




TECHNICAL NOTE

MesKit: a tool kit for dissecting cancer evolution of multi-region tumor biopsies through somatic alterations

Mengni Liu^{1,2,†}, Jianyu Chen^{1,†}, Xin Wang¹, Chengwei Wang¹, Xiaolong Zhang², Yubin Xie¹, Zhixiang Zuo ², Jian Ren ^{1,2,*} and Qi Zhao ^{2,*}

¹School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, China and ²State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, 651 E Dongfeng Road, Guangzhou, Guangdong 510060, China

*Correspondence address. Jian Ren, School of Life Sciences, Sun Yat-Sen University, Guangzhou 510006, Guangdong, China. E-mail: renjian.sysu@gmail.com  <http://orcid.org/0000-0002-4161-1292>; and Qi Zhao, Sun Yat-sen University Cancer Center, 651 E Dongfeng Road, Guangzhou 510060, China. Tel/Fax: +86-20-87342209; E-mail: zhaoyi@sysucc.org.cn  <http://orcid.org/0000-0002-8683-6145>

[†]Contributed equally.

Abstract

Background: Multi-region sequencing (MRS) has been widely used to analyze intra-tumor heterogeneity (ITH) and cancer evolution. However, comprehensive analysis of mutational data from MRS is still challenging, necessitating complicated integration of a plethora of computational and statistical approaches. **Findings:** Here, we present MesKit, an R/Bioconductor package that can assist in characterizing genetic ITH and tracing the evolutionary history of tumors based on somatic alterations detected by MRS. MesKit provides a wide range of analysis and visualization modules, including ITH evaluation, metastatic route inference, and mutational signature identification. In addition, MesKit implements an auto-layout algorithm to generate phylogenetic trees based on somatic mutations. The application of MesKit for 2 reported MRS datasets of hepatocellular carcinoma and colorectal cancer identified known heterogeneous features and evolutionary patterns, together with potential driver events during cancer evolution. **Conclusions:** In summary, MesKit is useful for interpreting ITH and tracing evolutionary trajectory based on MRS data. MesKit is implemented in R and available at <https://bioconductor.org/packages/MesKit> under the GPL v3 license.

Keywords: multi-region sequencing; somatic alterations; intra-tumor heterogeneity; metastatic routes; phylogenetic tree

Introduction

Cancer evolves through a process of somatic alterations [1], of which spatial and/or temporal changes can be detected by multi-region sequencing (MRS). Currently, MRS has become an effective and affordable way to trace the evolutionary history of carcinogenesis and metastasis. Cancer evolution research is focused on the identification and estimation of intra-tumor het-

erogeneity (ITH), phylogenetic reconstruction, mutational signature analysis, and so forth. Numerous MRS studies have identified extensive ITH among many solid tumors originating in the liver, prostate, esophagus, breast, and lung [2–7]. In addition, increased ITH has been implicated in dismal cancer prognosis [8–10]. While recent studies have largely generated descriptive summaries of ITH, a quantitative understanding of the hetero-

Received: 8 January 2021; Revised: 23 March 2021; Accepted: 23 April 2021

© The Author(s) 2021. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

geneity within and between tumors from the same patient is more informative for personal therapeutics.

Recently, plenty of MRS studies have used phylogenetic trees to show the temporal sequence and heterogeneous divergence between samples [2, 11, 12]. There are also increasing efforts to reconstruct subclonal phylogenies via a “clone tree,” which summarizes lineage relationships between cellular subpopulations [13–15]. Phylogenetic reconstruction over the cancer cell fraction (CCF) estimates has identified both monoclonal and multiclonal seeding patterns in several cancers [3, 16–18]. The distinction between these 2 patterns may have important clinical implications [19]; it is thus necessary to infer metastatic routes and to explore potential metastasis drivers.

Moreover, MRS provides insights into the dynamics of mutational processes during tumor progression. A previous study indicated that DNA damage repair dysfunction might be crucial for mutation accumulation during osteosarcoma evolution [20]. Recently, Yan et al. [21] performed MRS of tumors from 39 patients with esophageal squamous cell carcinoma and identified several potential actionable targets, such as *EGFR* and *FGFR1*. They also showed that APOBEC mutations and aging predominated in the early stage of tumorigenesis of esophageal squamous cell carcinoma. These findings suggest that the MRS strategy has the potential to reveal mutational mechanisms and thereby could improve both diagnosis and treatment.

The downstream analysis of MRS data focuses on somatic alterations, including somatic single-nucleotide variants (sSNVs), small insertions and deletions (INDELs), and copy number alterations (CNAs). At present, many tools are available to analyze somatic alterations, which has greatly promoted the development of cancer genomics. For example, Maftools [22] provides multiple functions for pathway annotation and *de novo* signature and enrichment analysis. MutationalPatterns [23] and deconstructSigs [24] are powerful tools for exploring mutational patterns and identifying mutational signatures of a single tumor sample. Besides, Palimpsest [25] enables the identification of different mutational signatures between clonal and subclonal mutations. In addition, lots of methods infer tumor heterogeneity by assessing the complex subclonal structure of tumors. Approaches such as SciClone [26], PhyloSub [27], and PyClone [28] are based solely on point mutations (sSNVs or INDELs), while SCHISM [29], DPclust [30], and PhyloWGS [31] adjust for CNAs in their models in different ways. In general, high-depth sequencing improves the accuracy of subclonal reconstruction and resolution [32]. However, performing integrated mutational analysis of MRS using these tools is inconvenient because different preprocessing steps and inconsistent input formats are required. On the other hand, it is laborious and time-consuming to generate publication-quality images such as mutational profiles and phylogenetic trees, which necessitates manual modifications using extra graphic editors.

To address these concerns, we present MesKit, an R/Bioconductor package that provides commonly used analysis and visualization modules for MRS studies. MesKit was designed as an easy-to-use R package that only requires a MAF file and a clinical file as inputs, enabling researchers to evaluate the contribution of point mutations to heterogeneity within/between tumors from the same patient. MesKit can also be used to depict mutational profiles, track evolutionary dynamics, and characterize mutational patterns at different levels. Notably, we implemented an auto-layout algorithm to visualize rooted phylogenetic trees with annotations. In addition, MesKit enables easy integration and analysis of segmentation data and CCF data and a Shiny application is provided to

facilitate interactive analysis. Finally, we applied MesKit on 2 high-quality MRS datasets of hepatocellular carcinoma (HCC) [2] and colorectal cancer (CRC) [12] (Supplementary Table S1). We reproduced well-known heterogeneous features and evolutionary patterns, together with potential driver events of HCC and CRC, demonstrating the robustness of MesKit in interpreting ITH and for inferring evolutionary trajectories based on MRS data.

Materials and Methods

Data collection and preprocessing

We used 2 cohorts in our analysis. The HCC cohort included tumor tissue ($n = 52$) and matched blood (germline, $n = 11$) samples from 11 patients, which were collected before treatment [2]. All samples were sequenced using whole-exome sequencing (mean depth of 158 \times) and re-analyzed with a uniform pipeline described below. In brief, we performed sSNV calling for each tumor/normal pair with Mutect (version 1.1.7) [33], while INDELs were detected with Strelka v2.7.1 [34]. Additionally, we adopted the “force calling” method [35] to rescue potential real mutations for each sample based on the aggregate set of somatic events in each patient using Samtools mpileup (version 1.2) [36]. Both sSNVs and INDELs were annotated through ANNOVAR (v.20191024) [37]. The following filters were further applied to identify the sSNVs and INDELs: (i) Mutations with <15 total reads or 5 variant reads were discarded. (ii) Mutations listed in dbSNP147 were removed unless they were documented in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. (iii) Mutations listed in the National Heart, Lung, and Blood Institute Exome Sequencing Project were removed. Copy number analysis based on exome sequencing was performed using Sequenza v3.0.0 [38]. Segments smaller than 500 kb were filtered and only autosomes were used in copy number analysis. CCFs of mutations were estimated by PyClone (v0.13.0) [28], which adjusted the variant allele frequencies (VAFs) of somatic mutations on the basis of local copy numbers of the mutated loci and tumor purity (Sequenza v3.0.0) [38]. The second cohort (the CRC cohort) comprised 6 patients processed with MRS for paired primary tumors and metastases (3–5 regions each) [12]. We obtained somatic mutation calls for sSNVs and INDELs, copy number segment data, and CCF estimates of mutations from the original study. Driver genes of HCC and CRC were defined by IntOGen (v.2020.2). The GISTIC2 results of The Cancer Genome Atlas (TCGA) HCC and TCGA CRC projects were obtained from the Broad Genome Data Analysis Center (GDAC, <http://gdac.broadinstitute.org>) repository (analysis stamp: 2016.01.28).

Clonal status of somatic mutations

Because multiple samples collected from a single tumor collectively reflect its clonal composition, these regions should be considered as a whole to capture the overall tumor dynamics. Here, we assume that each tumor follows neutral exponential growth in a well-mixed population [39, 40]. When MRS data are available, the merged CCF (CCF_{merged}) of each mutation is computed by integrating multiple regions as previously described [12, 41, 42]:

$$CCF_{\text{merged}} = \begin{cases} \frac{\sum_{i=1}^k CCF_i \times d_i}{\sum_{i=1}^k d_i} & CCF < 1 \\ 1 & CCF \geq 1 \end{cases} \quad (1)$$

where d_i and CCF_i refer to the sequencing depth and CCF estimation in region i , respectively. The clonal status of sSNVs/INDELS are determined based on CCFs. A CCF value of 1 indicates that the mutation is present in 100% of the cancer cells in a sample, while a CCF value < 1 indicates that the mutation is present in a subset of the cancer cells in a sample and thus is subclonal. In each sample, a mutation is classified as clonal when the upper bound of the 95% confidence interval (CI) of the CCF is ≥ 1 and subclonal otherwise [43]. For MRS data, a mutation is considered subclonal when all of the following criteria are satisfied: (i) ≥ 1 region with upper bound of 95% CI of the CCF < 1 , (2) ≥ 1 region with CCF < 0.5 , and (iii) CCF_{merged} of mutation $m < 0.5$ (the cut-off was chosen for its good performance in defining subclonality in simulated virtual tumors [44, 45]).

Estimation of ITH

MesKit includes several measures of ITH defined by recent genomic studies. For a single region/tumor, it is common to infer subpopulations of tumor cells by clustering VAFs or CCFs [26, 30]. To implement this process, we used Gaussian finite mixture models for 1D clustering of VAFs or CCFs using mclust R package [46]. Because copy number gains and losses can alter the fraction of reads bearing a mutation, we only focused on heterozygous mutations within copy number-neutral and loss-of-heterozygosity-free regions when clustering VAFs. More recently, Mroz et al. [47] developed the mutant-allele tumor heterogeneity (MATH) index, which corresponds to the ratio of the median absolute deviation (MAD) and the median of the VAF values among tumor-specific mutated loci. Generally, a more heterogeneous tumor with a higher MATH score tends to have a wider distribution of VAFs among all mutation loci and centers at a lower fraction.

$$MATH = 1.4826 \times \frac{MAD(VAF)}{\text{Median}(VAF)}. \quad (2)$$

Another approach to estimate ITH is calculating the area under the curve (AUC) of the cumulative density function based on the CCFs per tumor, and tumors with higher AUC values are considered to be more heterogeneous [48]. Moreover, to quantify the genetic divergence of ITH between regions or tumors, we introduced 2 classical metrics derived from population genetics, Wright fixation index (F_{ST}) [49] and Nei genetic distance [50]. Calculations of between-region genetic heterogeneity within tumors only consider subclonal mutations because clonal mutations present in all regions do not contribute to ITH. For pairwise comparisons of heterogeneity between tumors, both clonal and subclonal mutations were taken into consideration. The F_{ST} index estimating between-region ITH for k regions was computed as described previously [44]:

$$F_{ST} = \frac{1}{r} \times \sum_{j=1}^r F_{STj}^{\text{Hudson}}, \quad r = \binom{k}{2}, \quad (3)$$

$$F_{STj}^{\text{Hudson}} = \frac{\sum_{m=1}^{m_t} (f_a^m - f_b^m)^2 - \frac{f_a^m \times (1 - f_a^m)}{d_a^m - 1} - \frac{f_b^m \times (1 - f_b^m)}{d_b^m - 1}}{\sum_{m=1}^{m_t} f_a^m \times (1 - f_b^m) + f_b^m \times (1 - f_a^m)}, \quad (4)$$

where m_t represents the total number of sSNVs in regions a and b , f_a^m denotes the VAF for sSNV m , and d_a^m denotes the sequencing depth for sSNV m in region a .

The Nei genetic distance for k regions within the same tumor was defined as follows [50]:

$$D_{\text{Nei}} = \frac{1}{r} \times \sum_{j=1}^r D_{\text{Nei}j}, \quad r = \binom{k}{2}, \quad (5)$$

$$D_{\text{Nei}j} = -\log \frac{\sum_{m=1}^{m_t} ccf_a^m \times ccf_b^m + (1 - ccf_a^m)(1 - ccf_b^m)}{\sqrt{[\sum_{m=1}^{m_t} ccf_a^{m2} + (1 - ccf_a^m)^2] \times [\sum_{m=1}^{m_t} ccf_b^{m2} + (1 - ccf_b^m)^2]}} \quad (6)$$

where m_t represents the total number of sSNVs in regions a and b . ccf_a^m and ccf_b^m represent the CCF values in region a and region b for mutation m , respectively.

Inference of metastatic routes

For spatially separated lesions from the same patient, the potential metastatic route can be determined by comparing subclonal architecture between paired lesions. Here, MesKit integrated a Jaccard similarity index (JSI)-based method to identify seeding patterns based on the CCFs of sSNVs for paired lesions [42]. The Jaccard coefficient for a lesion pair (a , b) is calculated as follows:

$$JSI = \frac{SS_{ab}}{PC_a + PC_b + SS_{ab}}, \quad (7)$$

where SS_{ab} and PC_a/PC_b represent shared subclonal sSNVs of lesion pair (a , b) and private clonal sSNVs of lesion a/b , respectively. The mean SS_{ab} , and PC_a/PC_b of all sample pairs from lesion a and lesion b are used to compute the JSI for lesions with MRS data.

Construction and visualization of phylogenetic trees

MesKit reconstructs the phylogeny of multiple specimens from individual patients on the basis of the presence or absence of somatic mutations. This process is implemented in getPhyloTree function via utilization R implementations of several standard phylogenetic approaches from the APE [51] and PHANGORN [52] R packages, including distance-based methods (neighbor-joining [NJ] [53] and minimum evolution [54]), as well as character-based methods (maximum parsimony [MP] [55] and maximum likelihood [ML] [56]). Notably, we implemented an auto-layout algorithm via the plotPhyloTree function to generate customizable images of phylogenetic trees with annotations (Supplementary File S1). Furthermore, by using the treedist function from the PHANGORN [52] R package, MesKit enables the comparison of phylogenetic trees constructed by different methods via the compareTree function.

Mutational signature analysis

To illustrate the dynamic mutational spectrum during tumor progression, we implemented mutational signature analysis based on phylogenetic trees. The process starts with the construction of a mutation matrix accounting for 96 trinucleotide changes, where the sequence context of the base substitutions can be retrieved from the corresponding reference genome using the BSgenome R package. Six types of base substitution types are distinguished by convention: C>A, C>G, C>T, T>A, T>C, and T>G. As methylated cytosine at CpG sites with the attendant risk of spontaneous deamination are mutagenic hot spots in the human genome [57], C>T mutations can be divided into C>T at CpG sites and other sites [23]. Genomic mutations are temporally dissected into truncal (shared among all samples from the same patient) and branch mutations of phylogenetic trees. For each mutational type, the Fisher exact test is implemented to assess the difference between the truncal and branch mutations. Once the signature matrix is provided, the fitSignatures function estimates the optimal contributions of known signatures to reconstruct a mutational profile, which minimizes the residual sum of squares (RSS) between the original and reconstructed muta-

tional profiles. This process was implemented by integrating a non-negative least-squares algorithm using the `pracma` R package [77], as previously described in `MutationalPatterns` [23]. For convenience, we included known signature matrices (published by Alexandrov et al. [73] and `Cosmic` version 2, 3) along with the proposed etiology in `MesKit`. The similarity between mutational profiles A and B is calculated by cosine similarity as follows:

$$\text{sim}(A, B) = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}, \quad (8)$$

where mutational profiles A and B are non-zero vectors with n mutational types. Cosine similarity value can be used to test how well each mutational profile can be explained by the provided mutational signatures. Two mutational profiles are identical when the cosine similarity is 1 and are independent when the cosine similarity is 0.

Results

Overview of MesKit functions and implementation

`MesKit` was implemented as an open source R/Bioconductor package. With a MAF file and a clinical data file as standard inputs, `MesKit` provides a series of analysis and visualization functions to interpret mutational data from MRS experiments (Fig. 1). In addition, we implemented a Shiny application to facilitate the use of the package. Moreover, we built a Docker image that enables the deployment of the Shiny-based `MesKit` GUI in a C/S mode.

Mutational landscape of MRS studies

Generally, somatic mutations identified from MRS in a single tumor are classified as “public mutations” (existing in all regions of the tumor), “shared mutations” (existing in part of all regions), or “private mutations” (existing in a single region) [20, 44, 58]. Such spatial-mutation categories largely correspond to the temporal order of mutation genesis during tumor evolution: most public mutations occur early in tumor-initiating cells and are inherited by their offspring, whereas private mutations accumulate sporadically and markedly increase the ITH among different patients [59]. In `MesKit`, we implemented the `classifyMut` function to help categorize somatic mutations based on regional distribution, and/or to identify clonal and subclonal mutations according to their estimated CCFs (Methods). Analysis of the HCC and CRC cohorts showed significant inter-individual heterogeneity but much less intra-individual heterogeneity (Fig. 2 and Supplementary Fig. S1). In line with previous findings [13, 60, 61], the primary tumors and metastases of the CRC cohort exhibited high genomic concordance (Fig. 2A). As expected, public mutations harbor higher CCFs than private mutations (Supplementary Fig. S2), which were more likely to be clonal events. Recurrent mutations in putative driver genes of CRCs (defined by `IntOGen v.2020.2`), such as *KRAS* and *APC*, were clonal and shared between paired primary tumors and metastases, indicating their early occurrence in colorectal carcinogenesis (Fig. 2A and Supplementary Fig. S2). Interestingly, heterozygous *BRCA2* mutations were private to distant metastases, including the lung metastases (LU) and brain metastases (BM) of 2 patients (V824 and V930), while there is currently no strong evidence that shows that *BRCA2* mutations are associated with CRC metastasis. In addition, the `plotCNA` function of `MesKit` can be used to characterize the CNA landscape across samples on the basis of copy number data. Consistent with TCGA projects and other

previous studies of HCC [62, 63], a number of copy number alterations were observed in our HCC cohort, such as gains of 1q, 6p, 8q, and 13q, as well as losses of 1p, 4q, 9q, and 11q (Fig. 2B). Taken together, these data suggest that `MesKit` can easily characterize the mutational landscape and potential driver genes during cancer evolution.

ITH estimation

Understanding the degree and development of ITH is clinically important because ITH has been associated with treatment resistance and the prognosis of patients with cancer [64]. `MesKit` integrates several approaches to estimate ITH within and between regions/tumors from the same patient. In `MesKit`, the `mutCluster` function deduces distinct subpopulations of a sample/tumor by clustering VAFs/CCFs in 1 dimension based on Gaussian finite mixture models [46]. It should be noted that low-frequency clusters might be a mixture of subclones that contain mutations coming from numerous parallel lineages growing neutrally [32, 65]. Another approach is calculating MATH score, which is positively correlated with tumor heterogeneity and metastatic potential [66, 67]. Besides, we integrated an index described by Charoentong et al. [48], to assess ITH by calculating the AUC of the cumulative density function from all CCFs per sample/tumor. Samples/tumors with higher AUCs are considered to be more heterogeneous than those with lower AUCs. Applying these measures on HCC8010 showed that samples with wider distributions of VAFs tended to have higher MATH scores, and VAF-based ITH was comparable to that calculated by CCFs (Fig. 3A and B). Moreover, we introduced 2 measures from population genetics [44, 49, 50], named F_{ST} and Nei genetic distance, to enable pairwise comparisons between regions/lesions. Comparison of ITH between primary tumors and paired metastases in the CRC cohort showed no significant difference using these 2 indices (Wilcoxon signed-rank test, F_{ST} : $P = 0.5781$, Nei distance: $P = 0.1094$, Fig. 3C). Similarly, this observation supports the conclusion that primary and metastatic tumors of CRC exhibit a high degree of mutational discordance.

Inferring the clonality of metastatic seeding

Because metastasis is the major cause of cancer-related death, it is particularly important to gain a systematic understanding of how tumor cells disseminate and the scale of ongoing parallel evolution in metastatic and primary sites [68]. Given that mutations with similar CCFs tend to cluster into the same subpopulation [69, 70], many studies inferred the potential metastatic routes between different lesions from the same patient by plotting CCFs of mutations [12, 42, 71]. By this means, Xue et al. [71] identified both monoclonal and multiclonal origins of separate type combined hepatocellular and intrahepatic cholangiocarcinoma (cHCC-ICC). Here, we developed the `compareCCF` function to calculate the merged CCFs of distinct lesions with MRS data. To visualize the seeding patterns between lesions in a more intuitive way, the results of this function can be further used to plot CCF plots, where the clusters at (1, 1) correspond to the clonal mutations present in all cells in both lesions (CCF = 1), while those on axes refer to lesion-private subclones. In addition, `MesKit` integrated a JSI-based method to calculate mutational similarity between lesions [42]. Pairs following polyclonal seeding generally achieve higher JSI values because of their higher proportion of shared subclonal sSNVs and fewer lesion-private sSNVs (Methods). Analysis of the CRC cohort with these functions revealed that all BMs exhibited enrichment of metastasis-

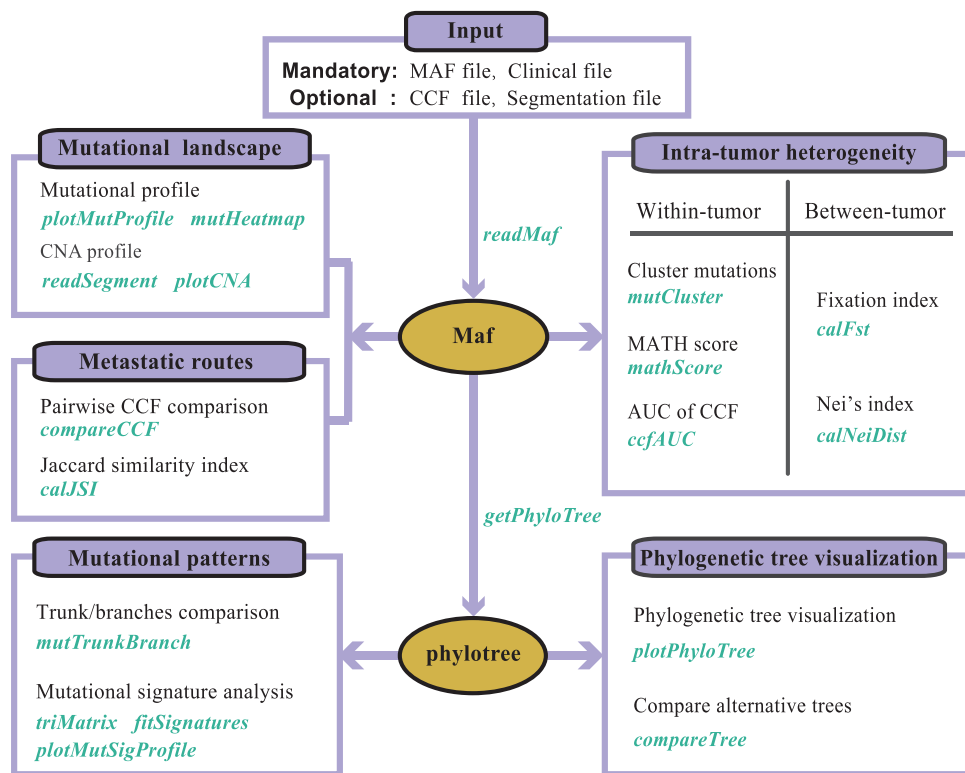


Figure 1: Overview of the MesKit package A. Overview of MesKit. MesKit consists of 5 major modules: characterizing mutational landscape, estimating ITH, inferring metastatic routes, exploring mutational patterns, and visualizing phylogenetic trees automatically. Corresponding functions for each module are displayed separately.

private clonal sSNVs and shared clonal sSNVs but lacked shared subclonal sSNVs (Fig. 3D). Moreover, all BMs comprised a single phylogenetic clade in the phylogenetic trees (Supplementary Fig. S4). These observations jointly indicated that the BMs of this CRC cohort followed a monoclonal seeding manner, consistent with the original study [12]. Besides, in both paired primaries and metastases of most CRCs, the merged CCFs of mutations in CRC driver genes including APC, KRAS, and TP53 were >0.6 , suggesting that they may contribute to CRC tumorigenesis and metastasis. Notably, lymph nodes showed higher JSI values than distant metastases in V750 and V824, indicating that polyclonal seeding was more prevalent in lymph node metastases (Fig. 3D). In summary, these results demonstrated the ability and efficiency of MesKit to identify distinct patterns of seeding between paired lesions.

Construction and visualization of phylogenetic trees

A systematic understanding of the evolutionary relationships among tumor regions from a single patient plays a fundamental role in MRS studies, with the phylogenetic tree being a primary tool for delineating the relationship between tumor regions and interpreting ITH [2, 11, 44]. Consistent with original studies, we applied the MP method to reconstruct the tumor phylogeny of the CRC cohort using the *getPhyloTree* function in MesKit. Phylogenetic trees were further visualized with the function *plotPhyloTree*, which provides options to color the branches according to the classification of mutations or putative known signatures. We consistently reproduced tree structures of most CRCs from the original study [12], in which the primary regions and metastatic regions were clearly separated (Fig. 4). Inspection of the phylogeny indicated early divergence

of the metastatic lineage in V402, V824, V930, V953, and V974, whereas divergence occurred during diversification of the primary tumor in V750. Moreover, we compared the MP-based phylogenetic trees with those constructed by the NJ method and ML method for each patient with CRC. Phylogenetic trees inferred through the 3 methods shared the same topology and clades for V402, V924, and V953 (Supplementary Fig. S3). When considering branch lengths, the MP-based trees were more similar to the NJ-based trees than the ML-based trees according to KF-branch distance [72] and weighted RF distance [74] (Supplementary Table S2). Collectively, these results demonstrate the functionality and efficiency of MesKit for analyzing and visualizing tumor phylogeny.

Temporal dissection of mutational signatures

Analysis of mutational signatures can be used to understand the mechanisms of transformation of normal cells to malignant cells and to identify underlying risk factors for tumor development. First, Alexandrov et al. [73] utilized $>7,000$ cancer genomes and exomes to identify 21 signatures across 30 tumor types. More recently, the Wellcome Trust Sanger Institute [78] published 30 mutational signatures (Version 2) in primary cancer and an expanded 67 single-base substitution signatures (Version 3). Considering the limited number of tumor samples assessed by MRS and thus the limited number of identified mutations, it is not amenable to conduct *de novo* signature extraction. Therefore, we developed the *fitSignatures* function to calculate the contribution of well-established signatures to mutations at different levels. By reconstructing the mutational profiles of the HCC and CRC cohorts using 30 COSMIC mutational signatures, we demonstrated that the signature contributions estimated by

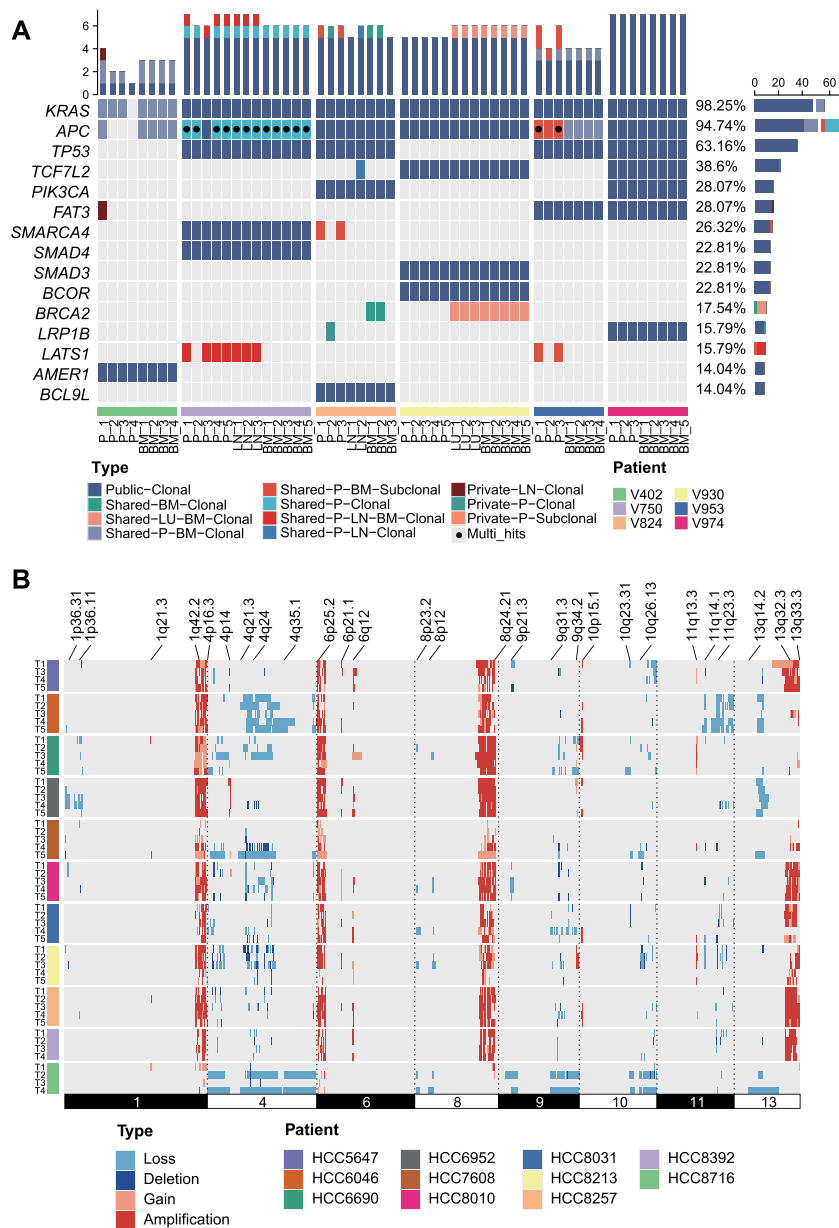


Figure 2: Mutational landscape of the HCC and CRC cohorts **A.** Mutational profile of the CRC cohort. The oncoprint of the top 15 most frequently mutated driver genes of CRC grouped by public, shared, or private mutations, including both clonal and subclonal drivers. Genes were sorted by mutational frequency, and those with multiple mutations were annotated as Multi-Hit. Samples were split by patients as indicated by the annotation bar (bottom). BM: brain metastasis; LN: lymph node metastasis; LU: lung metastasis; P: primary tumor. The stacked bar charts on the top and right show the number of different types of mutations per sample and per driver gene, respectively. **B.** The consistent CNAs of the HCC cohort with significant recurring CNAs were identified from the TCGA hepatocellular carcinoma project by GISTIC2.0 (obtained from the Broad GDAC website). Each track represents 1 tumor sample. Dark red indicates amplifications ($CN \geq 4$), light red indicates gains ($2 < CN < 4$), dark blue indicates deletions ($CN = 0$), and light blue indicates losses ($0 < CN < 2$).

fitSignatures function were highly similar to those calculated by 3 other signature deconvolution tools (average Pearson correlation: 1, MutationalPatterns [23]; 0.997, SignatureEstimation [74]; 0.948, deconstructSigs [24]) (Supplementary Fig. S4A and Table S3). The similarities (indicated by cosine similarity) and discrepancies (indicated by RSS) between the original and reconstructed mutational profiles generated with MesKit were also comparable to those generated from other tools (Supplementary Figure S4B). As shown in Fig. 5B, hierarchical clustering via Euclidean distance of the patients based on their cosine similarity values clearly separated the HCCs from the CRCs. These re-

sults demonstrate the ability of MesKit to reliably estimate signature contributions. We further applied the fitSignatures function with 30 COSMIC signatures to truncal and branch sSNVs of HCC5647, HCC7608, and HCC8716 (other HCCs were excluded because their truncal/branch sSNVs were < 50). All 3 HCCs exhibited a prominent decrease of the contribution of Signature 22 (exposures to aristolochic acid) in branch mutations compared with truncal mutations (Fig. 5A and Supplementary Table S4). Among them, HCC5647 and HCC8716 showed significantly higher percentages of T>A ($P < 0.01$) in truncal mutations than branch mutations (Supplementary Fig. S5), which is consistent

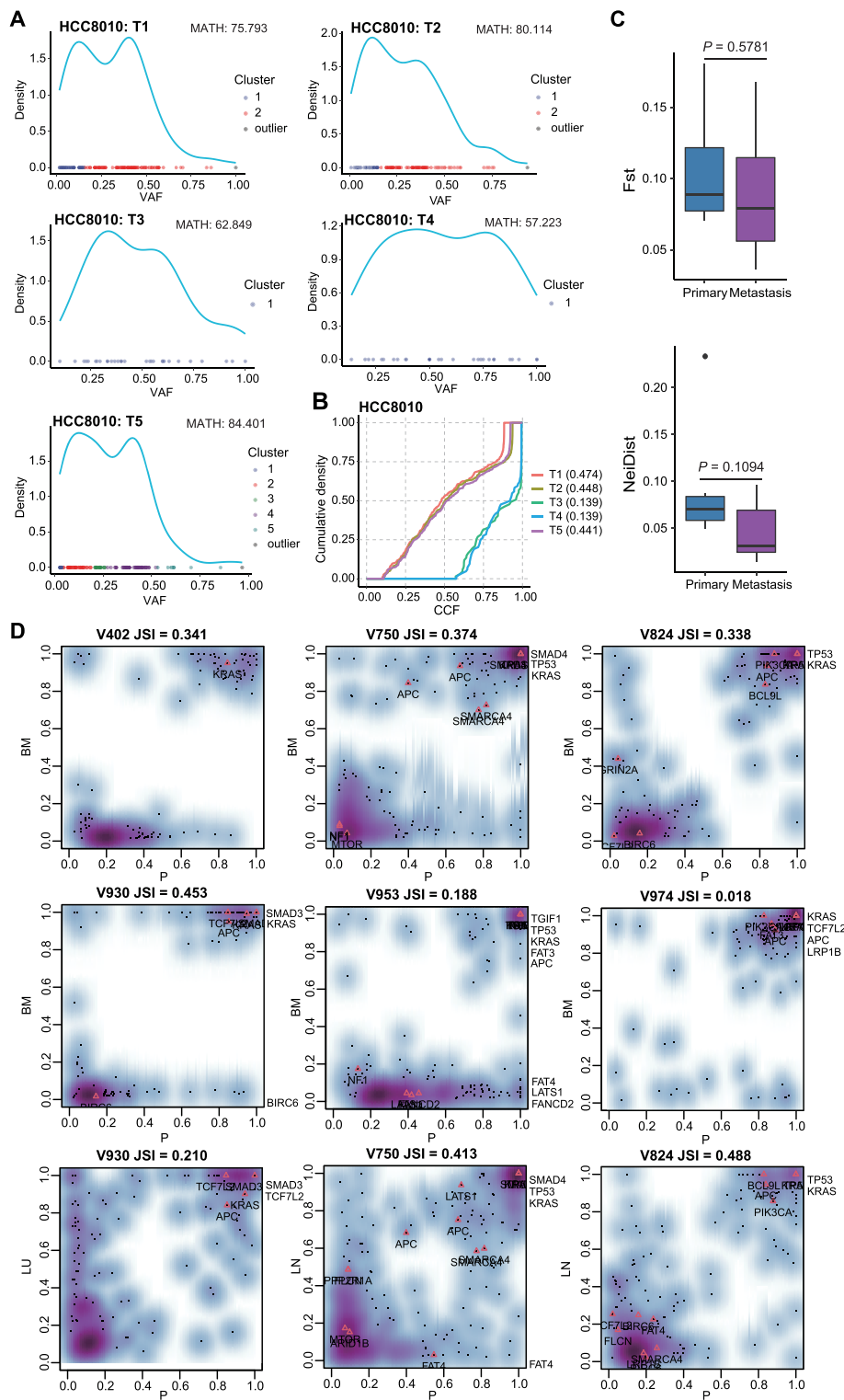


Figure 3: ITH estimation and the clonality of metastatic seeding **A.** Clustering mutations by VAFs of each tumor sample from HCC8010 based on a Gaussian finite mixture model. MATH scores are indicated above. **B.** CCF density plot of tumor samples from HCC8010. **C.** F_{ST} - and Nei distance-based quantification of ITH in paired primary tumors and metastases of the CRC cohort ($n = 7$). P-value, Wilcoxon rank-sum test (2-sided). **D.** Density plots of merged CCF values in paired primary tumors and metastases of the CRC cohort. For each pair, the JSI was computed according to equation (7). Putative CRC driver genes are indicated on the plot. BM: brain metastasis; LN: lymph node metastasis; LU: lung metastasis; P: primary tumor.

with the characteristic patterns of Signature 22 (characterized by $T > A$). Considering these observations, we hypothesized that exposure to aristolochic acid contributed significantly to muta-

genic process in the early stage of tumorigenesis for these HCCs. This analysis suggests the utility of MesKit to reveal the dynamic mutational processes.

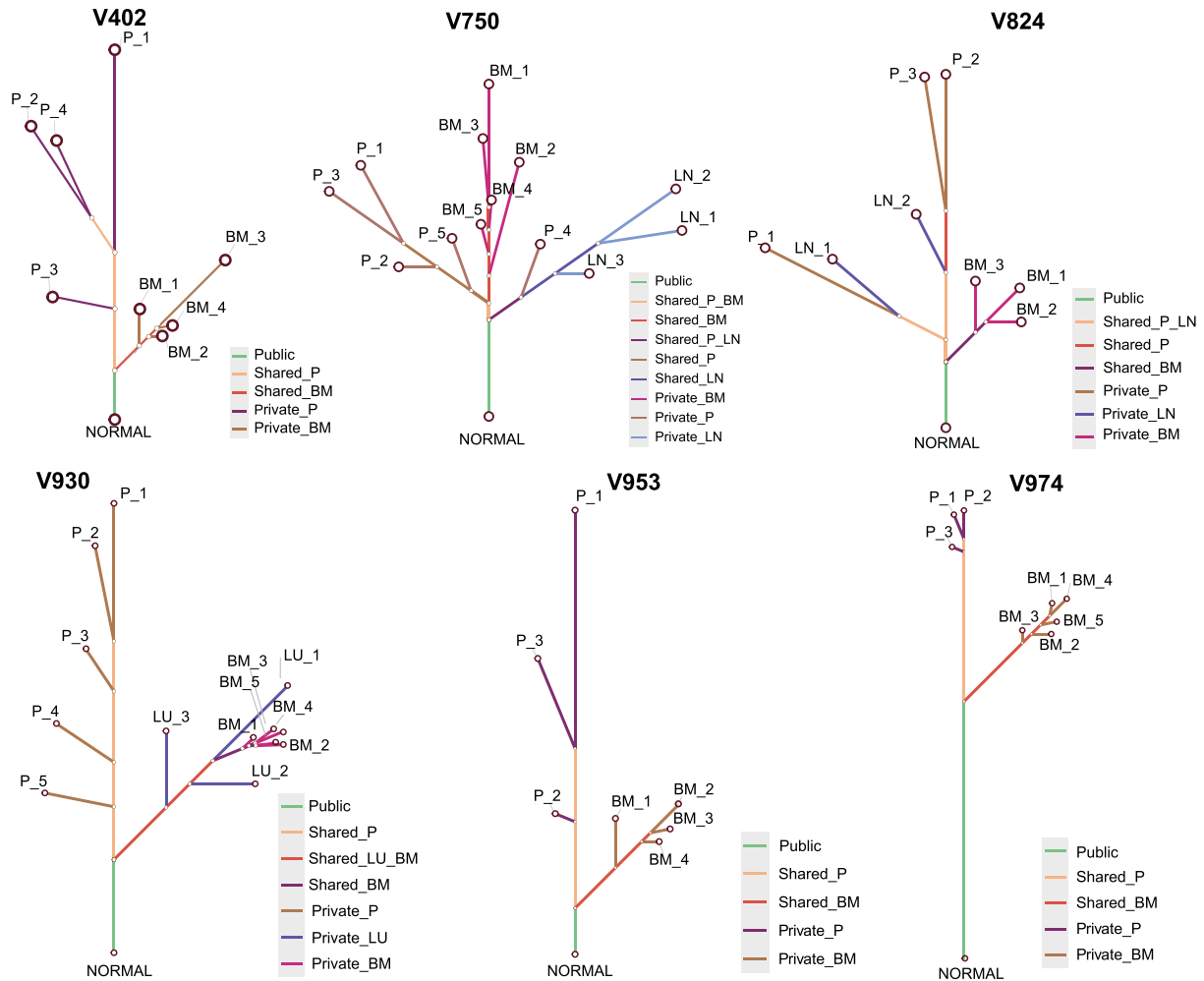


Figure 4: Phylogenetic trees of the CRC cohort. Phylogenetic trees of the CRC cohort were constructed from all sSNVs and INDELS using the MP algorithm. Branches were colored according to the regional distribution of mutations. The branch lengths are proportional to the number of mutations.

Discussion

Multi-region sequencing has become an affordable and effective way to investigate genetic heterogeneity and trace tumor evolutionary trajectory. Multiple spatial snapshots of tumors can help reduce sampling bias and detect minor subclones. Despite these advantages, there are few tools available to systematically analyze mutational data of multi-region samples from a single patient so far. In this regard, we present MesKit, an R/Bioconductor package, which incorporates a diversity of essential analysis and visualization functions for MRS studies. MesKit quantifies ITH based on somatic mutations by integrating several approaches described in recent cancer genome studies [47–50]. Besides, MesKit can be used to infer metastatic routes, characterize mutational patterns at different levels, and generate publication-quality images such as mutational profiles and phylogenetic trees. Via implementation of the Shiny application, MesKit enables researchers with minimal informatics skills to effortlessly interpret and visualize the intricate mutational data from MRS. Furthermore, we demonstrated the utility and efficiency of MesKit in interpreting ITH and inferring evolutionary trajectory using 2 published MRS datasets of HCC and CRC. Collectively, we believe that MesKit is a handy and feature-rich tool, which will greatly facilitate the exploration of mutational data from MRS experiments.

Because MesKit takes a MAF file and a clinical data file as standard inputs, it primarily evaluates ITH based on somatic mutations, and its assessment of contributions of CNAs is still limited. At present, several subclonal reconstruction methods are available to infer the relative order of occurrence between an SNV and its associated CNA. In future updates, we plan to implement the integration of results from these methods to provide insights into the clonality and temporal dynamics of ITH. On the other hand, as ITH arises through various mechanisms, it is invaluable to perform investigations at the genetic, transcriptomic, phenotypic, and cellular levels.

Availability of Source Code and Requirements

Project name: MesKit
 Project home page: <https://github.com/Niinleslie/MesKit>
 Operating system(s): Platform independent
 Programming language: R
 Other requirements: R \geq 4.0
 License: GPL-3
 RRID: SCR_020959
 biotools: meskit

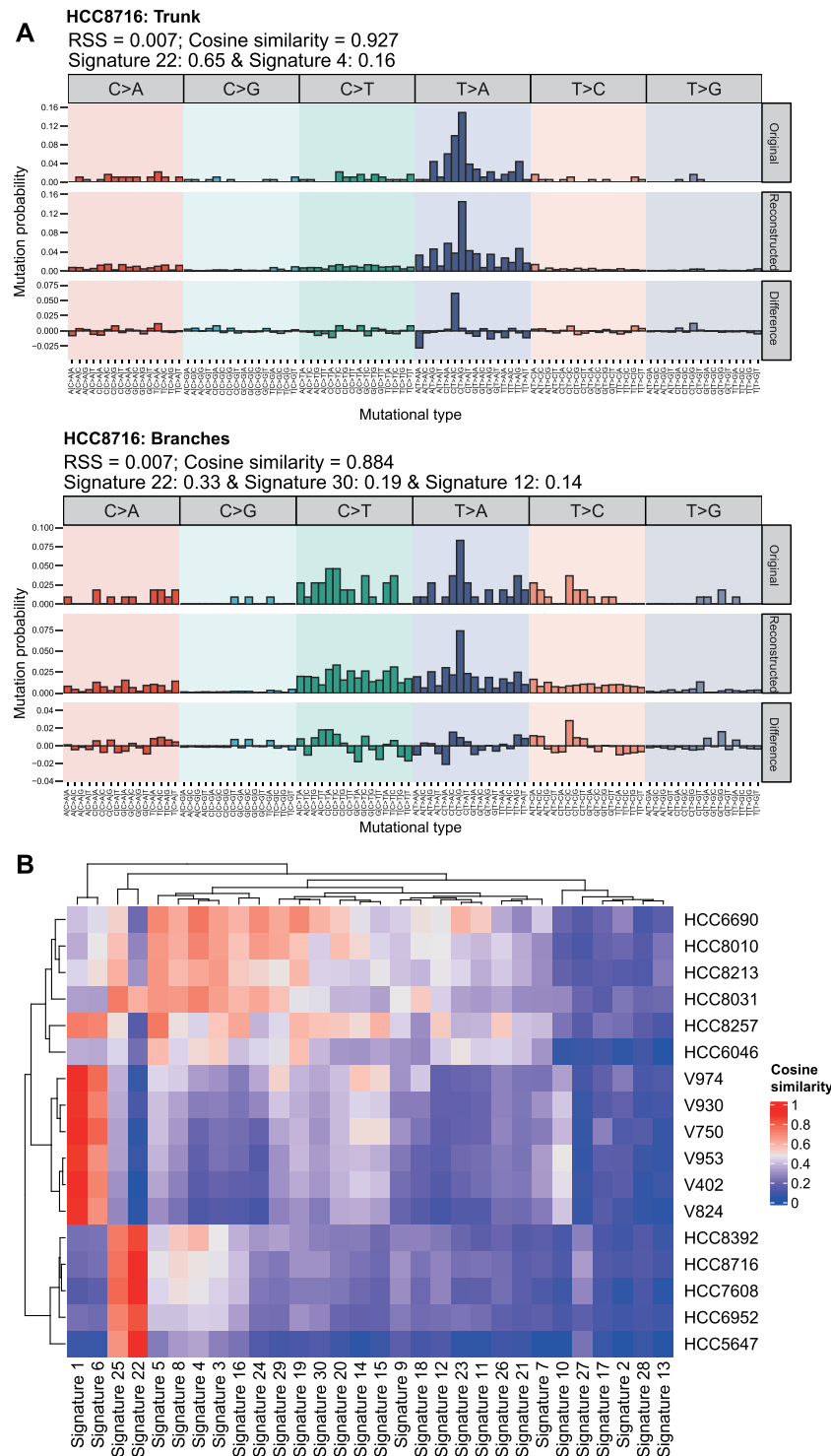


Figure 5: Temporal dissection of mutational signatures. **A.** Relative contribution of the 96 trinucleotide changes to the original mutational profile (upper panel), the reconstructed mutational profile (middle panel), and the difference between these profiles for truncal mutations and branch mutations from HCC patient HCC8716. The RSS, cosine similarity between the original and the reconstructed mutational profile and proposed etiology for mutational processes underlying the signature are indicated on the top. **B.** Heat map of cosine similarities between the 30 COSMIC signatures and the mutational profiles of the HCC and CRC cohorts. Patients were hierarchically clustered between the vectors of cosine similarities of signatures using the Euclidean distance methods. The signatures were ordered according to hierarchical clustering based on the cosine similarity between signatures.

Data Availability

The code for creating the figures in this article can be found and re-executed in a Code Ocean capsule [75]. Supporting data and an archival copy of the code are also available via the GigaScience database, GigaDB [76].

Additional Files

Figure S1: Mutational landscape of HCC and CRC cohorts **A.** Mutational profile of HCC cohort. Oncoprint of top 15 most frequently mutated driver genes of HCC were grouped by public, shared, or private mutations including both clonal and subclonal drivers. Stacked bar charts on the top and right show the number of mutations for different types per sample and per driver gene, respectively. Genes were sorted by mutational frequency and samples were split by patients as indicated by the annotation bar (bottom). **B.** The consistent CNAs of CRC cohort with significant recurring CNAs identified from TCGA Colorectal Adenocarcinoma project by GISTIC2.0 (obtained from Broad GDAC website). Each track represents 1 tumor sample. BM: brain metastasis; LN: lymph node metastasis; LU: lung metastasis; P: primary tumor. Dark red indicates amplifications ($CN \geq 4$); light red, gains ($2 < CN < 4$); dark blue, deletions ($CN = 0$); and light blue, losses ($0 < CN < 2$).

Figure S2: CCF heat maps of CRC cohort The heat maps of CCF values of tumor samples from the same patient. The color bar next to the heat map indicates the classification of mutations shared amongst different samples. The proportion of each classification is indicated in the legend. Putative CRC driver genes are labeled on the right.

Figure S3: Comparison of phylogenetic trees constructed by different methods of the CRC cohort. Comparison of the MP-based phylogenetic trees against those constructed by NJ method and ML method for each patient with CRC. For each pair, the different clades between 2 phylogenetic trees are highlighted in red (the first tree) or blue (the second tree).

Figure S4: Comparison of signature contributions measured by MesKit, MutationalPatterns, SignatureEstimation, and deconstructSigs. **A.** Relative contributions of all 30 COSMIC signatures for each patient in the HCC and CRC cohorts. **B.** Cosine similarity and RSS between the original and the reconstructed mutational profiles.

Figure S5: Mutation spectra of truncal and branch mutations of HCC5647, HCC7608, and HCC8716. Stacked bar plots show the proportions of truncal and branch mutations accounted for by each of the 6 mutation types in HCC5647, HCC7608, and HCC8716. The number of analyzed mutations is displayed on top of each bar. A Fisher exact test was used to compare truncal and branch mutations for each mutation type (2-sided test: $*P < 0.01$).

Figure S6: Schematic diagram of visualizing phylogenetic trees. Node N refers to a non-mutated normal sample: node 0 represents the starting node. In tree T_0 : $K = \{\text{node 0, node 2, node 4, node 5, node 8}\}$, $K^{[1]}$ is node 0; $B = \{\text{node 1, node 3, node 6, node 7}\}$, $B^{[1]}$ is node 1; $R = \{\text{node 1, node 7}\}$, $R^{[1]}$ is node 1; $L = \{\text{node 3, node 6}\}$, $L^{[1]}$ is node 3.

Table S1: Clinical features of the HCC and CRC cohorts

Table S2: Distance between the MP-based phylogenetic tree and the NJ-/ML-based phylogenetic tree for each patient in CRC cohort

Table S3: Relative contributions of all 30 COSMIC signatures for each patient in HCC and CRC cohorts, as measured by MesKit, MutationalPatterns, SignatureEstimation, and deconstructSigs

Table S4: Signature contributions of truncal and branch mutations of HCC5647, HCC7608, and HCC8716

Supplementary File S1: Phylogenetic visualization auto-layout algorithm

Abbreviations

AUC: area under the curve; BM: brain metastasis; CCF: cancer cell fraction; CI: confidence interval; CNAs: copy number alterations; COSMIC: Catalogue of Somatic Mutations in Cancer; CRC: colorectal cancer; F_{ST} : fixation index; GDAC: Genome Data Analysis Center; GUI: graphical user interface; HCC: hepatocellular carcinoma; INDELS: small insertions and deletions; ITH: intra-tumor heterogeneity; JSI: Jaccard similarity index; LN: lymph node; LU: lung metastasis; MAD: median absolute deviation; MAF: mutation annotation format; MATH: mutant-allele tumor heterogeneity; ML: maximum likelihood; MP: maximum parsimony; MRS: multi-region sequencing; NJ: neighbor-joining; RSS: residual sum of squares; sSNV: somatic single-nucleotide variant; TCGA: The Cancer Genome Atlas; VAF: variant allele frequency; WES: whole-exome sequencing.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Q.Z. and J.R. conceived the project. M.L., J.C., X.W. and C.W. developed the methodology and implemented the method. X.Z., Z.Z. and Y.X. helped test the software. M.L., Q.Z., and J.R. wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (Grant Nos. 91753137, 31471252, 31771462, 81772614, U1611261, and 31801105); National Key R&D Program of China (Grant No. 2017YFA0106700); Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No. 2017ZT07S096); Guangdong Natural Science Foundation (Grant No. 2018A030313323); Fundamental Research Funds for the Central Universities (SYSU: 19ykpy184) and China Postdoctoral Science Foundation (2017M610573).

References

- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;144(5):646–74.
- Lin DC, Mayakonda A, Dinh HQ, et al. Genomic and epigenomic heterogeneity of hepatocellular carcinoma. *Cancer Res* 2017;77(9):2255–65.
- Gundem G, Van Loo P, Kremeyer B, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* 2015;520(7547):353–7.
- Hong MK, Macintyre G, Wedge DC, et al. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat Commun* 2015;6(1):6605.
- Hao JJ, Lin DC, Dinh HQ, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nat Genet* 2016;48(12):1500–7.

6. Yates LR, Gerstung M, Knappskog S, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med* 2015;**21**(7):751–9.
7. de Bruin EC, McGranahan N, Mitter R, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* 2014;**346**(6206):251–6.
8. Zhang J, Fujimoto J, Zhang J, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multi-region sequencing. *Science* 2014;**346**(6206):256–9.
9. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 2014;**344**(6190):1396–401.
10. Jamal-Hanjani M, Wilson GA, McGranahan N, et al. Tracking the evolution of non-small-cell lung cancer. *N Engl J Med* 2017;**376**(22):2109–21.
11. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 2012;**366**(10):883–92.
12. Hu Z, Ding J, Ma Z, et al. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nat Genet* 2019;**51**(7):1113–22.
13. Kim TM, Jung SH, An CH, et al. Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. *Clin Cancer Res* 2015;**21**(19):4461–72.
14. El-Kebir M, Oesper L, Acheson-Field H, et al. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* 2015;**31**(12):i62–70.
15. Gerlinger M, Horswell S, Larkin J, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* 2014;**46**(3):225–33.
16. Liu W, Laitinen S, Khan S, et al. Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med* 2009;**15**(5):559–65.
17. Huang Y, Gao S, Wu S, et al. Multilayered molecular profiling supported the monoclonal origin of metastatic renal cell carcinoma. *Int J Cancer* 2014;**135**(1):78–87.
18. Cheung KJ, Padmanaban V, Silvestri V, et al. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. *Proc Natl Acad Sci U S A* 2016;**113**(7):E854–63.
19. Beltran H, Demichelis F. Prostate cancer: Inpatient heterogeneity in prostate cancer. *Nat Rev Urol* 2015;**12**(8):430–1.
20. Wang D, Niu X, Wang Z, et al. Multiregion sequencing reveals the genetic heterogeneity and evolutionary history of osteosarcoma and matched pulmonary metastases. *Cancer Res* 2019;**79**(1):7–20.
21. Yan T, Cui H, Zhou Y, et al. Multi-region sequencing unveils novel actionable targets and spatial heterogeneity in esophageal squamous cell carcinoma. *Nat Commun* 2019;**10**(1):1670.
22. Mayakonda A, Lin D-C, Assenov Y, et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* 2018;**28**(11):1747–56.
23. Blokzijl F, Janssen R, van Boxtel R, et al. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* 2018;**10**(1):33.
24. Rosenthal R, McGranahan N, Herrero J, et al. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* 2016;**17**(1):31.
25. Shinde J, Bayard Q, Imbeaud S, et al. Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics* 2018;**34**(19):3380–1.
26. Miller CA, White BS, Dees ND, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol* 2014;**10**(8):e1003665.
27. Jiao W, Vembu S, Deshwar AG, et al. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* 2014;**15**(1):35.
28. Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* 2014;**11**(4):396–8.
29. Niknafs N, Beleva-Guthrie V, Naiman DQ, et al. Subclonal hierarchy inference from somatic mutations: automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *PLoS Comput Biol* 2015;**11**(10):e1004416.
30. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers. *Cell* 2012;**149**(5):994–1007.
31. Deshwar AG, Vembu S, Yung CK, et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol* 2015;**16**(1):35.
32. Caravagna G, Heide T, Williams MJ, et al. Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat Genet* 2020;**52**(9):898–907.
33. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;**31**(3):213–9.
34. Saunders CT, Wong WS, Swamy S, et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 2012;**28**(14):1811–7.
35. Stachler MD, Taylor-Weiner A, Peng S, et al. Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat Genet* 2015;**47**(9):1047–55.
36. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**(16):2078–9.
37. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**(16):e164.
38. Favero F, Joshi T, Marquard AM, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 2015;**26**(1):64–70.
39. Williams MJ, Werner B, Barnes CP, et al. Identification of neutral tumor evolution across cancer types. *Nat Genet* 2016;**48**(3):238–44.
40. Durrett R. Population genetics of neutral mutations in exponentially growing cancer cell populations. *Ann Appl Probab* 2013;**23**(1):230–50.
41. Zhang C, Zhang L, Xu T, et al. Mapping the spreading routes of lymphatic metastases in human colorectal cancer. *Nat Commun* 2020;**11**(1):1993.
42. Hu Z, Li Z, Ma Z, et al. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. *Nat Genet* 2020;**52**(7):701–8.
43. McGranahan N, Favero F, de Bruin EC, et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med* 2015;**7**(283):283ra54.
44. Sun R, Hu Z, Sottoriva A, et al. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat Genet* 2017;**49**(7):1015–24.

45. Caswell-Jin JL, McNamara K, Reiter JG, et al. Clonal replacement and heterogeneity in breast tumors treated with neoadjuvant HER2-targeted therapy. *Nat Commun* 2019;**10**(1):657.
46. Scrucca L, Fop M, Murphy TB, et al. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J* 2016;**8**(1):289–317.
47. Mroz EA, Tward AD, Pickering CR, et al. High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer-Am Cancer Soc* 2013;**119**(16):3034–42.
48. Charoentong P, Finotello F, Angelova M, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep* 2017;**18**(1):248–62.
49. Bhatia G, Patterson N, Sankararaman S, et al. Estimating and interpreting FST: the impact of rare variants. *Genome Res* 2013;**23**(9):1514–21.
50. Lee JK, Wang J, Sa JK, et al. Spatiotemporal genomic architecture informs precision oncology in glioblastoma. *Nat Genet* 2017;**49**(4):594–9.
51. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 2004;**20**(2):289–90.
52. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics* 2011;**27**(4):592–3.
53. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;**4**(4):406–25.
54. Desper R, Gascuel O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol* 2002;**9**(5):687–705.
55. Yang Z. Phylogenetic analysis using parsimony and likelihood methods. *J Mol Evol* 1996;**42**(2):294–307.
56. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 1981;**17**(6):368–76.
57. Youssoufian H, Kazazian HH Jr, Phillips DG et al. Recurrent mutations in haemophilia A give evidence for CpG mutation hotspots. *Nature* 1986;**324**(6095):380–2.
58. Hu Z, Sun R, Curtis C. A population genetics perspective on the determinants of intra-tumor heterogeneity. *Biochim Biophys Acta Rev Cancer* 2017;**1867**(2):109–26.
59. Liu M, Liu Y, Di J, et al. Multi-region and single-cell sequencing reveal variable genomic heterogeneity in rectal cancer. *BMC Cancer* 2017;**17**(1):787.
60. Brannon AR, Vakiani E, Sylvester BE, et al. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biol* 2014;**15**(8):454.
61. Tan IB, Malik S, Ramnarayanan K, et al. High-depth sequencing of over 750 genes supports linear progression of primary tumors and metastases in most patients with liver-limited metastatic colorectal cancer. *Genome Biol* 2015;**16**(1):32.
62. Wang K, Lim HY, Shi S, et al. Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma. *Hepatology* 2013;**58**(2):706–17.
63. Guichard C, Amaddeo G, Imbeaud S, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat Genet* 2012;**44**(6):694–8.
64. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* 2018;**15**(2):81–94.
65. Williams MJ, Werner B, Heide T, et al. Quantification of sub-clonal selection in cancer from bulk sequencing data. *Nat Genet* 2018;**50**(6):895–903.
66. Rajput A, Bocklage T, Greenbaum A, et al. Mutant-allele tumor heterogeneity scores correlate with risk of metastases in colon cancer. *Clin Colorectal Cancer* 2017;**16**(3):e165–70.
67. Mroz EA, Rocco JW. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncol* 2013;**49**(3):211–5.
68. Campbell PJ, Yachida S, Mudie LJ, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* 2010;**467**(7319):1109–13.
69. Ding L, Ley TJ, Larson DE, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 2012;**481**(7382):506–10.
70. Griffith M, Miller CA, Griffith OL, et al. Optimizing cancer genome sequencing and analysis. *Cell Syst* 2015;**1**(3):210–23.
71. Xue R, Chen L, Zhang C, et al. Genomic and transcriptomic profiling of combined hepatocellular and intrahepatic cholangiocarcinoma reveals distinct molecular subtypes. *Cancer Cell* 2019;**35**(6):932–947.e8.
72. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* 1994;**11**(3):459–68.
73. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature* 2013;**500**(7463):415–21.
74. Huang X, Wojtowicz D, Przytycka TM. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* 2018;**34**(2):330–7.
75. Liu M. Supporting code and data for “MesKit: a tool kit for dissecting cancer evolution of multi-region tumor biopsies through somatic alterations” [Source Code]. *Code Ocean*. 2021; <https://doi.org/10.24433/CO.6811520.v2>.
76. Liu M, Chen J, Wang X, et al. Supporting data for “MesKit: a tool kit for dissecting cancer evolution of multi-region tumor biopsies through somatic alterations.” *GigaScience Database* 2021. <http://dx.doi.org/10.5524/100891>.
77. Hans W. Borchers. *pracma: Practical Numerical Math Functions*. R package version 2.3.3. <https://CRAN.R-project.org/package=pracma>. Accessed 1 January 2021.
78. COSMIC Mutational Signatures. <http://cancer.sanger.ac.uk/cosmic/signatures>. Accessed 1 July 2020.