The Cancer Genome Atlas (TCGA) tumors, it is paramount to evaluate the potential clinical impact of the different subclonal structures across cancers with low, moderate, and high TMB. The most recent multi-year international consortium effort to perform a pan-cancer subclonal reconstruction used 2,658 tumor samples, most of which lacked annotation of clinical outcomes[12]. Previous work to address this need with available clinical data was conducted on 1,165 tumors almost a decade ago[13], or since then from a combined total of ~800 tumors from several individual cancer-type studies on leukemia[23], prostate[24], lung[25], kidney[26], and most recently neuroblastoma[27]. A major hurdle in realizing such efforts has been the computational burden and poor scalability of existing methods for the reconstruction of subclonal structures using sequencing data generated from heterogeneous bulk tumor samples[12].

In this study, we overcome these challenges by: 1) developing a statistical method called Clonal structure identification through Pairwise Penalization, or CliPP, in order to improve the computational speed per sample by 100-1,000 fold, as compared to existing methods[28–30], while maintaining equal accuracy; 2) generating pan-cancer subclonal structures of pre-treatment tumors with available clinical follow-up data of at least 5 years. We have evaluated the clinical impact of subclonal driver mutations and subclonal mutational load from 7,708 patient samples in TCGA, 98 esophageal adenocarcinoma samples from the International Cancer Genome Consortium Pan-cancer Analysis of Whole Genomes (ICGC-PCAWG) study, and 21 metastatic castration resistant prostate cancer (mCRPC) samples from an immunotherapy clinical trial. We find that subclonal mutational load, a surrogate measure for the latency time between the most recent common ancestor and the latest subclonal event[11], is an essential feature of cancer and may serve as a robust prognostic and therapeutic biomarker, particularly in low- to moderate-TMB cancers.

Results

Overview of subclonal reconstruction by CliPP

For a given tumor sample, the variant allele frequency (VAF) spectrum (**Fig. 1b**) exhibits a mixture of multiple peaks corresponding to varying cellular prevalence (CP) of single-nucleotide variants (SNVs). Subclonal reconstruction refers to a procedure to identify cancer cell clones by clustering variant read counts, adjusting for allele copies and tumor purity, to identify groups of variants with similar CPs. Available tools employ a Bayesian modeling framework[29–32], which comes with a very high computational cost, particularly for running WGS data, e.g., it takes 1-2 days to analyze one sample. This challenge was recently alleviated by a variational Bayes[28,33] approach, yet it is still not fast enough to process tens of thousands of tumor samples. The accuracy of subclonal reconstruction is best achieved by a consensus approach across callers with distinct modeling strategies to account for different errors while sharing the true signals[12]. We therefore propose an orthogonal approach to existing methods called Clonal structure identification through Pairwise Penalization (CliPP), which is rooted in the domain of regularized regression in machine learning. Imposing a pairwise penalty in the form of smoothly clipped absolute deviation (SCAD)[34] to the parameter (e.g., CP) estimation per data point (e.g., mutation) is, in general terms, a modeling advancement over the fused LASSO regression[35] and a specialized extension of the recently proposed CARDS algorithm[36], which searches for homogeneity in regression parameters. Consequently, in this novel application of the penalized likelihood statistical framework to subclonal reconstruction, our resulting sparse and homogeneous pattern of CPs across mutations

corresponds to a clustering procedure (**Fig. 1c**, **Methods**). **Fig. 1d** illustrates how homogeneity increases as the penalty term (λ) increases, resulting in fewer clusters. CliPP produces the clustered mutations along with their associated homogenized CP values as the output. CliPP offers exceptional computational efficiency owing to three key factors: 1) the subclonal reconstruction problem is reformulated as a constrained problem that can be efficiently solved by existing optimization techniques; 2) Markov chain Monte Carlo methods are generally time-consuming, while CliPP circumvents the need for such procedures; and 3) CliPP utilizes parallel computing of matrix multiplications, which further expedites the process (see **Methods**).

Using three simulation datasets (total n=5,515) covering a wide range of values for key features that are known to influence accuracy[12] including tumor purity, percent of genome with copy number alterations (CNA), sequencing read depth, and the number of mutation clusters, CliPP performs similarly or slightly better than PhyloWGS, a popular method in subclonal reconstruction (**Methods**, **Figs. 2a,b**, **Extended Data Fig. 2a-d, Supplementary Information Section 1**). Using WGS data from 1,548 patient tumor samples in the PCAWG study, we further evaluated results from 10 existing methods[12,28] (see **Methods**). When compared with these 10 individual methods and their consensus calls in terms of the estimated fraction of clonal mutations, CliPP showed the second highest concordance to the consensus (concordance correlation coefficient = 0.96, **Fig. 2c, Extended Data Fig. 2e**), confirming its high accuracy.

We next compare the computational speed of CliPP with those of PhyloWGS, PhylogicNDT[30], PyClone[32] as well as PyClone-VI[28], a recent fast version of PyClone, on TCGA and PCAWG samples (see **Methods**). Overall, CliPP finishes analyzing 2,778 PCAWG samples within 16 hours, and 9,654 TCGA samples in one hour, presenting at least a 2,000-fold improvement in speed as compared to PyClone and PhyloWGS (**Figs. 2d, e**). While closer to CliPP than PyClone and PhyloWGS, PyClone-VI's speed remains largely lower than CliPP in samples with more than 5,000 mutations in WGS data, with CliPP showing a median of 4-fold speedup (median absolute deviance [MAD]=2.7 fold). WES data covers only 1% of the genome, and as such, most samples (75%) in TCGA present with less than 200 mutations. It is in samples with lower numbers of SNVs that CliPP has the largest speed advantage **(Fig. 2e)**. Using the processing speeds in PCAWG on samples with a similar distribution of mutation numbers, we project in TCGA the speed advantage of CliPP versus PyClone-VI and PhylogicNDT to be ~660 times (**Fig. 2f, Methods**) and ~1,200 times (by extension using **Fig. 2e, Supplementary Information Section 2**), respectively.

In summary, the comparable accuracy and substantial speedup achieved by CliPP has allowed us to process the large pan-cancer patient cohorts of WES data in TCGA, which has otherwise been prohibited by the very long computing times of other methods.

The subclonal landscape of 7,708 tumors

Here, we provide an updated portrait of the extent of genetic intra-tumor heterogeneity (ITH) across and within tumor types with a larger sample size in TCGA, as compared to the previous two pan-cancer studies with ~1,200 tumors[13] and ~2,700 tumors[12]. One bottleneck issue with using TCGA data is that WES data surveys only the coding portion of the cancer genome **(Fig. 3a)** and whether WES can substitute for WGS to provide useful information for cancer evolution remains unclear for researchers designing new studies. With 465 samples from 21 cancer types analyzed by CliPP on both WES and WGS data (profiled by TCGA and PCAWG, **Extended Data Fig. 1a**), we find CliPP-based subclonal reconstruction results are highly consistent between the two platforms, with a median inter-rater agreement (B-statistic, see **Methods**) of 0.81, indicating close to perfect agreement (**Fig. 3b**). More mutations changed identity from being clonal in WGS to subclonal in WES, as compared to the opposite direction (McNemar's test *P*-values < 0.001, **Extended Data Fig. 3a**). Mutations shifting from WGS-clonal to WES-subclonal are expected, since the higher read coverage provided by WES is more powered to detect deviations from clonality (median ratio of read depth = 1.5 across 49,597 overlapping mutations, **Extended Data Figs. 3b,c**, **Supplementary Information Section 3**). In conclusion, WES provides adequate data for subclonal reconstruction; therefore it is appropriate to evaluate subclonality in TCGA, which has a wealth of clinical annotations as compared to PCAWG.

Across 7,708 tumors from 32 cancer types in TCGA, we observe that 93% (n=7,188) have between 1 and 4 clusters. As TMB increases, samples tend to have fewer mutation clusters, indicating less potential for a diverse subclonal structure (**Fig. 3c**). We quantify ITH via subclonal mutational load (sML), the proportion of mutations that are subclonal, merging mutations from multiple subclonal clusters (see **Methods**). Consistent with the reported PCAWG results[12], in TCGA across cancer types, the median sML is moderately negatively correlated with the median TMB (Pearson correlation = -0.55, *P*-value = 0.001, **Fig. 3c**), but not correlated with the median number of reads per chromosome copy (nrpcc[37]) (Pearson correlation = 0.28, *P*-value = 0.12**)**. However, across samples in each cancer type, sML is not correlated with TMB (median Pearson correlation = -0.14) but shows weak to moderate correlations with nrpcc (median Pearson correlation = 0.29, **Supplementary Information**). Thyroid and thymus cancers, which have low

median TMB, exhibit the highest levels of sML. Melanoma, lung squamous, bladder and B-cell lymphoma, which have high TMB, exhibit the lowest levels of sML.

The landscape of subclonal driver mutations across 16 cancer types

Next, we update subclonal driver annotations in TCGA and further investigate their association with clinical outcomes, a question that has only been evaluated anecdotally[38] thus far.

Driver mutations are genetic alterations that contribute to (or "drive") the development and progression of cancer[39]. Much effort has been made to identify and catalog these mutations from the background "passenger" mutations by identifying signals of positive selection in these mutations[40]. Due to their role in tumor initiation and the fitness advantage they confer, most driver mutations are clonal. However, these same driver mutations can occur within subclones and confer a selective advantage later in tumor evolution[12]**.** Using a robustly annotated set of 8,708 driver mutations in 574 genes across 2,568 samples from 16 cancer types[41], we find that 1,773 (20.4%) driver mutations were subclonal (**Fig. 4a**, **Supplementary Table 1**). Across cancer types, prostate cancer and acute myeloid leukemia (AML) exhibited the highest proportion of subclonal driver mutations at 47% (77/164 and 21/45), while ovarian and colorectal cancers exhibited the lowest proportion at 12% (37/305 and 40/323). Among the 78 driver genes with more than 20 mutations observed (**Methods, Fig. 4a, Supplementary Information Section 4**), mutations in *SMARCA4* (95% clonal), *CDKN2A* (94% clonal), and *TP53* (92% clonal) were most often clonally present, while mutations in *SMAD4* (47% subclonal), *CHD3* (39% subclonal), and *FGFR2* (38% subclonal) were frequently subclonally present. Additionally, we observe cancer type-specific variations in driver mutation subclonality. For example, subclonal *PIK3CA* mutations are observed at frequencies between 10-40% across cancers, with the exception of lung squamous cell carcinoma, where it is almost exclusively clonal (21/22 mutations). Similar to a previous study[42], we also found subclonal *IDH1* mutations in 25 out of 99 *IDH*-mutated lower grade glioma (LGG) samples, although no difference in overall survival (OS) was observed between LGG patients with clonal or subclonal *IDH1* mutations. Overall, driver mutations were reported in 10% of subclonal mutation clusters in TCGA, again in line with the 11% observed in PCAWG[12] (**Fig. 4b**).

Studies of specific single cancer types have found mixed evidence for survival being associated with clonal and subclonal drivers[43,44], yet this question has not been addressed across cancer types in a comprehensive manner. Out of all 15 driver gene/cancer-type combinations (**Supplementary Table 2**) with more than 10 samples bearing subclonal mutations, we find 4 combinations (*PTEN* in UCEC, *KRAS* in UCEC, *BRAF* in THCA, *FAT1* in HNSC) to present

significant association with prognosis, all of which are consistent in direction, i.e., we observe poorer progression-free (PFS) and/or OS in patients presenting subclonal driver mutations (**Fig. 4c**). These effects, with the exception of *KRAS*, remain statistically significant in a Cox model after adjusting for age and sex (**Fig. 4d**). Other than *BRAF*-THCA being a re-discovery from a recent study[45], the other three pairs are newly identified associations. When limited to the most frequently occurring mutation sites, such as *BRAF* V600E, *PTEN* mutations to R130 and R233, and *KRAS* mutations to G12, the association with PFS or OS remains significant (**Extended Data Fig. 4a-d**). In summary, our pan-cancer large-scale subclonal annotation using CliPP has provided new insights into the potential impact of the timing of driver mutations on patient outcomes.

Subclonal mutation load (sML) is predictive of cancer prognosis

Following TMB as a widely accepted clinical marker[19] and the recent endeavors in further delineating the timing of mutations[11], we hypothesize that subclonal mutation load (sML), a surrogate measure for the latent time between the most recent common ancestor and the latest subclonal event[11], is a biologically and clinically relevant feature for cancer.

Across TCGA, we find that sML is significantly associated with clinical or pathological features/subtypes, with phenotypes associated with poorer prognosis consistently exhibiting lower sML, such as in triple-negative breast cancer[46,47], high Gleason score in prostate cancer[48], smoking status or low PD-L1 expression in lung cancer[49], and HPV status in head and neck cancer[50] (**Extended Data Fig. 5a, Supplementary Information Section 5.1**). In contrast, we observe no significant associations between sML and age, sex, TMB, clinical stage, or technical factors such as the size of the tumor biopsy or nrpcc (**Supplementary Information Section 5.2**). Together, our observations suggest a unique role of sML as a clinically relevant feature, independent of TMB, stage, or other possible confounders.

Using clinical outcomes (OS and PFI) in TCGA across 28 cancer types and their subtypes (see **Methods, Supplementary Table 3**), we find that high sML is associated with better OS/PFI at a pan-cancer level (**Extended Data Fig. 5b**) and within 16 (sub)types, with an additional 3 cancer (sub)types demonstrating the opposite trend (**Methods**, **Fig. 5a**, **Extended Data Fig. 5c, Supplementary Tables 4, 5**). The three cancer (sub)types where high sML is instead associated with poorer survival are ovarian, lung squamous-smokers and melanoma, which present the highest levels of genome instability or the highest median TMB among all cancers (**Fig 5a:** top panel, **Extended Data Fig. 5d**). These effects remain unchanged after adjusting for potential technical confounders (**Supplementary Information Section 5.3**). The significant predictive

effect of sML persisted for 9 cancer (sub)types as we moved from dichotomizing sML to using it as the continuous variable in a Cox model (**Methods**, **Fig. 5b**, **Supplementary Table 6**). Among these cancer types, AML presents an opposite trend for the continuous sML, which may be explained by a different biological underpinning for hematological cancers. As expected, adding TMB as a covariate did not replace the impact of sML (**Methods**, **Fig. 5c**). Using variable selection, we find significant main effects of sML in 7 cancer (sub)types (**Fig. 5c**) and a significant interaction term of sML and TMB in one cancer type, thyroid carcinoma (**Methods**, **Fig. 5d, Supplementary Table 7**). By stratifying the thyroid cancer patients using both TMB and sML, we find the prognostic effect of sML is prominent within samples with higher TMBs (**Fig. 5e**). This is consistent with the idea that low TMB in thyroid cancers is associated with indolent behavior, thus separation of outcomes is more identifiable in higher TMB tumors, where outcomes are more variable. Replacing sML-based patient classification with TMB-based classification did not provide the same statistical significance (**Supplementary Information Section 5.3**), supporting the unique contribution of sML.

We also evaluate sML patient stratification with the addition of Shannon Index (SI)[51], a diversity measure that increases when there are more mutation clusters and when mutations are more equally divided in each cluster. Given the collinear relationship between sML and SI (**Fig. 5f**, **Extended Data Fig. 6a**), SI only contributes significant additional information in high sML tumors, which represent the minority of TCGA (**Fig3c**). For samples with a high subclonal mutation load, a high SI indicates that subclonal mutations belong to multiple mutation clusters with distinct CPs, whereas a low SI indicates all mutations belong to the same or a small number of clusters. As these behaviors represent different evolutionary trajectories, we indeed find the addition of SI partitions high sML tumors into further distinct survival groups for 11 cancer types (**Extended Data Figs. 6b-c**, **Supplementary Information Section 6**). In summary, with low- and moderate-TMB cancers, SI can sometimes be used to complement sML, but overall sML remains an easily interpretable and representative feature that encodes cancer evolution.

Validating sML in low to moderate-TMB cancers using independent patient cohorts

Finally, to further validate our clinical observations above, we obtained sequencing data from patient cohorts outside of TCGA, hence not limited to pitfalls in the clinical follow-up data from TCGA[52]. These data were generated from patient samples that present a range of TMB not meeting the criterion of >10 (high-TMB) for immunotherapy (**Fig. 6a**).

We applied CliPP to calculate sML in 98 whole-genome sequenced esophageal adenocarcinoma samples from the International Cancer Genome Consortium PCAWG study[53] (**Fig. 6b**). Both

TCGA and PCAWG esophageal adenocarcinoma datasets share similar fractions of high versus low sML samples (**Extended Data Figs. 7a, b**), with the latter achieving a higher level of statistical significance in distinguishing patients' OS outcomes (log-rank test *P*-value = 0.004, **Fig. 6b**), likely due to the larger sample size. Both the continuous and dichotomized sML yielded a significant hazard ratio, even after adjusting for TMB (**Fig. 6c, Supplementary Tables 4-7**), and were not associated with nrpcc (**Extended Data Figs. 7c, d**). We have therefore validated the association of high sML with better OS in esophageal cancer in an independent, non-WES dataset.

We obtained WES data from 21 men with mCRPC participating in a clinical trial of ipilimumab (NCT02113657, **Fig. 6d**, **Methods**). mCRPC is a tumor type with low TMB in which benefit from immune checkpoint blockade is restricted to patient subgroups from whom identifying biomarkers remain to be defined[54]. By comparing a subset of patients previously categorized as favorable and unfavorable outcomes (n=7 vs. 8), we observed a trend in sML to be higher in the favorable group, although not statistically significant (Wilcoxon test *P*-value=0.19, **Fig. 6e**). The distribution of sML is not correlated with TMB (Pearson correlation=-0.19, **Fig. 6f**). Using radiographic/clinical PFS (rcPFS) as an endpoint for all patients, including the previously indeterminate group (n=6), we find that a natural cutoff of 0.5 on sML can separate patients into distinct rcPFS outcome groups (n=9 vs. 12, log rank *P*-value=0.005, **Fig. 6g**), who continued to demonstrate distinct survival outcomes up to 8 years of follow-up (log rank *P*-value=0.03, **Fig. 6h**). These clinical associations are independent of nrpcc (**Extended Data Figs 7e**) but could be confounded with tissue sites of sample origin, e.g., prostate versus lymph node or liver (**Extended Data Figs 7f**). We followed up with a prostate sample-specific analysis and still observed significant associations of sML with rsPFS and OS (n=6 vs. 5, log rank *P*-values=0.009 and 0.01, respectively, **Extended Data Figs 7g, h**, **Methods**). The high versus low sML tumor samples present significantly different levels of CD8 T-cell densities, as measured by immunohistochemistry (Wilcoxon test *P*-value=0.008, **Fig. 6i**). CIBERSORTx[55] based immune cell subtype profiling further suggests differences in the myeloid cell populations (**Extended Data Fig. 7i**). In summary, these data support our hypothesis that beyond its prognostic value, sML may be a useful marker to identify patients that benefit from immunotherapy.

Discussion

Accurate characterization of the clonality of single-nucleotide variants across large-scale patient cohorts (7,827 tumors) using CliPP has generated important clinical and biological insights. Among the majority of cancers in TCGA (16 out of 19 cancer types, **Fig. 6j**), unfavorable clinical

outcomes were associated with a low subclonal mutational load (sML), found in tumors that have presumably undergone recent clonal sweeps[5,7,14], i.e., positive selection. Favorable clinical outcomes were associated with a high sML, found in tumors that may be slow-growing and have a longer lag time, i.e., long latency, since the last clonal sweep event. A recent report noted that tumors with a high number of clonal mutations should have a higher likelihood for neoantigen response[20]. This is consistent with our results in the three remaining TCGA cancer types, i.e., melanoma, lung squamous cell carcinoma (smoker) and ovarian cancer. With this cancer group, who share the highest TMB or the highest genome instability (as measured by percent copy number aberrations), we indeed found that a low sML, corresponding to a high number of clonal mutations, is associated with favorable clinical outcomes. Therefore, our comprehensive evaluation provides new evidence supporting distinct evolutionary paths between cancers with high versus low mutational load, and how their divergent evolutionary dynamics may impact clinical outcomes.

Because of the established links between TMB, neoantigen response, and immunotherapy, we used an mCRPC clinical trial study to directly evaluate the relationship between sML and immune activities in tumors. We found signal in sML to identify mCRPC patients with improved survival and progression times after receiving ipilimumab. Therefore, as compared to patients with high-TMB tumors that are expected to benefit from immunotherapy[17–20], our study points to a new target population: patients with low to moderate-TMB tumors (~90% of all cancers) that may also benefit from immunotherapy, when they present a high subclonal mutational load.

Finally, using CliPP data, we observe a pan-cancer adverse effect of subclonal driver mutations on clinical outcomes, suggesting a potential of subclonal annotation to inform the development of new therapeutic strategies for the well-known and targetable driver gene events, such as *KRAS*, *PTEN*, and *BRAF*. When driver mutations are found in subclones, the main clone likely has other driver alterations; these tumors are thus less addicted to any single oncogenic signal and likely more resistant to treatment. Clinical testing that does not ascertain the clonality of driver mutations may fail to adequately capture the prognostic significance of some driver mutations. This shortcoming in the clinic suggests that future sequencing platforms will need to become more sophisticated to address subclonal tumor mutations to adequately address patient needs.

In conclusion, a high-level summary measure of cancer evolution such as the clonal/subclonal status of single-nucleotide variants, when computed across a large number of patient samples, sheds new light on the relationship between genetic intra-tumor heterogeneity and therapeutic resistance. Highly efficient and accurate methods like CliPP represent a new generation of

techniques that bridge the gap between studying tumor biology and developing precision treatment strategies.

Methods

Additional details and results are described in **Supplementary Information**. Here, we summarize the key aspects of the analysis.

The CliPP Model

In the wake of the genomic revolution, several approaches have been developed to call and quantify cancer cell clones through clustering of variant read counts[56]. Many of the existing methods address subclonal reconstruction with mixture models, as implemented via Dirichlet processes, which are computationally costly, especially when the underlying subclonal structure is complex. CliPP is a novel statistical framework that addresses this task from the field of penalized regression, drastically reducing the computational burden while maintaining high accuracy. It ensures homogeneity by introducing a pairwise penalty on parameter estimation without the need to pre-order coefficients, an advancement over preceding methods such as fused LASSO and CARDS algorithm. The full list of mathematical notations is provided in **Supplementary Information Section 7.** First, we assume that each sample is composed of two populations of cells with respect to the i-th SNV: one population containing normal cells or cancer cells without the SNV, and the other population of cancer cells harboring this SNV. We adopt the infinite sites assumption[57] to assume that there is a single origin for each SNV. Within cancer cells, SNV i may occur in one or a few cell populations (**Fig 1a**), therefore we introduce β_i to denote the cancer cell fraction (CCF) of SNV i, i.e., the fraction of cancer cells that carry SNV i. A clonal SNV exhibits a CCF of 1.0, indicating 100% presence in tumor cells. We further define the cellular prevalence (CP) of SNV i as ϕ_i, so that $\phi_i = \rho\beta_i$, where ρ denotes tumor purity, i.e., the proportion of cancer cells among all cells. Given S SNVs, we assume that there exist K groups of SNVs presenting distinct CCFs, where K is significantly smaller than S. We expect SNVs from a common cancer cell population or subclone to have identical CP (or CCF), and correspondingly share the same parameter ϕ (or β). We follow an established model for observed variant read counts[12], and assume that the observed number of variant reads r_i at each SNV follows a binomial distribution $Binomial(n_i, \theta_i)$, where the total number of reads n_i follow a Poisson distribution $Possion(D)$. We further express the expected proportion of the variant allele θ_i as:

$$\theta_i = f(\phi_i) = \frac{\phi_i b_i^V}{(1-\rho)c_i^N + \rho c_i^T},\ i = 1, \ldots, S. \qquad (1)$$

Hence the observed log-likelihood function, after canceling some constant, follows:

$$l(\boldsymbol{\phi}) = \sum_{i=1}^{S}\left[r_i \log f(\phi_i) + (n_i - r_i)\log\left(1 - f(\phi_i)\right)\right] \qquad (2)$$

We use b_i^V to denote the SNV-specific copy number at the i-th SNV. This measurement is distinct from allele-specific copy number when the SNV does not occur in all tumor cells. We then use c_i^N and c_i^T to denote the total copy number for normal and tumor cells, respectively.

Input data. CliPP requires information of c_i^N, c_i^T, m_i^T, r_i, n_i, and ρ as its input. Here m_i^T denotes the tumor cell major allele count at the i-th SNV. Parameters c_i^T, m_i^T, and ρ can be obtained using CNA-based deconvolution methods such as ASCAT[58] and ABSOLUTE[59], whereas c_i^N is typically set as 2. Parameter b_i^V is not directly observed or estimatable using existing CNA software tools. Following a current convention that proved to be effective[56] we assume b_i^V as:

$$b_i^V = \min\left(\left[\frac{r_i}{n_i}\frac{1}{\rho}\left(\rho c_i^T + c_i^N(1-\rho)\right)\right]_0, m_i^T\right),\ i = 1, \ldots, S, \qquad (3)$$

where $[\cdot]_0$ rounds x to the nearest positive integer. Here m_i^V represents the copy number of major allele covering SNV i, which enforces the logical constraint that even if an SNV happened before any copy number event, the number of alleles carrying said SNV cannot exceed the total number of alleles.

Output data. CliPP provides the inferred subclonal structure, including the total number of clusters, the total number of SNVs in each cluster, the estimated CP for each cluster, and the mutation assignment, i.e., cluster ID for each mutation. This output can then serve as the basis for various downstream analysis, including the inference of phylogenetic trees[56].

Parameter estimation using a regularized likelihood

The CPs, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_S)$ may be estimated by maximizing the corresponding log-likelihood for one SNV at a time. However, our primary goal is to identify a homogeneous and sparse structure, i.e., clusters, of the CPs across all SNVs. Penalized estimation is a canonical tool to achieve homogeneity detection and parameter estimation simultaneously[60]. Therefore, we introduce a pairwise penalty to seek such homogeneity in $\boldsymbol{\phi}$. To facilitate computation, we employ a normal approximation of the binomial random variable, i.e., $\sqrt{n_i}\left(\frac{r_i}{n_i} - \theta_i\right) \sim N\left(0,\ \theta_i(1-\theta_i)\right)$, which lead to the following approximated log-likelihood

$$\hat{\ell}(\boldsymbol{\phi}) = \sum_{i=1}^{S}\log\frac{1}{\sqrt{2\pi n_i f(\phi_i)(1-f(\phi_i))}} - \frac{(n_i f(\phi_i) - \mathrm{r_i})^2}{2n_i f(\phi_i)\left(1-f(\phi_i)\right)}. \qquad (4)$$

To promote homogeneity in $\boldsymbol{\phi}$, we equip the negative log-likelihood $-\hat{\ell}(\boldsymbol{\phi})$ with a pairwise shrinkage penalty, thereby solving the following optimization problem

$$\min_{\boldsymbol{\phi}} -\hat{\ell}(\boldsymbol{\phi}) + \sum_{1\le i<j\le S} p_\lambda(|\phi_i - \phi_j|), \tag{5}$$

where $\lambda > 0$ is a tuning parameter that controls the degree of the penalization, and $p_\lambda(\cdot)$ denotes a sparsity-inducing penalty function to identify the homogeneity structure in $\boldsymbol{\phi}$, such as LASSO[61], SCAD[34], and MCP[62] to name a few. Here, we focus on the SCAD penalty which is defined as

$$p(x|\lambda,\gamma) = \begin{cases} \lambda|x|, & if\ |x| \le \lambda \\ \dfrac{2\gamma\lambda|x| - x^2 - \lambda^2}{2(\gamma-1)}, & if\ \lambda < |x| < \gamma\lambda \\ \dfrac{\lambda^2(\gamma+1)}{2}, & if\ |x| \ge \gamma\lambda \end{cases} \tag{6}$$

where γ is a hyper-parameter that controls the concavity of the SCAD penalty. Building upon the work of Fan and Li[34], we set γ to 3.7. These concave penalties offer sparsity similar to the LASSO penalty, allowing them to automatically yield sparse estimates. It is proven that the SCAD-penalized estimators possess the oracle property in identifying sparsity structure of the data[34]. One can choose another penalty instead of SCAD, such as LASSO, and the formulation is straightforward.

Optimizing the objective function in Equation (5) is nontrivial because the target variable $\boldsymbol{\phi}$ is bounded, *i.e.*, $\phi_i = \rho\beta_i \in [0,1]$, and the SCAD penalty is not convex. We employ a re-parametrization of ϕ_i, $\omega_i = \log\frac{\phi_i}{1-\phi_i}$, $i = 1, \dots, S$, to remove the box constraint, which yields

$$\theta_i = g(\omega_i) = \frac{b_i^V e^{\omega_i}}{(1+e^{\omega_i})(2 - 2\rho + c_i^T\rho)},\ i = 1, \dots, S. \tag{7}$$

Note that ω's are monotonic with respect to ϕ's; $\omega_i = \omega_j$ implies $\phi_i = \phi_j$ and vice versa, thus the homogeneity pursuit of $\boldsymbol{\phi}$ can be achieved by the homogeneity pursuit of ω. We perform a transformation on the loss function from Equation (5), with an updated penalty function that identifies the homogeneity structure in ω. Given that our goal is to pursue $\phi_i = \phi_j$, we could alternatively enforce $\omega_i = \omega_j$ when $|\omega_i - \omega_j| < \epsilon$ for some error term ϵ. Therefore, we reformulate the loss function of $\boldsymbol{\omega}$ as follows:

$$\min_{\boldsymbol{\omega}} -\hat{\ell}(\boldsymbol{\omega}) + \sum_{1\le i<j\le S} p_\lambda(|\omega_i - \omega_j|). \tag{8}$$

This is a non-convex optimization problem with respect to $\boldsymbol{\omega}$, and its computation is non-trivial due to the complex log-likelihood. To address this challenge, we employ the alternating direction

method of multipliers (ADMM)[63], where we approximate the negative log-likelihood at each iteration by a quadratic function, thereby simplifying the computation. Note that the SCAD penalty possesses the unbiasedness property, ensuring that it does not shrink large estimated parameters throughout iterations. This property is particularly important in ADMM algorithms, as biases during the iterations could significantly impact the search for subgroups. Further mathematical derivations and the computational construction of ADMM can be found in the **Supplementary Information Sections 7.1- 7.3**.

Finally, it is crucial for the regularized likelihood-based approach to select a proper λ, in order to balance between over- and under-fitting. Here, the choice of λ determines the final number of clusters, with higher values yielding fewer clusters. We currently implement an *ad hoc* selection approach, which focuses on the interpretability of the outcome by ensuring that clonal mutations have an estimated CCF around 1. Our automated λ selection pipeline operates as follows: for each sample, CliPP will run on the data with 11 different $\lambda's$ spanning from 0 to 0.25, specifically: 0.01, 0.03, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25. For each sample, we compute a score $A = \frac{\max(CP) - purity}{purity}$. If there are one or more results that satisfy $A < 0.05$, we choose the largest λ associated with those results. If all scores A are greater than 0.01, we choose the λ associated with the smallest A. Note with applying the CliPP software to any new datasets, the choice of λ can be decided by users.

CliPP has an automatic post-processing pipeline for better biological interpretability. It also allows for down-sampling when processing samples with huge amounts of SNVs. Please find the details in **Supplementary Information Section 7.4.**

Parallel computing

CliPP utilizes the Eigen (v3.449)[64] C++ linear algebra library, enabling efficient multi-threading matrix operations. This approach addresses substantial computational demands highlighted in the **Supplementary Information Section 7.3**, removing a key bottleneck. Moreover, by integrating the OpenMP library, CliPP parallelizes multiple runs, optimizing resource utilization across modern multi-core systems while exploring different tuning parameters denoted by λ.

Benchmarking CliPP performance in simulated datasets

We assess the ability of CliPP to correctly reconstruct the subclonal organization of a tumor on three simulated datasets. We generated an in-house simulation dataset CliPPSim4k to thoroughly benchmark the performance in samples with fewer copy number aberrations (CNAs) and higher read depth to cover the scenario of both whole-genome and whole-exome sequencing (WGS,

WES) data (**Extended Data Fig. 2a**). We further included PhylogicSim500, a simulation dataset containing 500 samples where the copy number profiles were sampled from the PCAWG WGS data, with other parameters independently sampled from fixed distributions (**Extended Data Fig. 2b**). We also included the SimClone1000 dataset, a dataset containing 965 samples that were simulated based on a grid design to cover the spectrum of scenarios encountered in the PCAWG WGS data (**Extended Data Fig. 2c**). Details for both datasets can be found in Dentro et al[12]. CliPP was run on all datasets using the default setting, and PhyloWGS v1.0-rc2 was run on CliPPSim4k using the recommended setting. We used available PhyloWGS results on the other two simulation datasets[12]. We evaluated the accuracy of each method using three metrics that measure bias in estimated number of clusters (rdNC), fraction of clonal mutations (rdCF), and CP across all variants (RMSE), respectively, and a total error score that averages over the three metrics to represent an overall performance. Please refer to **Supplementary Information Section 1** for further details.

Subclonal reconstruction in PCAWG

The Pan-Cancer Analysis of Whole Genomes (PCAWG) study[12,53] dataset comprises whole-genome sequencing (WGS) data obtained from a cohort of tumor samples spanning 38 cancer types. PCAWG-related data were downloaded from the ICGC data portal, using Bionimbus Protected Data Cloud (PDC).

The PCAWG dataset contains WGS data obtained from 2,778 tumor samples (2,658 distinct donors), including 2,605 primary tumors and 173 metastases or recurrences (minimum average coverage at 30x in the tumor)[53]. As input to CliPP, we use the consensus mutation calls derived from 4 distinct callers[53] and the consensus CNA calls from 6 distinct callers[12]. In adherence to established best practices[37], we employ the metric of reads per tumor chromosomal copy (nrpcc) to filter out samples with insufficient read coverage for subclonal reconstruction. We calculate $nrpcc = \frac{\rho}{\psi} d$, where d is the depth of sequencing, ρ is the purity and ψ is the average tumor sample ploidy, i.e., $\psi = \rho\psi_T + (1-\rho)\psi_N$, where ψ_T is tumor cell ploidy and $\psi_N = 2$ is normal cell ploidy. We applied a stringent threshold of $nrpcc \geq 10$ to obtain a high-confidence cohort of 1,993 samples, following the PCAWG study[12]. For 149 samples with >50,000 SNVs, we deployed the down-sampling strategy described in **Supplementary Information Section 7.4**.

To benchmark the accuracy of CliPP in real data, we used subclonal reconstruction results from 9 individual methods (Bayclone[65], Ccube[66], cloneHD[67], CTPsingle[68], DPClust[31], PhylogicNDT[30], PhyloWGS[29], PyClone[32], Sclust[69]) and the consensus calls, which were available for 1,548 samples. We further ran CliPP and PyClone-VI[28] on these samples. The comparison was

performed based on the proportion of subclonal mutations estimated by each method and summarized by calculating the concordance correlation coefficient (CCC) for each method to each other method and to the consensus subclonal mutation proportion.

Subclonal reconstruction in TCGA

The Cancer Genome Atlas (TCGA) contains whole-exome sequencing (WES) data from a cohort of tumor samples spanning 32 TCGA cancer types. The TCGA dataset consists of 9,654 WES samples with consensus mutation calls[70] from 5 variant callers and 2 indel callers with matched copy number segments and tumor purity estimates from ASCAT[58]. The CliPP analysis pipeline for TCGA samples includes additional steps for quality control and filtering (**Extended Data Fig. 1a**). In summary, samples were excluded if they had either too few SNVs (resulting in insufficient statistical power) or too many SNVs (indicating potential artifacts). Samples with low read coverage ($nrpcc < 10$) were excluded. We further implemented filters and additional CliPP runs using its alternative mode for samples with negligible CNA events hence no reliable purity estimates using copy number-based methods. These steps aimed to reduce the potentially large negative impact from inaccurate CNA profiling in WES data. The resulting cohort size in TCGA was $n = 7{,}708$ from 32 cancer types.

Computational speed benchmark

We measured the elapsed real time for each method from initiation to completion on the same machine equipped with Intel(R) Xeon(R) Gold 6,132 CPU @ 2.60GHz, using 28 CPU cores for consistency. In order to compute the speed for PhyloWGS, PyClone, and PhylogicNDT, which are time-consuming to study, we employed a grid-based sampling approach on the PCAWG dataset by selecting five samples across seven predefined SNV grids (100, 500, 1,000, 2,000, 3,000, 4,000 and 5,000 SNVs). In instances of sample scarcity at these exact SNV counts, the nearest equivalent (within a +/- 5% window) was selected. For the computational speed benchmark in TCGA WES data, we employed a strategy to match the distribution of relatively low SNV numbers to already obtained PCAWG data without further running these methods on more samples. Please refer to **Supplementary Information Section 2** for further details.

PhyloWGS v1.0-rc2 was executed using the same settings applied in the simulation study. PyClone v0.13.1 was run with both recommended and default configurations[1][footnote]. PyClone-VI v0.1.1 was applied to all PCAWG samples using recommended and default settings,

[1] https://github.com/Roth-Lab/pyclone

incorporating 40 clusters for fitting (-c 40) and 100 restarts of variational inference (-r 100)[2] [footnote]. PhylogicNDT v1.0[3] [footnote] was executed for 1,000 iterations (-ni 1000), with the command to calculate CCF histograms (--maf_input_type calc_ccf).

Analysis of matched WGS and WES data from the same tumor samples

We identified a set of tumor samples (n=510) that contributed to the generation of both PCAWG WGS and TCGA WES data with neighboring slides through matching study IDs. To ensure a fair comparison of the two sequencing platforms, we removed samples with outlier TMBs, i.e. the 3rd quartile+1.5x interquartile range (IQR) of all samples for each cancer type. The cohort comprised 510 samples with 124,901 SNVs before filtering, and after filtering, it consisted of 465 samples with a total of 49,597 SNVs. We then compared the CliPP outputs on these samples using Bangdiwala's B statistic[71], to quantify the agreement of clone/subclonal mutation assignment from WGS and WES data. This statistic can address zero counts in the contingency table better than the commonly used Kappa statistic, which makes it a better fit for this comparison. Interpretations of thresholds for Bangdiwala's B statistic were obtained from Munoz and Bangdiwala[72]. Bias in disagreement was assessed in the substantial to perfect agreement samples ($\mathrm{B} \geq 0.49$) and the fair to moderate agreement samples ($0.09 \leq \mathrm{B} < 0.49$) using McNemar's test. Please refer to **Supplementary Information Section 3** for further details.

Subclonal mutational load (sML)

We introduce subclonal mutational load (sML) as the percentage of subclonal SNVs out of all SNVs.

$sML = \frac{\#\ of\ subclonal\ SNVs}{\#\ of\ SNVs}$. We calculate tumor mutation burden (TMB) as the total number of SNVs per mega-base (Mb). With WES data, we restrict the TMB and sML calculations to the coding sequence (CDS). We annotate CDS as defined by GENCODE[73] Release 19 (annotated as "gene_type" = "protein_coding" and "feature" = "CDS"). We use 38Mb as the exome size[74].

Driver mutation analysis

Driver mutations were identified and annotated by OncodriveMut[41] for TCGA samples from 16 cancer types. After merging with the samples that passed CliPP filtering criteria, the resulting driver annotation set consisted of 2,568 samples with subclonal estimates for 8,708 SNV driver mutations in 574 unique genes. Two-sided Wilcoxon rank-sum test was used to compare mutation position distribution between clonal and subclonal drivers within genes, with no significant

[2] https://github.com/Roth-Lab/pyclone-vi
[3] https://github.com/broadinstitute/PhylogicNDT

observed bias for any gene. To address sequence similarity issues affecting variant allele frequency (VAF) accuracy, mapping quality scores were computed for the 80 genes with at least 20 driver mutations. This process involved generating perfect 100-base read fastq files for every genome position and aligning them with BWA-mem. Subsequently, mapping quality scores ranging from 0 to 60 were extracted from the resulting BAMs. We then compare the distribution of mapping quality scores between clonal and subclonal mutations for the 9 genes containing annotated mutation sites with a mapping quality score less than 60 using two-sided Wilcoxon rank sum tests. We found evidence supporting significantly lower mapping quality scores in only 2 of these genes (*MLL3* and *NCOR1*, see **Supplementary Information Section 4**). These two genes were subsequently filtered out from our analyses and visualizations. Log-rank tests comparing overall survival and progression-free survival between clonal and subclonal mutations were performed for each gene-cancer pair with 10 or more subclonal driver mutations (15 pairs in total, **Supplementary Table 2**). Cox proportional hazard models were used to confirm associations between subclonal status and survival outcomes, adjusting for age and sex covariates. The survival analysis in *FAT1* mutated HNSC samples focused on HPV-negative samples, excluding a single HPV-positive case with a *FAT1* mutation.

Association of sML with survival outcomes

To assess clinical relevance of sML through its association with clinical outcomes, we characterized the relationship between sML and the patient survival data (overall survival (OS) and progression-free interval (PFI) on all 28 cancer types that present more than 50 samples, see a sample size summary in **Supplementary Table 3**). For well-defined clinical cancer subtypes and features, we further evaluated the effect of sML within these patient subgroups, including hormone receptor status in breast cancer, CMS subtype 2 (most frequently occurring) and MSI status in colorectal cancer, adenocarcinoma and squamous cell carcinoma in esophageal cancer, HPV status in head and neck cancer, *IDH1* mutation status in low-grade glioma, smoking status in lung cancer, Gleason score in prostate cancer, and endometroid status in uterine and endometrial cancer. Please refer to **Supplementary Information Section 5.1** for further details.

We followed the recommendation by Liu et al.[52] to identify appropriate survival measures. We first associated sML with clinical outcomes by splitting samples into high and low sML using a recursive partitioning survival tree model (rpart)[75] for each cancer type, with the maximum tree depth constrained to 1 (sML cutoffs are provided in **Supplementary Table 4**). To ensure well-balanced grouping, we required at least 10 samples and ≥10% representation per sML group (“high” versus “low”) for each cancer type. To analyze survival outcomes, we applied Kaplan-

Meier (KM) analysis, generating KM curves through the survminer R package[76]. These curves illustrate survival probability across time, stratified by high vs. low sML. Patient outcome stratification in cancer type or subgroups with log-rank test *P*-values <= 0.1 was deemed significant or marginally significant.

We furthered our analysis on how sML impacts clinical outcomes along with other patient characteristics, such as age and sex, by utilizing multivariate Cox proportional hazard (PH) models. The PH assumption was deemed appropriate by testing the Schoenfeld residuals for survival data across cancer types. We employed multivariate CoxPH models in three iterations for all 28 cancer types: 1) using sML as a binary variable, high versus low (**Supplementary Table 5**); 2) using sML as a continuous variable (**Supplementary Table 6**); 3) using both sML and TMB as continuous variables, plus a model selection step, e.g., with Akaike information criterion (AIC), to determine whether to include their interaction term (**Supplementary Table 7**). We report cancer (sub)types where the grouping is balanced as described above for high/low sML, and the estimated hazard ratio for sML and/or its interaction with TMB is statistically significantly different from 1 (*P*-value < 0.1), excluding cancer types with wide confidence intervals on the hazard ratio. When both OS and PFI show significant effects, or both cancer type and cancer (sub)type show significant effects of sML, we only include one outcome in the figures, but all results are provided in **Supplementary Tables 5-7**.

Please refer to **Supplementary Information Section 5** for further details on the technical considerations in the clinical relevance of sML, and **Supplementary Information Section 6** for further details on evaluation of Shannon Index (SI) in TCGA.

Analysis of WES data from NCT02113657

We obtained raw WES data generated from 30 tumor samples of patients with metastatic castration-resistant prostate cancer (mCRPC) before they received anti-CTLA-4 therapy (ipilimumab) as part of the clinical trial NCT02113657[54]. The paired-end FASTQ files were generated on the Illumina HiSeq 2000 after adapter trimming, underwent sequence quality assessment using FASTQC v0.11.8. Subsequently, the reads were aligned to the GRCh38/hg38 build through BWA mem 0.7.15-r1140[77]. The resulting BAM files were processed according to the Genome Analysis Toolkit (GATK) Best Practices[78,79], including steps such as marking duplicates, joint realignment of paired tumor-normal BAMs, and recalibration of base quality scores. Consensus mutation calls were derived as those from 2 out of 3 callers: MuSE 2.0[80,81] calls and PASS-filtered results from MuTect2[82] and Strelka2[83]. Copy number aberrations (CNAs), tumor purity and ploidy were called using ASCAT[58] (v3.1.2) configured for WES data processing.

Patients' response to immunotherapy outcome data as well as immunohistochemistry (IHC) staining for CD8 T cell densities were obtained from the previous report of the trial[54]. Following the pre-processing pipeline used for TCGA (**Extended Data Fig. 1a**), we removed samples with nrpcc<10 and TMB>100, and further removed samples with less than 10 coding SNVs (i.e., TMB > 0.25) which would otherwise inflate the subclonal fraction estimates. See more details in **Fig. 6d**.

Data availability

Raw read counts of bulk DNA sequencing data, clinical data and somatic mutations from 10,039 tumor samples across 32 TCGA cancer types are available for download from the Genomic Data Commons data portal (https://portal.gdc.cancer.gov/). The tumor purity and copy number profiling of TCGA samples were obtained using ASCAT v3[58] and the results were released on https://github.com/VanLoo-lab/ascat/tree/master/ReleasedData.

PhylogicSim500 and SimClone1000 dataset can be downloaded from https://data.mendeley.com/datasets/compare/by4gbgr9gd.

The PCAWG dataset is available through the ICGC data portal: https://dcc.icgc.org/pcawg. Specifically, the SNV sequencing reads can be downloaded through https://dcc.icgc.org/repositories, while the CNV and purity data can be accessed through https://dcc.icgc.org/releases/PCAWG/consensus_cnv. Sample information, including mapping from ICGC identifiers to TCGA identifiers was downloaded through the ICGC portal. Additionally, the survival/clinical data for the UK-ESAD cohort was downloaded through the ICGC portal.

Driver mutation annotations were downloaded from IntOGen at https://www.intogen.org/download?file=intogen_driver_mutations_catalog-2016.5.zip

All other relevant data are available from the corresponding author upon reasonable request.

Code availability

The CliPP model used for reconstructing tumor subclonal structure is freely available and can be downloaded from https://github.com/wwylab/CliPP. CliPP version 1.3.2 was used to generate the results in this work.

Author contributions

Y.J., M.D.M, K.Y., S.J.S. conceived the project. Y.J., K.Y., S.J.S and W.W. developed the mathematical method. Y.J., and K.Y. built the software tool, Y.J. and M.D.M performed the data analysis and interpretation, planned the figure design, and wrote the manuscript collaborating with

all the other authors. S.J. helped develop the analysis pipeline. S.J., S.G., Q.T. and S.C. assisted on data analysis and interpretation. R.L. assisted with data maintenance and manuscript writing. Y.T. contributed to the code optimization. T.L. assisted with the copy number data. S.K., J.P.S., J.A., A.S., J.R.W. and P.M. advised on cancer-type specific clinical outcome analysis. S.S. and P.S. contributed the clinical trial dataset for validation. S.S., P.S. and A.A advised on data analysis for prostate cancer. M.T. and P.V.L assisted in data interpretation, manuscript writing, and copy number profiling. M.K. and H.Z. advised on the mathematical modeling. W.W. conceived the study, planned and supervised the work, performed the analysis, and wrote the manuscript collaborating with all the other authors. All authors contributed to the interpretation of results and commented on and approved the final manuscript.

Acknowledgements

Y.J. is supported by in part by RP210006. Y.J., M.D.M., S.J. and W.W. are supported by R01CA268380. S.G., Q.T. and W.W. are supported by DoD PC210079. Q.T. is supported in part by R01CA275990. R.L. is supported in part by R01CA283402. J.A. is supported by the Caporella Family fund, the Kohn family fund, the Shipman Family fund, and Stupid Strong Foundation. P.M. was supported by the Andrew Sabin Family Foundation Fellowship, Gateway for Cancer Research, a Translational Research Partnership Award (KC200096P1) by the United States Department of Defense, an Advanced Discovery Award by the Kidney Cancer Association, a Translational Research Award by the V Foundation, the MD Anderson Physician-Scientist Award, donations from the Renal Medullary Carcinoma Research Foundation in honor of Ryse Williams, as well as philanthropic donations by the Chris “CJ” Johnson Foundation, and by the family of Mike and Mary Allen. T.L. was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (CC2008), the UK Medical Research Council (CC2008), and the Wellcome Trust (CC2008). A.K.S. was funded by NCI (CA281701 and CA209904), Ovarian Cancer Research Fund Alliance, American Cancer Society Research Professor Award, and the Frank McGraw Memorial Chair in Cancer Research. S. J. Shin was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT), grant number 2022M3J6A1063595, and 2023R1A2C1006587. Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Number RF1AG082938. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. J.P.S is supported by the Cancer Prevention & Research Institute of Texas as a CPRIT Scholar in Cancer Research (RR180035) and Conquer Cancer (CDA-7604125121). J.R.W. was supported by the Mark

Foundation for Cancer Research ASPIRE award. M.K. is supported in part by R01CA268380. P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. P.V.L. is a CPRIT Scholar in Cancer Research and acknowledges CPRIT grant support (RR210006). This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (CC2008), the UK Medical Research Council (CC2008), and the Wellcome Trust (CC2008). For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

The author(s) acknowledge the support of the High Performance Computing for research facility at the University of Texas MD Anderson Cancer Center for providing computational resources that have contributed to the research results reported in this paper.

Declaration of interests

S.S. consults for Bristol Myers Squibb. P.M. has received honoraria for service on a Scientific Advisory Board for Mirati Therapeutics, Bristol Myers Squibb, and Exelixis; consulting for Axiom Healthcare Strategies; non-branded educational programs supported by DAVA Oncology, Exelixis and Pfizer; and research funding for clinical trials from Takeda, Bristol Myers Squibb, Mirati Therapeutics, Gateway for Cancer Research, and the University of Texas MD Anderson Cancer Center. A.K.S. Consults for GSK, Astra Zeneca, Merck, ImmunoGen, Onxeo, Iylon, and Kiyatec, and is a shareholder of BioPath. P.S. is a scientific advisory committee member for Achelois, Afinni-T, Apricity, Asher Bio, BioAlta LLC, Candel Therapeutics, Catalio, Carisma, C-Reveal Therapeutics, Dragonfly Therapeutics, Earli Inc, Enable medicine, Glympse, Henlius/Hengenix, Hummingbird, ImaginAB, InterVenn Biosciences, Lava Therapeutics, Lytix Biopharma, Marker Therapeutics, Oncolytics, PBM Capital, Phenomic AI, Polaris Pharma, Trained Therapeutix Discovery, Two Bear Capital, Xilis, Inc. Additionally, P.S. has private investments in Adaptive Biotechnologies, BioNTech, JSL Health, Sporos, and Time Bioventures. The other authors declare no competing interests.

References

1. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
2. Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature 1982 3005888* **300**, 143–149 (1982).
3. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **340**, 1546–1558 (2013).
4. Garraway, L. A. & Lander, E. S. Lessons from the Cancer Genome. *Cell* **153**, 17–37 (2013).
5. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature 2012 4817381* **481**, 306–313 (2012).
6. De Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
7. Maley, C. C. *et al.* Classifying the evolutionary and ecological features of neoplasms. *Nature Reviews Cancer 2017 1710* **17**, 605–619 (2017).
8. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
9. McPherson, A. *et al.* Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics 2016 487* **48**, 758–767 (2016).
10. Black, J. R. M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nature Reviews Cancer* **21**, 379–392 (2021).
11. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
12. Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239-2254.e39 (2021).
13. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine* **22**, 105–113 (2016).

14. Persi, E., Wolf, Y. I., Leiserson, M. D. M., Koonin, E. V. & Ruppin, E. Criticality in tumor evolution and clinical outcome. *Proceedings of the National Academy of Sciences* **115**, (2018).
15. Caravagna, G. *et al.* Subclonal reconstruction of tumors by using machine learning and population genetics. *Nature Genetics* **52**, 898–907 (2020).
16. Persi, E. *et al.* Mutation-selection balance and compensatory mechanisms in tumour evolution. *Nature Reviews Genetics* **22**, 251–262 (2021).
17. Samstein, R. M. *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature Genetics* **51**, 202–206 (2019).
18. Segal, N. H. *et al.* Epitope Landscape in Breast and Colorectal Cancer. *Cancer Research* **68**, 889–892 (2008).
19. FDA approves pembrolizumab for adults and children with TMB-H solid tumors. *FDA* (2020).
20. Litchfield, K. *et al.* Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596-614.e14 (2021).
21. Tilk, S., Tkachenko, S., Curtis, C., Petrov, D. A. & McFarland, C. D. Most cancers carry a substantial deleterious load due to Hill-Robertson interference. *eLife* **11**, e67790 (2022).
22. Martínez-Jiménez, F. *et al.* Genetic immune escape landscape in primary and metastatic cancer. *Nature Genetics* **55**, 820–831 (2023).
23. Landau, D. A. *et al.* Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell* **152**, 714–726 (2013).
24. Espiritu, S. M. G. *et al.* The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. *Cell* (2018)
25. Jamal-Hanjani, M. *et al.* Tracking the evolution of non-small-cell lung cancer. *New England Journal of Medicine* **376**, 2109–2121 (2017).
26. Turajlic, S. *et al.* Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. *Cell* **173**, 581-594.e12 (2018).

27. Körber, V. *et al.* Neuroblastoma arises in early fetal development and its evolutionary duration predicts outcome. *Nature Genetics* **55**, 619–630 (2023).
28. Gillis, S. & Roth, A. PyClone-VI: scalable inference of clonal population structures using whole genome data. *BMC Bioinformatics* **21**, 571 (2020).
29. Deshwar, A. G. *et al.* PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology* **16**, 35 (2015).
30. Leshchiner, I. *et al.* Inferring early genetic progression in cancers with unobtainable premalignant disease. *Nature Cancer* **4**, 550–563 (2023).
31. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149**, 994–1007 (2012).
32. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nature Methods* **11**, 396–398 (2014).
33. Miller, C. A. *et al.* SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Computational Biology* **10**, e1003665 (2014).
34. Fan, J. & Li, R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association* **96**, 1348–1360 (2001).
35. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. Sparsity and Smoothness Via the Fused Lasso. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **67**, 91–108 (2005).
36. Ke, Z. T., Fan, J. & Wu, Y. Homogeneity Pursuit. *Journal of the American Statistical Association* **110**, 175–194 (2015).
37. Tarabichi, M. *et al.* A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nature Methods* **18**, 144–155 (2021).
38. Guerra, A. *et al.* The Primary Occurrence of *BRAF* ^{V600E} Is a Rare Clonal Event in Papillary Thyroid Carcinoma. *Journal of Clinical Endocrinology and Metabolism* **97**, 517–524 (2012).
39. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

40. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nature Reviews Cancer* **20**, 555–572 (2020).
41. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine* **10**, 25 (2018).
42. Luo, S. *et al.* IDH clonal heterogeneity segregates a subgroup of non-1p/19q codeleted gliomas with unfavourable clinical outcome. *Neuropathology and Applied Neurobiology* **47**, 394–405 (2021).
43. Lan, Y. *et al.* Revealing clonality and subclonality of driver genes for clinical survival benefits in breast cancer. *Breast Cancer Research and Treatment* **175**, 91–104 (2019).
44. Shi, J. *et al.* Comprehensive characterization of clonality of driver genes revealing their clinical relevance in colorectal cancer. *Journal of Translational Medicine* **20**, 362 (2022).
45. Finkel, A. *et al.* Subclonality for BRAF Mutation in Papillary Thyroid Carcinoma Is Associated With Earlier Disease Stage. *Journal of Clinical Endocrinology and Metabolism* **101**, 1407–1413 (2016).
46. Parker, J. S. *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology* **27**, 1160–1167 (2009).
47. Pu, M. *et al.* Research-based PAM50 signature and long-term breast cancer survival. *Breast Cancer Research Treatment* **179**, 197–206 (2020).
48. Egevad, L., Granfors, T., Karlberg, L., Bergh, A. & Stattin, P. Prognostic value of the Gleason score in prostate cancer. *BJU International* **89**, 538–542 (2002).
49. Liu, L. *et al.* Favorable outcome of patients with lung adenocarcinoma harboring POLE mutations and expressing high PD-L1. *Molecular Cancer* **17**, 81 (2018).
50. Ragin, C. C. R. & Taioli, E. Survival of squamous cell carcinoma of the head and neck in relation to human papillomavirus infection: Review and meta-analysis. *International Journal of Cancer* **121**, 1813–1820 (2007).

51. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423 (1948).

52. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).

53. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

54. Subudhi, S. K. *et al.* Neoantigen responses, immune correlates, and favorable outcomes after ipilimumab treatment of patients with prostate cancer. *Science Translational Medicine* **12**, eaaz3577 (2020).

55. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37**, 773–782 (2019).

56. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harbor Perspectives in Medicine* **7**, a026625 (2017).

57. Kimura, M. The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations. *Genetics* **61**, 893–903 (1969).

58. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences* **107**, 16910–16915 (2010).

59. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology* **30**, 413–421 (2012).

60. Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. (CRC press, 2015).

61. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **58**, 267–288 (1996).

62. Zhang, C. H. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894–942 (2010).

63. Boyd, S. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning* **3**, 1–122 (2010).
64. Guennebaud, G. & Jacob, B. Eigen. *URl: http://eigen. tuxfamily. org* (2010).
65. Sengupta, S. *et al.* Bayclone: Bayesian nonparametric inference of tumor subclones using NGS data. in *Biocomputing 2015* 467–478 (WORLD SCIENTIFIC, Kohala Coast, Hawaii, USA, 2014).
66. Yuan, K., Macintyre, G., Liu, W., PCAWG-11 working group & Markowetz, F. *Ccube: A Fast and Robust Method for Estimating Cancer Cell Fractions*. http://biorxiv.org/lookup/doi/10.1101/484402 (2018).
67. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Reports* **7**, 1740–1752 (2014).
68. Donmez, N. *et al.* Clonality Inference from Single Tumor Samples Using Low-Coverage Sequence Data. *Journal of Computational Biology* **24**, 515–523 (2017).
69. Cun, Y., Yang, T.-P., Achter, V., Lang, U. & Peifer, M. Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. *Nature Protocols* **13**, 1488–1501 (2018).
70. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Systems* **6**, 271-281.e7 (2018).
71. Bangdiwala, S. A graphical test for observer agreement. *45th International Statistical Institute Meeting* (1985).
72. Munoz, S. R. & Bangdiwala, S. I. Interpretation of Kappa and B statistics measures of agreement. *Journal of Applied Statistics* **24**, 105–112 (1997).
73. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Research* **49**, D916–D923 (2021).
74. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine* **9**, 34 (2017).

75. Therneau, T. M., Atkinson, E. J. & Foundation, M. An Introduction to Recursive Partitioning Using the RPART Routines. Bioconductor (2023).

76. Kassambara, A., Kosinski, M., Biecek, P. & Scheipl, F. survminer: Drawing Survival Curves using ‘ggplot2’. (2021).

77. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

78. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011).

79. Van Der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics* **43**, (2013).

80. Ji, S., Zhu, T., Sethia, A. & Wang, W. Accelerated somatic mutation calling for whole-genome and whole-exome sequencing data from heterogenous tumor samples. *Genome Research* **34**, 633–641 (2024).

81. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology* **17**, 178 (2016).

82. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31**, 213–219 (2013).

83. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods* **15**, 591–594 (2018).

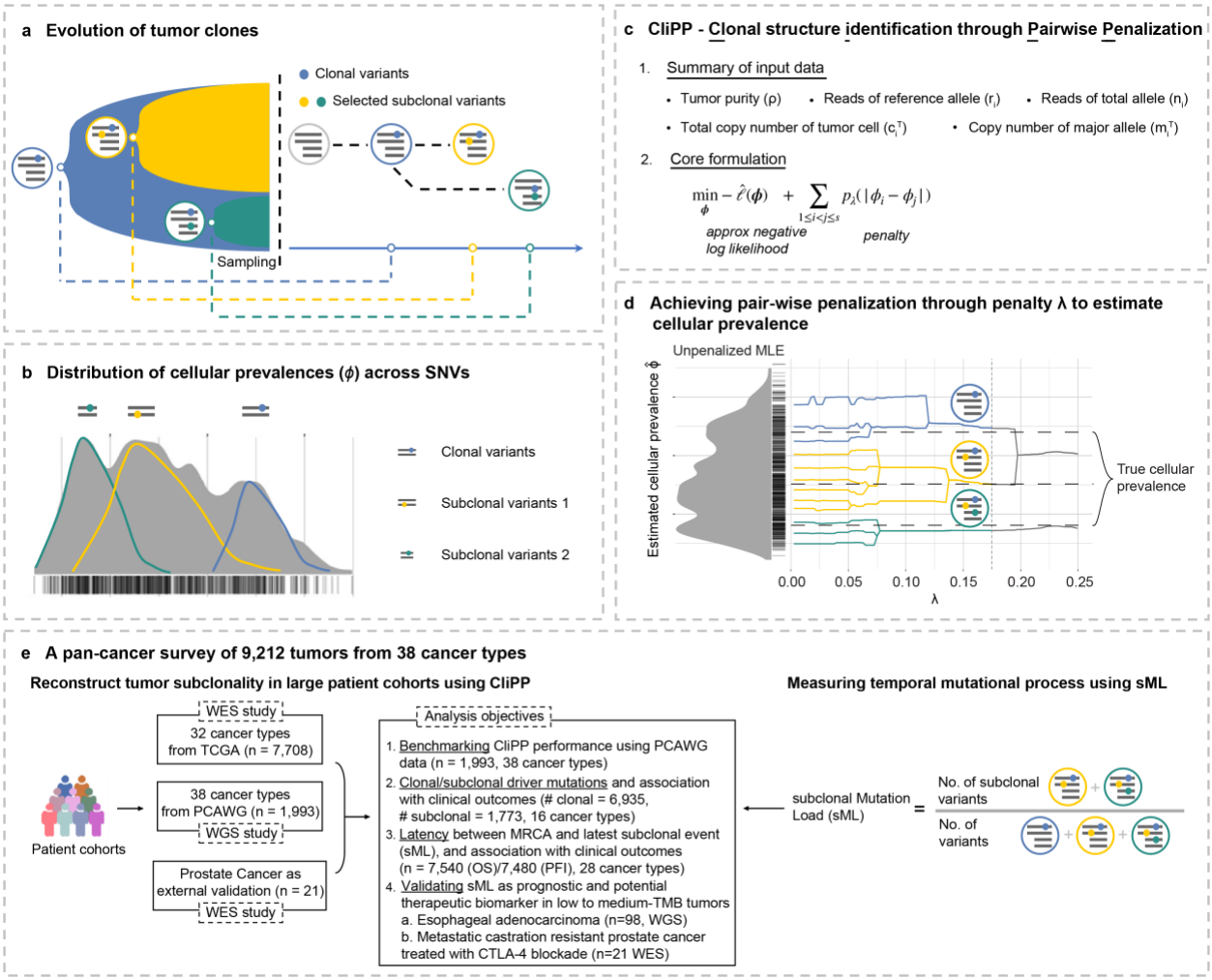


Fig. 1: Subclonal reconstruction and CliPP model overview. (a) A visual representation of tumor evolution with genotype circles indicating representative somatic single-nucleotide variants (SNVs). Colored dots represent SNVs. **(b)** Toy example of the cellular prevalence (CP) distribution of SNVs characterized by one clonal cluster (blue) and two subclonal clusters (yellow and green). **(c)** Input data and the core regularization function for the CliPP model, a novel computational tool for subclonal reconstruction. **(d)** CliPP infers clusters of SNVs and CP values using a penalty function that varies over a penalty parameter λ. The y-axis corresponds to CP estimates (0 to 1) connecting the nearest CP estimates over 100 evenly spaced λ values (0.01 to 0.25). Vertical grey lines indicate 10 default λ choices used in CliPP for finding the optimal solution efficiently, with the vertical dotted line indicating the chosen λ (see **Methods**). Somatic SNV clusters distinguishing parental cancerous cells (blue) from new cancerous cells (yellow and green) converge for homogeneity pursuit in CPs. **(e)** Objectives for utilizing the CliPP-based tumor subclonal reconstruction results from 9,212 tumor samples across 38 cancer types in both the TCGA and ICGC-PCAWG.

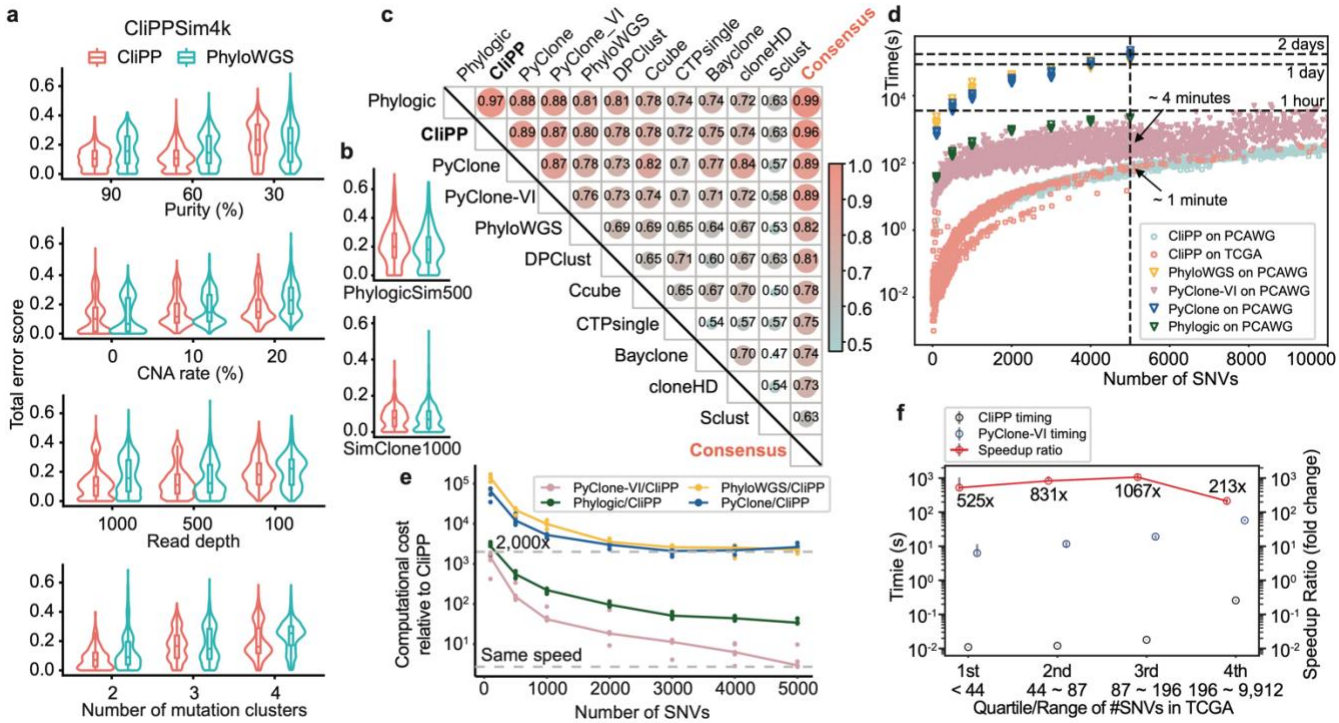


Fig. 2: CliPP benchmarking using simulation data and real patient cohorts. (a) CliPP vs. PhyloWGS performance on CliPPSim4k samples shown through paired violin and boxplots representing total error score distributions. These scores reflect combined metrics for accuracy in estimating the number of clusters, fraction of clonal mutations, and cellular prevalence. Subplots illustrate error score variations across 4 key features (purity, CNA rate, read depth and number of mutation clusters). A lower score indicates better performance. **(b)** CliPP vs. PhyloWGS performance on the PhylogicSim500 and SimClone1000 samples. Violin plots paired with boxplots display the total error score distributions. **(c)** Pairwise Concordance Correlation Coefficient (CCC) heatmap of the estimated fraction of clonal mutations by 11 individual reconstruction methods, including CliPP, and the consensus calls on the PCAWG dataset (n=1,548). Color intensity reflects CCC values. **(d)** Computational times for subclonal reconstruction on 2,146 PCAWG WGS samples and 9,843 TCGA WES samples. The vertical dashed line corresponds to samples with 5,000 SNVs. Labels indicate average computing times for 5 samples with ~5,000 SNVS, CliPP (1 minute) and PyClone-VI (4 minutes). Other methods' timings: PhylogicNDT (37 minutes), PyClone (2,765 minutes), and PhyloWGS (2,450 minutes). **(e)** Computational time comparison relative to CliPP in fold change on PCAWG samples with <= 5,000 SNVs. Each dot represents one of the 35 samples where all five methods were applied. **(f)** Projected time comparison between CliPP and Pyclone-VI in TCGA samples categorized into four quartiles based on the number of SNVs. Open dots correspond to the median time per quartile

for each method (see Methods). Error bars depict the 95% confidence intervals. CliPP's error bars are too small to show at this scale.

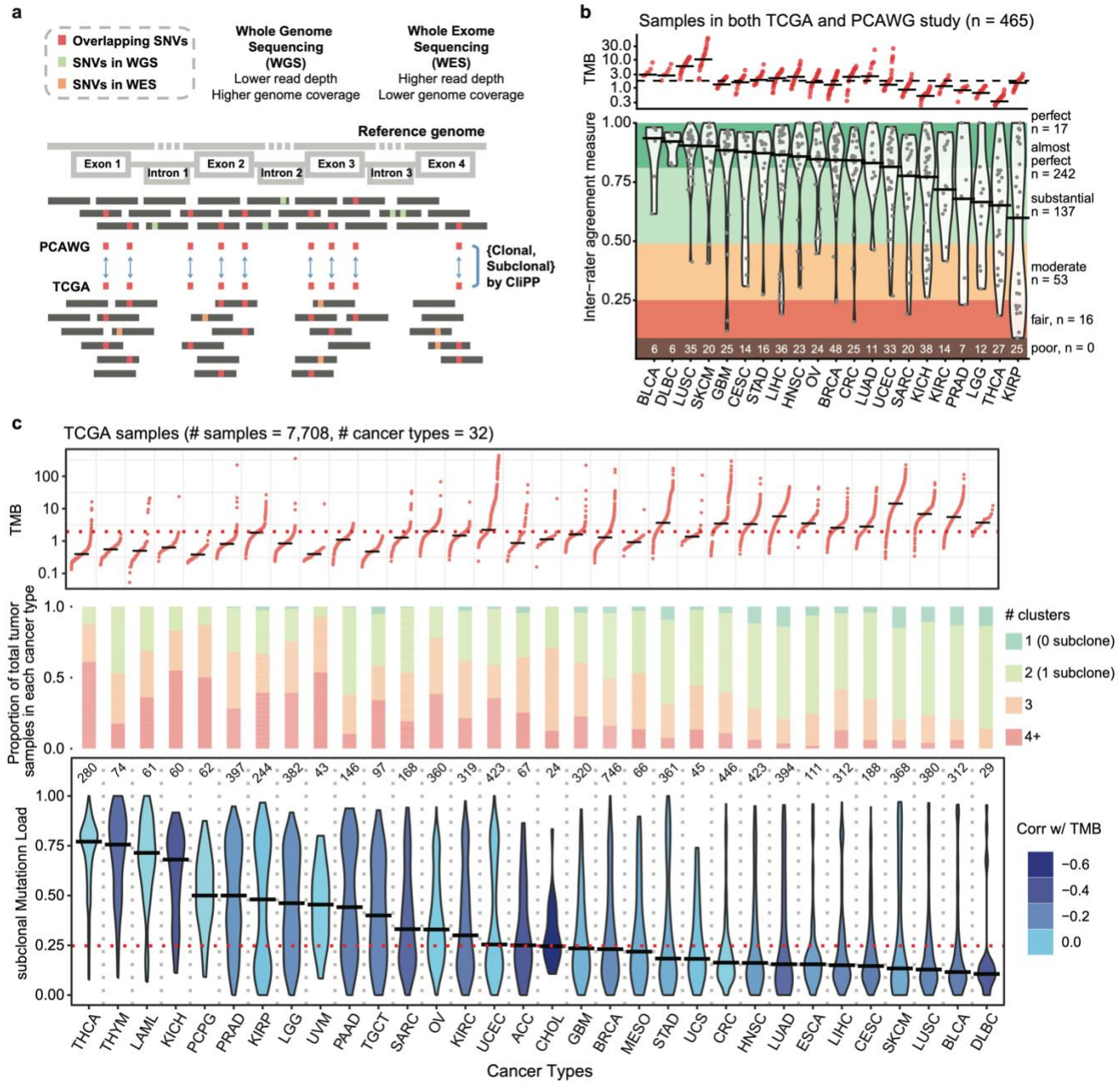


Fig. 3: The subclonal landscape of 7,708 tumors in TCGA. (a) Illustration of coverage and read depth differences between WGS and WES data. SNVs observed in samples sequenced by both methods are shown in red for cross-platform subclonal reconstruction comparison. **(b)** Violin plots display inter-rater agreement per sample of subclonal reconstruction results by CliPP from WES in TCGA and WGS in PCAWG, ordered by median agreement score (black bar). Cancer types with higher TMB (shown in red in top panel) tend to have higher agreement. All samples

exhibit fair or better level of agreement. (See **Extended Data Fig. 1b** for the reference of cancer-type acronyms.) **(c)** Distributions of the subclonal mutation load (sML) across 32 cancer types. The number of tumor samples for each cancer type is indicated above each violin plot. Median values for each cancer type are represented by black bars, while the median across all samples is represented by a red dotted line. Additionally, the sigmoid plot depicts **Fig. 3 (cont'd):** the distributions of TMB in all tumor samples, and the stacked bar plots show the number of clusters for each sample estimated by CliPP across all tumor samples. Violin plots are color-coded in accordance with the Pearson correlation coefficients of sML versus TMB.

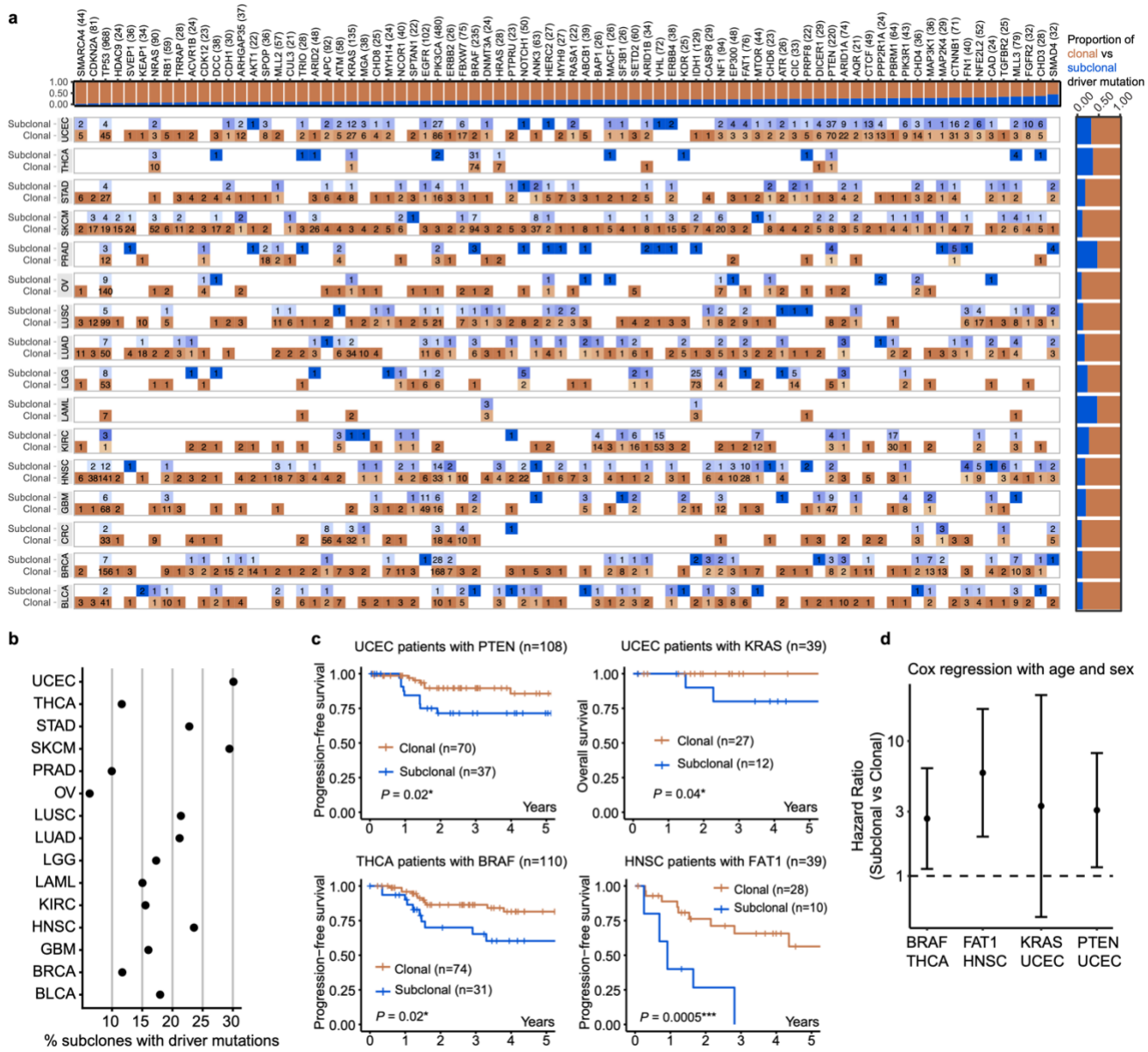


Fig. 4: Subclonality of driver mutations across most frequently mutated genes in TCGA. **(a)** Heatmap depicting the distribution of clonal (orange squares) and subclonal (blue squares) driver single nucleotide variants (SNVs) across top 78 genes in 16 cancer types (see **Methods**). Orange denotes clonal and blue denotes subclonal mutations. Numbers within squares represent SNV counts. Darker shade represents a higher proportion of clonal or subclonal mutations. Proportions of subclonal and clonal mutations are summarized by gene (top) and cancer type (right). **(b)** Percentage of subclonal mutation clusters containing a driver mutation, by cancer type. **(c)** Kaplan-Meier curves showing pairs of driver gene and cancer type with statistically significant survival differences between those with clonal (orange) versus subclonal (blue) driver mutations of the gene. Log-rank test *P*-values are shown. Significance levels are denoted as follows:

*$P < 0.05$, **$P < 0.01$ and ***$P < 0.001$. (d) Hazard ratios and 95% confidence intervals for the driver gene/cancer pairs from (c), after adjusting for age at clinical diagnosis, and sex (not applicable in UCEC).

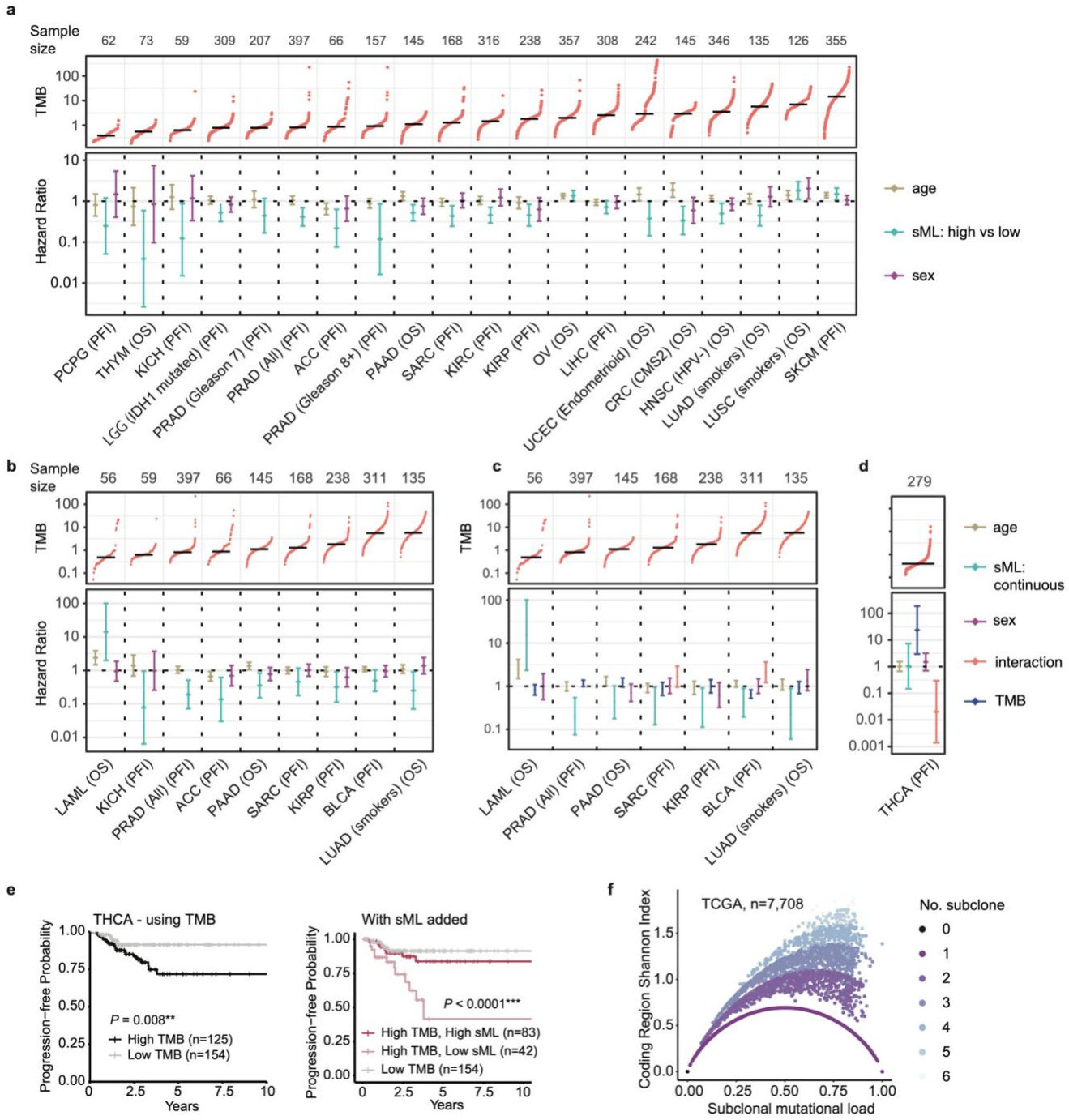


Fig. 5: Biological and clinical implications of sML in TCGA. (a) Forest plot of hazard ratios (HRs) and 95% confidence intervals (CIs) from multivariate Cox proportional hazard models for Overall Survival (OS) or Progression-Free Interval (PFI) in TCGA for binarized (low/high) sML.

Models are adjusted for age, sML (high versus low), and sex if applicable. Cancer types are ordered by median tumor mutational burden (TMB), for which a sigmoid plot is shown on top. The number of tumor samples for each cancer type is shown above each sigmoid plot. **(b)** Forest plot of HRs and 95% CIs of multivariate Cox proportional hazard models for OS or PFI in TCGA for continuous sML. Models are adjusted for age, sML as a continuous variable, and sex if applicable. Cancer types are ordered by median TMB. **(c)** Forest plot **Fig. 5 (cont'd)**: of HRs and 95% CIs of multivariate Cox proportional hazard models for OS or PFI in TCGA for continuous sML and TMB. Models are adjusted for age, sML as a continuous variable, TMB, and sex if applicable. An interaction term between sML and TMB is included if at least marginally significant (α=0.1). Cancer types are ordered by median TMB. **(d)** Forest plot of HR and 95% CI of multivariate Cox proportional hazard model for PFI in thyroid cancer in TCGA. The model is adjusted as (**c**). **(e)** Kaplan-Meier curves of PFI for TCGA thyroid cancer samples which are categorized first by TMB alone, and then by TMB and sML. *P*-values are obtained by log-rank tests. (**f**) The scatter plot of Shannon Index versus sML in TCGA, where the color variations represent samples with different numbers of subclones.

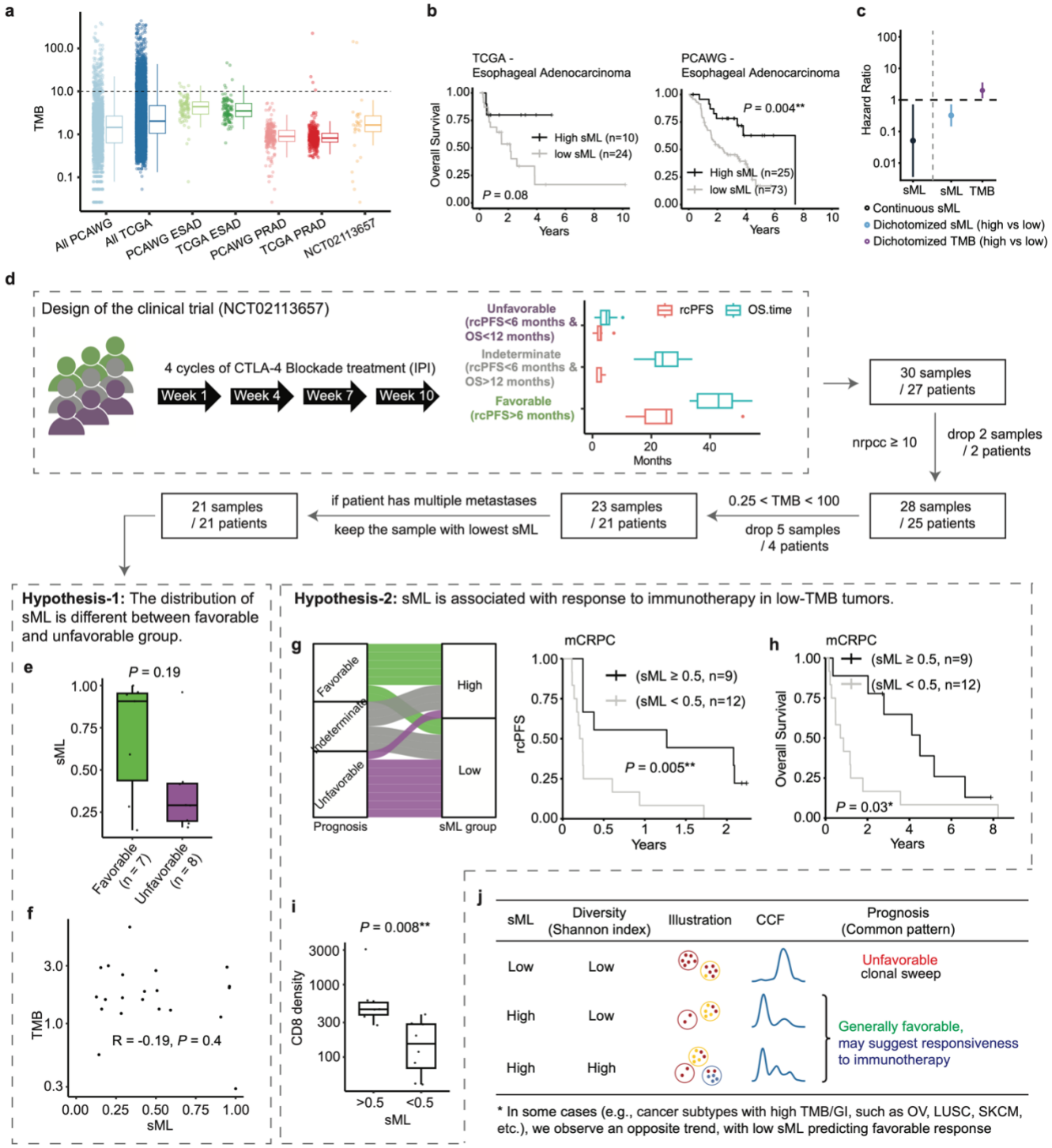
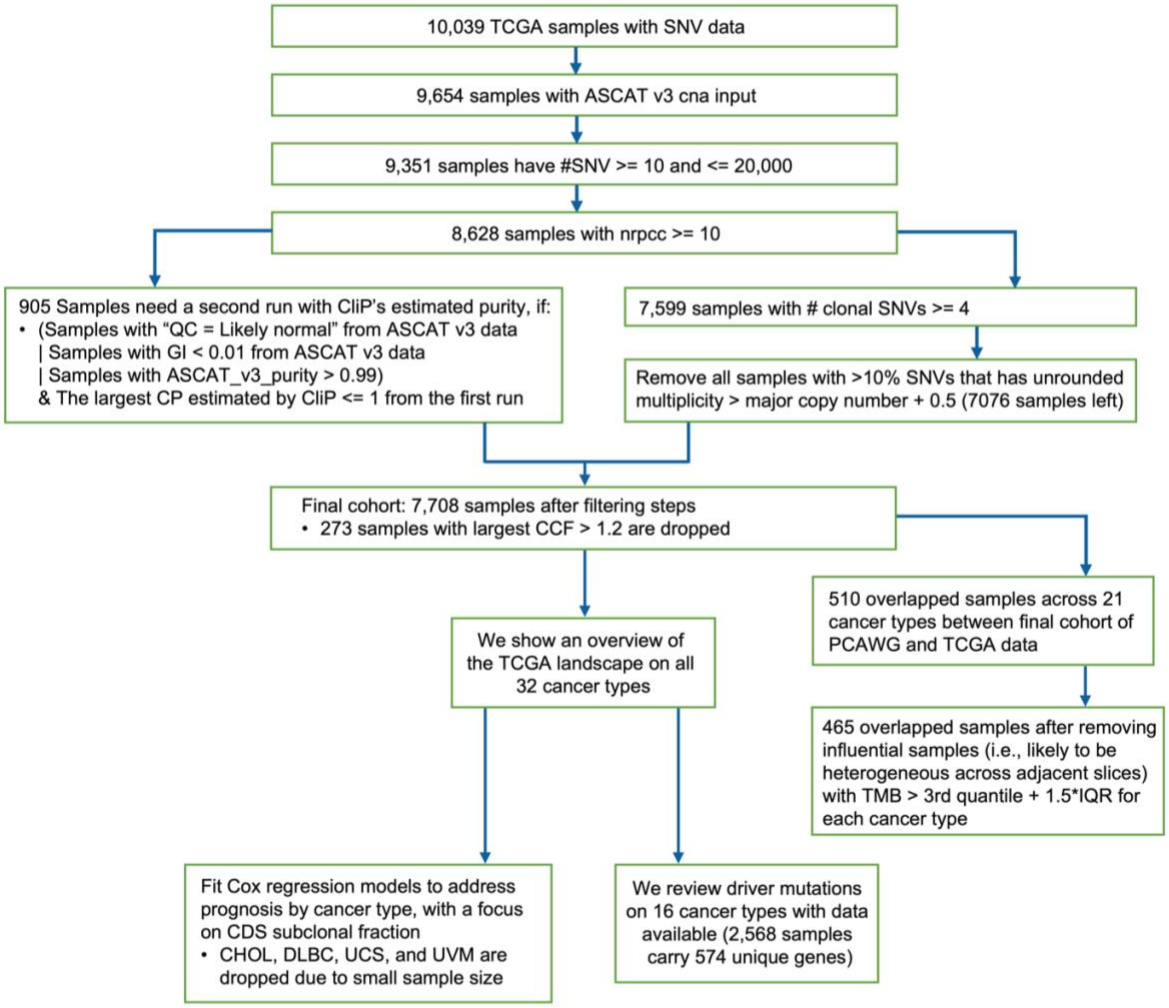


Fig. 6: Additional validation with non-TCGA patient cohorts and a data-driven hypothesis. **(a)** Boxplots with paired data points show tumor mutational burden (TMB) distribution for all PCAWG and TCGA samples, along with their esophageal cancer and prostate adenocarcinoma subsets and the mCRPC samples in the clinical trial NCT02113657 (unfiltered). Dashed line

represents the TMB threshold commonly used for receiving immunotherapy **(b)** Kaplan-Meier curves depict Overall Survival (OS) in esophageal adenocarcinoma patients from TCGA (left) and PCAWG (right), divided by high and low sML. **Fig. 6 (cont'd): (c)** Hazard ratio and 95% confidence interval for the significant terms in a Cox proportional hazard model for the PCAWG ESAD samples. The first model (black) was fit with age, sex, TMB (as a continuous variable) and sML (as a continuous variable), with only continuous sML remaining significant. The second model (blue) was fit with age, sex, TMB (dichotomized), and sML (dichotomized), with both sML and TMB showing significant effects. **(d)** Design of the mCRPC clinical trial (NCT02113657). Patients with rcPFS > 6 months are classified as favorable; patients with rcPFS < 6 months and OS < 12 months are considered unfavorable; all other patients are indeterminate. **(e)** Boxplot showing distributions of sML in the favorable versus the unfavorable group. **(f)** Scatter plot showing a weak, non-signficant negative correlation between sML and TMB (log10 transformed). **(g)** Clinical outcomes in patients stratified by sML=0.5. Alluvial plot comparing the sML category assignment (high vs low) and annotated treatement response (favorable, indeterminate, unfavorable), and the corresponding Kaplan-Meier curves of rcPFS. *P*-values are obtained by log-rank tests. **(h)** Overall survival in patients stratified by sML=0.5. (**i**) Boxplot showing distributions of CD8 density (cells/mm^2) in high versus low sML patient groups (n=9 vs. 12). *P*-value of two-sided Wilcoxon rank sum test is shown. **(j)** A data-driven hypothesis to summarize the implications of the CliPP-based large-scale subclonal analysis on cancer evolution. Significance levels are denoted as follows: **P* < 0.05, ***P* < 0.01 and ***P < 0.001.

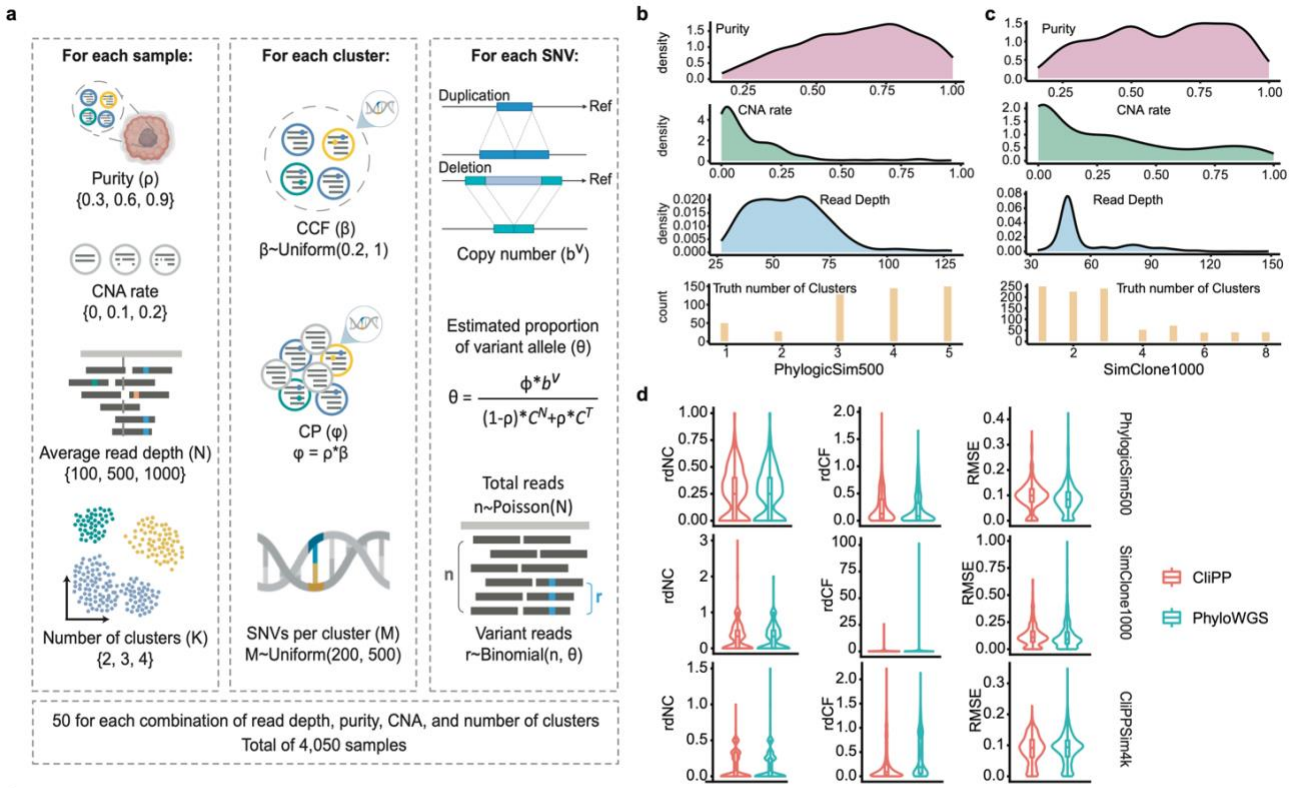
a

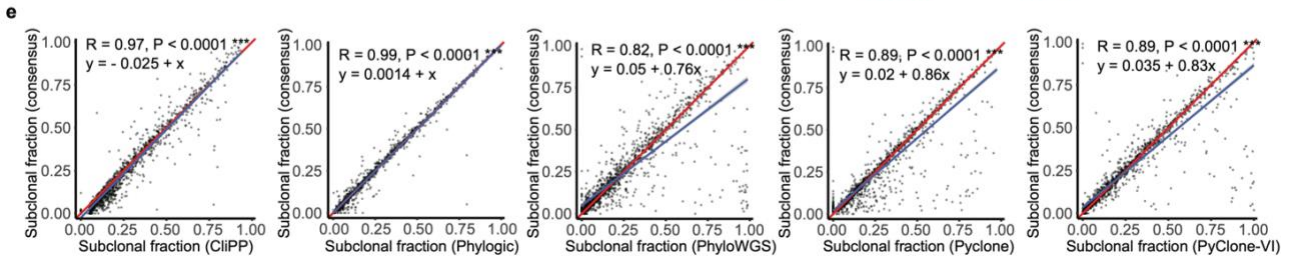


b

Abbreviation	Cancer type	Abbreviation	Cancer type
ACC	Adrenocortical carcinoma	MALY	Malignant lymphoma
BLCA	Bladder urothelial carcinoma	MESO	Mesothelioma
BOCA	Bone cancer	OV	Ovarian serous cystadenocarcinoma
BRCA	Breast invasive carcinoma	PAAD	Pancreatic adenocarcinoma
BTCA	Biliary tract cancer	PBCA	Pediatric brain cancer
CESC	Cervical squamous cell carcinoma & adenocarcinoma	PCPG	Pheochromocytoma and paraganglioma
CHOL	Cholangiocarcinoma	PRAD	Prostate adenocarcinoma
CLLE	Chronic lymphocytic leukemia	SARC	Sarcoma
CMDI	Chronic myeloid disorders	SKCM	Melanoma
CRC	Colorectal carcinoma	STAD	Stomach adenocarcinoma
DLBC	Diffuse large B-cell lymphoma	TGCT	Testicular germ cell tumors
ESCA	Esophageal carcinoma	THCA	Thyroid papillary carcinoma
GBM	Glioblastoma multiforme	THYM	Thymoma
HNSC	Head & neck squamous cell carcinoma	UCEC	Uterine corpus endometrial carcinoma
KICH	Renal chromophobe	UCS	Uterine carcinosarcoma
KIRC	Renal clear cell carcinoma	UVM	Uveal melanoma
KIRP	Renal papillary carcinoma		
LAML	Acute myeloid leukemia		
LGG	Brain lower grade glioma		
LIHC	Hepatocellular carcinomas		
LUAD	Lung adenocarcinoma		
LUSC	Lung squamous cell carcinoma		

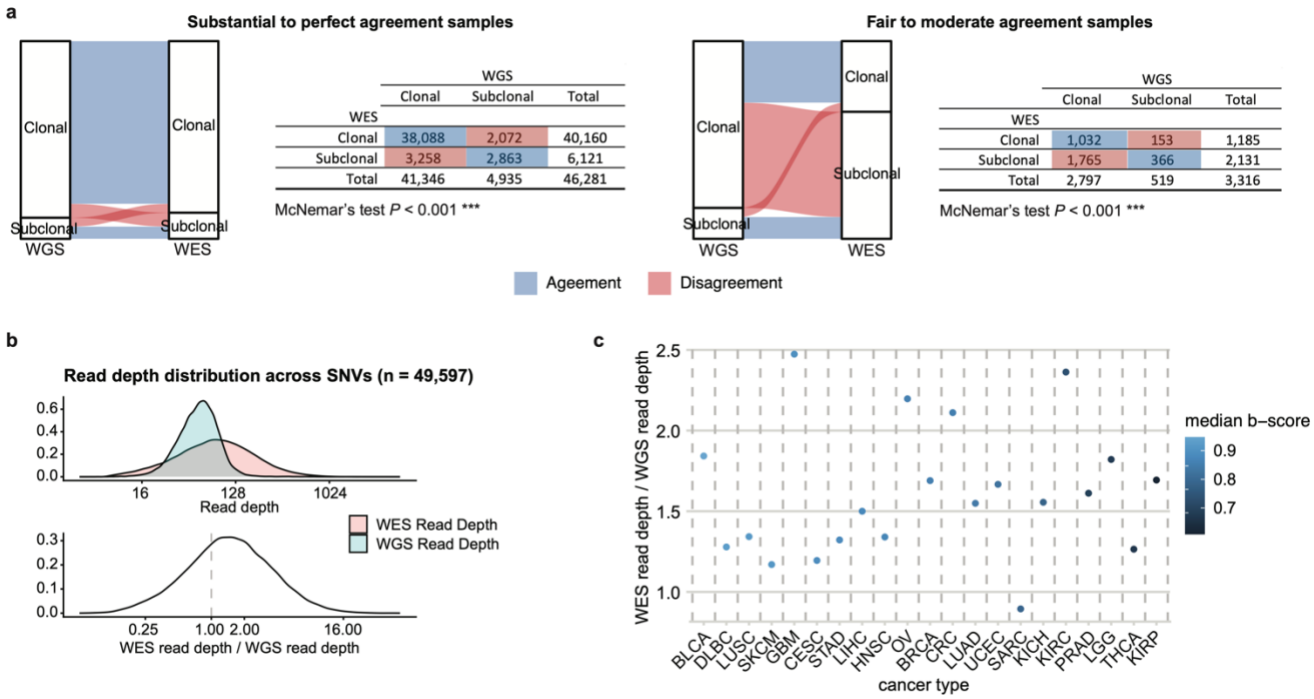
Extended Data Fig. 1: Overview of the TCGA data analysis. (a) CONSORT diagram for data processing in TCGA. **(b)** Acronyms for all 38 cancer types analyzed in this study.



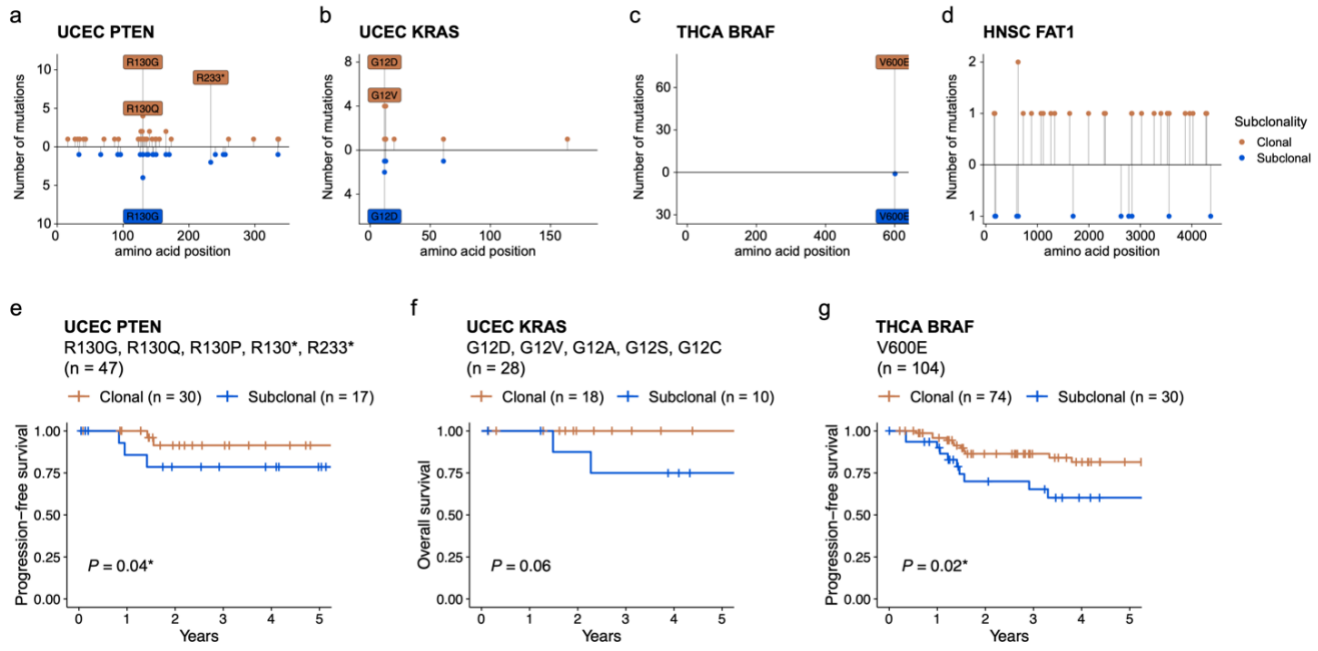


Extended Data Fig. 2: Performance benchmarking using both simulated and patient cohort datasets. (a) Overview of the simulation design to benchmark subclonal reconstruction performance. 50 samples are generated for each of 81 total combinations of sample-level parameters (purity, average read depth, CNA rate, and number of clusters), resulting in a simulated cohort of 4,050 samples. Within each sample, further parameters are simulated per cluster (CCF, CP, number of SNVs), and per SNV (copy number, proportion of the variant allele and total reads) to generate variant and reference reads (see **Methods**). **(b-c)** Distributions of purity, CNA rate, read depth, and number of clusters across samples from PhylogicSim500 and SimClone1000. **(d)** Performance comparisons between CliPP and PhyloWGS on PhylogicSim500, SimClone1000, and CliPPSim4k. For PhylogicSim500, the median values for (CliPP, PhyloWGS) are: (0.25, 0.25) for rdNC, (0.13, 0.084) for rdCF, (0.099, 0.084) for RMSE.

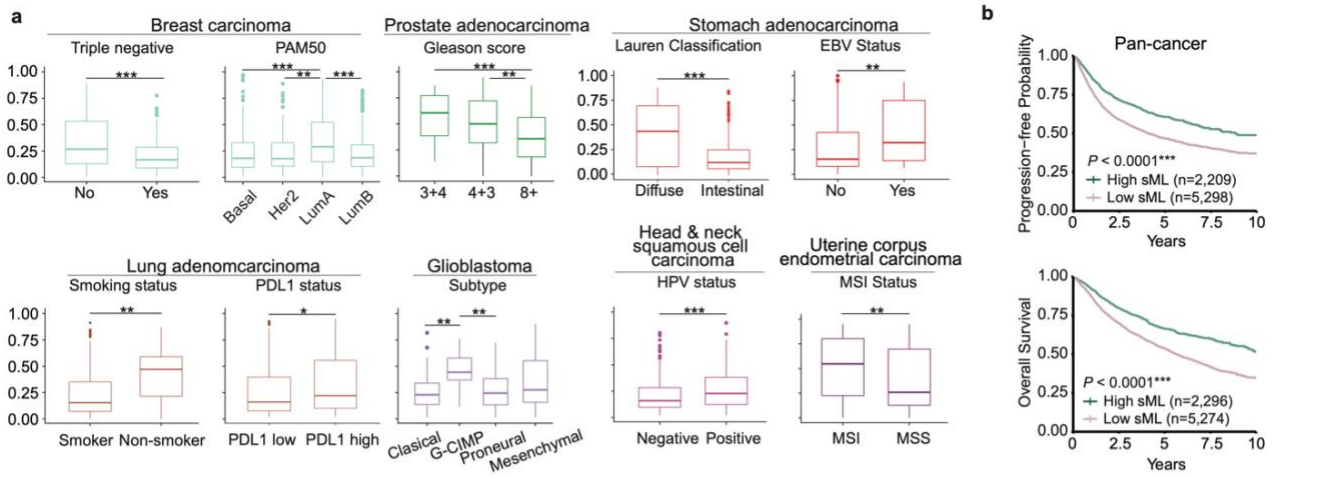
For SimClone1000, the median values are: (0.33, 0.33) for rdNC, (0.15, 0.15) for rdCF, (0.11, 0.091) for RMSE. For CliPPSim4k, the median values are: (0, 0) for rdNC, (0.088, 0.192) for rdCF, (0.091, 0.094) for RMSE. **(e)** Scatter plots of subclonal fraction estimations from individual method versus consensus calls in PCAWG. Red denotes the diagonal line. Blue denotes a linear regression line.

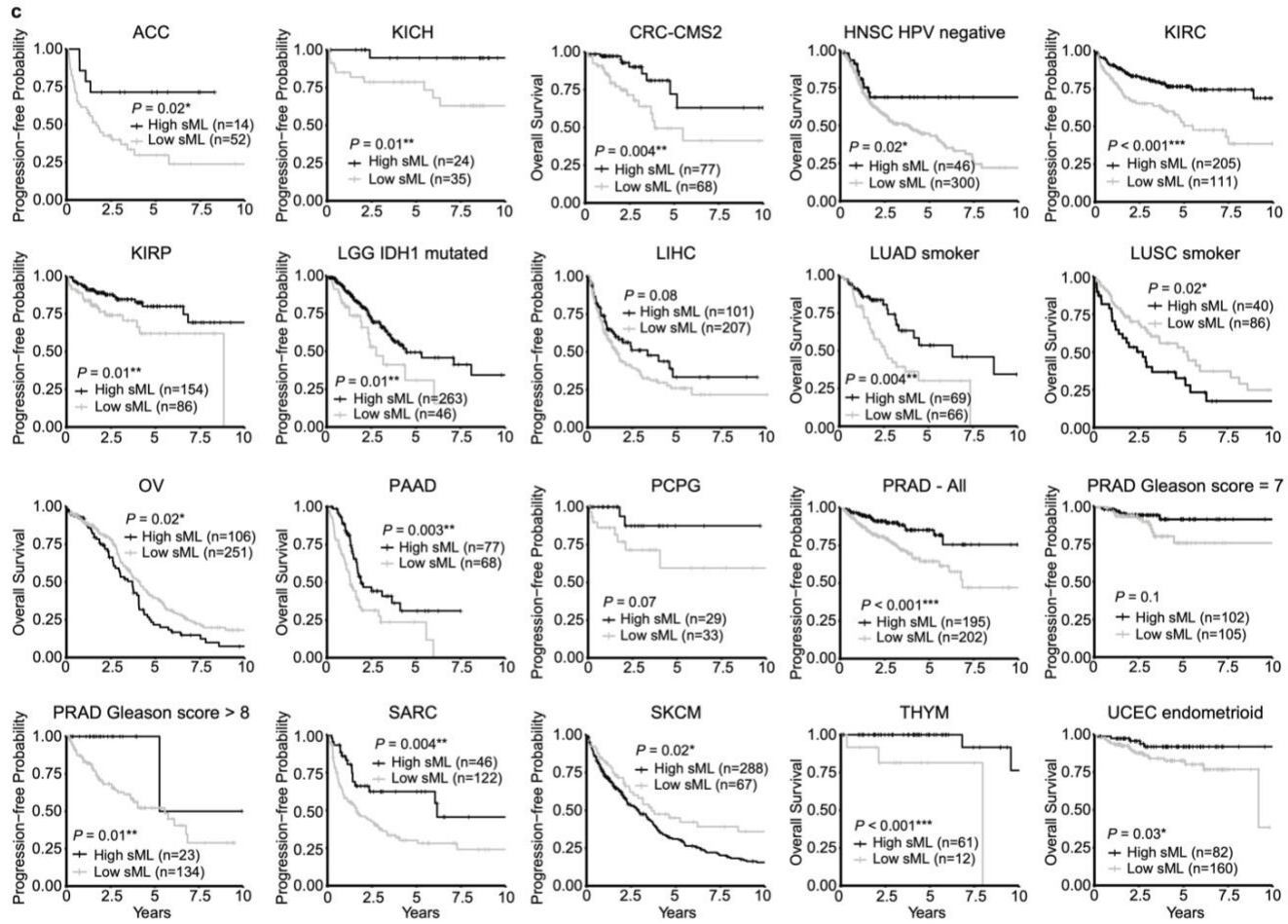


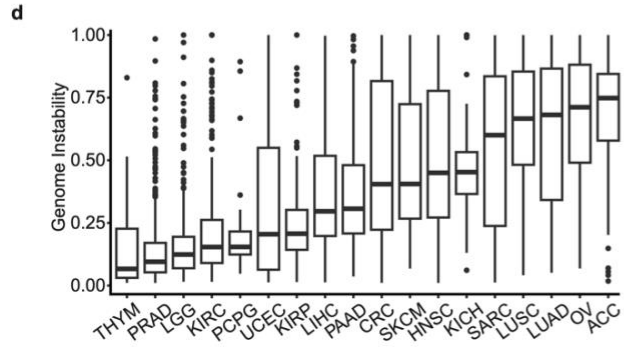
Extended Data Fig. 3: Benchmarking subclonal reconstruction pan-cancer using WES data. (a) Alluvial plots comparing the clonality assignment per mutation from WGS to WES. Mutations in samples with higher agreement (substantial or perfect agreement) are shown on the left. Mutations in samples with lower agreement (fair or moderate agreement) are shown on the right. Each alluvial plot is paired with a two-way contingency table with detailed numbers, with *P*-values indicated for McNemar's test for a biased shift between the red discordant squares. **(b)** Read depth distribution across SNVs for WES and WGS data (top) and the distribution of ratios of read depths between WES and WGS across SNVs (bottom). Vertical dotted line at 1 indicates equal read depth between WES and WGS. **(c)** Median of the read depth ratio of each overlapped sample from TCGA WES dataset and PCAWG WGS dataset by cancer type, ordered by median B statistic (see **Methods**).



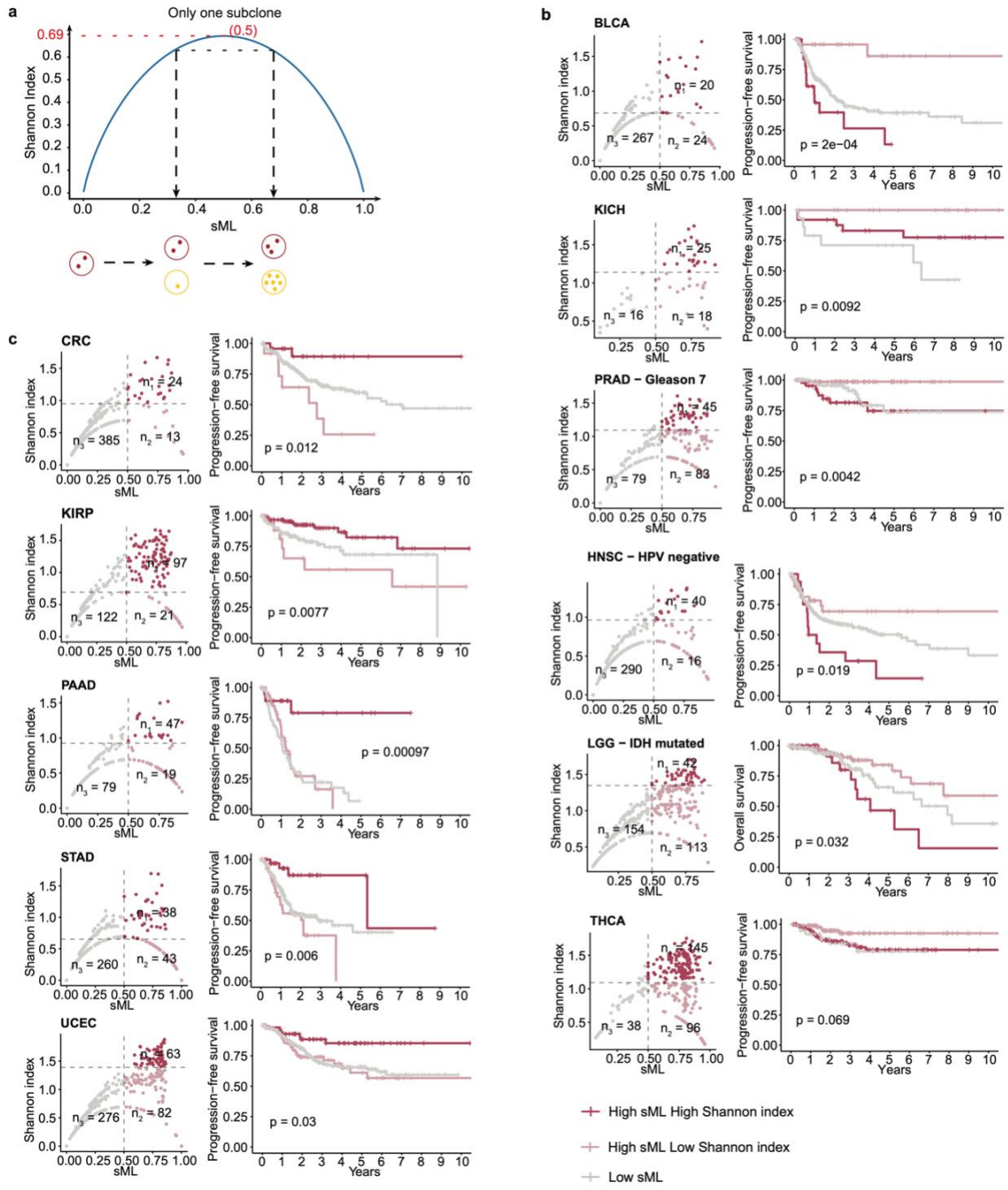
Extended Data Fig. 4: (**a-d**) Lollipop plots showing mutation position distributions within significant cancer-driver gene pairs. Frequencies of clonal mutations in orange are indicated by the vertical height of the "lollipops", while frequencies of subclonal mutations are shown in blue. Frequently recurrent mutations (n>4) are labeled. (**e**) Kaplan-Meier curves showing progression-free survival for clonal (orange) and subclonal (blue) UCEC patients with driver mutations in *PTEN*, restricting to the most recurrent sites (all mutations at R130 and R233). **(f**) Kaplan-Meier curves showing overall survival for clonal (orange) and subclonal (blue) UCEC patients with driver mutations in *KRAS*, restricting to the most recurrent sites (all mutations at G12). (**g**) Kaplan-Meier curves showing progression-free survival for clonal (orange) and subclonal (blue) THCA patients with driver mutations in *BRAF*, restricting to the most recurrent site (V600E). Log-rank test *P*-value are shown.


a
Breast carcinoma
Triple negative
PAM50
Prostate adenocarcinoma
Gleason score
Stomach adenocarcinoma
Lauren Classification
EBV Status
No
Yes
Basal
Her2
LumA
LumB
3+4
4+3
8+
Diffuse
Intestinal
Lung adenomcarcinoma
Smoking status
PDL1 status
Glioblastoma
Subtype
Head & neck squamous cell carcinoma
HPV status
Uterine corpus endometrial carcinoma
MSI Status
Smoker
Non-smoker
PDL1 low
PDL1 high
Clasical
G-CIMP
Proneural
Mesenchymal
Negative
Positive
MSI
MSS
b
Pan-cancer
Progression−free Probability
P < 0.0001***
High sML (n=2,209)
Low sML (n=5,298)
Years
Overall Survival
P < 0.0001***
High sML (n=2,296)
Low sML (n=5,274)
Years


c
ACC
P = 0.02*
High sML (n=14)
Low sML (n=52)
KICH
P = 0.01**
High sML (n=24)
Low sML (n=35)
CRC-CMS2
P = 0.004**
High sML (n=77)
Low sML (n=68)
HNSC HPV negative
P = 0.02*
High sML (n=46)
Low sML (n=300)
KIRC
P < 0.001***
High sML (n=205)
Low sML (n=111)
KIRP
P = 0.01**
High sML (n=154)
Low sML (n=86)
LGG IDH1 mutated
P = 0.01**
High sML (n=263)
Low sML (n=46)
LIHC
P = 0.08
High sML (n=101)
Low sML (n=207)
LUAD smoker
P = 0.004**
High sML (n=69)
Low sML (n=66)
LUSC smoker
P = 0.02*
High sML (n=40)
Low sML (n=86)
OV
P = 0.02*
High sML (n=106)
Low sML (n=251)
PAAD
P = 0.003**
High sML (n=77)
Low sML (n=68)
PCPG
P = 0.07
High sML (n=29)
Low sML (n=33)
PRAD - All
P < 0.001***
High sML (n=195)
Low sML (n=202)
PRAD Gleason score = 7
P = 0.1
High sML (n=102)
Low sML (n=105)
PRAD Gleason score > 8
P = 0.01**
High sML (n=23)
Low sML (n=134)
SARC
P = 0.004**
High sML (n=46)
Low sML (n=122)
SKCM
P = 0.02*
High sML (n=288)
Low sML (n=67)
THYM
P < 0.001***
High sML (n=61)
Low sML (n=12)
UCEC endometrioid
P = 0.03*
High sML (n=82)
Low sML (n=160)
Progression−free Probability
Overall Survival
Years

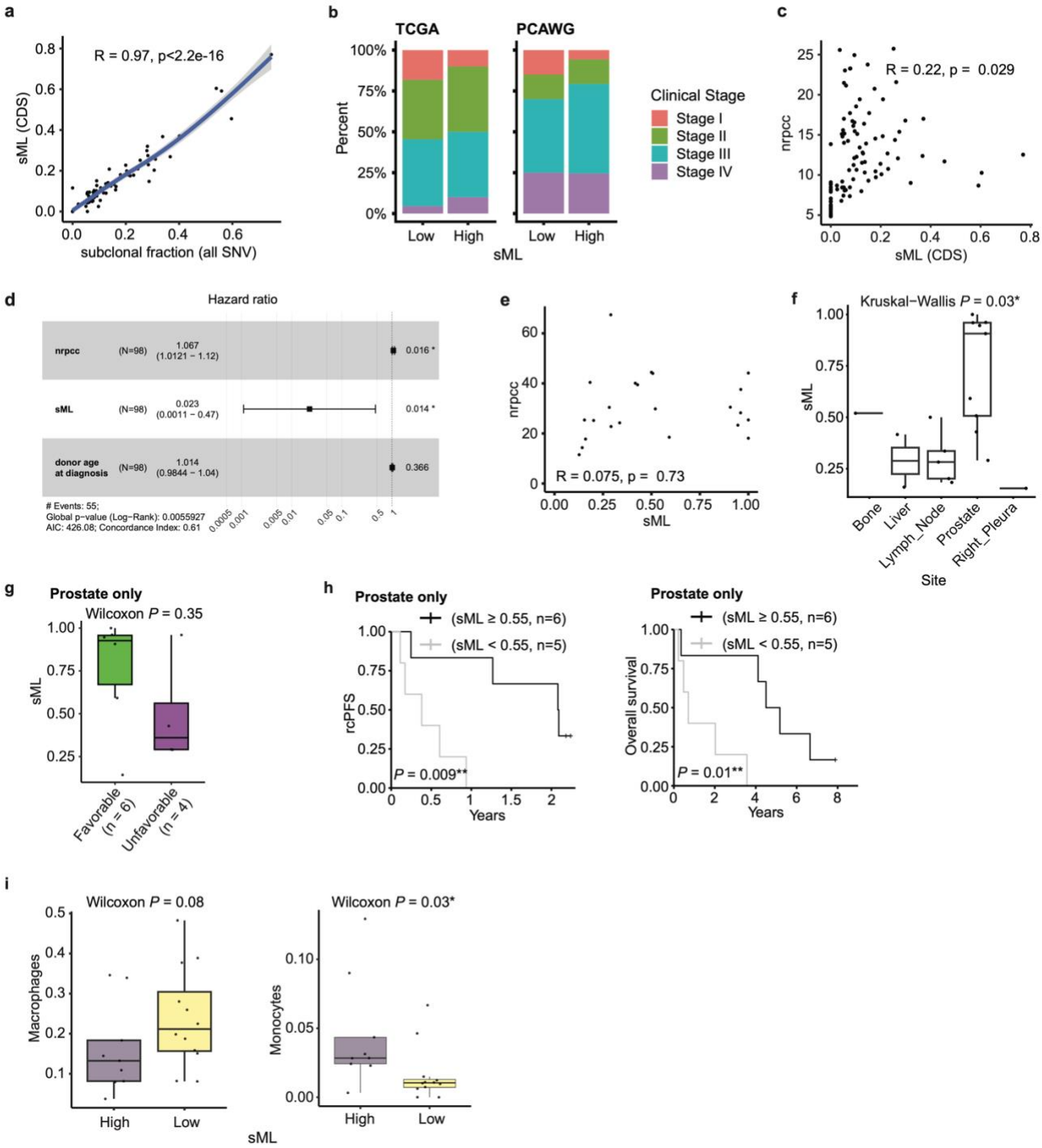

d
Genome Instability
THYM
PRAD
LGG
KIRC
PCPG
UCEC
KIRP
LIHC
PAAD
CRC
SKCM
HNSC
KICH
SARC
LUSC
LUAD
OV
ACC

Extended Data Fig. 5 (cont'd): (**a**) Boxplots comparing sML between clinically utilized subtypes and cancer phenotypes. *P*-values for pairwise comparisons were calculated using Wilcoxon rank-sum test. (**b**) Kaplan-Meier curves of progression-free interval (PFI, n=7,507) and overall survival (OS, n=7,570) for the entire pan-cancer cohort of TCGA samples. Dark green lines indicate patients with high sML, and light purple lines indicate patients with low sML. Log-rank test *P*-values between high- and low-sML groups are shown. (**c**) Kaplan–Meier curves of OS or PFI for cancer types/subtypes shown in **Fig. 6a**. Black lines denote high sML and grey lines denote low sML. *P*-values of log-rank tests between high- versus low-sML patient groups are shown. **(d)** Distributions of genome instability (GI) scores provided by ASCAT, a measure of the proportion of the genome displaying copy number aberrations, across cancer types featured in Fig. 5a.



Extended Data Fig. 6: Subclonal diversity associated with subclonal mutational load. (**a**) Mathematical relationship between Shannon Index and sML using the simple case of one subclone as illustration. Black dashed lines demonstrate that at one subclone despite sharing the same Shannon Index (SI), samples can exhibit different sML values. The red dashed line denotes the highest SI score and the corresponding sML value of 0.5. (**b**) & (**c**) Pairs of scatter plots

displaying the relationship between Shannon Index and sML and KM plots. Gray dots indicate low sML, while the red dots indicate **Extended Data Fig. 6 (cont'd):** high sML with high Shannon index, and the pink dots represent high sML with low Shannon index. Additionally, for each cancer type, using the same cutoffs shown in the scatter plots and the same color scheme, KM curves are shown illustrating Progression-Free Interval (PFI) or overall survival (OS) differences between the three groups. Log-rank test *P*-values are provided for comparing the patient groups. Cancer types are grouped as follows: (**b**) Cancer types where high sML and low Shannon index present most favorable outcomes (BLCA, HPV negative HNSC, KICH, *IDH1* mutated LGG, PRAD with Gleason Score 7, THCA). (**c**) Cancer types where high sML and high Shannon index present most favorable outcomes (CRC, KIRP, PAAD, STAD, UCEC).



Extended Data Fig. 7: Analysis of the non-TCGA validation cohorts. (a) Scatter plot demonstrates a strong correlation between sML (on coding region) and subclonal fraction (on all SNVs) in PCAWG ESAD dataset, suggesting that either estimate can be used in downstream analyses. For consistency with the TCGA results, however, which exclude poorly sequenced SNVs outside of exons by focusing solely on the coding region, we use CDS sML for all

subsequent analysis of the PCAWG ESAD samples. (**b**) Stacked bar chart showing comparable distributions of clinical/pathological stages between the 34 TCGA and 98 PCAWG samples from esophageal adenocarcinoma. **(c)** Scatter plot of nrpcc versus sML in **Extended Data Fig. 7 (cont'd):** PCAWG-ESAD. **(d)** Forest plot of a Cox regression model showing sML remains significant after adjusted by nrpcc. **(e)** Scatter plot of sML versus nrpcc in the mCRPC samples. **(f)** Distribution of sML among different tumor sites within the mCRPC cohort. sML level is significantly different between different sites. **(g)** Boxplots showing the distribution of sML of prostate tissue samples within the favorable and unfavorable group. Kruskal-Wallis *P*-value is shown. **(h)** Kaplan-Meier curves showing comparing patient survival outcomes when stratifying patients using only prostate tissue derived samples, with high sML still showing significantly better survival (rcPFS & OS) among this reduced sample set.. Log-rank test *P*-values are shown. **(i)** Boxplots showing the distributions of selected immune cell types (macrophages & monocytes), with estimated cell proportions from CIBERSORTx (LM22 as reference), between samples with high and low sML (n=9 vs. 12).

Supplementary Tables – All supplementary tables are provided as separate excel files, and are additionally available at https://drive.google.com/drive/folders/1Lz-wP_wt_h8tHdetjGLZTiHwNEhS31zq

Supplementary Table 1. Clonal status of driver mutations for 16 TCGA cancer types.

Supplementary Table 2. Log-rank tests for cancer-type driver gene pairs with more than 10 subclonal mutations.

Supplementary Table 3. Sample size and number of survival events among 30 cancer types.

Supplementary Table 4. Summary of clinical outcomes and cutoffs used to dichotomize sML as a biomarker.

Supplementary Table 5. Multivariate Cox proportional hazard models with age, sex, and dichotomized sML as candidate predictors for OS and PFI across cancer types in TCGA and PCAWG-ESAD

Supplementary Table 6. Multivariate Cox proportional hazard models with age, sex, and continuous sML as candidate predictors for OS and PFI across cancer types in TCGA and PCAWG-ESAD

Supplementary Table 7. Multivariate Cox proportional hazard models with age, sex, continuous sML, TMB, and sML x TMB as candidate predictors for OS and PFI across cancer types in TCGA and PCAWG-ESAD.

Supplementary Table 8. CliPP performance benchmarking results in CliPPSim4k.