From copy number alterations to structural variants: the evolutionary cascade of papillary renal cell carcinomas

Bin Zhu[1†], Maria Luana Poeta[2†], Manuela Costantini[2,3†], Tongwu Zhang[1†], Jianxin Shi[1], Steno Sentinelli[4], Vincenzo Pompeo[3], Maurizio Cardelli[5], Boian S. Alexandrov[6], Burcak Otlu[7], Xing Hua[1], Kristie Jones[8], Seth Brodie[8], Jorge R. Toro[9], Meredith Yeager[8], Mingyi Wang[8], Belynda Hicks[8], Ludmil B. Alexandrov[7], Kevin M. Brown[1], Stephen Chanock[1#], Vito Michele Fazio[8,9#], Michele Gallucci[3#], Maria Teresa Landi[1*#]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, MD 20892

[2]Department of Bioscience, Biotechnology and Biopharmaceutics, University of Bari, 70126 Bari, Italy.

[3]Department of Urology, "Regina Elena" National Cancer Institute, 00144, Rome, Italy

[4]Department of Pathology, "Regina Elena" National Cancer Institute, 00144, Rome, Italy

[5] Advanced Technology Center for Aging Research, Scientific Technological Area, Italian National Research Center on Aging (INRCA), 60121 Ancona, Italy.

[6]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA

[7]Department of Cellular and Molecular Medicine and Department of Bioengineering and Moores Cancer Center, University of California, San Diego, La Jolla, CA, 92093, USA

[8]Cancer Genomics Research Laboratory (CGR), Frederick National Laboratory for Cancer Research, Frederick, MD

[9]Washington, DC Veteran Affairs Medical Center, Washington, DC 20422

[10]Laboratory of Molecular Medicine and Biotechnology, University Campus Bio-Medico of Rome, 00128 Rome, Italy

[11]Laboratory of Oncology, IRCCS H. "Casa Sollievo della Sofferenza", 71013 San Giovanni Rotondo (FG), Italy

[†]These authors contributed equally to this work; [#]These authors jointly supervised this work

***Corresponding author:**
Maria Teresa Landi, M.D., Ph.D.
Division of Cancer Epidemiology and Genetics
National Cancer Institute, NIH
9609 Medical Center Drive, Room 7E106
Bethesda, MD 20892-9769

Phone: (240) 276-7236
landim@mail.nih.gov

ABSTRACT

Background

Intratumor heterogeneity (ITH) and tumor evolution have been described for clear cell renal cell carcinomas (ccRCC), but only limited data are available for other kidney cancer subtypes. Moreover, previous ITH studies predominately focused on single nucleotide variants (SNVs); little is known of the stepwise process in which additional genomic alterations such as copy number alterations (SCNAs) or structural variants (SV) are acquired.

Results

We investigated ITH and clonal evolution of papillary renal cell carcinoma (pRCC) and rarer kidney cancer subtypes using whole-genome sequencing and multi-omics analyses in 124 samples from 29 subjects. We collected multiple samples from the center of the tumor to the periphery and matched metastatic lesions to capture changes occurring along the physical tumor expansion. We used phylogenetic analysis to order the impact of SCNAs, SNVs, and SVs along the evolutionary trajectory of these tumors. While the few mutations in cancer driver genes were clonal, pRCC ITH was lowest for SCNAs, intermediate for SNVs, and highest for SVs. The phylogenetic analysis confirmed a clonal expansion cascade along these genomic alteration types. Moreover, while SNVs and SCNAs were similar, SVs were >20 times more frequent in pRCC type 2 than pRCC type 1, suggesting a role for SVs in pRCC type 2 aggressive behavior.

Conclusions

Unlike ccRCC or other cancer types, pRCCs tumorigenesis appears to begin from SCNAs and/or rare mutations in cancer driver genes. No effective treatment is available for this tumor. Our work highlights the need for tailored intervention against large-scale somatic alterations beyond SNVs.

Keywords
papillary renal cell carcinoma, clonal evolution, whole-genome sequencing, intratumor heterogeneity, structural variants

BACKGROUND

Kidney cancer includes distinct subtypes (Moch et al., 2016) based on cytoplasmic (e.g., clear cell renal cell carcinoma, ccRCC), architectural (e.g., papillary renal cell carcinoma, pRCC), or mesenchymal (e.g., renal fibrosarcomas, rSRC) features; each of these subtypes has distinct implications for clinical prognosis. Rarer subtypes have also been defined by anatomic location (e.g., collecting duct renal cell carcinoma, cdRCC). Within subtypes, there can be further differences in both tumor characteristics and prognoses; for example, pRCC type 1 is more benign compared to the aggressive type 2. Recent cancer genomic characterization studies have revealed that the genomic landscape of major kidney cancer subtypes can be complex (Cancer Genome Atlas Research, 2013; Cancer Genome Atlas Research et al., 2016; Davis et al., 2014). In this regard, patterns of intratumor heterogeneity (ITH) and tumor evolution have become the focus of intense investigation, primarily through multi-region whole-exome or whole-genome sequencing studies in ccRCC (Gerlinger et al., 2014; Gerlinger et al., 2012; Mitchell et al., 2018). However, understanding the importance of ITH in other kidney cancer subtypes is either limited, such as for pRCC, the second most common kidney cancer subtype, where only four tumors have been characterized (Kovac et al., 2015) or completely lacking, such as for cdRCC or rSRC. Moreover, previous ITH studies predominately focused on single nucleotide variants (SNVs); little is known of the stepwise process in which additional genomic alterations (e.g., structural variants, SVs) are acquired, thus providing new clues into the genomic events critical for different sub-types of kidney cancer. Herein, we report on the genomic characterization of pRCC and rarer kidney cancer subtypes, specifically examining both the core and periphery of selected tumors and, when available, metastatic lesions in order to investigate ITH and clonal evolution.

RESULTS

We conducted an integrative genomic and epigenomic ITH analysis of pRCC and rarer kidney cancer subtypes and provide new insights into clonal evolution, which is distinct from clear cell renal cell carcinoma (Ricketts et al., 2018). We examined multiple consecutive samples from the core center of the tumor through the tumor's periphery as well as a normal sample ~5 cm distant from the tumor, and, when feasible, metastatic regions in the adrenal gland (**Figure 1a, Online Methods**). We performed 60X multi-region whole-genome sequencing (mWGS, **Table S1**) in 124 primary tumor and metastasis samples from 29 treatment-naive kidney cancers (**Table S2**), as well as genome-wide methylation and SNP array profiling, and deep targeted sequencing (average 500X coverage) (**Table S3)** of 254 known cancer driver genes (Lawrence et al., 2014) (**Table S4**). WGS data included 13 pRCC type 1 (pRCC1) tumors, 12 pRCC type 2 (pRCC2) tumors, and rarer subtypes (one each of cdRCC, rSRC, mixed pRCC1/pRCC2 and pRCC2/cdRCC) (**Figure 1b, Online Methods).** A section of each sampled region was histologically examined: tumor samples included in the analyses had to exceed 70% tumor nuclei by pathologic assessment by a senior pathologist and the normal samples had no evidence of tumor nuclei. We also estimated the sample purity based on somatic copy number alterations (SCNA) or, in copy neutral genomic regions, based on variant allele fraction (VAF) of single nucleotides (**Figure S1).** The estimated purity based on WGS data was used to properly infer cancer cell fractions (CCF) and to construct lineage phylogenetic trees. Data on genome-wide methylation levels provided further information on sample cell type composition.

Mutation rates, frequency of driver mutations, and germline variants in known cancer susceptibility genes

The average SNV and indel rates were 1.25/Mb and 0.18/Mb, respectively, with differences observed across histological subtypes: on average, 0.99/Mb and 0.18/Mb for pRCC1, 1.54/Mb and 0.21/Mb for pRCC2, and 0.90/Mb and 0.11/Mb for the other subtypes (**Figure 1c**). Among the published kidney cancer driver genes, we observed clonal driver SNVs (definition of driver mutation in Online Methods) in *MET* in four pRCC1 tumors; *SMARCB1* in two pRCC2; *TERT* promoter in two pRCC2; *NFE2L2* in one pRCC1; *SETD2*, *PBRM1* and *NF2* in one pRCC2 tumor each. We also found clonal indels in *NF2* in two tumors (cdRCC and mixRCC), and *MET* (mixRCC), *SMRCB1* (pRCC1) and *ROS1* (pRCC2) indels in one tumor each. We found no mutations in *TP53*, a gene found mutated in a very high proportion of cases across cancer types (Ding et al., 2018), and no mutations in the 5'UTR region of *TERT,* which was recently reported as mutated in a fraction of ccRCC (Mitchell et al., 2018) (**Figure 1c** and **Figure S2** and **Table S5 and Table S6**). Eight pRCC1 (31%) and three pRCC2 (25%) had no detected SNVs or indels in previously reported driver genes, even after deep targeted sequencing, suggesting that SNVs in other genes or other genomic alterations are the likely driver events.

An analysis of the germline sequencing data provided evidence of rare, potentially deleterious, germline variants in known cancer susceptibility genes (minor allele frequency <0.1% in an Italian whole exome sequencing data from 1,368 subjects with no cancer(Landi et al., 2008) and the GnomAD European-Non Finnish-specific data from 12,897 subjects(Lek et al., 2016)). These include two different variants in *POLE* in two different tumors; two different variants in *CHEK2* in two different tumors*;* one variant in *PBRP1* and *PTCH1* both in a single tumor; and additional rare variants, one per tumor (e.g., *TP53, MET, EGFR*, among others, **Table S7**). This is consistent with a recent report on the relative high frequency of germline mutations in cancer susceptibility genes in non-clear cell renal cell carcinomas(Carlo et al., 2018).

Intratumor heterogeneity based on SNV multi-regional trees (MRTs)

To explore ITH and spatial concurrence of SNVs, we first constructed multi-regional trees (MRTs) using the parsimony ratchet method (Nixon, 1999) based on present or absent SNVs. We used 14 tumors with at least three regional samples per tumor, including three pRCC1, eight pRCC2, and single tumors from three rarer subtypes. The root of the MRTs represents normal cells without somatic SNVs. The longer the trunk length in an MRT, the lower the level of ITH. On average 30.9% of pRCC SNVs were in branches, with low heterogeneity across histological subtypes and within each subtype (**Figure S3**). This contrasts with what has been shown for clear-cell renal cell (Gerlinger et al., 2014; Gerlinger et al., 2012; Moore et al., 2018), where approximately two-third of somatic mutations are in branches. Among the genomic regions, a few pRCC tumors showed higher ITH in the promoters, 5'UTR and the first exon regions (**Figure S3**). Examples from two pRCC1 and pRCC2 SNV MRTs are shown in **Figure 2a/c** (the remaining are shown in in **Figure S5**). The metastatic samples in pRCC2_1824_13 (**Figure 2c**), which may have originated from direct invasion of the primary tumors, share the same driver mutations in *PBRM1, SMARCB1,* and *BIRC3.*

Lineage phylogenetic trees (LPTs) based on SNVs

Each tumor region likely contains a mixture of different cell lineages (Alves et al., 2017). Thus, to infer the evolutionary history of SNVs in these tumors, we constructed lineage phylogenetic trees (LPTs). We used PyClone (Roth et al., 2014) to define subclones based on clusters of SNVs sharing similar CCF, adjusting for SCNAs and purity. On average, we identified 4, 3.4 and 2 subclone lineages in pRCC1, pRCC2 and the rarer subtypes, respectively, which were estimated based on approximately 300 SNVs/tumor with coverage >100x (**Table S8**). We cannot exclude that, with deeper coverage across a larger number of SNVs and with more regions sampled from some of the tumors, the PyClone algorithm could identify more subclones.

With the identified subclones, SCHISM (Niknafs et al., 2015) was applied to construct a phylogenetic tree for each region of a given tumor. This analysis allowed us to infer which lineage is prevalent in the tumor. For example, in tumor sample pRCC2_1824_13 (**Figure 2d**): a lineage consisting of MRCA subclone C1, 1st-generation subclone C2 and 2nd-generation subclone C3 was present in all regions. In contrast, other 1st-generation subclones were present only in a subset of tumor regions, particularly those in closer physical proximity, and did not lead to descendant subclones. Metastatic samples, M01, M02 and M03 in pRCC2_1824_13 shared the same subclones with region T02, with the exception of a metastatic sample-specific subclone (C8), suggesting that the metastasis likely originated from this region. The likelihood that the metastatic samples were spread from region T02 was further supported by the SNV MRT of the same tumor (**Figure 2c**), in which the metastatic samples shared the highest proportion of SNVs with T02 and hence were closest to T02. In addition to the two LPTs shown in **Figure 2**, LPTs of other tumors showed similar shallow and umbrella-shaped branching evolution, with one MRCA clone leading to multiple 1st-generation subclones and in some cases, 2nd- or 3rd-generation subclones, with a dominant lineage (**Figure S6**).

Dominance of mutational signatures 5 and 40 in both trunk and branches and shorter telomere length in metastatic samples

De novo extraction of SNV mutational signatures identified the patterns of four distinct mutational signatures, termed, signatures A through D (**Figure S7**). Comparison of these four *de novo* deciphered signatures to the global consensus set of mutational signatures(Alexandrov et al., 2018) revealed that signatures A through D are linear combinations of six previously known SNV mutational signatures (**Table S9a**): single base signatures (SBS) 1, 2, 5, 8, 13, and 40. Signatures 5 and 40 are both with unknown etiology and were found across all examined RCC subtypes (mean contributions 32.6% and 59.9%, respectively, **Table S10**). We also observed a small proportion of mutations attributed to the clock-like (Alexandrov et al. (2015) mutational signature 1 (3.5% of total SNVs) and signature 8 (1.4%), which has an unknown etiology. Further, a low mutational activity was detected for signature 2 (0.6%) and signature 13 (0.7%), both attributed to the activity of the APOBEC family of deaminases (**Figure S8**). All signatures were found in both MRTs trunk and branches (**Figure S9**) and varied only slightly between primary and metastatic samples (**Figure S10**). The dominance of mutational signatures 5 and 40 in both the trees' trunks and branches could reflect a continuous exposure to a metabolic mutagen, which may be actively reabsorbed in the kidney proximal tubule and contribute to the

tumor initiation and/or progression. Alternatively, similar to many other cancers, the predominance of signatures 5 and, possibly signature 40, could reflect the temporal progression of mutational events (Alexandrov et al., 2015; Alexandrov et al., 2013a).

Examining the effect of genomic architecture on SNV mutational signatures revealed that mutations attributed to signature 1 were enriched in both early replicating regions of the genome and the parts of the genome with higher nucleosome occupancy (**Figure S11**). In contrast, the mutations attributed to signatures 5 and 40 were mostly unaffected by replication timing and nucleosome occupancy (**Figure S11**). Interestingly, signatures 1, 5, and 40 exhibited a statistically significant transcriptional strand-bias (**Figure S11**). Statistically significant replication strand-bias was observed for signature 40 (**Figure S11**). There were not sufficient number of mutations to evaluate the transcription and replication strand biases for signatures 2, 8, and 13.

De novo extraction of indel mutational signatures was performed across all samples using an indel classification scheme outlined in Alexandrov et al. (2018). Three distinct indel signatures were identified and termed signatures INDEL-A, INDEL-B, and INDEL-C (**Figure S12**). Comparison of these three *de novo* deciphered signatures to the global consensus set of mutational signatures revealed that signatures INDEL-A through INDEL-C are linear combinations of six previously known indel mutational signatures (**Table S9b**): signatures ID-1, ID-2, ID-3, ID-5, ID-6 and ID-8. Signatures INDEL-A and INDEL-B were found in most samples, while signature INDEL-C was found only in a subset of samples. The indels attributed to the indel mutational signatures varied from a dozen to more than 800 (**Table S11**) indicating that different indel mutational processes have been active in papillary renal cell carcinoma.

We estimated telomere length (TL) based on the numbers of telomere sequence $(TTAGGG/CCCTAA)_4$ using TelSeq (Ding et al., 2014). The normal and metastatic tissue samples on average had longer (8.51 kb, $p = 1.16 \times 10^{-06}$) and shorter (4.4 kb, $p = 1.96 \times 10^{-03}$) TL, respectively than the primary tumor tissue samples (6.12 kb) (**Figure S13, Table S12**).

Somatic copy number alterations are mostly clonal in pRCCs

We analyzed SCNAs from mWGS data by considering both total copy number and minor copy number (**Table S13-14**). pRCC1 and, to a lesser extent, pRCC2 showed frequent amplification of chromosomes 7 (which includes the *MET* gene), 17, 12, and 16 (**Figure S14**). We observed no genome doubling. The majority of SCNAs in papillary renal cell carcinomas were shared across all regions of a given tumor (**Figure 1c** and **3a**, and **Figure S15**), with very few region-specific SCNAs (e.g. 13q in pRCC2_1782_08, **Figure S15**). This suggests that most SCNAs are clonal (**Figure 3a**) as previously suggested(Mitchell et al., 2018). Hierarchical clustering showed that the samples from the same tumors tended to cluster together (**Figure S16**), suggesting a higher inter-tumor heterogeneity than ITH. Metastatic lesions shared most SCNAs with their primary tumors, but also held metastasis-specific SCNAs (e.g., hemizygous deletion loss of heterozygosity in 4q of pRCC2_1824_13, **Figure 3c**), indicating ongoing SCNA clonal evolution during metastasis. Among the rarer subtypes, both rSRC and cdRCC had clonal focal homozygous deletions of *CDKN2A* at 9p21.3 (**Figure 1c** and **Figure S17-18, Table S15**).

Cancer cell fractions (CCF) of SCNAs reveal that SCNAs are early events in tumor evolution

To infer the evolutionary history of SCNAs in these tumors, we estimated CCF of SCNAs at each region and calculated the average CCF of SCNAs across the primary and (if available) metastatic regions. Most average SCNA CCFs were higher than 0.5, a cut-off previously used to define clonal events(Turajlic et al., 2018), suggesting that SCNAs are early events in pRCC evolution. CCF clusters are shown in **Figure 3b.** We validated SCNA findings using our SNP array data (**Figure S19**) and confirmed the largely clonal nature of these alterations.

Structural variant frequency differs between pRCC1 and pRCC2

Somatic SVs were called by the Meerkat algorithm (Yang et al., 2013), which distinguishes a range of SVs and suggest plausible underlying mechanisms, including retrotransposition events. pRCC2 had significantly more SV events per tumor, averaging 23.6, as compared to 1.2 events per tumor in pRCC1 (p-value=1.07 x 10^{-3}, Wilcoxon rank test, **Table S16**). Tandem duplications, chromosomal translocations, and deletions were the most prevalent types of variants (36.4%, 34.0%, and 29.4%, respectively, **Figure 4a**). Some SVs involved known cancer driver genes (**Figure 1c**), including a deletion within *MET* in one pRCC2, and several fusions involving genes previously reported in renal cancer or other tumors. These included *ALK*/*STRN* (Kelly et al., 2014) and *MALAT1*/*TFEB* (Kauffman et al., 2014) in two different pRCC2 and *EWSR1*/*PATZ1* (Cantile et al., 2013) in rSRC (**Supplemental Material**). We had high quality RNA material to validate the latter two SVs (**Figure S20**).
SVs were distributed unevenly between tumors and across the genome (**Figure 4a**). Some tumors, particularly among pRCC1s, had almost no SVs (e.g. pRCC1_1671_08 in **Figure 4b**); some had SVs clustered in a hotspot (**Figure S21**), while still others had many SVs, like pRCC2_1824_13 (**Figure 4c**) or pRCC2_1782_08 (**Figure 4d**), the latter showing high genomic instability. Interestingly, pRCC2_1782_08 had a high number of LINE-1 clonal retrotransposition events detected by TraFiC (Tubio et al., 2014) (**Figure 4a**, **Figure S22**), while somatic retrotransposition events were rarely detected in the rest of samples **(Table S17)** as expected (Rodriguez-Martin et al., 2017). At least three transposon insertions could have potentially affected the expression of proteins involved in chromatin regulation and chromosome structural maintenance and (in turn) the maintenance of genome integrity in this tumor (**Supplemental Material**). As previously observed in other malignancies (Huang et al., 2015; Kadoch et al., 2013; Sausen et al., 2013; Shain and Pollack, 2013), this same tumor also harbored a clonal driver mutation in *ARID1B* (**Figure 1c**), a gene coding a subunit of the SWI/SNF complex, suggesting that alteration of chromatin-remodeling complexes may have contributed to this tumor's overall genomic instability.

Cancer cell fractions (CCF) of SVs reveal that SVs are late events in tumor evolution

In contrast to SCNAs (**Figure 3a**), most SVs were subclonal or late events within the tumors (**Figure S23**). Specifically, on average 60% of SVs were in internal or terminal branches. This is consistent with the average CCF of SVs across regions; in most of the tumors with more than three sampled regions, CCF was less than 0.5/tumor (**Figure 4e**). In contrast, cdRCC_1929,

pRCC2_1429 and pRCC2_1552, tumors with three sampled regions only, >50% SV CCF was larger than 0.5, possibly an illusion of clonality due to sampling bias(McGranahan and Swanton, 2017). We validated 88% of the WGS-Meerkaat detected SV events and their subclonal nature using a PCR-based sequencing methodology (Ampliseq; **Figure S24**, **Supplemental Material**). It is notable that 83% of the SVs also showed concordance at the clonal/subclonal status, confirming that SVs in pRCC have high ITH.

Methylation ITH varies across genomic regions

To generate methylation MRTs, we chose the top 1% of methylation probes in CpG islands with the greatest intratumoral methylation range and constructed the methylation MRTs based on the Euclidean distances between regions, following the minimum evolution method (Brocks et al., 2014; Desper and Gascuel, 2002). In general, methylation MRTs varied significantly across tumors and histological subtypes (**Figure S25**). Two trees are shown in **Figure 5b:** the trunks are short (**Figure 2 a/c**), indicating substantial methylation ITH. Unlike SNV ITH, methylation ITH analysis showed greater ITH in enhancer regions, and no ITH in promoter/5'UTR/1st exons or CpG island regions (**Figure 5a**), suggesting a possible role of methylation ITH in shaping regulatory function, but tight control during the tumor evolution on the genome regions directly affecting gene expression.

Methylation ITH likely reflects different cell types

Unsupervised clustering analysis based on the top 1% of the most variable methylation probes showed that the samples with purity <30% clustered together but separately from the normal or the tumor tissue samples (**Figure 5c**), likely because they were enriched with stromal, immune or other non-epithelial cells. Differing cell-type compositions in the tumor peripheral samples could, at least in part, explain the discrepancies between SNV MRTs and methylation MRT. For example, the T10 sample of pRCC2_1824_13 (**Figure 5b**), the most distant from the tumor center with copy neutral genome and VAF-based purity = 0.106 (**Figure S26**), appeared distant from the other tumor samples or from the normal sample in the methylation MRT, but very close to the normal sample in the corresponding SNV MRT (**Figure 2c**). Similarly, although metastasis samples in this tumor appear to arise from the T02 region based on the SNV MRT and LPT (**Figure 2c/d**), they are distant from the T02 region in the methylation MRT, likely because methylation reflects the different tissue type (adrenal gland). This finding is comparable to what has been recently reported in the TCGA pan-can analyses, where methylation profiles have been used to infer cell-of-origin patterns across cancer types (Hoadley et al., 2018).

DISCUSSION

Our study performed detailed characterization of the intra-tumor heterogeneity of the second most common type of renal cell carcinoma, papillary renal cell carcinoma, as well as rarer kidney cancer subtypes. For these subtypes, we have described the branching and multi-clonal architecture of tumor evolution, including the metastatic phase. We analyzed samples from the tumor center to tumor periphery at a precise physical distance from each other to gauge the tumor progression. Notably, we integrated multiple types of genomic alterations, beyond SNV

analyses. We thoroughly characterized sample purity relying on the pathology-based evaluation and molecular-based estimates from SCNA profiles, SNV VAFs, and methylation levels; validated SCNAs, SNVs and SVs with alternative laboratory assays; and conducted both multi-regional trees and lineage phylogenetic analyses across all tumor samples.

The number of SV events and, to a lesser extent, SNVs in cancer driver genes was higher in pRCC2 compared to pRCC1. In both tumor subtypes, the multi-regional trees were remarkably different across sets of genomic alterations with ITH increasing from SCNAs to SNVs and from SNVs to SVs (**Figure 6a**), suggesting that multiple events underlie the tumor trajectory. Specifically, the percentage of subclonal events in both pRCC1 and pRCC2 was lowest for SCNAs (0.3% and 3.2%, in pRCC1 and pRCC2, respectively), intermediate for SNVs (28.3% and 31.9%, respectively), and highest for SVs (66.7% and 60.4%, respectively, **Figure 6b**). A phylogenetic analysis, estimating cancer cell fraction of SNVs, SCNAs, and SVs confirmed this evolutionary pattern of genomic changes across pRCCs (an example is shown in **Figure 6c)**. In contrast, the few SNVs, indels and fusions we identified in known cancer driver genes were clonal in both tumor subtypes. Thus, our data indicate that papillary renal cell carcinomas initiate through a combination of large clonal SCNAs and a few mutations in driver genes, while tumorigenesis is further promoted by additional SNVs and SVs. Although ITH is generally correlated with the number of samples/tumor, the increasing ITH across genomic alteration types was consistent in both pRCC subtypes and irrespective of the number of tumor samples.

Based on these findings, we hypothesize that various forms of genomic alterations occur successively in different evolutionary stages as a clonal expansion cascade in the papillary renal cell carcinoma genome. The SNV lineage phylogenetic trees of both pRCC subtypes show a rather shallow branching evolution, suggesting a gradual accumulation of mutations leading to successive waves of subclone expansion. In contrast, the observed clonal status of SCNAs may be the result of an early punctuated burst of large-scale genomic alterations, providing growth advantage to a few cells as initiation clones that expand stably. At the time of diagnosis, the descendants of these cells, which would have accumulated additional mutations, would appear to be characterized by a single or few major SCNA events, which can expand through the metastatic process. In support of this hypothesis, bulk- and single-cell based copy number and sequencing studies of breast and prostate cancers (Baca et al., 2013; Newburger et al., 2013; Wang et al., 2014) have suggested that complex aneuploid copy number changes may occur in only a few cell divisions at the earliest stages of tumor progression, reflecting a punctuated evolution. The large SV ITH we observed particularly in pRCC2, was possibly driven by fusions or retrotransposon insertion affecting driver genes, and may reflect the early stages of additional punctuated evolution when there is still competition and progressive divergence of subclones, or the late stage of a branching evolution. Although these genomic and epigenomic changes may follow different evolutionary models, they may operate concurrently to various degrees (Davis et al., 2017). Further investigation of a larger number of less common kidney cancer types could provide further insight into the evolutionary processes of these tumors.

Understanding the clonal expansion dynamics of these cancers has potentially important implications also for the diagnosis and therapeutic treatment. Based on the observed clonal patterns of both SCNAs and somatic alterations in driver genes, a single tumor biopsy would be sufficient to characterize these changes. However, although targeted therapies against the few,

mostly not recurrent, driver gene mutations or rare germline variants we identified (e.g., *MET*, *VHL*, *PBMR1*, *ARID1B*, *SMARCA4*, *ALK, TFEB*) are either available or presently being evaluated in clinical trials, therapies against SCNAs are critically needed. Compounds that inhibit the proliferation of aneuploid cell lines (Tang et al., 2011) or impact the more global stresses associated with aneuploidy in cancer (Canovas et al., 2018; Tang and Amon, 2013) or notably target the bystander genes that are deleted together with tumor suppressor genes (collateral lethality) (Dey et al., 2017; Kryukov et al., 2016; Muller et al., 2015) are encouraging and should be further explored. Further therapeutic challenges for the renal cell tumors we studied are provided by the subclonal nature of SVs, and to a lesser extent SNVs, as well as the low mutation burden (Rizvi et al., 2015) and the notable lack of TP53 mutations (Munoz-Fontela et al., 2016; Owada et al., 2017) that may hinder response to immune checkpoint inhibitors. Notably, while the numbers of SCNAs and SNVs were similar between pRCC1 and pRCC2, the number of SV events clearly distinguished the two subtypes. The higher SV events parallel the more aggressive tumor behavior of pRCC2, emphasizing the importance of further investigating SVs in light of a possible therapeutic intervention for this subtype.

CONCLUSIONS

We characterized in detail the intra-tumor heterogeneity of the second most common type of renal cell carcinoma, papillary renal cell carcinoma, as well as rarer renal cancer subtypes. We integrated multi-region whole genome sequencing data and other multi-omics analyses with laboratory validation of major findings and described the branching and multi-clonal architecture of tumor evolution, including the metastatic phase. Based on our work, we propose a model of clonal expansion dynamics for these tumors, which can have important implications for our understanding of tumorigenesis and its application to precision medicine. It will be important to study whether the clonal expansion cascade we observed in papillary renal cell carcinomas are common to other cancer types and whether studies of paired precursor and tumor lesions from the same subjects can confirm and expand our model of clonal cascade. These studies would inform both our knowledge of tumorigenesis and the development of new treatment modalities targeting these key events.

ACKNOWLEDGEMENTS

This work utilized the computational resources of the NIH high-performance computational capabilities Biowulf cluster (http://hpc.nih.gov) and DCEG CCAD cluster. We are grateful to the patients and families who contributed to this study and the many investigators who are involved in the NCI-sponsored GEPIKID study of kidney cancer. We also thank the NCI TCGA Program Office for organizational and logistical support, Ms. Preethi Raj for graphical support, and The National Cancer Institute Cancer Genomics Research Laboratory (CGR) and Ms. Margherita Dabrosvka for sample preparation and quality control laboratory analyses.

AUTHOR CONTRIBUTIONS

M.L.P. and M.Co. conceived the surgical sampling design, collected all samples and organized field activities. B.Z. performed the statistical analysis of phylogenetic trees and supervised the genomic analyses. T.Z. conducted all bioinformatics analyses. J.S. and X.H. helped with the statistical analyses. S.S. reviewed the histological diagnosis of all tumors. M.Co. and V.P. conducted all clinical examinations and collected clinical data. M.Ca. analyzed the retrotransposition events. B.S.A., B.O., L.B.A. analyzed mutational signatures and related topography characteristics; K.J., S.B., M.Y., M.W. and B.H. confirmed all laboratory validations. J.T. participated in data interpretation. K.M.B, J.S., and S.C. participated in study conception and data interpretation. S.C. provided resources for the genomics analyses. V.M.F. supervised the field activities and data collection. M.G. performed all surgeries and supervised the sampling collection. M.T.L. conceived the study. B.Z. and M.T.L. discussed the results and implications and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests

REFERENCES:

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet *Chapter 7*, Unit7 20.
Alaei-Mahabadi, B., Bhadury, J., Karlsson, J.W., Nilsson, J.A., and Larsson, E. (2016). Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. Proc Natl Acad Sci U S A *113*, 13768-13773.
Alexandrov, L., Kim, J., Haradhvala, N.J., Huang, M.N., Ng, A.W.T., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E., Lopez-Bigas, N.*, et al.* (2018). The Repertoire of Mutational Signatures in Human Cancer. bioRxiv.
Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat Genet *47*, 1402-1407.
Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L.*, et al.* (2013a). Signatures of mutational processes in human cancer. Nature *500*, 415-421.
Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013b). Deciphering signatures of mutational processes operative in human cancer. Cell Rep *3*, 246-259.
Alves, J.M., Prieto, T., and Posada, D. (2017). Multiregional Tumor Trees Are Not Phylogenies. Trends Cancer *3*, 546-550.
Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363-1369.
Assenov, Y., Muller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods *11*, 1138-1140.
Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M.*, et al.* (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666-677.
Brocks, D., Assenov, Y., Minner, S., Bogatyrova, O., Simon, R., Koop, C., Oakes, C., Zucknick, M., Lipka, D.B., Weischenfeldt, J.*, et al.* (2014). Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. Cell Rep *8*, 798-806.
Cancer Genome Atlas Research, N. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature *499*, 43-49.
Cancer Genome Atlas Research, N., Linehan, W.M., Spellman, P.T., Ricketts, C.J., Creighton, C.J., Fei, S.S., Davis, C., Wheeler, D.A., Murray, B.A., Schmidt, L.*, et al.* (2016). Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. N Engl J Med *374*, 135-145.
Canovas, B., Igea, A., Sartori, A.A., Gomis, R.R., Paull, T.T., Isoda, M., Perez-Montoyo, H., Serra, V., Gonzalez-Suarez, E., Stracker, T.H.*, et al.* (2018). Targeting p38alpha Increases DNA Damage, Chromosome Instability, and the Anti-tumoral Response to Taxanes in Breast Cancer Cells. Cancer Cell *33*, 1094-1110 e1098.
Cantile, M., Marra, L., Franco, R., Ascierto, P., Liguori, G., De Chiara, A., and Botti, G. (2013). Molecular detection and targeting of EWSR1 fusion transcripts in soft tissue tumors. Med Oncol *30*, 412.

Carlo, M.I., Mukherjee, S., Mandelker, D., Vijai, J., Kemel, Y., Zhang, L., Knezevic, A., Patil, S., Ceyhan-Birsoy, O., Huang, K.C., *et al.* (2018). Prevalence of Germline Mutations in Cancer Susceptibility Genes in Patients With Advanced Renal Cell Carcinoma. JAMA Oncol.

Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B., Olshen, A.B., *et al.* (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol *34*, 155-163.

Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., Fan, X., Ding, L., Lee, A.Y., Boutros, P., Chen, J., *et al.* (2017). novoBreak: local assembly for breakpoint detection in cancer genomes. Nat Methods *14*, 65-67.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol *31*, 213-219.

Cmero, M., Ong, C.S., Yuan, K., Schröder, J., Mo, K., Corcoran, N.M., Papenfuss, A.T., Hovens, C.M., Markowetz, F., and Macintyre, G. (2017). SVclone: inferring structural variant cancer cell fraction. bioRxiv.

Davis, A., Gao, R., and Navin, N. (2017). Tumor evolution: Linear, branching, neutral or punctuated? Biochim Biophys Acta *1867*, 151-161.

Davis, C.F., Ricketts, C.J., Wang, M., Yang, L., Cherniack, A.D., Shen, H., Buhay, C., Kang, H., Kim, S.C., Fahey, C.C., *et al.* (2014). The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell *26*, 319-330.

Desper, R., and Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol *9*, 687-705.

Dey, P., Baddour, J., Muller, F., Wu, C.C., Wang, H., Liao, W.T., Lan, Z., Chen, A., Gutschner, T., Kang, Y., *et al.* (2017). Genomic deletion of malic enzyme 2 confers collateral lethality in pancreatic cancer. Nature *542*, 119-123.

Ding, L., Bailey, M.H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D.L., Weerasinghe, A., Huang, K.L., Tokheim, C., *et al.* (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell *173*, 305-320 e310.

Ding, Z., Mangino, M., Aviv, A., Spector, T., Durbin, R., and Consortium, U.K. (2014). Estimating telomere length from whole genome sequence data. Nucleic Acids Res *42*, e75.

Fortin, J.P., Labbe, A., Lemire, M., Zanke, B.W., Hudson, T.J., Fertig, E.J., Greenwood, C.M., and Hansen, K.D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol *15*, 503.

Freed, D., Pan, R., and Aldana, R. (2018). TNscope: Accurate Detection of Somatic Mutations with Haplotype-based Variant Candidate Detection and Machine Learning Filtering. bioRxiv.

Gao, J., Chang, M.T., Johnsen, H.C., Gao, S.P., Sylvester, B.E., Sumer, S.O., Zhang, H., Solit, D.B., Taylor, B.S., Schultz, N., *et al.* (2017). 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med *9*, 4.

Gao, R., Davis, A., McDonald, T.O., Sei, E., Shi, X., Wang, Y., Tsai, P.C., Casasent, A., Waters, J., Zhang, H., *et al.* (2016). Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat Genet *48*, 1119-1130.

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A.J., Salm, M.P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C.R., *et al.* (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat Genet *46*, 225-233.

Gerlinger, M., Rowan, A.J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., *et al.* (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med *366*, 883-892.
Hao, J.J., Lin, D.C., Dinh, H.Q., Mayakonda, A., Jiang, Y.Y., Chang, C., Jiang, Y., Lu, C.C., Shi, Z.Z., Xu, X., *et al.* (2016). Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. Nat Genet *48*, 1500-1507.
Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., *et al.* (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291-304 e296.
Huang, H.T., Chen, S.M., Pan, L.B., Yao, J., and Ma, H.T. (2015). Loss of function of SWI/SNF chromatin remodeling genes leads to genome instability of human lung cancer. Oncol Rep *33*, 283-291.
Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet *91*, 839-848.
Kadoch, C., Hargreaves, D.C., Hodges, C., Elias, L., Ho, L., Ranish, J., and Crabtree, G.R. (2013). Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. Nat Genet *45*, 592-601.
Kauffman, E.C., Ricketts, C.J., Rais-Bahrami, S., Yang, Y., Merino, M.J., Bottaro, D.P., Srinivasan, R., and Linehan, W.M. (2014). Molecular genetics and cellular features of TFE3 and TFEB fusion kidney cancers. Nat Rev Urol *11*, 465-475.
Kelly, L.M., Barila, G., Liu, P.Y., Evdokimova, V.N., Trivedi, S., Panebianco, F., Gandhi, M., Carty, S.E., Hodak, S.P., Luo, J.H., *et al.* (2014). Identification of the transforming STRN-ALK fusion as a potential therapeutic target in the aggressive forms of thyroid cancer. P Natl Acad Sci USA *111*, 4233-4238.
Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Kallberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., *et al.* (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods *15*, 591-594.
Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310-315.
Kovac, M., Navas, C., Horswell, S., Salm, M., Bardella, C., Rowan, A., Stares, M., Castro-Giner, F., Fisher, R., de Bruin, E.C., *et al.* (2015). Recurrent chromosomal gains and heterogeneous driver mutations characterise papillary renal cancer evolution. Nat Commun *6*, 6336.
Kryukov, G.V., Wilson, F.H., Ruth, J.R., Paulk, J., Tsherniak, A., Marlow, S.E., Vazquez, F., Weir, B.A., Fitzgerald, M.E., Tanaka, M., *et al.* (2016). MTAP deletion confers enhanced dependency on the PRMT5 arginine methyltransferase in cancer cells. Science *351*, 1214-1218.
Landi, M.T., Consonni, D., Rotunno, M., Bergen, A.W., Goldstein, A.M., Lubin, J.H., Goldin, L., Alavanja, M., Morgan, G., Subar, A.F., *et al.* (2008). Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. BMC Public Health *8*, 203.
Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495-501.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.
Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.
McGranahan, N., and Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell *168*, 613-628.
Mitchell, T.J., Turajlic, S., Rowan, A., Nicol, D., Farmery, J.H.R., O'Brien, T., Martincorena, I., Tarpey, P., Angelopoulos, N., Yates, L.R., *et al.* (2018). Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. Cell.
Moch, H., Cubilla, A.L., Humphrey, P.A., Reuter, V.E., and Ulbright, T.M. (2016). The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part A: Renal, Penile, and Testicular Tumours. Eur Urol *70*, 93-105.
Moore, A.L., Kuipers, J., Singer, J., Burcklen, E., Schraml, P., Beisel, C., Moch, H., and Beerenwinkel, N. (2018). Intra-tumor heterogeneity and clonal exclusivity in renal cell carcinoma. bioRxiv.
Morganella, S., Alexandrov, L.B., Glodzik, D., Zou, X., Davies, H., Staaf, J., Sieuwerts, A.M., Brinkman, A.B., Martin, S., Ramakrishna, M., *et al.* (2016). The topography of mutational processes in breast cancer genomes. Nat Commun *7*, 11383.
Muller, F.L., Aquilanti, E.A., and DePinho, R.A. (2015). Collateral Lethality: A new therapeutic strategy in oncology. Trends Cancer *1*, 161-173.
Munoz-Fontela, C., Mandinova, A., Aaronson, S.A., and Lee, S.W. (2016). Emerging roles of p53 and other tumour-suppressor genes in immune regulation. Nat Rev Immunol *16*, 741-750.
Newburger, D.E., Kashef-Haghighi, D., Weng, Z., Salari, R., Sweeney, R.T., Brunner, A.L., Zhu, S.X., Guo, X., Varma, S., Troxell, M.L., *et al.* (2013). Genome evolution during progression to breast cancer. Genome Res *23*, 1097-1108.
Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res *31*, 3812-3814.
Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., *et al.* (2012). The life history of 21 breast cancers. Cell *149*, 994-1007.
Niknafs, N., Beleva-Guthrie, V., Naiman, D.Q., and Karchin, R. (2015). SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing. PLoS Comput Biol *11*, e1004416.
Nixon, K.C. (1999). The Parsimony Ratchet, a new method for rapid parsimony analysis. Cladistics *15*, 407-414.
Owada, Y., Muto, S., Takagi, H., Inoue, T., Watanabe, Y., Yamaura, T., Fukuhara, M., Okabe, N., Matsumura, Y., Hasegawa, T., *et al.* (2017). Correlation between mutation burden of tumor and immunological/clinical parameters in considering biomarkers of immune checkpoint inhibitors for non-small cell lung cancer (NSCLC). J Clin Oncol *35*.
Ricketts, C.J., De Cubas, A.A., Fan, H., Smith, C.C., Lang, M., Reznik, E., Bowlby, R., Gibb, E.A., Akbani, R., Beroukhim, R., *et al.* (2018). The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. Cell Rep *23*, 313-326 e315.
Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., Ho, T.S., *et al.* (2015). Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science *348*, 124-128.

Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A., Demeulemeester, J., Ju, Y.S., Zamora, J., Detering, H., Li, Y., Contino, G., Dentro, S.C., *et al.* (2017). Pan-cancer analysis of whole genomes reveals driver rearrangements promoted by LINE-1 retrotransposition in human tumours. bioRxiv.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Cote, A., and Shah, S.P. (2014). PyClone: statistical inference of clonal population structure in cancer. Nat Methods *11*, 396-398.

Sausen, M., Leary, R.J., Jones, S., Wu, J., Reynolds, C.P., Liu, X., Blackford, A., Parmigiani, G., Diaz, L.A., Jr., Papadopoulos, N., *et al.* (2013). Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. Nat Genet *45*, 12-17.

Schliep, K.P. (2011). phangorn: phylogenetic analysis in R. Bioinformatics *27*, 592-593.

Shain, A.H., and Pollack, J.R. (2013). The spectrum of SWI/SNF mutations, ubiquitous in human cancers. PLoS One *8*, e55119.

Shen, R., and Seshan, V.E. (2016). FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res *44*, e131.

Tang, Y.C., and Amon, A. (2013). Gene copy-number alterations: a cost-benefit analysis. Cell *152*, 394-405.

Tang, Y.C., Williams, B.R., Siegel, J.J., and Amon, A. (2011). Identification of aneuploidy-selective antiproliferation compounds. Cell *144*, 499-512.

Tubio, J.M.C., Li, Y., Ju, Y.S., Martincorena, I., Cooke, S.L., Tojo, M., Gundem, G., Pipinikas, C.P., Zamora, J., Raine, K., *et al.* (2014). Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science *345*, 1251343.

Turajlic, S., Xu, H., Litchfield, K., Rowan, A., Chambers, T., Lopez, J.I., Nicol, D., O'Brien, T., Larkin, J., Horswell, S., *et al.* (2018). Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. Cell *173*, 581-594 e512.

Van Loo, P., Nordgard, S.H., Lingjaerde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., *et al.* (2010). Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A *107*, 16910-16915.

Wang, P.P., Parker, W.T., Branford, S., and Schreiber, A.W. (2016). BAM-matcher: a tool for rapid NGS sample matching. Bioinformatics *32*, 2699-2701.

Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., *et al.* (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature *512*, 155-160.

Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.H., Zhang, C., Ren, X., Protopopov, A., Chin, L., *et al.* (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. Cell *153*, 919-929.

FIGURE LEGENDS

Figure 1: Study design and genomic landscape. (a) A schematic illustration of the dissection of multiple tumor samples from the center of the tumor towards the tumor's periphery, plus metastatic samples in the adrenal gland as well as normal samples. For the analysis, the normal sample more distant from the tumor and with absence of tumor nuclei was chosen as reference. (b) Summary of subjects and samples that underwent different analyses based on DNA availability: whole-genome sequencing (124 samples from 29 subjects), deep targeted sequencing of cancer driver genes (139 samples from 38 subjects), genome-wide methylation (139 samples from 28 subjects) or SNP array profiling (only tumor samples, 101 samples from 38 subjects). (c) Tumor genomic alterations across histological subtypes. Shown are genome level changes, such as mutation burden, numbers of structural variants (SV) and retrotransposition events (TE), and various forms of genomic alterations (denoted by different colors) in particular genomic regions or genes.

Figure 2: Examples of multi-regional and lineage phylogenetic trees from two tumors. In multi-regional trees (a,c), the branches represent the regions (primary tumor, T01-T10, and metastatic samples, M01-M03), with the lengths of branch or trunk proportional to the number of single nucleotide variants (SNVs) shared by the regions. The root of the trees are normal cells without somatic SNVs. Driver genes and recurrent somatic copy number alterations are noted on the trees. The portions of trees with internal and terminal branches are enlarged to the right of each. In the lineage phylogenetic trees (b,d), the evolution history is described for each region, with circle representing a subclone (colored: subclone present; blank: subclone absent). Arrows link the parent and descendant subclones. Subclone numbers and graduation of circle colors are based on SNV CCFs ranking, from the highest to the lowest. For example, from the most-recent common ancestor (MRCA) clone C1 in panel d, four descendant subclones emerged, and from subclone C2, one new subclone, C3. In turn, C3 continued expanding divergently into two descendant subclones, one of which, C8, was metastatic-specific and was shared by all adrenal gland metastatic samples.

Figure 3: Somatic copy number alterations (SCNAs). (a) Genome-wide SCNAs across tumors in a circos plot. The outermost circle denotes the genomic positions and the inner ones represent SCNAs for each tumor. Colored bars denote whether SCNAs occur in terminal branches, internal branches, or trunk. (b) The distribution of mean cancer cell fraction (CCF) of SCNAs across all pRCC tumors. (c) SCNAs of tumor pRCC2_1824_13. Top panel: genome-wide SCNAs on ten primary tumors (T01-T10) and three metastatic samples (M01, M02 and M03); T10 has low purity and has no SCNAs. Bottom panels: metastatic sample-specific SCNAs on chromosome 4 for total copy number log-ratio (red line: estimated total copy number log-ratio; green line: median; purple line: diploid state). DLOH: hemizygous deletion loss of heterozygosity; HET: diploid heterozygous; NLOH: copy neutral loss of heterozygosity; ALOH: amplified loss of heterozygosity; ASCNA: allele-specific copy number amplification; BCNA: balanced copy number amplification.

Figure 4: **Structural variants (SV) and retrotransposition events (TE).** (a) Frequency of SV events and TE insertions for each sample. (b, c, d) Circos plots for SV events for three tumors; involved driver genes are noted. Alu: elements originally characterized by the action of the

Arthrobacter luteus (Alu) restriction endonuclease; ERVK:mouse endogenous retrovirus K; L1: Long interspersed element-1. (e) The distribution of mean cancer cell fraction (CCF) of SVs across tumors.

Figure 5: Intratumor heterogeneity (ITH) of methylation profiles. (a) methylation ITH on genomic regions for each sample and tumor subtype. (b) examples from methylation multi-regional trees from two tumors. The branches represent the regions and the lengths of branch or trunk are proportional to the distances of methylation profiles between regions. (c) Unsupervised hierarchical clustering of methylation profiles, based on the top 1% most variable methylation probes. Sample IDs are followed by the purity estimated by SCNAs or SNV VAF in parentheses. Boxed profiles denote the clustered samples with low purity (<30%). The background colors of the sample IDs represent different histological subtypes and tumor or normal tissue samples.

Figure 6: **Clonality of different genomic alteration types.** The proportion of somatic copy number alterations (SCNAs), single nucleotide variants (SNVs) and structural variants (SVs) in trunk, internal branches and terminal branches for each tumor (a) and across histological subtypes (b). The *p*-values in panel (b) were given by Wilcoxon rank test to compare subclone proportions between different types of genomic alterations. Tumor-level lineage phylogenetic tree (LPT) of pRCC2_1824 (same as the region-specific LPTs in Fig 2d) is depicted based on CCF of SNVs (CCF for each clone is labeled) (d). The occurrence of individual SCNAs (wavy pink lines) and SVs (blue lines) events are marked on the SNV LPT based on their respective SCNA and SV CCFs. Note that the SNV CCF of the metastasis-specific subclone C8 is the average of SNV CCFs across both primary tumor and metastatic regions.

METHODS

Patients and specimens

This study was based on archived samples collected at the Regina Elena Cancer Institute, Rome, Italy, under patients' written informed consent to allow banking of biospecimens for future scientific research. This work was excluded from IRB Review per 45 CFR 46 and NIH policy for the use of specimens/data by the Office of Human Subjects Research Protections (OHSRP) of the National Institutes of Health.

The study population included 29 patients with kidney cancers, including 13 with papillary type 1 (pRCC1); 12 with papillary type 2 (pRCC2); and one each with collecting duct tumor (cdRCC); renal fibrosarcoma rSRC (with negative stain for AE1/AE3, PAX8, CD99, FLI-1, WT1, actine ml, desmine, Myod-1, and HMB45; and positive staining for vimentine and S-100 (focal)); mixed pRCC1/pRCC2 and an unclassified renal cancer with mixed features of pRCC2 and cdRCC. The histological diagnosis was reviewed by an expert uropathologist (Dr. Steno Sentinelli) based on the 2016 World Health Organization (WHO) classification of renal tumors(Moch et al., 2016).

Based on DNA sample availability, we conducted whole genome sequencing (WGS) on 124 samples from 29 subjects, deep targeted sequencing on 139 samples from 38 subjects, SNP array genotyping on 101 samples from 38 subjects, and genome-wide methylation profiling on 139

samples from 28 subjects (**Figure1b,** more details in **Figure S27**). All assays were performed on tumor, metastasis and normal tissue samples, with the exception of the SNP array genotyping, which was conducted only on tumor samples.

Study Design

All tumors were treatment-naive. We used a study design with multiple tumor samples taken at a distance of ~1.5 cm from each other starting from the center of the tumor towards the periphery, plus multiple samples from the most proximal to most distant area outside the tumor. When present, we also collected multiple samples from metastatic regions outside the kidney (adrenal gland) (**Figure 1a**). For the analyses presented here, we analyzed all multiple tumor and metastatic samples/tumor with at least 70% tumor nuclei at histological examination. As a reference, we used the furthest "normal" sample from each tumor, with histologically-confirmed absence of tumor nuclei.

Sequencing platforms:

Whole genome sequencing

Genomic DNA was extracted from fresh frozen tissue using the QIAmp DNA mini kit (Qiagen) according to the manufacturer's instructions. Libraries were constructed and sequenced on the Illumina HiSeqX at the Broad Institute, Cambridge, MA with the use of 151-bp paired-end reads for whole-genome sequencing (mean depth= 65.7x and 40.1x, for tumor and normal tissue, respectively). Output from Illumina software was processed by the Picard data-processing pipeline to yield BAM files containing well-calibrated, aligned reads to genome-build hg19. All sample information tracking was performed by automated LIMS messaging. More details are included in the Supplemental Material.

Genome-wide SNP genotyping

Genome-wide SNP genotyping, using Infinium HumanOmniExpress-24-v1-1-a BeadChip technology (Illumina Inc. San Diego, CA), was performed at the Cancer Genomics Research Laboratory (CGR). Genotyping was performed according to manufacturer's guidelines using the Infinium HD Assay automated protocol. More details are included in the Supplemental Materials.

Targeted Sequencing

A targeted driver gene panel was designed for 254 candidate cancer driver genes(Lawrence et al., 2014). For each sample, 50 ng genomic DNA was purified using Agencourt AMPure XP Reagent (Beckman Coulter Inc, Brea, CA, USA) according to manufacturer's protocol, prior to the preparation of an adapter-ligated library using the KAPA JyperPlus Kit (KAPA Biosystems, Wilmington, MA) according to KAPA-provided protocol. Libraries were pooled, and sequence capture was performed with NimbleGen's SeqCap EZ Choice (custom design; Roche NimbleGen, Inc., Madison, WI, USA), according to the manufacturer's protocol. The resulting post-capture enriched multiplexed sequencing libraries were used in cluster formation on an Illumina cBOT (Illumina, San Diego, CA, USA) and paired-end sequencing was performed using an Illumina HiSeq 4000 following Illumina-provided protocols for 2x150bp paired-end sequencing at The National Cancer Institute Cancer Genomics Research Laboratory (CGR). More details are included in the Supplemental Materials.

Methylation analysis

400 ng of sample DNA, according to Quant-iT PicoGreen dsDNA quantitation (Life Technologies, Grand Island, NY), was treated with sodium bisulfite using the EZ-96 DNA Methylation MagPrep Kit (Zymo Research, Irvine, CA) according to manufacturer-provided protocol. Bisulfite conversion modifies non-methylated cytosines into uracil, leaving 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) unchanged. High-throughput epigenome-wide methylation analysis, using Infinium MethylationEPIC BeadChip (Illumina Inc., San Diego, CA) which uses both Infinium I and II assay chemistry technologies was performed according to manufacturer-provided protocol at CGR. More details are included in the Supplemental Materials.

Bioinformatics pipelines:

Whole-Genome data processing and alignment

The WGS FASTQ files were processed and aligned through an in-house computational analysis pipeline, according to GATK best practice for somatic short variant discovery (https://software.broadinstitute.org/gatk/best-practices/). First, quality of short insert paired-end reads was assessed by FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Next, paired-end reads were aligned to the reference human genome (build hg19) using BWA-MEM aligner in the default mode(Li and Durbin, 2009). The initial BAM files were post-processed to obtain analysis-ready BAM files. In particular, sequencing library insert size and sequencing coverage metrics were assessed, and duplicates were marked using Picard tools (https://broadinstitute.github.io/picard/); indels were realigned and base quality scores were re-calibrated according to GATK best practice; In addition, BAM-matcher was used to determine whether two BAM files represent samples from the same tumor(Wang et al., 2016); VerifyBamID was used to check whether the reads were contaminated as a mixture of two samples(Jun et al., 2012).

Somatic mutation calling from whole genome sequencing data

The analysis-ready BAM files from tumor, metastasis, and matched normal samples were used to call somatic variants by MuTect2 (GATK 3.6, https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_gatk_tools_walkers_cancer_m2_MuTect2.php) with the default parameters. In the generated VCF files, somatic variants notated as "Somatic" and "PASS" were kept. A revised method described by Hao, et al.(Hao et al., 2016) was used to further filter the somatic variants. More details are included in the Supplemental Materials. For indels, we reported those that overlapped across three different software, mutect2(Cibulskis et al., 2013), strelka2(Kim et al., 2018), and tnscope(Freed et al., 2018). Indels were left-aligned and normalized using bcftools. The intersection of "PASS" indels from all three calling tools were combined by GATK "CombineVariants". Additional filters were applied to the final set before downstream analysis: tumor alternative allele fraction >0.04; normal alternative allele fraction <0.02; tumor total read depth>=8; normal total read depth>=6; and tumor alternative allele read depth >3.

Identification of putative driver mutations and driver genes

To create putative cancer driver gene and mutation lists, we first listed the putative cancer driver genes on the basis of recent large-scale TCGA Pan-kidney cohort (KICH+KIRC+KIRP) sequencing data (http://firebrowse.org), i.e. the significantly mutated genes identified by MutSig2CV algorithm with q value less than 0.1. In addition, we included the genes from the COSMIC cancer gene census list (May 2017, http://cancer.sanger.ac.uk/census) in the putative kidney driver gene set. Putative driver mutations were defined if they met one of the following requirements: (i) if the variant was predicted to be deleterious, including stop-gain, frameshift and splicing mutation, and had a SIFT(Ng and Henikoff, 2003) score < 0.05 or a PolyPhen(Adzhubei et al., 2013) score >0.995 or a CCAD(Kircher et al., 2014) score >0.99; or (ii) If the variant was identified as a recurrent hotspot (statistically significant, http://cancerhotspots.org) or a 3D clustered hotspot (http://3dhotspots.org) in a population-scale cohort of tumor samples of various cancer types using a previously described methodology(Chang et al., 2016; Gao et al., 2017).

Mutational signature analysis from whole genome sequencing data

Mutational signatures were extracted using our previously developed computational framework SigProfiler(Alexandrov et al., 2013b). A detailed description of the workflow of the framework can be found in Refs(Alexandrov et al., 2018; Alexandrov et al., 2013b) , while the code can be downloaded freely from: https://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler. Topography analysis of mutational signatures was performed using our previously developed methodology(Morganella et al., 2016). Detailed description of the methodology can be found in Supplemental Methods.

Somatic copy-number alteration analysis

Allele-specific copy-number alteration (SCNA) analysis was performed using FACETS(Shen and Seshan, 2016) v0.5.6 (https://github.com/mskcc/facets) with the following parameters to increase the strictness of the segments: normal read filter depth=15; window size = 5000. The 'snp-pileup' command with argument '--min-map-quality 20 --min-base-quality 20 --min-read-counts 15,0' was used to extract read counts of the reference and alternate alleles from tumor and normal tissue samples BAM files separately. SCNA events were allocated to different genotypes according to the total copy number and minor copy numbers. SCNA events were defined as chromosome-arm if the same SCNA level overlapped with at least 90% of the chromosome arm's coordinates. Focal SCNA events were defined if their individual lengths were less than half the length of a chromosome arm. The fraction of the copy-number-altered genome was defined as the fraction of the genome with either non-diploid copy-number or evidence of loss of heterozygosity. Besides SCNA events, tumor purity and ploidy were estimated using FACETS. For downstream bioinformatic analysis, we excluded seven samples, for which the purity could not be estimated by FACETS, including pRCC1_1689_06_T02, pRCC1_1472_01_T01, pRCC2_1824_13_T10, pRCC2_1782_08_T08, pRCC2_1552_03_T03, pRCC2_1410_02_T01, mixRCC_2028_03_T01.

Clonality analysis and subclonal hierarchy inference

Subclones were defined by single nucleotide variants with clustered cancer cell fractions (CCF) using PyClone(Roth et al., 2014). Pyclone is a Bayesian clustering method for grouping sets of deeply sequenced somatic mutations into putative clonal clusters while estimating their cellular prevalences and accounting for allelic imbalances introduced by segmental copy number changes

and tumor tissue sample contamination. Input data were somatic single-nucleotide variants (SNVs) and corresponding copy number alterations as well as tumor purity estimated based on WGS data. The top 300 SNVs with sequencing depth >100x for each sample were selected and clustered according to similar CCF, after which sets of clustered mutations were identified as a subclones. To obtain reliable estimates of mutation cellularity, we clustered the mutations that were present in the majority of the tumor regions.

The tumor subsclonality phylogenetic reconstruction algorithm SCHISM(Niknafs et al., 2015) (SubClonal Hierarchy Inference from Somatic Mutations) was used to construct lineage phylogenetic trees. SCHISM was run with PyClone output (subclone clusters and mutations to each cluster) and default parameter settings to infer the order of somatic alterations and thus define subclonal hierarchy in each patient. The lineage phylogenetic tree was plotted using R DiagrammeR package (http://rich-iannone.github.io/DiagrammeR/) with cellular prevalence for each cluster.

Somatic structural variant calling

We used the Meerkat algorithm(Yang et al., 2013) to call somatic SVs and estimate the corresponding genomic positions of breakpoints from recalibrated BAM files. Merkaat has been found to perform better than other previous software in a large analysis across different cancer types(Alaei-Mahabadi et al., 2016). We used parameters adapted to the sequencing depth for both tumor and normal tissue samples and the library insert size. In summary, candidate breakpoints were first found based on soft-clipped and split reads, which requires identifying at least two discordant read pairs, with one read covering the actual breakpoint junction, and then confirmed to be the precise breakpoints by local alignments ('meerkat.pl'). Mutational mechanisms were predicted based on homology and sequencing features ('mechanism.pl). SVs from tumor genomes were filtered by those in normal genomes. SVs found in simple or satellite repeats were also excluded from the output ('somatic_sv.pl'). The final somatic SVs were annotated as a uniformed format for all breakpoints ('fusions.pl'). We compared the results obtained by Meerkat with those obtained by Novobreak(Chong et al., 2017) (v1.1.3rc) (Supplemental Material). We opted to retain Meerkat-derived results because they were more conservative and were largely confirmed by laboratory testing. The CCF of SVs in each region was estimated by Svclone(Cmero et al., 2017); the copy-number subclone information generated by the Battenberg algorithm(Nik-Zainal et al., 2012) was used as input for the filter step. To substantially increase the number of variants available for clustering, we applied the coclustering mode to estimate CCF for both SVs and SNVs simultaneously and calculated the average CCF of SVs across regions. SVs with CCF > 0.5 were defined as clonal.

Validation of somatic structural variants

We selected four in-frame fusions *MALAT-TFEB, MET-MET* deletion, *STRN-ALK*, and *EWSR1-PATZ1,* for validation by reverse transcription and PCR-based sequencing. The *MALAT-TFEB* and *EWSR1-PATZ1* fusions were validated and confirmed by Sanger sequencing. The other two fusions were not validated because of poor RNA quality from FFPE samples (RIN=2.6). We selected 381 additional structural variants from pRCC tumors for validation by Ion Torrent PGM Sequencing using a custom AmpliSeq primer pool. We were able to successfully design compatible primers for 303 of them. These included: 87 trunk SVs, 115 internal branch SVs, and 101 terminal branch SVs. 5 SVs failed QC. Among the remaining 298 SVs, 263

(263/298=88.3%) were validated at the tumor level and 217 (217/263=83%) were validated at clonal level as trunk, internal, or terminal branches. Further details are in the Supplemental Material.

Somatic mutation calling from deep targeted sequencing data

We utilized the WGS pipeline to process raw reads, align reads to the reference human genome hg19, and to call somatic SNVs by GATK MuTect2. We then performed multiple mutation filtering and mutation annotation. Given the deep sequencing coverage, we used strict filtering criteria, retaining variants with read depth >=30 in tumor samples and the number of variant supporting reads>=8. Among the 254 targeted candidate cancer driver genes, we found 67 genes with non-synonymous single nucleotide variant detected by targeted sequencing, 93.6% of which were SNVs called based on WGS data. In contrast, 78.6% of SNVs detected by WGS data were validated by targeted sequencing. High correlation was observed for the variant allele fraction between target sequencing and whole genome sequencing (Correlation coefficient= 0.87, P $=8.54x10^{-88}$).

Copy-number analysis from genome-wide SNP genotyping data

Genome Studio (Illumina, Inc.) was used to cluster and normalize raw genotyping data. Both BAF and LogR data were generated and exported for downstream analysis. ASCAT(Van Loo et al., 2010) (https://www.crick.ac.uk/peter-van-loo/software/ASCAT) was used to estimate the allele-specific copy numbers without matched normal data. Purity, ploidy, and segmentation data generated by ASCAT were compared to those generated by FACETS (**Figure S14**).

Analysis for DNA methylation profiling

Genome-wide DNA methylation was profiled on Illumina Infinium methylation EPIC arrays (Illumina, San Diego, USA). Methylation of tumor and normal samples was measured according to the manufacturer's instruction at CGR. Raw methylation densities were analyzed using the RnBeads pipeline(Assenov et al., 2014) and the minfi package(Aryee et al., 2014). In total, we retained 814,408 probes for the downstream analysis. Duplicated samples were selected based on probe intensity, SNP calling rate, and the percentage of failed probes. No batch effects were identified and there were no plating issues. “Functional Normalization”(Fortin et al., 2014), implemented in the minfi R package was used to perform normalization to obtain the final methylation levels (beta value). Hyper- and hypo-methylation were arbitrarily defined by at least 20% in-/decrease relative to the matched normal samples, respectively (Further details in the **Supplemental Material**).

Unsupervised clustering of SCNAs and methylation profiles

We estimated the SCNAs by considering both total copy numbers and minor copy numbers as in **Table S14**. Unsupervised hierarchical clustering was performed using Euclidean distance and Ward’s linkage method. For the methylation profiles, we selected the top 1% of probes with the greatest difference between maximum and minimum methylation levels within each tumor. For hierarchical clustering, a Euclidean distance was calculated and Ward’s linkage was performed. Normal samples were excluded for the calculation of intratumoral DNA methylation range. Heatmaps were drawn using the superheat (https://github.com/rlbarter/superheat) and ComplexHeatmap R package.

Measuring intratumoral heterogeneity of SNVs and methylation in genomic regions
We measured genomic region-specific intratumoral heterogeneity (ITH) of each tumor with at least three samples for different genomic profiles: SNV and Methylation. For SNVs, ITH was measured by the median mutation variability for each tumor across different genomic regions/contexts, including intergenic, 1to5kb, promoters, 5'-UTRs, first exon, exon-intron boundaries, exons, introns, intron-exon boundaries, 3'-UTRs, lncrna_gencode and enhancers_fantom defined in R annotatr package (https://github.com/hhabra/annotatr). For each mutation and each patient, the mutation variability was measured as 1-N (mutated samples) /M (sample size). The higher the mutation variability, the more ITH.

Similar to SNVs, DNA methylation variability was calculated between normal samples and within samples in each tumor. Interindividual variability was analyzed by comparing normal samples from all subjects. The genomic region-specific methylation inter- and intra-tumor heterogeneity was measured by the median methylation variability of involved CpG sites.

Multi-regional analysis of SNVs
All SNVs that passed the filtering criteria were considered for constructing multi-regional trees. Trees were built using binary presence or absence matrices built from the regional distribution of variants within the tumor. The R Bioconductor package phangorn(Schliep, 2011) was utilized to perform the parsimony ratchet method(Nixon, 1999), generating unrooted trees. Branch lengths were determined using the acctran function.

Multi-regional analysis of SCNAs
Allelic somatic copy number alterations were identified by FACETS. Similar to the method described by Gao, et al.(Gao et al., 2016), maximum-parsimony trees were created using SCNA matrices, based on the parsimony ratchet algorithm implemented in the R package phangorn. Amplifications, neutral changes, and deletions were treated as characters and missing values were treated as ambiguous sites. Events occurring on the sex chromosomes were not analyzed. Branch lengths and ancestral character probability distributions were inferred using the Acctran algorithm. The CCF of each SCNA at chromosome cytoband level for each tumor region was obtained from FACETS. Specifically, the CCF was estimated based on the Expectation-Maximization (EM) cellular fraction (cf.em) of each segmentation divided by the max cf.em in the same sample. The average CCF of each SCNA was calculated and SCNAs with average CCF > 0.5 were defined as clonal.

Multi-regional analysis of methylation
The top 1% (n=8,144) of CpG sites with the greatest intratumoral methylation range (excluding normal samples) were used to generate DNA methylation Euclidean distance matrices. Methylation-based multi-regional trees were inferred by the minimum evolution method(Brocks et al., 2014; Desper and Gascuel, 2002) using the fastme.bal function in the R package ape (https://cran.r-project.org/web/packages/ape/index.html). Confidence for clades on multi-regional trees was assessed by bootstrapping using the boot.phylo function from the R package ape (1000 bootstrap replicates).

Visualization of multi-regional trees

Multi-regional trees were exported in the Newick format from R and converted to data.frame using ggplot R package (https://cran.r-project.org/web/packages/ggplot2/index.html). The trees were rebooted by the bottom node annotated as normal/germline sample. Enlarged trees were created manually by revising the tree branch length for some tumors. Each tumor region was represented as a tip of the tree. Tree data were output for visualization using modified function "plot.tree.clone.as.branch" in clonvol package (https://github.com/hdng/clonevol). Potential driver events including driver genes, structural variants, and focal copy number alterations were annotated on the branch side of SNV RPTs. Finally, trees were manually adjusted for better visualization.

Statistical analysis

Statistical analyses were performed using R software (https://www.r-project.org/). Categorical variables were compared using the Fisher's Exact test. Group variables were compared using Wilcoxon rank sum and signed rank test. *P* values were derived from two-sided tests and those less than 0.05 were considered as statistically significant.

Accession codes

The genomic data have been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession code phs001573.v1.p1.

Websites

Picard tools
https://broadinstitute.github.io/picard/

oncotator
https://github.com/broadinstitute/oncotator

COSMIC cancer gene census
http://cancer.sanger.ac.uk/census

Cancer hotspots
http://cancerhotspots.org

3D clustered hotspot
http://3dhotspots.org

ASCAT
https://www.crick.ac.uk/peter-van-loo/software/ASCAT

FACETS
https://github.com/mskcc/facets

GATK (Genome Analysis Toolkit)
https://software.broadinstitute.org/gatk/

Bam-readcount (Count DNA sequence reads in BAM files)

https://github.com/genome/bam-readcount

PHYLIP
http://evolution.genetics.washington.edu/phylip.html

fpFilter Perl script,
https://github.com/ckandoth/variant-filter

SCHISM (SubClonal Hierarchy Inference from Somatic Mutations)
http://karchinlab.org/apps/appSchism.html

PyClone (Probabilistic model for inferring clonal population structure from deep NGS sequencing)
https://bitbucket.org/aroth85/pyclone/wiki/Home

RnBeads (Comprehensive analysis of DNA methylation data)
http://rnbeads.mpi-inf.mpg.de/

TCGA firebrowse
http://firebrowse.org

COSMIC Mutational Signatures
http://cancer.sanger.ac.uk/cosmic/signatures

Svclone (To infer and cluster cancer cell fractions (CCFs) of structural variant (SV) breakpoints)
https://omictools.com/svclone-tool

gnomAD (The Genome Aggregation Database)
http://gnomad.broadinstitute.org/

R packages:

deconstructSigs
https://github.com/raerose01/deconstructSigs

Phangorn (Phylogenetic Reconstruction and Analysis)
https://github.com/KlausVigo/phangorn

Clonevol (inferring and visualizing clonal evolution in multi-sample cancer sequencing)
https://github.com/hdng/clonevol

DiagrammeR
http://rich-iannone.github.io/DiagrammeR/

ape (analysis of phylogenetics and evolution)
http://ape-package.ird.fr

Minfi (analyze illumina infinium DNA methylation arrays)
http://bioconductor.org/packages/release/bioc/html/minfi.html

ComplexHeatmap (Making complex heatmaps)
https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html

Superheat (generating beautiful and customizable heatmaps)
https://github.com/rlbarter/superheat

Annotatr (Annotation of Genomic Regions to Genomic Annotations)
http://bioconductor.org/packages/release/bioc/html/annotatr.html

Factoextra (Extract and Visualize the Results of Multivariate Data Analyses)
https://github.com/kassambara/factoextra

SUPPLEMENTAL TABLE LEGENDS

Table S1. Coverage information for whole-genome sequencing data. The annotations for each column in tab "WGS_metrics" are included in tab "Note".

Table S2. Patient characteristics. They include age at diagnosis, gender, tumor size, stage and survival status.

Table S3. List of sample IDs. IDs for samples in each genomic and epigenomic analysis, including whole-genome sequencing, genome-wide methylation and SNP array genotyping, and deep targeted sequencing.

Table S4. Gene list for targeted sequencing (from Lawrence, et al. 10).

Table S5. Non-synonymous single nucleotide variants and related functional annotation. Non-synonymous single nucleotide variants were based on whole genome sequencing and/or deep target sequencing.

Table S6. Insertions and deletions (indels) in previously reported cancer driver genes with their functional annotation.

Table S7. Potentially deleterious germline variants in cancer susceptibility genes

Table S8. Cancer cell fraction (CCF) estimates for each tumor subclone. Estimates are based on the PyClone algorythm. SD: standard deviation.

Table S9: Comparison between four identified SNV mutational signatures (a) and three identified indel mutational signatures (b) with previously known SNV and indels mutational signatures, respectively.

Table S10. Number of SNVs contributing to SNV mutational signatures in each sample. Mutational signatures are estimated based on the Sigprofile.

Table S11. Number of indels contributing to indel mutational signatures.

Table S12. Telomere length estimates based on whole-genome sequencing data. Telomere length is estimated using the Telseq algorithm. The annotation of columns in tab "Final-telseq" is included in tab "Note".

Table S13. Segmentation of copy number alterations based on whole genome sequencing. Segmentation estimates are estimated using the FACETS algorithm. The annotation of columns in tab "Facets_cncf_info" is included in tab "Note".

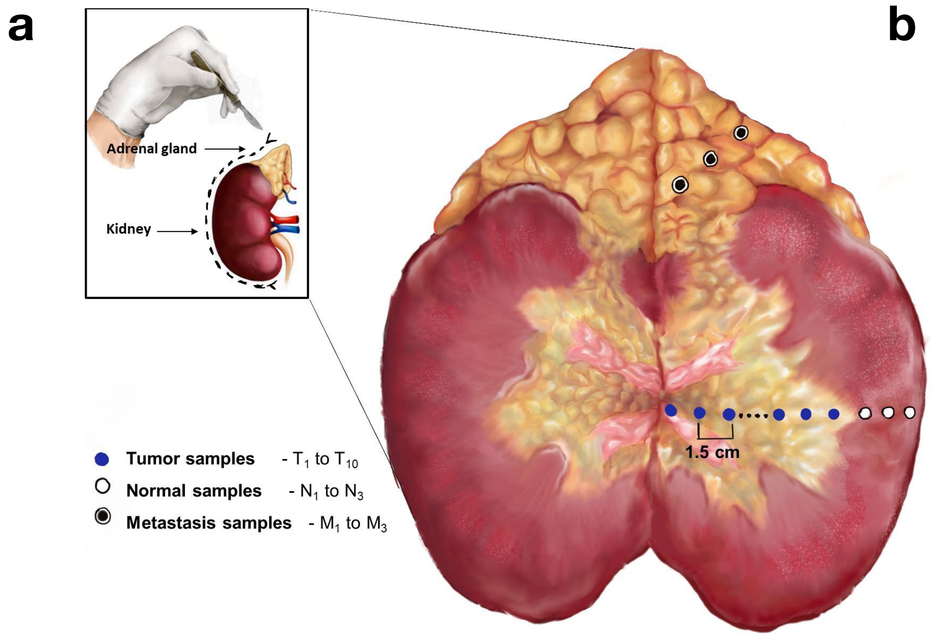
Table S14. Annotations of allele-specific copy number alteration types.

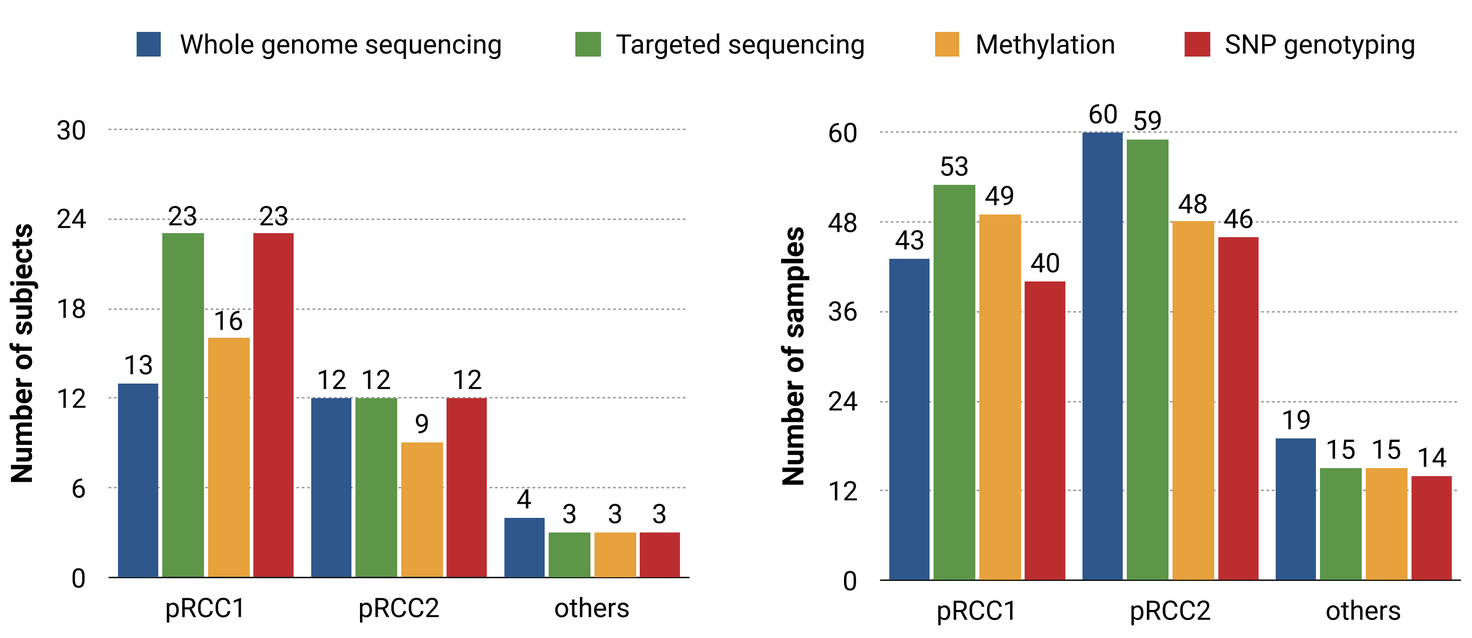
Table S15. Annotation of CDKN2A deletion segment. The locations of the deletion segment are listed for each sample, based on whole genome sequencing and genotyping data.

Table S16. Structural variants (SVs) identified by the Meerkat algorithm. The annotation of columns in tab fusion_list_all" is provided in the Meerkat User Manual (http://gensoft.pasteur.fr/docs/Meerkat/0.185/Manual_0.185.pdf).

Table S17. Retrotransposition events identified by TraFiC. The annotation of columns in tab "TE_trafic" is included in tab "Note".

Figure 1





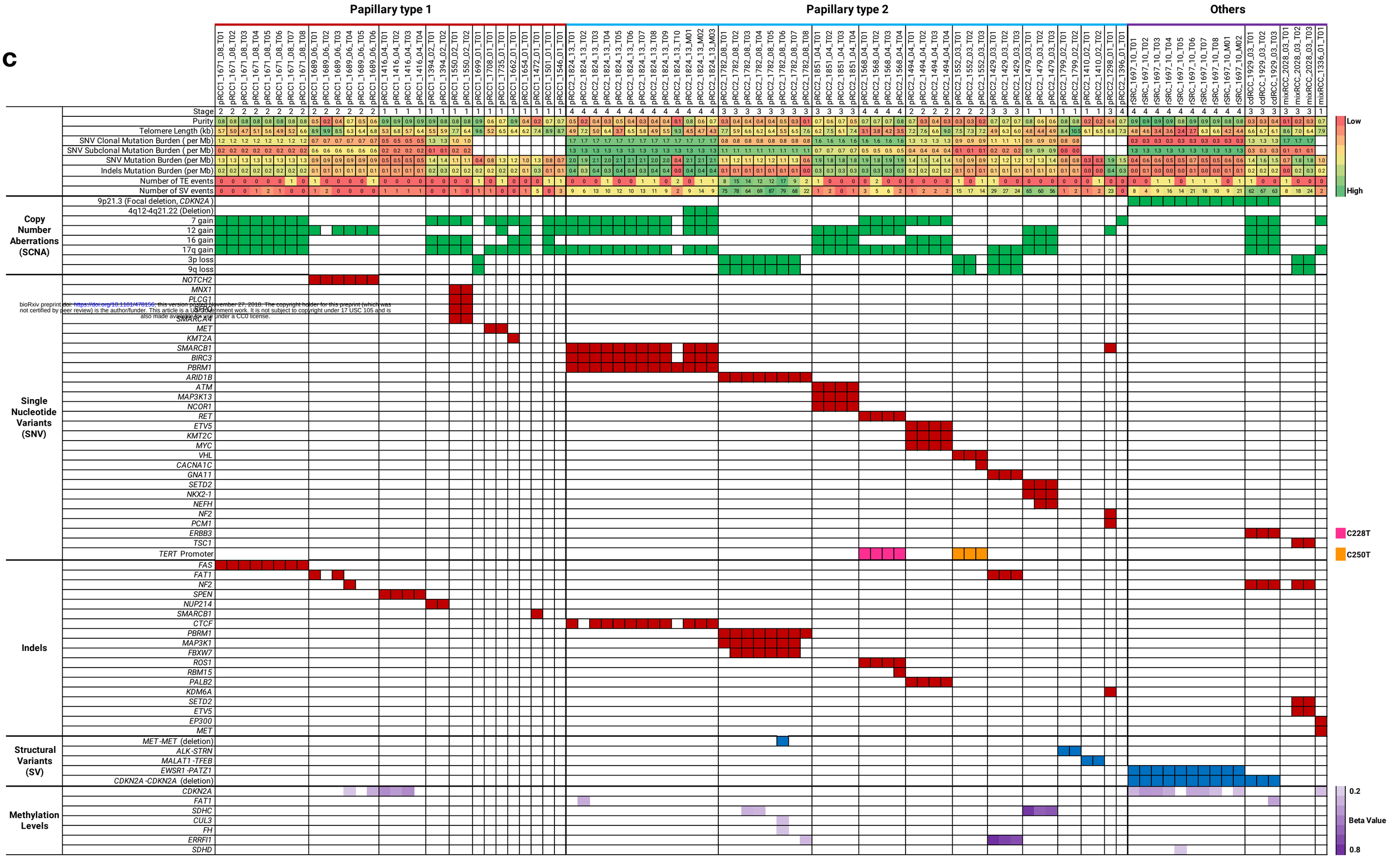
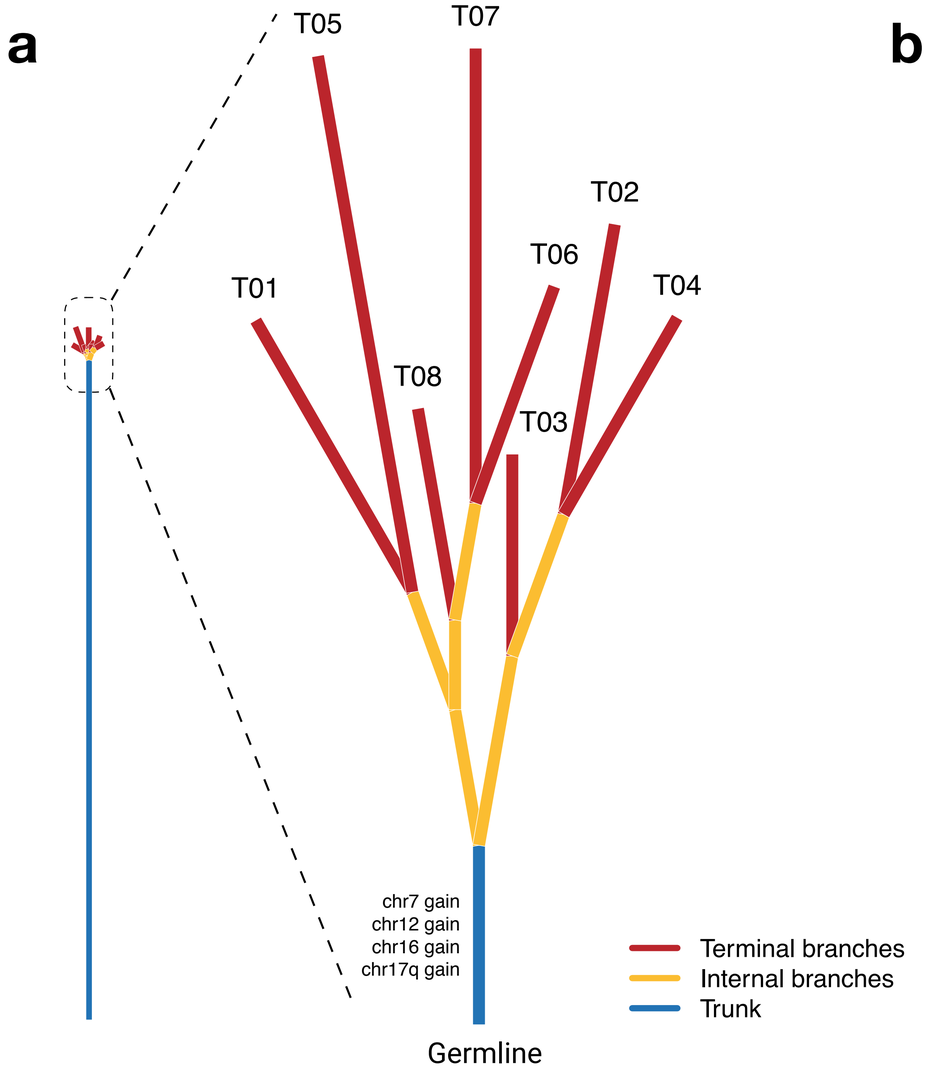
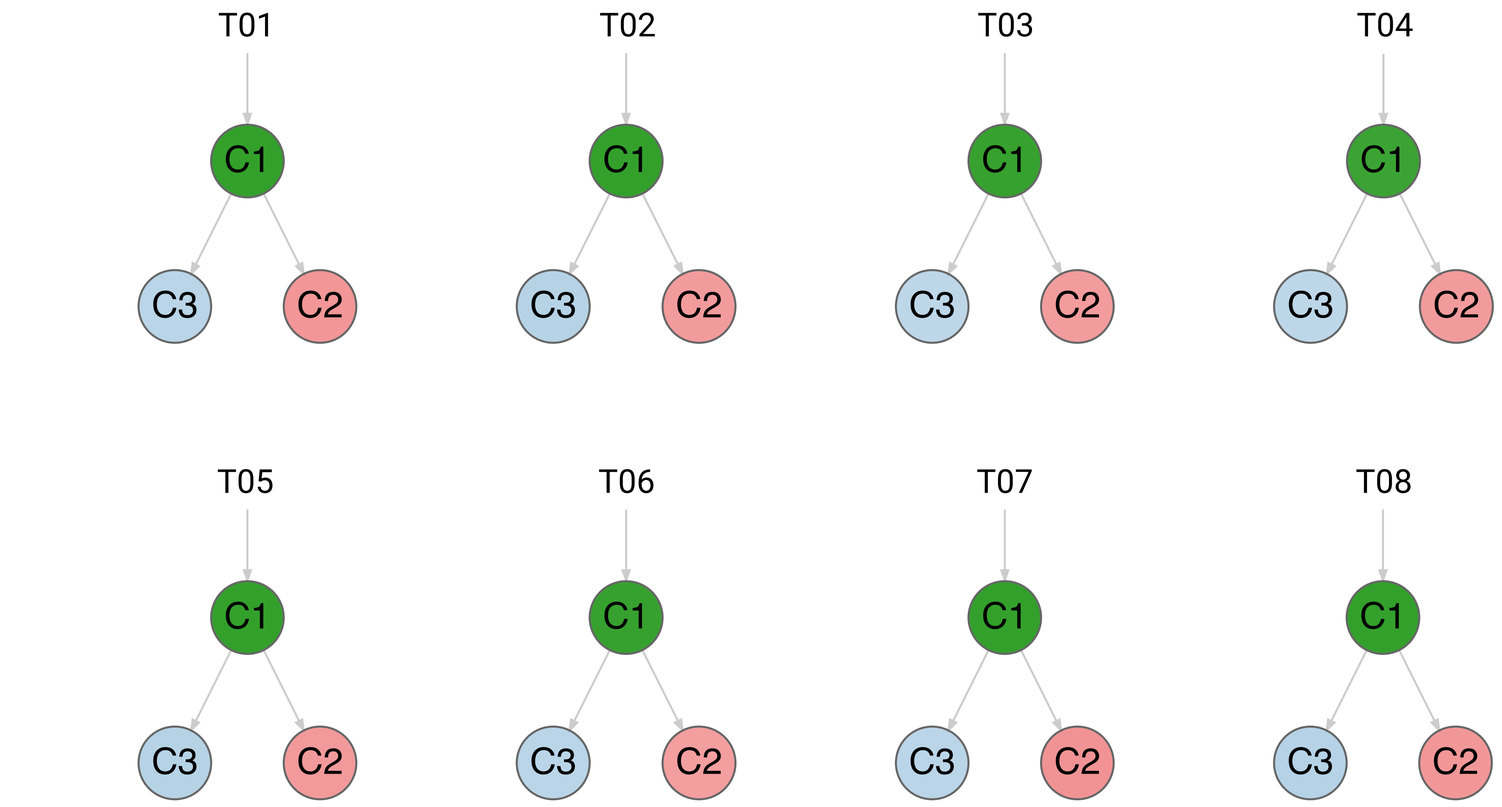
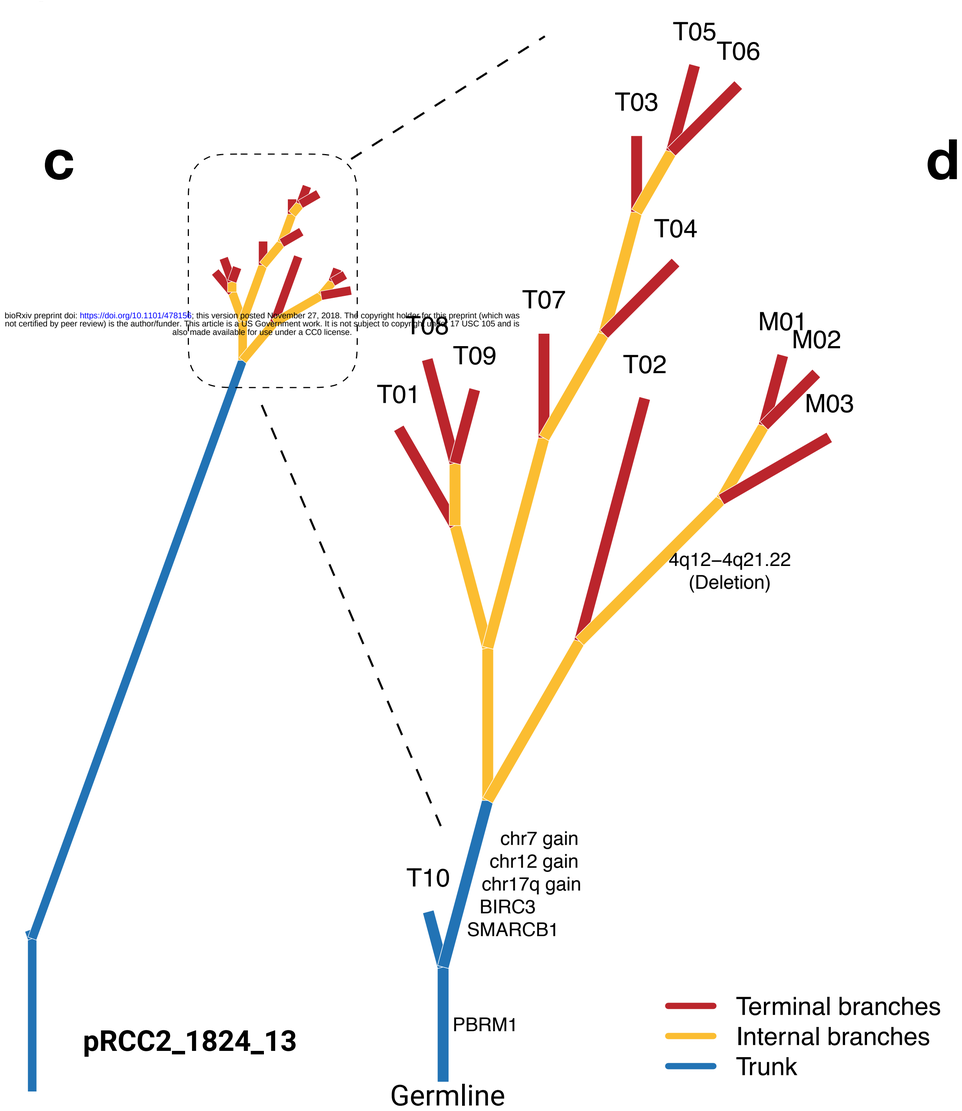


Figure 2







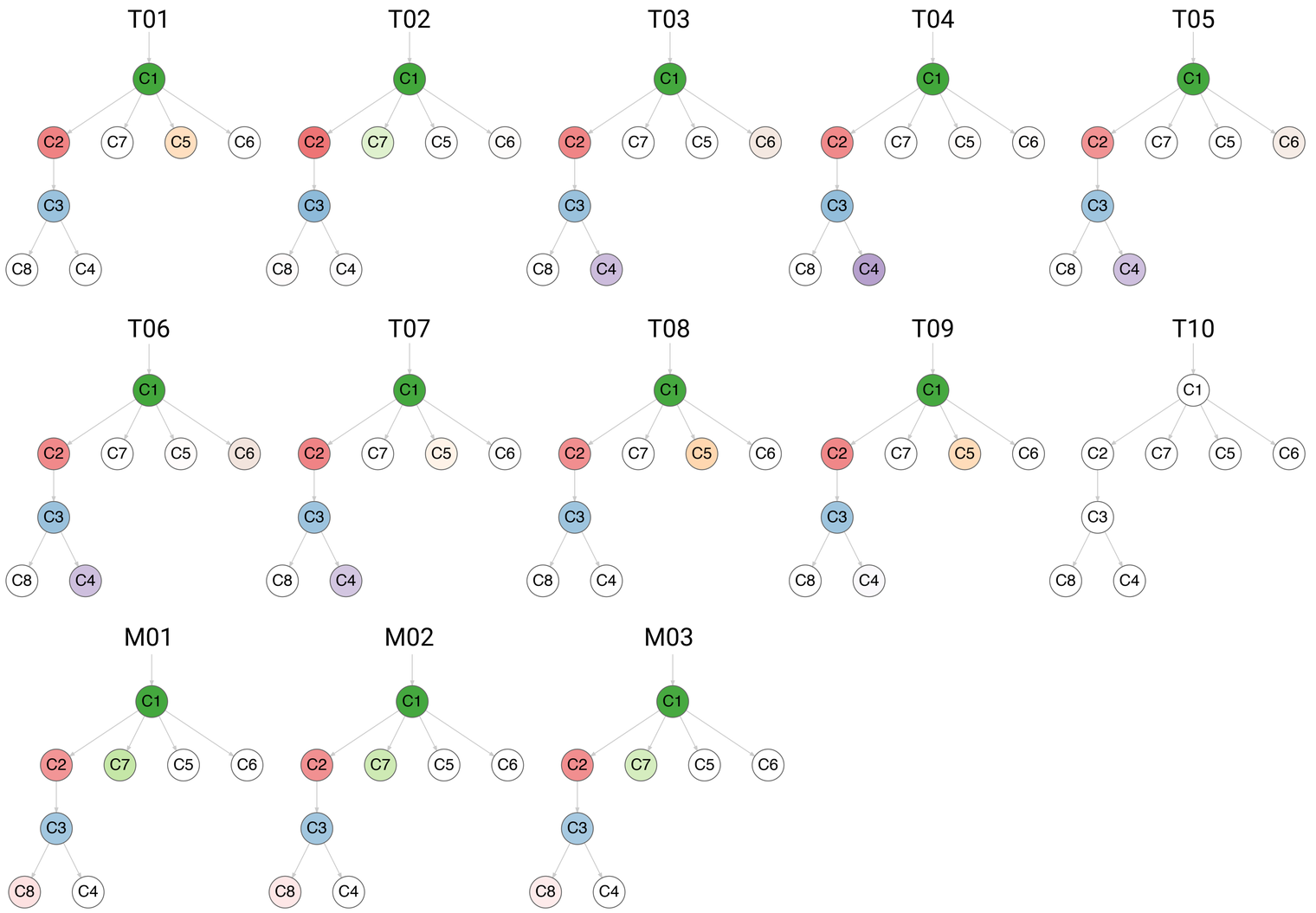


Figure 3

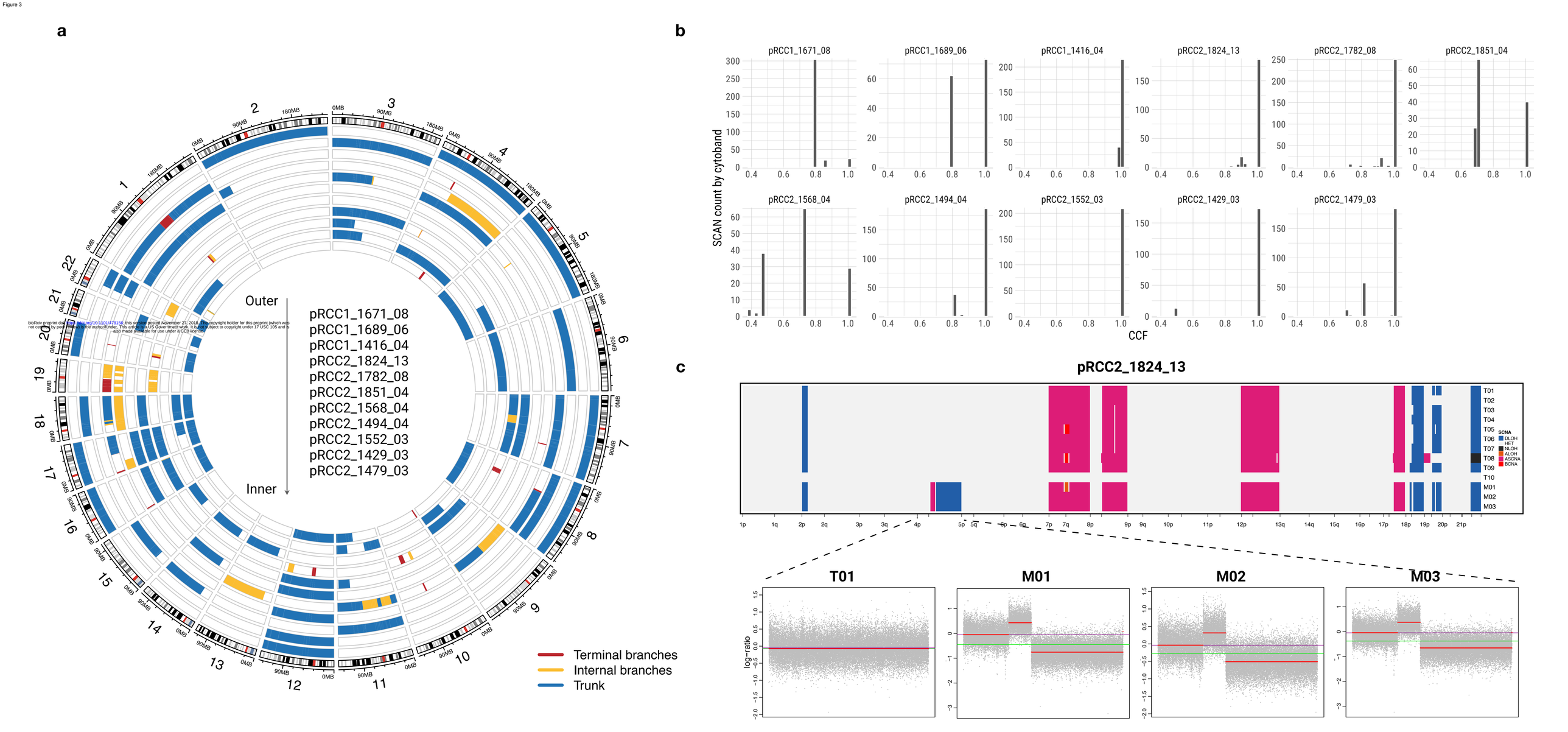
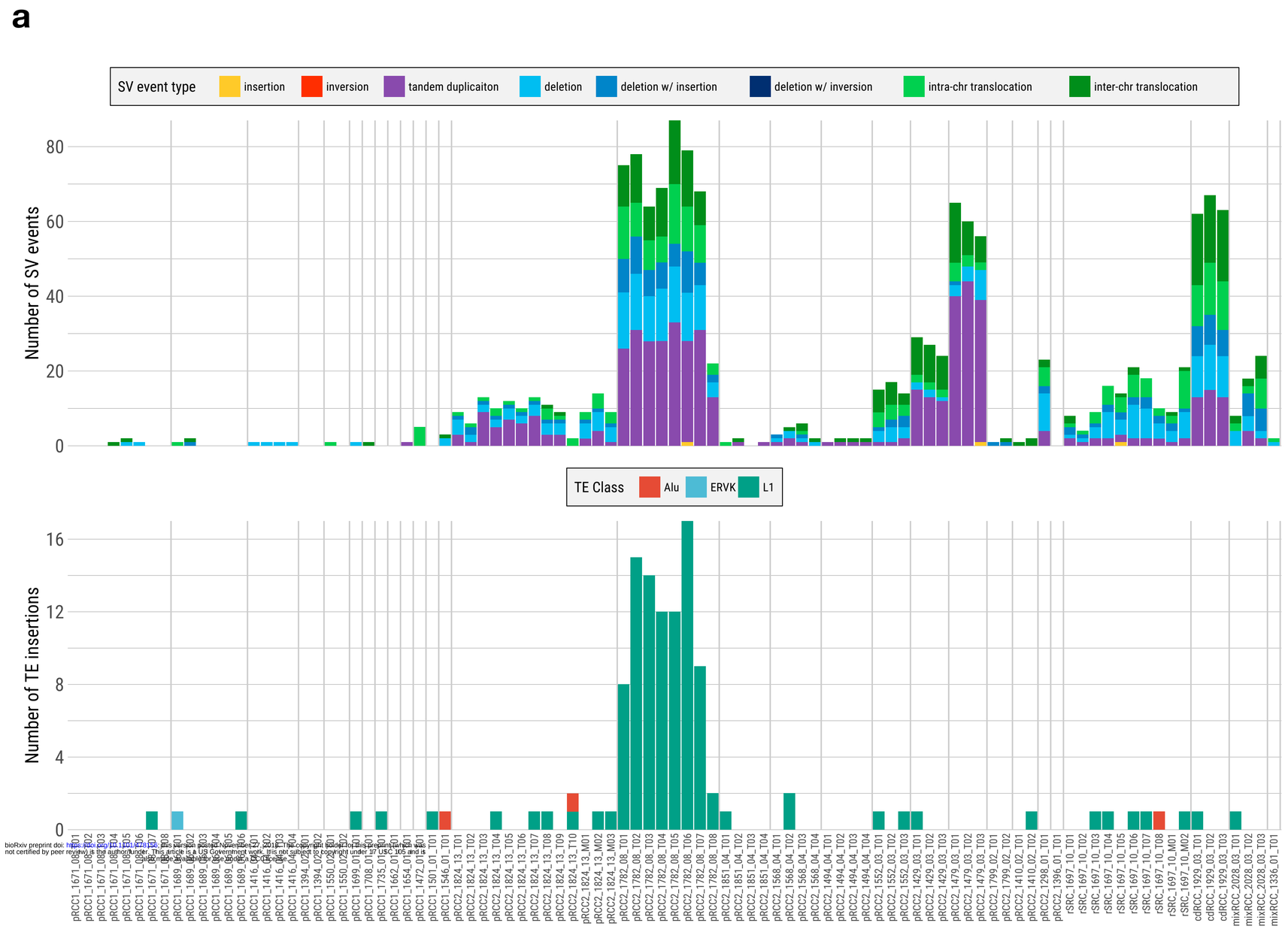
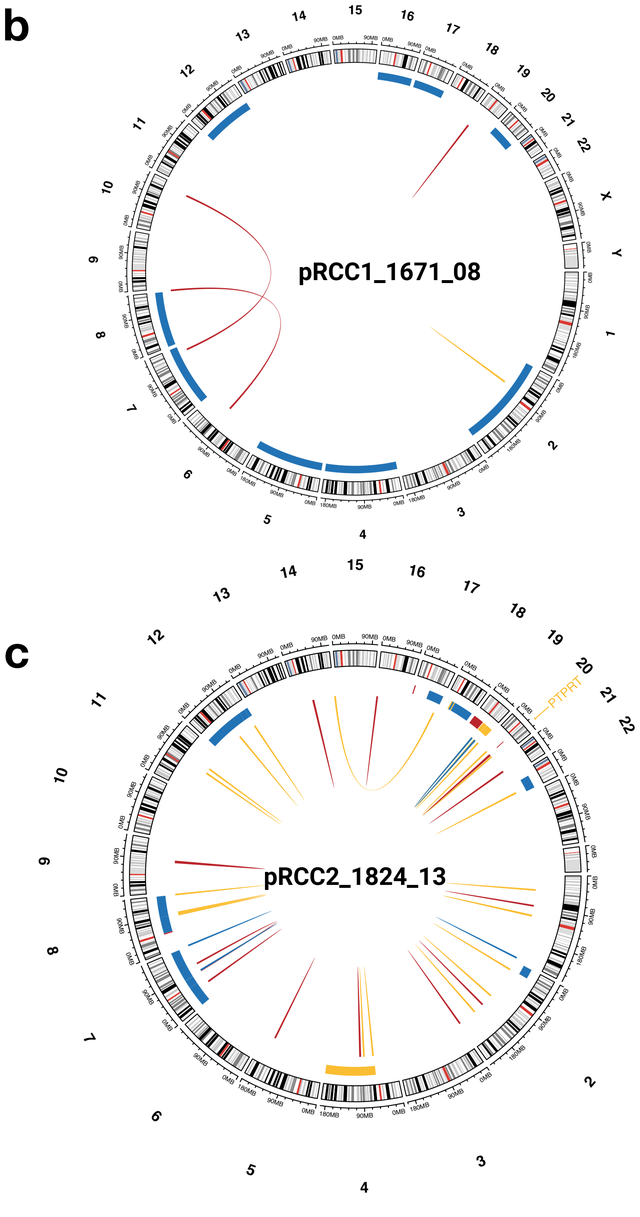
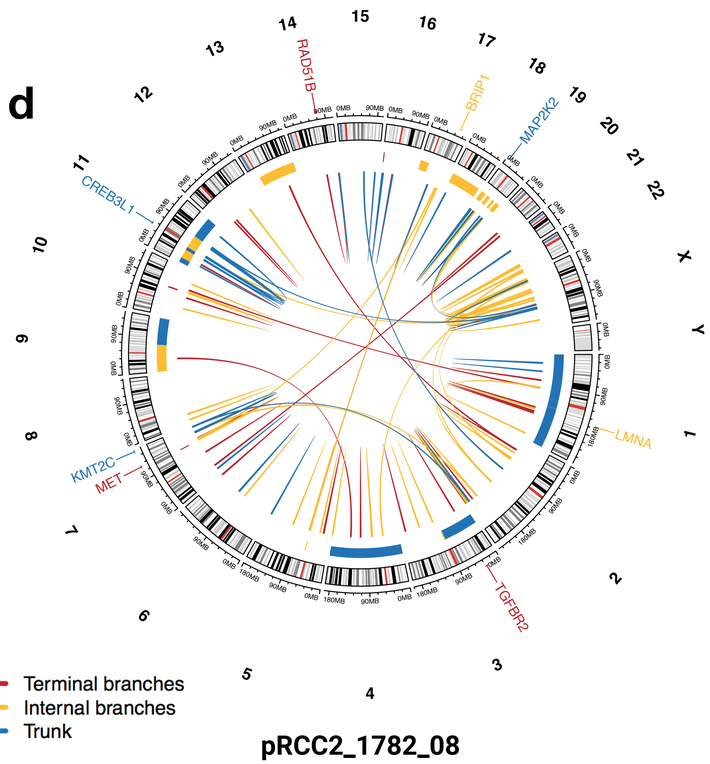


Figure 4







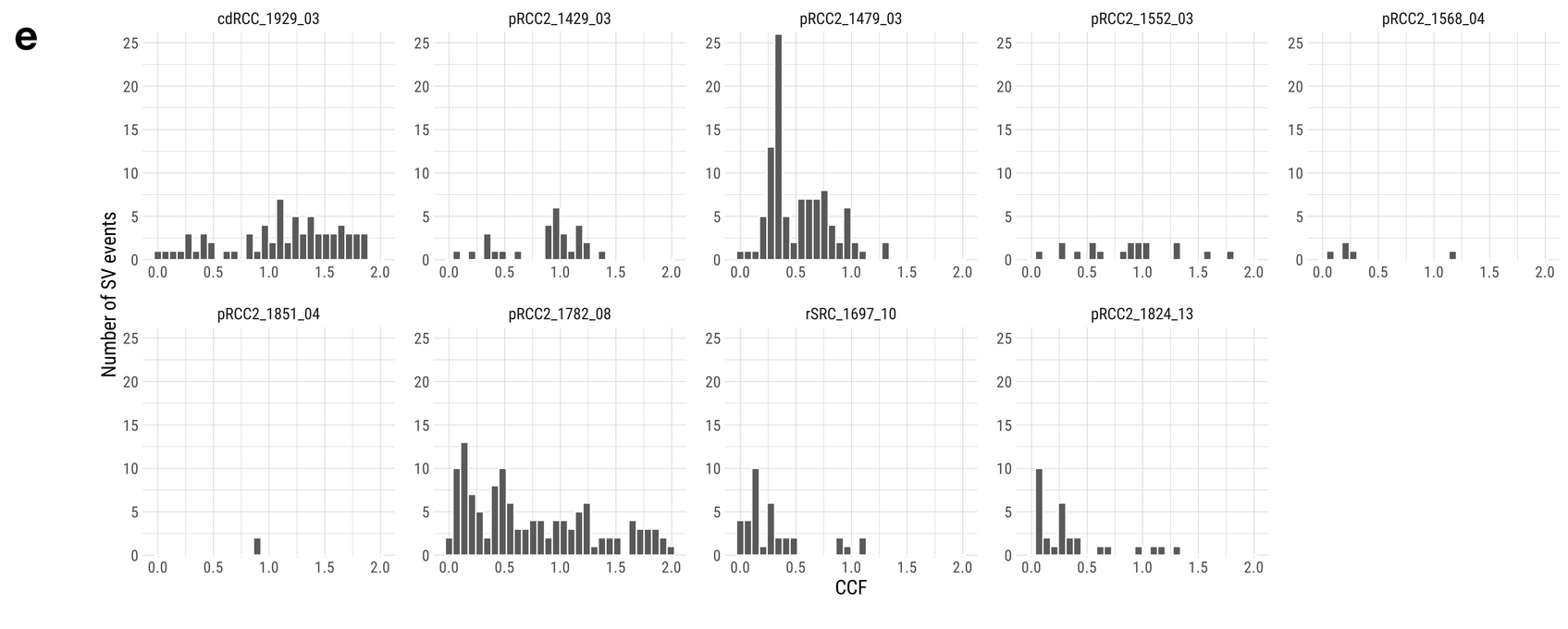
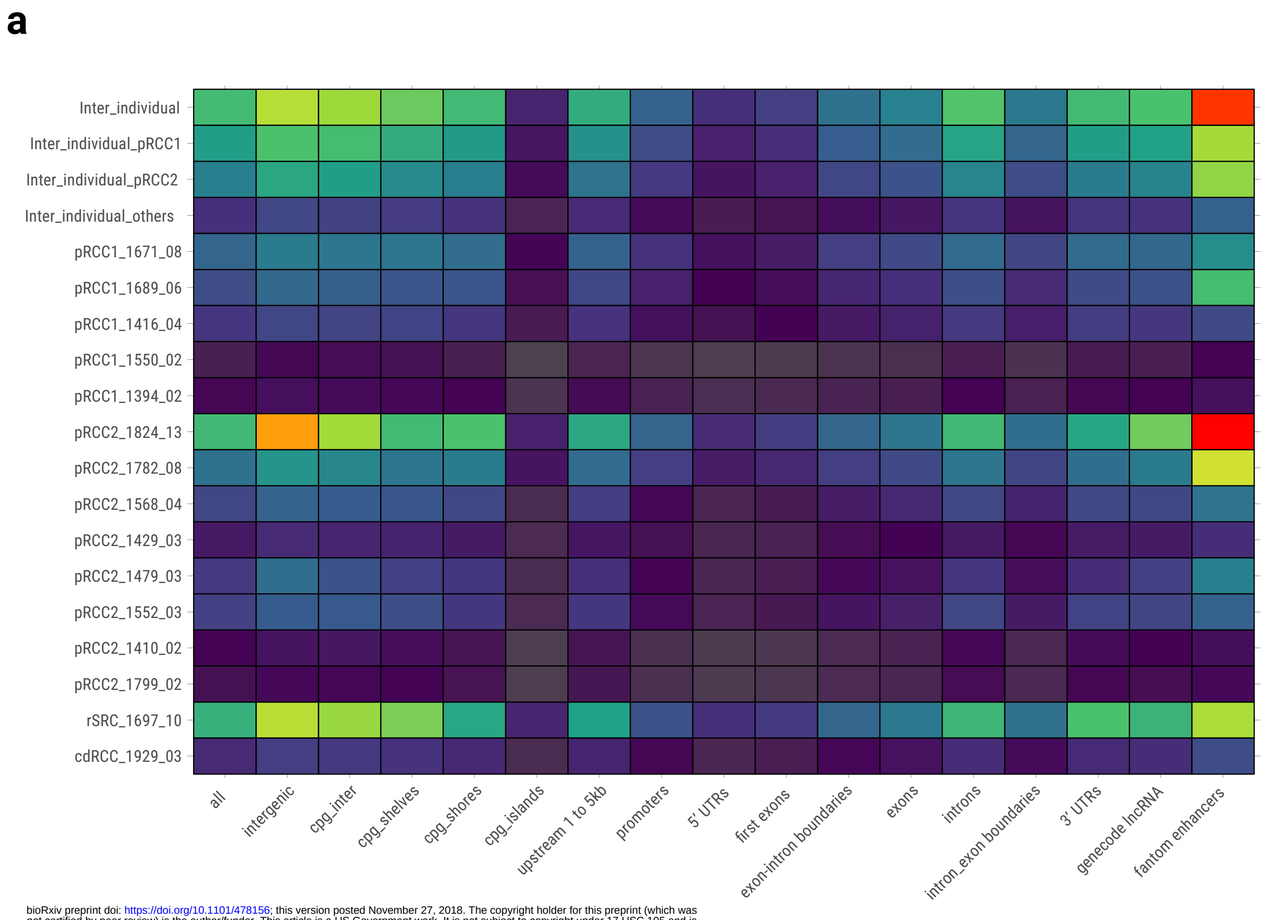
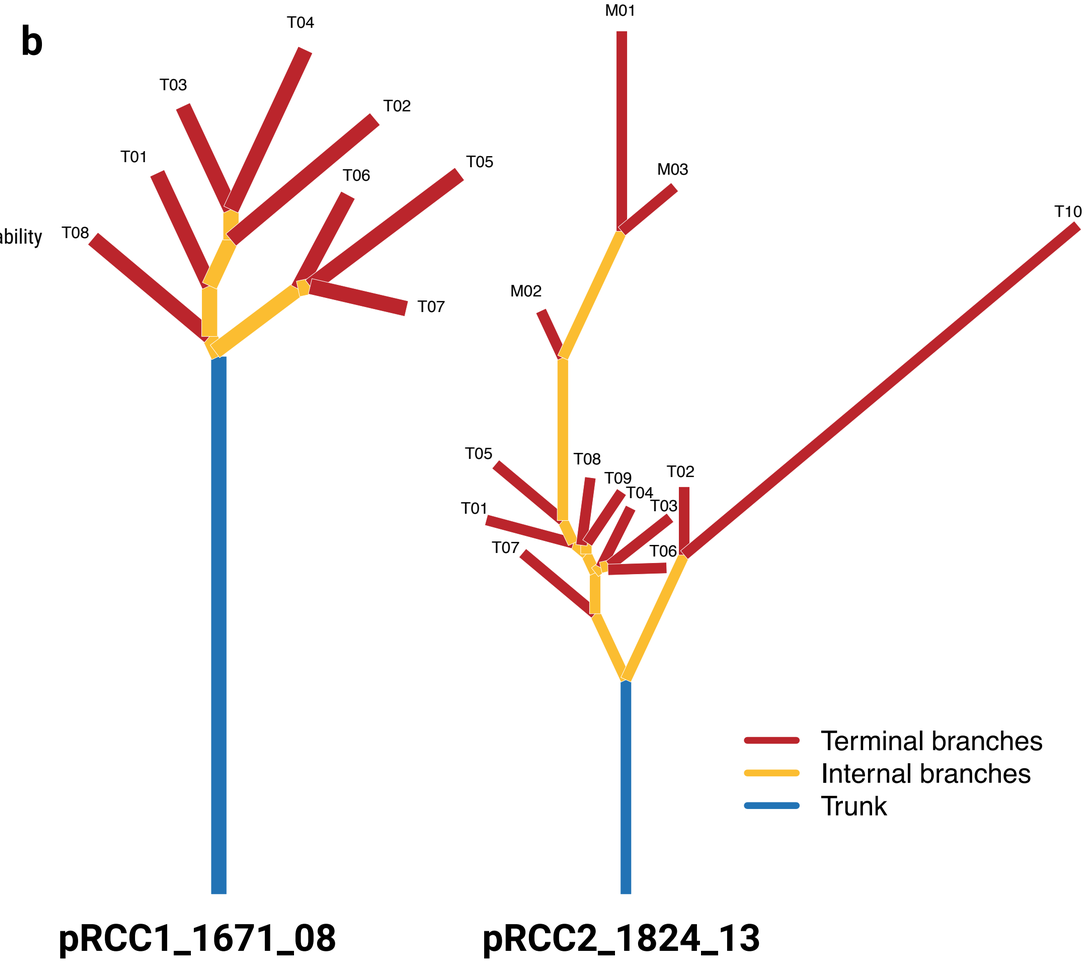


Figure 5





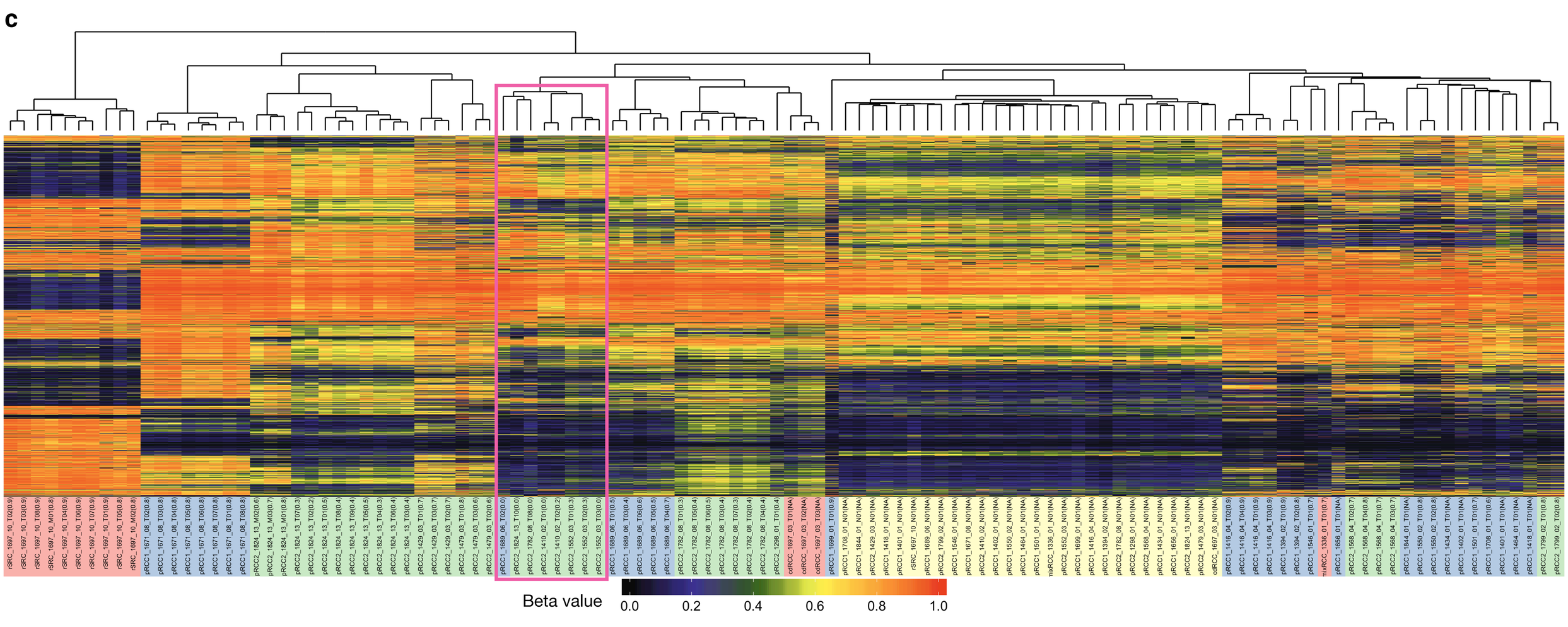
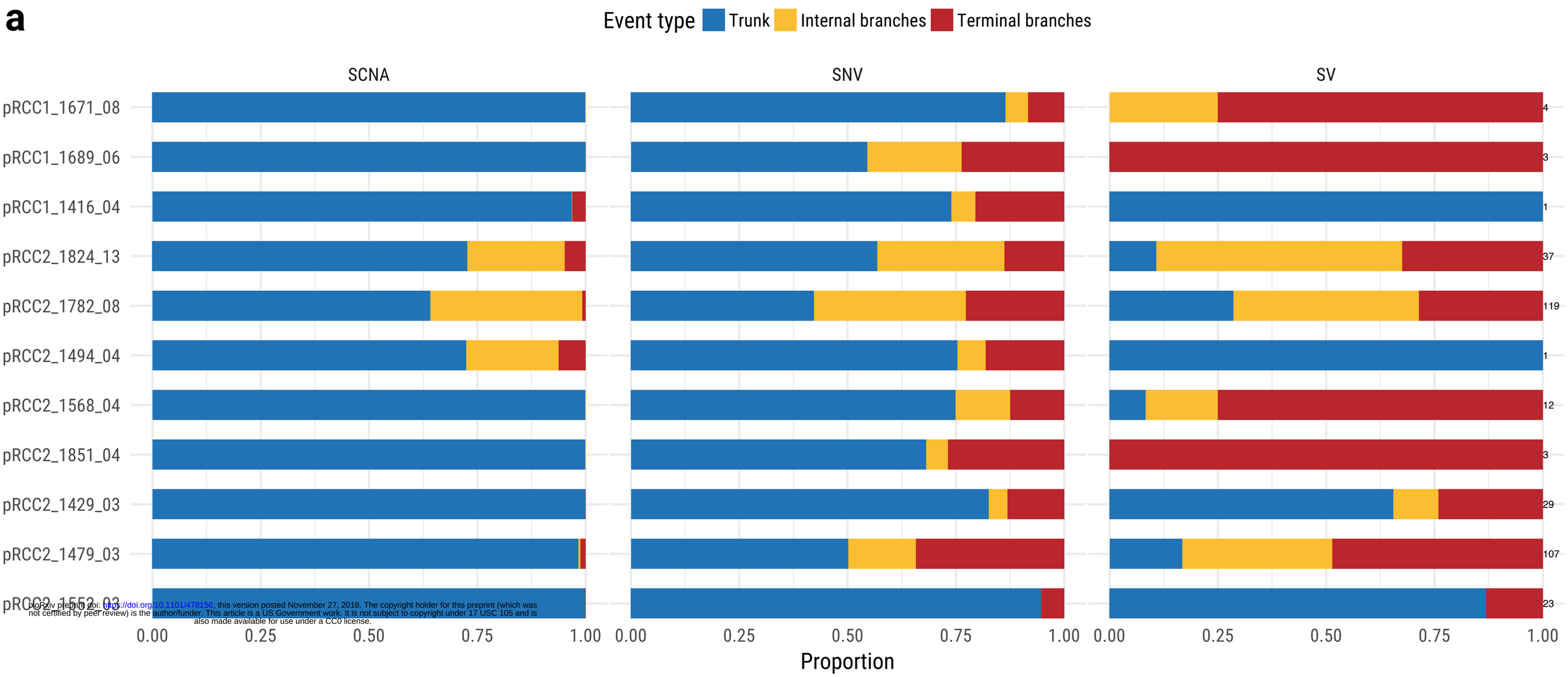
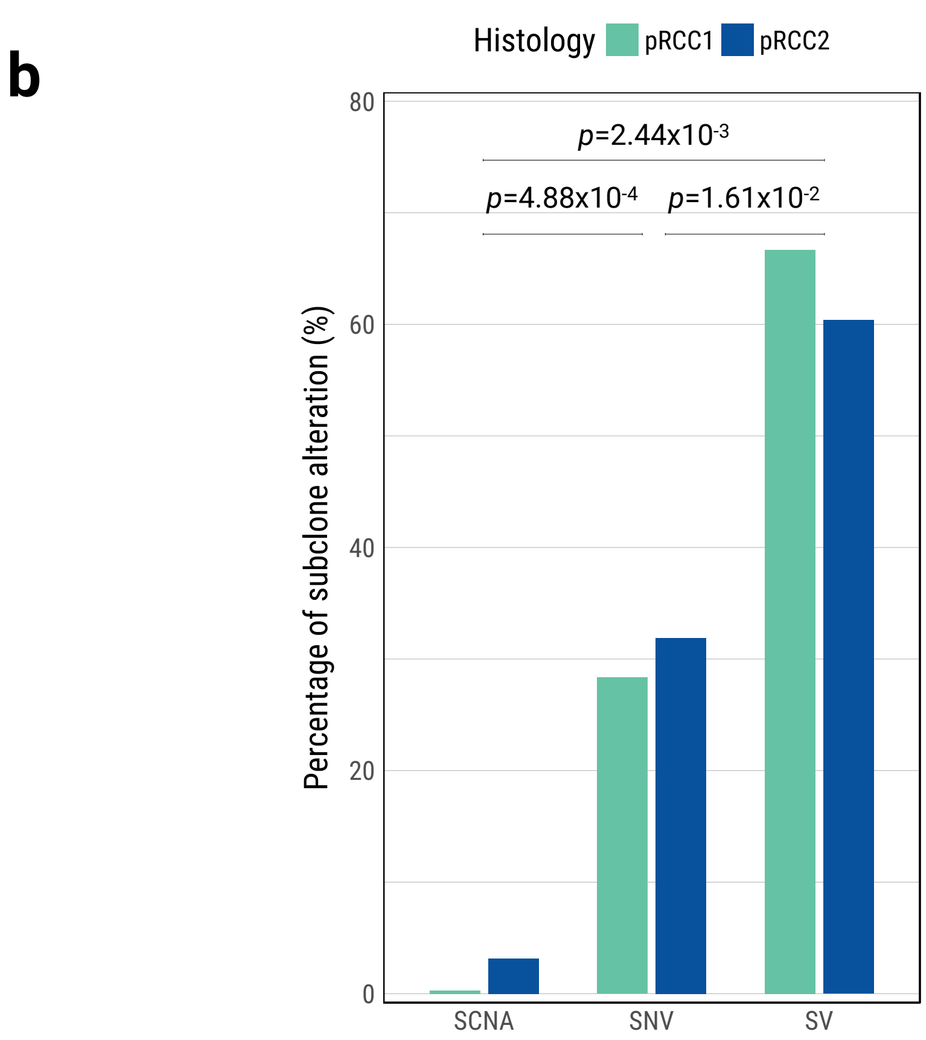
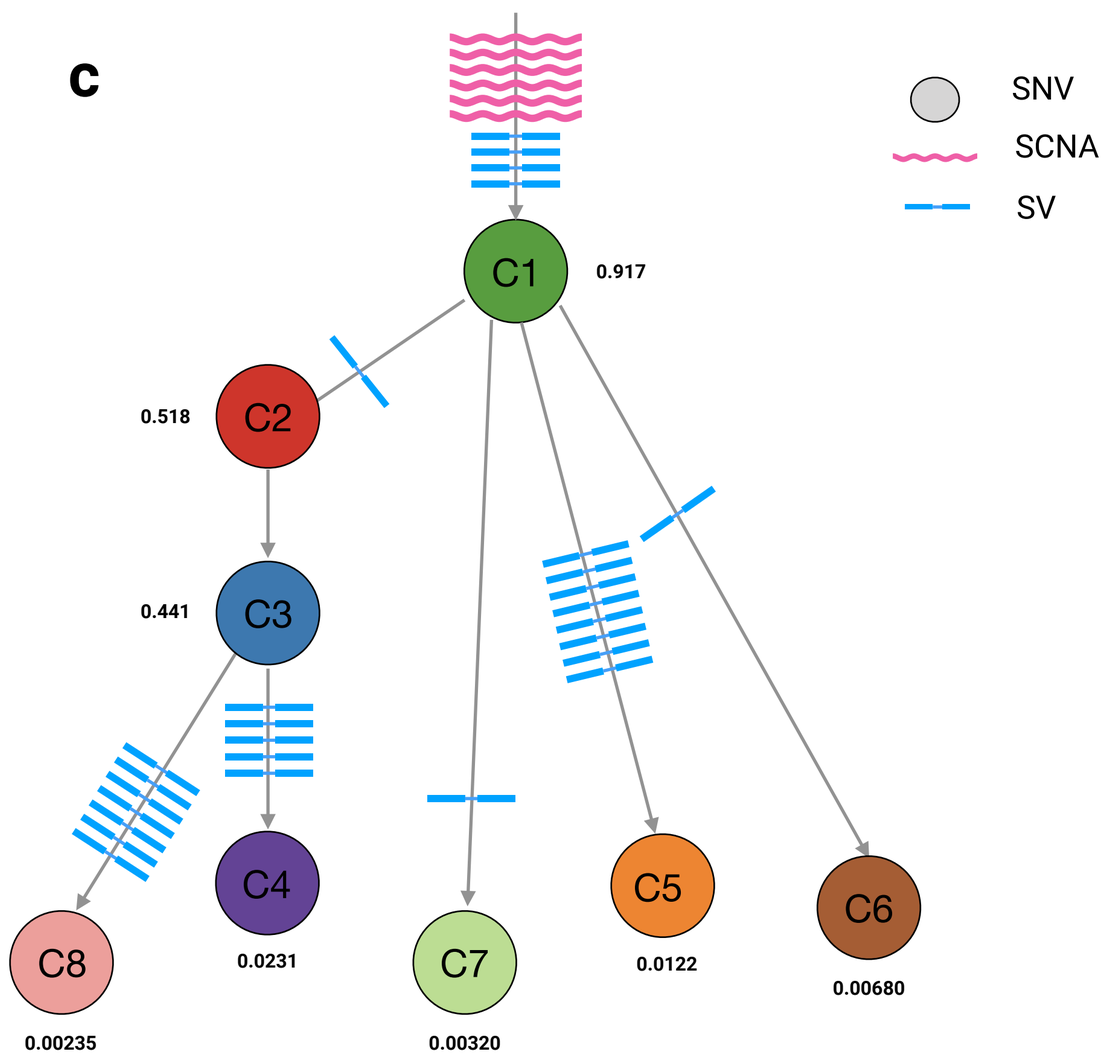


Figure 6







Supplemental Material

Extraction of genomic DNA from fresh frozen tissues specimens
After weight measurements, fresh frozen tissue samples (25 mg) were immediately put into 1 ml of 0.2 mg/ml Proteinase K (Qiagen) in DNA Lysis Buffer (10 mMTris-Cl (pH 8.0), 0.1 M EDTA (pH 8.0), and 0.5% (w/v) SDS) for 24 hrs at 56°C with shaking at 850 rpm in Thermomixer R (Eppendorf) until the tissue was completely lysed. Genomic DNA was extracted from fresh frozen tissue using the QIAmp DNA mini kit (Qiagen) according to the manufacturer's instructions. Each sample was eluted in volume of 200 µl AE buffer. DNA concentration was determined by Nanodrop spectrophotometer. All DNA samples were aliquoted and stored at −80 °C until use.

DNA Processing
DNA was quantified utilizing the QuantiFluor® dsDNA System (Promega Corporation, USA). DNA was normalized to 25ng/ul and underwent fragment analysis via AmpFLSTR™ Identifiler™ PCR Amplification Kit (ThermoFisher Scientific, USA). DNA samples are required to meet minimum mass and concentration thresholds for each assay, as well as show no evidence of contamination or profile discordance in the Identifiler assay. Samples meeting these requirements are aliquoted at the appropriate mass needed for downstream assay processing.

Whole-genome sequencing

Libraries were constructed and sequenced on the Illumina HiSeqX with the use of 151-bp paired-end reads for whole-genome sequencing.

Preparation of libraries for cluster amplification and sequencing
An aliquot of genomic DNA is taken from a stock sample at a target of 350ng in 50µL of solution to serve as the input into shearing. Samples undergo fragmentation by means of acoustic shearing using Covaris focused-ultrasonicator, targeting 385bp fragments. Following fragmentation, additional size selection is performed using a SPRI cleanup. Library preparation is performed using a commercially available kit provided by KAPA Biosystems (KAPA Hyper Prep without amplification module, product KK8505), and with palindromic forked adapters with unique 8 base index sequences embedded within the adapter (purchased from IDT). Following sample preparation, libraries were quantified using quantitative PCR (kit purchased from KAPA biosystems) with probes specific to the ends of the adapters. This assay was automated using Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 1.7nM. Samples are then pooled into 24-plexes and the pools are once again qPCRed. Samples were then combined with HiSeq X Cluster Amp Mix 1,2, and 3 into single wells on a strip tube using the Hamilton Starlet Liquid Handling System.

Cluster amplification and sequencing
Cluster amplification of the templates was performed according to the manufacturer's protocol (Illumina) using the Illumina cBot. Flowcells were sequenced on HiSeqX Sequencing-by-Synthesis Kits, then analyzed using RTA2.

Filtering criteria of somatic mutation calling from whole genome sequencing data
We used a revised method described by Jia-Jie Hao and colleagues[1] to filter the somatic variants. Specifically, a variant was kept if at least 8 reads covered this variant in the normal samples and 3 reads in the tumor samples. Somatic variants with variant allele frequency (VAF) less than 0.07 were discarded. In addition, somatic variants were filtered using the VarScan[2] 'processSomatic' command with arguments tailored to our WGS samples with --min-tumor-freq 0.07, --max- normal-freq 0.02 and --p-value 0.05. The resulting somatic variants were further filtered to reduce false positives using the fpFilter Perl script (https://github.com/ckandoth/variant-filter). To remove possible germline variants from the called somatic variants, somatic variants were filtered against the dbSNP135 (https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary_byOrg.cgi?build_id=135), the 1000 genomes (phase 3 v5, http://www.internationalgenome.org/category/phase-3/), the ExAC v0.3.1 database (http://exac.broadinstitute.org/), and an in-house germline variant database from Italian population for SNPs with MAF<0.001. The filtered variants were annotated with Oncotator (version 1.1.9.0, https://github.com/broadinstitute/oncotator). To increase the sensitivity of somatic mutation calling, disease-associated variants annotated in the ClinVar database and the COSMIC database were retained. In addition, following the approach described by Stachler, et al. [3], we leveraged the multiple region sequencing and salvaged somatic variants that were detected from at least one tumor region and were missed in other tumor regions due to low VAF. In brief, Bam-readcount (https://github.com/genome/bam-readcount) was used to obtain read counts for

unique somatic variants across all tumor regions. A somatic variant was considered to be absent if either its VAF was less than 0.02 or there were fewer than three reads.

Analysis of mutational signatures

Nucleosome Occupancy Analysis

Nucleosomes are the basic units of DNA packaging, consisting of eight core histone proteins wrapping DNA sequence of about 147 bp long around itself. Consecutive nucleosomes are connected to each other by stretches of DNA called "linker DNA". To explore the relationship between mutational signatures and nucleosome occupancy, we downloaded the nucleosome occupancy signal of K562 cell line generated by micrococcal nuclease sequencing (MNase-seq) from ENCODE project [4].To examine the average nucleosome occupancy signal around the mutations, we considered all single point mutations with probability >= 0.5 to be in that signature. First, we took a window ± 1 kb centered at the mutation start position and counted all the nucleosome occupancy signals separately for each base in this window. We repeated this procedure for all mutations, accumulating and counting the signals within this 2 kb window. We then calculated the average nucleosome occupancy signal at each base by dividing the accumulated nucleosome occupancy signal by the accumulated number of counts for each base.

To interpret this nucleosome analysis, if there are no relationships between the nucleosome occupancy and the mutations in a signature, a flat line would be seen. If mutations occur at nucleosome positions we would see a peak where the mutations are centered. If the mutations occur at linker DNA stretches, there would be a trough (valley) where the mutations are centered.

Replication Time Analysis

The ENCODE project provides genome-wide assessment of DNA replication timing in various cell lines using sequencing-based "Repli-seq" methodology. We used MCF-7 cell line for all analysis. We downloaded wavelet-smoothed replication time signal data as well as replication peaks and replication valleys. Replication peaks, corresponding to replication initiation zones (Peaks) and replication termination zones (Valleys) were determined from local maxima and minima, respectively, in the wavelet-smoothed replication time signal data [4]. We sorted wavelet-smoothed replication time signal in descending order, divided them into ten deciles, each containing equal number of signals. Single point mutations with probability >= 0.5 for each signature are distributed into the corresponding decile in which it falls into, so that the first decile contains the mutations that are replicated the earliest and the last decile contains the mutations that are replicated the latest. The number of mutations in each decile is divided by the number of attributable bases (including 'A's, 'T's,'C's, 'G's and excluding 'N's) in the corresponding decile which gives the mutation density. Then, these mutation densities are divided by the highest mutation density which results in normalized mutation densities.

Transcription and Replication Strand Bias Analysis

Single point mutations are called on the + strand of the reference genome and converted into pyrimidine context. We identified the transcribed and un-transcribed strands of the genome based on hg19 NCBI RefSeq curated genes obtained from UCSC Table Browser using transcription start and end positions. Gene containing strand was annotated as un-transcribed strand whereas complementary strand was annotated as transcribed strand. We searched for overlapping single point mutations and gene transcripts. If there were at least one gene transcript on the same strand as the single point mutation, then 'un-transcribed strand' count was increased otherwise 'transcribed count' was increased. We considered all possible gene transcripts. Among 33,067 gene transcripts, 16,863 (1,639,967,346 bp) of them were on the "+" strand and 16,194 (1,570,613,951 bp) of them were on the "-" strand.

To investigate replication strand bias, we leveraged on replication time data peaks and valleys. We ordered the peaks and valleys consecutively and found the consecutive regions with consistent positive slope in terms of replication time signal between each consecutive valley and peak. In a similar manner we found the consecutive region with consistent negative slope between each peak and valley and annotated the leading and lagging strands, respectively. Note that each strand with lagging/leading strand annotations implies leading/lagging on the opposite strand. To annotate a DNA stretch, as leading or lagging strand, we discarded the latest replication termination zones ± 25,000 bp from the valley's midpoint, and we required at least 10,000 bp long DNA stretch with consistent positive or negative slope, respectively. All statistical tests for significance of strand-bias are based on Fisher exact test after FDR correction.

Targeted Capture Sequencing

DNA Preparation: For each sample, 50 ng genomic DNA was purified using Agencourt AMPure XP Reagent (Beckman Coulter Inc, Brea, CA, USA) according to manufacturer's protocol. An adapter-ligated library was prepared with the KAPA HyperPlus Kit (KAPA Biosystems, Wilmington, MA) using Bioo Scientific NEXTflex™ DNA Barcoded Adapters (Bioo Scientific, Austin, TX, USA) according to KAPA-provided protocol.

Pre-Hybridization LM-PCR: Genomic DNA sample libraries were amplified prior to hybridization by ligation-mediated PCR consisting of one reaction containing 20 µL library DNA, 25 µL 2x KAPA HiFi HotStart ReadyMix, and 5µL 10x Library Amplification Primer Mix (includes two primers whose sequences are: 5'-AATGATACGGCGACCACCGA-3' and 5'-CAAGCAGAAGACGGCATACGA-3'). PCR cycling conditions were as follows: 98˚C for 45 seconds, followed by 7 cycles of 98˚C for 15 s, 60˚C for 30 s, 72˚C for 30 s. The last step was an extension at 72˚C for 1 minute. The reaction was kept at 4˚C until further processing. The amplified material was cleaned with Agencourt AMPure XP Reagent (Beckman Coulter, Inc., Brea, CA, USA) according to the KAPA-provided protocol. Amplified sample libraries were quantified using Quant-iT™ PicoGreen dsDNA Reagent (Life Technologies, Carlsbad, CA, USA).

Liquid Phase Sequence Capture: Prior to hybridization, amplified sample libraries with unique barcoded adapters were combined in equal amounts into 1.1 µg pools for multiplex sequence capture. Sequence capture was performed with NimbleGen's SeqCap EZ Choice Library, using a custom design comprised of 254 candidate cancer driver genes (Roche NimbleGen, Inc., Madison, WI, USA). Prior to hybridization the following components were added to the 1.1 µg pooled sample library, 4 µL of NEXTflex HE Universal Oligo 1, 250 µM (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'), 40 µL total 25 µM NEXTflex INV-HE blocking oligos, equal volumes of each blocking oligo complementary to the barcodes in the pool (5'-CAAGCAGAAGACGGCATACGAGAT**X**GTGACT GGAGTTCAGACGTGTGCTCTTCCGATCT/C3 Spacer/-3', where **X** is 8-bases of sequence specific to adapter barcode used for library construction), and 5 µL of 1 mg/mL COT-1 DNA (Invitrogen, Inc., Carlsbad, CA, USA). Samples were dried down by puncturing a hole in the plate seal and processing in an Eppendorf 5301 Vacuum Concentrator (Eppendorf, Hauppauge, NY, USA) set to 60˚C for approximately 1 hour. To each dried pool, 7.5 µL of NimbleGen Hybridization Buffer and 3.0 µL of NimbleGen Hybridization Component A were added, and placed in a heating block for 10 minutes at 95˚C. The mixture was then transferred to 4.5 µL of EZ Choice Probe Library and hybridized at 47˚C for 64 to 72 hours. Washing and recovery of captured DNA were performed as described in NimbleGen SeqCap EZ Library SR Protocol.

Post-Hybridization LM-PCR: Pools of captured DNA were amplified by ligation-mediated PCR consisting of one reaction for each pool containing 20µl captured library DNA, 25 µL 2x KAPA HiFi HotStart ReadyMix, and 5µL 10x Library Amplification Primer Mix (includes two primers whose sequences are: 5'-AATGATACGGCGACCACCGA-3' and 5'-CAAGCAGAAGACGGCATACGA-3'). PCR cycling conditions were as follows: 98˚C for 45 seconds, followed by 12 cycles of 98˚C for 15 s, 60˚C for 30 s, 72˚C for 30 s. The last step was an extension at 72˚C for 1 minute. The reaction was kept at 4˚C until further processing. The amplified material was cleaned with Agencourt AMPure XP Reagent (Beckman Coulter, Inc., Brea, CA, USA) according to NimbleGen SeqCap EZ Library SR Protocol. Pools of amplified captured DNA were then quantified via Kapa's Library Quantification Kit for Illumina (Kapa Biosystems, Woburn, MA, USA) on the LightCycler 480 (Roche, Indianapolis, IN, USA).

Sequencing: The resulting post-capture enriched multiplexed sequencing libraries were used in cluster formation on an Illumina cBOT (Illumina, San Diego, CA, USA) and paired-end sequencing was performed using an Illumina HiSeq 4000 following Illumina-provided protocols for 2x150bp paired-end sequencing.

Methylation analysis

DNA Preparation: 400 ng of sample DNA, according to Quant-iT PicoGreen dsDNA quantitation (Life Technologies, Grand Island, NY), was treated with sodium bisulfite using the EZ-96 DNA Methylation MagPrep Kit (Zymo Research, Irvine, CA) according to manufacturer-provided protocol. Bisulfite conversion modifies non-methylated cytosines into uracil, leaving 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) unchanged. For every 95 samples, an internal control, NA07057 (Coriell Cell Repositories, Camden, NJ) was utilized to confirm the efficiency of bisulfite conversion and subsequent methylation analysis.

Methylation Analysis: High-throughput epigenome-wide methylation analysis, using Infinium MethylationEPIC BeadChip (Illumina Inc., San Diego, CA) which uses both Infinium I and II assay chemistry technologies, was performed according to manufacturer-provided protocol. Bisulfite-treated samples were denatured and neutralized then whole genome amplified, isothermally, to increase the amount of DNA template. The amplified product was enzymatically fragmented, precipitated and resuspended in hybridization buffer. Eight samples were applied to each BeadChip and hybridized overnight where fragmented DNA samples anneal to locus-specific 50mers (covalently linked to one of over 800,000 bead types). Two beadtypes correspond to each CpG locus for Infinium I assays: one bead type corresponds to methylated, another bead type to unmethylated state of the CpG site, while one beadtype corresponds to each CpG locus for Infinium II assays. Single-base extension of the oligos on the BeadChip, using the captured DNA as template, incorporates tagged nucleotides on the BeadChip, which are subsequently fluorophore labeled during staining. The Illumina iScan scanned the BeadChips at two wavelengths to create image and intensity files.

Data Analysis: The intensity files from the Illumina methylation assay on the MethylationEPIC platform were processed and analyzed with the R programming language using the R package "minfi". Briefly, raw intensity file (idats) are loaded into R using minfi. Samples are excluded if the percent of probes with detection p value greater than 0.01 is greater than 4%. Concordance is checked for both expected and unexpected replicates using the ~60 polymorphic SNPs on the array. Raw methylation beta values are normalized according to previously published methods[5].

Genotyping analysis

Illumina Infinium Genotyping Array BeadChips: High-throughput, genome-wide SNP genotyping, using Infinium BeadChip technology (Illumina Inc. San Diego, CA), was performed at the Cancer Genomics Research Laboratory (CGR). Genotyping was performed per the manufacturer's guidelines using the Infinium automated protocol. For the Infinium HumanOmniExpress-24 chip, 20ng genomic DNA, quantitated using Quant-iT™ PicoGreen dsDNA Reagent (ThermoFisher Scientific, Waltham, MA, USA) is denatured and neutralized, then isothermally amplified by whole-genome amplification. The amplified product is enzymatically fragmented, then precipitated and re-suspended. Resuspended samples are denatured, then hybridized to locus-specific 50-mer oligonucleotides which are attached to 1-micron beads on the BeadChip. These 50-mer probes stop one base before the location of interest. Enzymatic single-base extension of the oligos on the BeadChip, using the captured DNA as a template, incorporates tagged nucleotides on the BeadChip, which are subsequently fluorophore-labeled during staining. The fluorescent label determines the genotype call for the sample. The Illumina iScan scans the BeadChips at two wavelengths to detect the fluorescent label, creating image files that are converted into genotype calls based on the detected fluorescence.

Illumina GenomeStudio Genotyping Module v2.0: The GenomeStudio is a Windows-based software that analyzes Illumnia genotyping data. Typically, a clustering project is created from a sample sheet to load sample intensities (*.idat files), and a SNP manifest (*.bpm file) for the type of array used. After scan, an initial clustering is performed to identify samples that need to be re-scanned from the laboratory. Once re-scan is complete, any sample that yields higher call rate by re-scan will have its intensity file substituted with a better one. Then the final clustering of all samples performed using a cluster position file (*.egt file). Once final clustering is finished, genotypes of all samples are exported through the 'Report Wizard'. Genotypes can be exported into 'Locus X DNA' format (LBD), then further converted into multiple different formats (eg. plink or glu format), or can be exported directly into PLINK format if a plug-in is installed for this function.

Validation of structural variants

Comparison between Meerkat and Novobreak SV calling
We compared the SV results obtained by Meerkat with those obtained by Novobreak[6](v1.1.3rc). "Filter_sv2.pl" was used to filter the SVs, and only SVs with calling quality "QUAL" above 30 were kept. We examined whether SVs called by Meerkat could be called by Novobreak and *vice versa*, by comparing both left and right breakpoints in a given window size from 0 to 10 Mbp. The results showed that Meerkat-based calling was more conservative than Novobreak regardless the window size; Novobreak detected almost twice the number of SVs detected by Meerkat. About 60% of SVs identified by Meerkat could be replicated by Novobreak, while 32% of SVs identified by Novobreak could be replicated by Meerkat. Thus, we decided to use the SVs identified by Meerkat.

Gene fusion validation

We selected four in-frame fusions: *MALAT-TFEB*, *MET-MET* deletion, *STRN-ALK*, and *EWSR1-PATZ1* for further validation. Briefly, total RNA from tumors with a fusion detected through our WGS pipeline were reverse transcribed using Superscript III (Thermo) primed with random hexamers. Subsequently, PCR using gene specific primers were performed with Platinum Taq supermix (Thermo) at 55C annealing temp. PCR products were agarose gel purified and sequenced on an ABI 3730xl. RNA samples to validate *EWSR-PATZ1* and *MALAT1-TFEB* fusions had relatively high RIN score (RIN=8 and 6, respectively). These fusions were confirmed by Sanger sequencing as valid fusion products. In contrast, RNA samples to validate the *MET-MET* deletion and *STRN-ALK* fusion were derived from FFPE samples with RIN=2.6. We were unable to validate these fusions by this method, potentially due to the poor RNA quality.

Validation of additional structural variants

381 structural variants, identified by Meerkat from whole genome sequencing data, were selected for validation. These events were found in pRCC1 and pRCC2 tumors with at least three samples. For each structural variant to be validated, 250bp of sequence was pulled from one side of each Meerkat breakpoint and spliced together, using the modified script "primers.pl" from the Meerkat package, such that the resulting 500bp Fasta file would serve as a reference for primer design of an amplicon that would only be generated in samples containing the structural variant. Alignment and orientation of the 500bp Fasta sequences were manually checked using UCSC BLAT. These 381 Fasta files were provided to the Ampliseq Designer v6.1.4 (ThermoFisher Scientific, Waltham, MA, USA) for custom Ampliseq panel design. 303 events had primer pairs successfully designed across the SV breakpoints and were included in a custom panel, with an average amplicon length of 352bp. A previously-designed custom panel, containing 74 amplicons, with primer pairs designed in hg19 rather than across custom SV breakpoints, was combined with this custom panel in order to ensure that samples containing only one or a few of the SV events would amplify sufficiently to generate quantifiable library. Sample DNA (30ng) was amplified using this custom AmpliSeq primer pool, and libraries were prepared following the manufacturer's Ion AmpliSeq Library Preparation protocol. Individual samples were barcoded, pooled, templated, and sequenced on the Ion Torrent PGM Sequencer using the Ion Chef and sequenced on a 318 chip per manufacturer's instructions. Sequence data was aligned to the custom reference used to generate the panel design, and reads were identified which spanned the SV breakpoint at 250bp.

Retrotranspon analysis

The LINE-1 insertions (**Fig. 4a**) identified by TraFiC[7] were located in introns and intergenic regions. At least five insertions were judged to be *bona fide* L1HS insertions due to simultaneous presence of the following features: 1) concordant insertion position confirmed by sequence reads of different amplicons; 2) a 5' end corresponding to a portion of the L1HS consensus sequence; 3) TSD (target site duplications) of 11-17 bp and an A/T-rich insertion site (typical features of target-primed retrotransposition). The precise length of these insertions was not determined, because only the 5' and 3' ends of the inserted fragments were present in the amplicons generated in the NGS libraries. However, based on the alignment of the sequenced 5' ends with the L1HS consensus sequence (obtained from the public database www.girinst.org), no full-length L1 insertion was present. The minimal estimated length of the observed LINE-1 insertions ranged from few hundreds to about 1200 base pairs. At least three of the insertions could potentially affect the expression of proteins involved in chromatin regulation and chromosome structural maintenance and (in turn) the maintenance of genome integrity: 1) Chromosome 9: 1929235. From ENCODE analysis, this insertion appears to be located in a gene-regulatory site. *SMARCA2*, a member of the SWI/SNF family of chromatin remodeling factors, is located about 80Kb away. 2) Chromosome 7: 18991805-18992490. The insertion is located in an intron of *HDAC9* (Histone deacetylase 9), a gene whose increased or ectopic expression is involved in carcinogenesis[8]. 3) Chromosome 18: 34786106-34786799. Located in an intron of *KIAA1328* (hinderin), which binds to SMC3 (structural maintenance of chromosomes 3). SMC3 knockdown triggers genomic instability[9].

Correlation of distance matrices

Euclidean distance matrices of SNV, SCNA, and methylation data were generated for phylogenetic tree construction as mentioned above. Genetic and epigenetic distance matrix comparisons were performed and visualized as the lower triangular matrix heatmap in modified fviz_dist function in factoextra R package (https://cran.r-project.org/web/packages/factoextra/index.html). Pearson's correlation coefficient was used to estimate the similarity between SNVs, CNAs, and methylation distance matrices. For bootstrapping analysis[10], the null

distribution for each patient was generated by randomly shuffling the labels of Euclidean distance matrices for 100,000 times and then calculated the corresponding correlation coefficient for each shuffle. An empirical *P*-value was calculated by comparing the observed correlation coefficient with the bootstrapped one under the null distribution.

In general, the similarity of methylation with SNV and SCNA ITH, measured by the Pearson's correlation coefficient of two regional distance matrices, was weak (PCC= 0.655 and 0.674, respectively).

ITH and driver mutations in rSRC_1697_10 fibrosarcoma

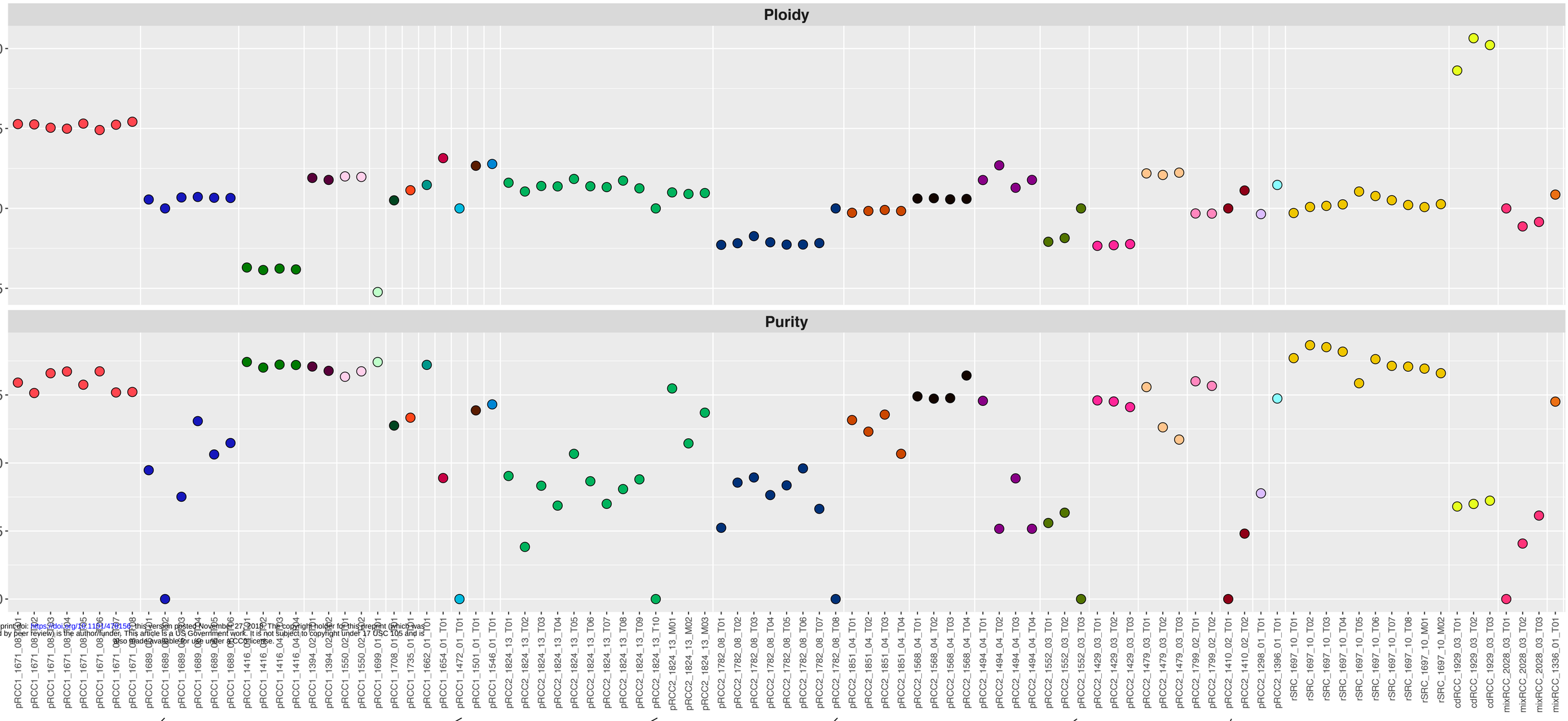
We did not find any mutations in the 254 driver genes for sarcomas rSRC_1697_10, except a nonsense mutation in *ZNF652*. In contrast, a *EWSR1*/*PATZ1* fusion and *CDKN2A* deletion were detected in all samples in this subject as clonal events (Fig. 1C). *CDKN2A* methylation was also present in almost all samples. The SNV MRT (**Figure S5**) shows a large proportion of branches, suggesting substantial ITH for SNVs. Five subclones were detected (**Figure S6**) and the subclones in metastatic samples were present in the primary tumor samples. Hotspots of SVs were observed on chromosome 1 (**Figure S21**) for particular tumor regions (T04, T05, T06, T07 and M02).

References:

1. Hao, J.J. *et al*. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nat Genet* **48**, 1500-1507 (2016).
2. Koboldt, D.C. *et al*. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-76 (2012).
3. Stachler, M.D. *et al*. Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat Genet* **47**, 1047-55 (2015).
4. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
5. Fortin, J.P. *et al*. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol* **15**, 503 (2014).
6. Chong, Z. *et al*. novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat Methods* **14**, 65-67 (2017).
7. Tubio, J.M.C. *et al*. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
8. Lapierre, M. *et al*. Histone deacetylase 9 regulates breast cancer cell proliferation and the response to histone deacetylase inhibitors. *Oncotarget* **7**, 19693-708 (2016).
9. Ghiselli, G. SMC3 knockdown triggers genomic instability and p53-dependent apoptosis in human and zebrafish cells. *Mol Cancer* **5**, 52 (2006).
10. Brocks, D. *et al*. Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep* **8**, 798-806 (2014).

Figure S1: Ploidy and purity of each sample. Ploidy and purity of each sample was estimated based on (a) copy number and (b) density plots of single nucleotide variant allele fraction (VAF) for low purity samples. The SNV-based purity = 2*(the mode of VAF density) and is labeled in red color in panel b.

a



b

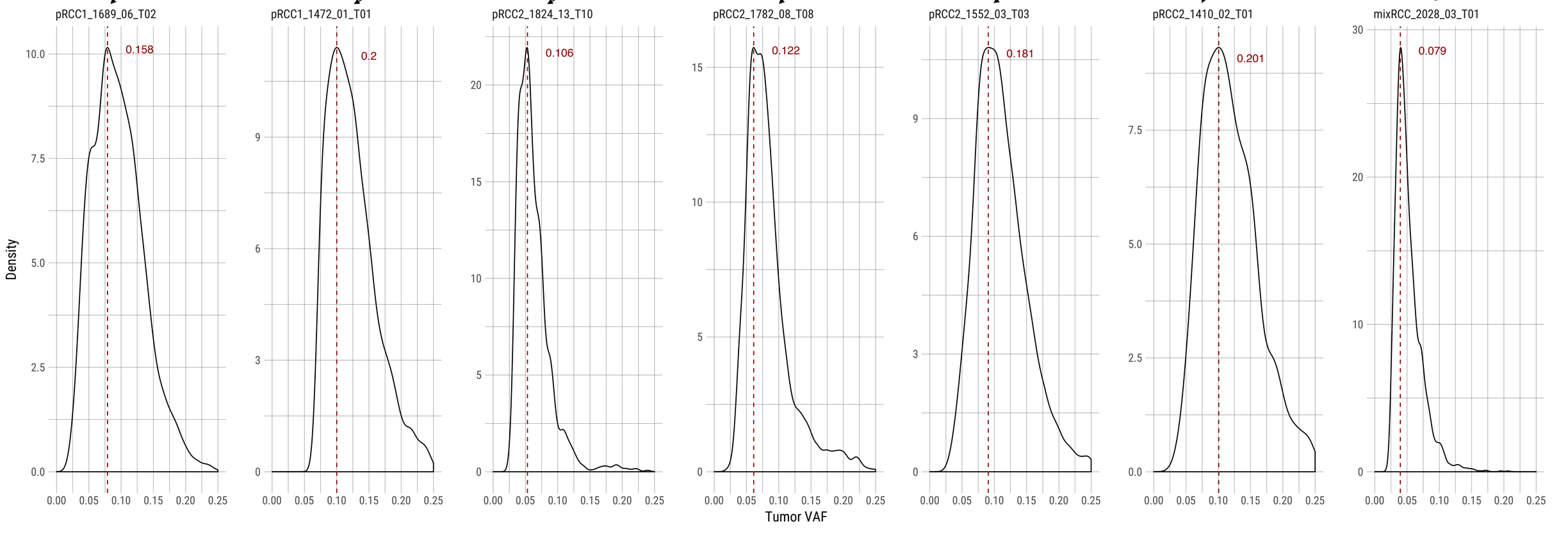


Figure S2: Single nucleotide variants (SNVs) in known cancer driver genes. Single nucleotide variants (SNVs) for each sample were called based on both whole-genome sequencing and deep targeted sequencing. Missense mutations, nonsense mutations and splice site mutations are annotated with different colors.

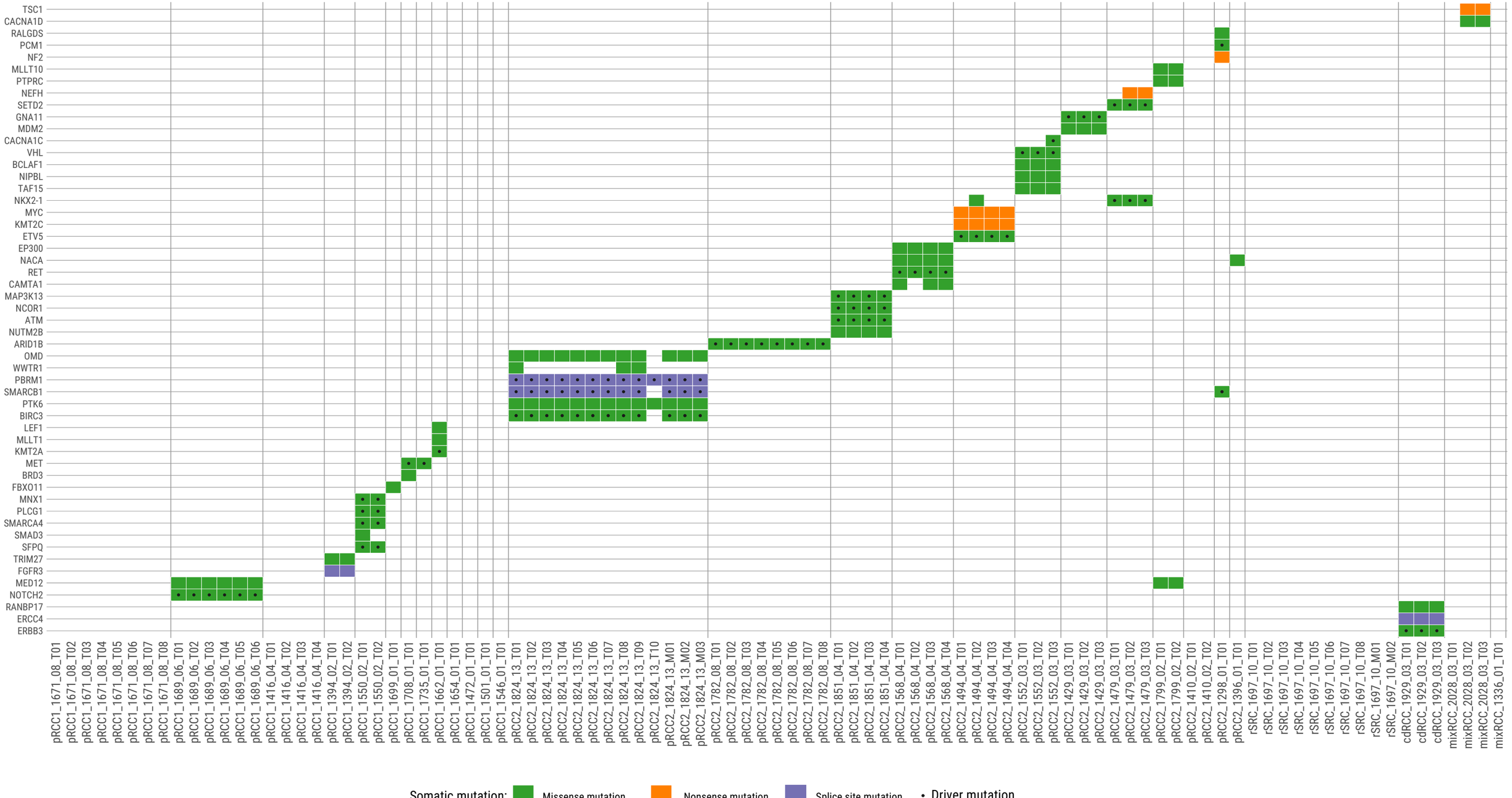

TSC1
CACNA1D
RALGDS
PCM1
NF2
MLLT10
PTPRC
NEFH
SETD2
GNA11
MDM2
CACNA1C
VHL
BCLAF1
NIPBL
TAF15
NKX2-1
MYC
KMT2C
ETV5
EP300
NACA
RET
CAMTA1
MAP3K13
NCOR1
ATM
NUTM2B
ARID1B
OMD
WWTR1
PBRM1
SMARCB1
PTK6
BIRC3
LEF1
MLLT1
KMT2A
MET
BRD3
FBXO11
MNX1
PLCG1
SMARCA4
SMAD3
SFPQ
TRIM27
FGFR3
MED12
NOTCH2
RANBP17
ERCC4
ERBB3
pRCC1_1671_08_T01
pRCC1_1671_08_T02
pRCC1_1671_08_T03
pRCC1_1671_08_T04
pRCC1_1671_08_T05
pRCC1_1671_08_T06
pRCC1_1671_08_T07
pRCC1_1671_08_T08
pRCC1_1689_06_T01
pRCC1_1689_06_T02
pRCC1_1689_06_T03
pRCC1_1689_06_T04
pRCC1_1689_06_T05
pRCC1_1689_06_T06
pRCC1_1416_04_T01
pRCC1_1416_04_T02
pRCC1_1416_04_T03
pRCC1_1416_04_T04
pRCC1_1394_02_T01
pRCC1_1394_02_T02
pRCC1_1550_02_T01
pRCC1_1550_02_T02
pRCC1_1699_01_T01
pRCC1_1708_01_T01
pRCC1_1735_01_T01
pRCC1_1662_01_T01
pRCC1_1654_01_T01
pRCC1_1472_01_T01
pRCC1_1501_01_T01
pRCC1_1546_01_T01
pRCC2_1824_13_T01
pRCC2_1824_13_T02
pRCC2_1824_13_T03
pRCC2_1824_13_T04
pRCC2_1824_13_T05
pRCC2_1824_13_T06
pRCC2_1824_13_T07
pRCC2_1824_13_T08
pRCC2_1824_13_T09
pRCC2_1824_13_T10
pRCC2_1824_13_M01
pRCC2_1824_13_M02
pRCC2_1824_13_M03
pRCC2_1782_08_T01
pRCC2_1782_08_T02
pRCC2_1782_08_T03
pRCC2_1782_08_T04
pRCC2_1782_08_T05
pRCC2_1782_08_T06
pRCC2_1782_08_T07
pRCC2_1782_08_T08
pRCC2_1851_04_T01
pRCC2_1851_04_T02
pRCC2_1851_04_T03
pRCC2_1851_04_T04
pRCC2_1568_04_T01
pRCC2_1568_04_T02
pRCC2_1568_04_T03
pRCC2_1568_04_T04
pRCC2_1494_04_T01
pRCC2_1494_04_T02
pRCC2_1494_04_T03
pRCC2_1494_04_T04
pRCC2_1552_03_T01
pRCC2_1552_03_T02
pRCC2_1552_03_T03
pRCC2_1429_03_T01
pRCC2_1429_03_T02
pRCC2_1429_03_T03
pRCC2_1479_03_T01
pRCC2_1479_03_T02
pRCC2_1479_03_T03
pRCC2_1799_02_T01
pRCC2_1799_02_T02
pRCC2_1410_02_T01
pRCC2_1410_02_T02
pRCC2_1298_01_T01
pRCC2_1396_01_T01
rSRC_1697_10_T01
rSRC_1697_10_T02
rSRC_1697_10_T03
rSRC_1697_10_T04
rSRC_1697_10_T05
rSRC_1697_10_T06
rSRC_1697_10_T07
rSRC_1697_10_T08
rSRC_1697_10_M01
rSRC_1697_10_M02
cdRCC_1929_03_T01
cdRCC_1929_03_T02
cdRCC_1929_03_T03
mixRCC_2028_03_T01
mixRCC_2028_03_T02
mixRCC_2028_03_T03
mixRCC_1336_01_T01
Somatic mutation:
Missense mutation
Nonsense mutation
Splice site mutation
Driver mutation

Figure S3: Clonality of single nucleotide variants (SNVs) for each tumor. The proportions of SNVs in trunks, internal branches and terminal branches of multi-regional trees (MRT) are presented.

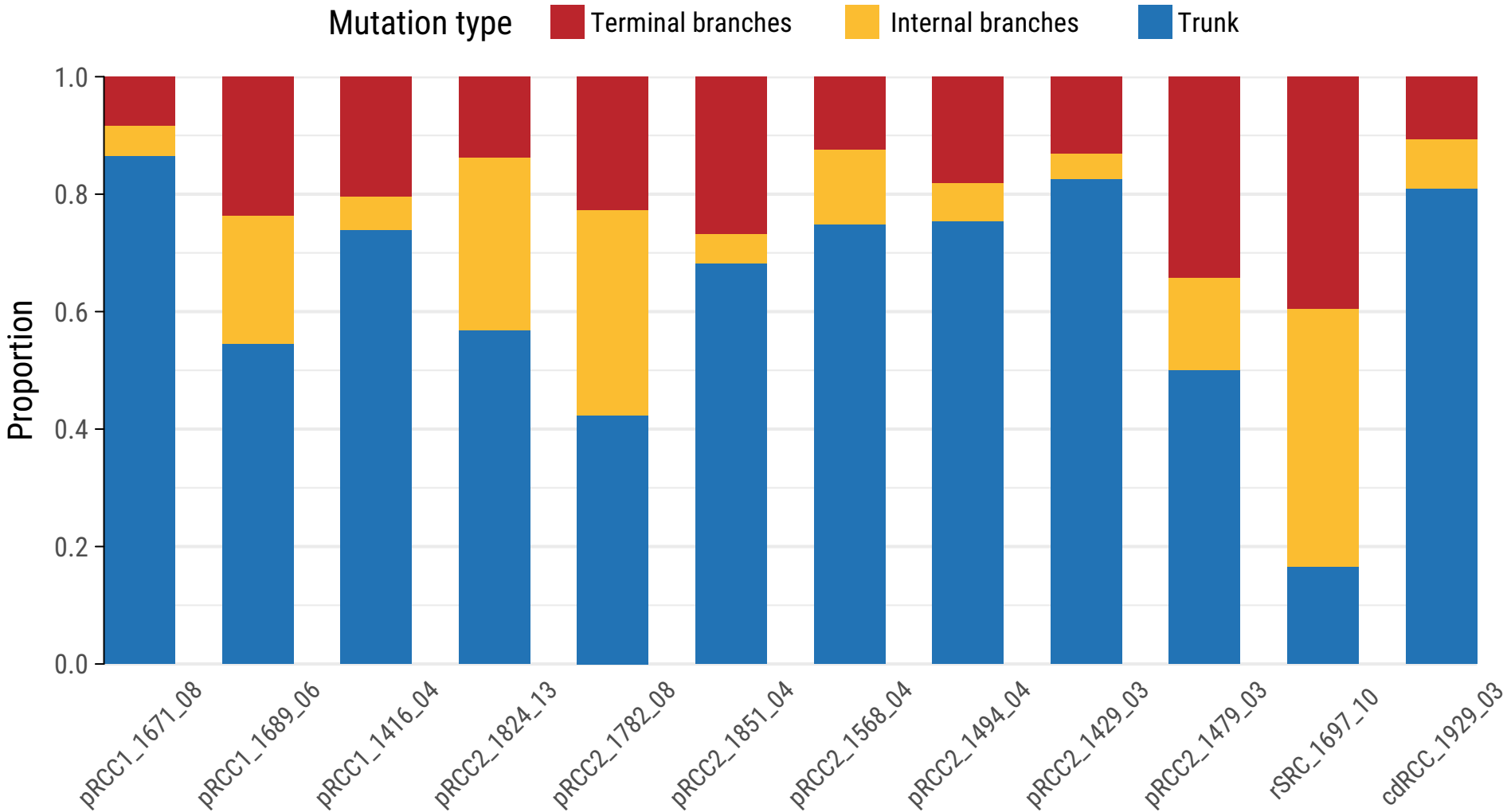

Mutation type
Terminal branches
Internal branches
Trunk
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
pRCC1_1671_08
pRCC1_1689_06
pRCC1_1416_04
pRCC2_1824_13
pRCC2_1782_08
pRCC2_1851_04
pRCC2_1568_04
pRCC2_1494_04
pRCC2_1429_03
pRCC2_1479_03
rSRC_1697_10
cdRCC_1929_03

Figure S4: Intra-tumor heterogeneity of single nucleotide variants by genomic regions. Genomic regions include intergenic, 1 to 5kb from the transcription starting site (TSS), promoters (0 to 1 kb from the TSS), 5'-UTRs, first exon, exon-intron boundaries, exons, introns, intron-exon boundaries, 3'-UTRs, long non-coding RNAs (lncRNAs) and enhancers (annotated by FANTOM) defined in R annotatr package (https://github.com/hhabra/annotatr).

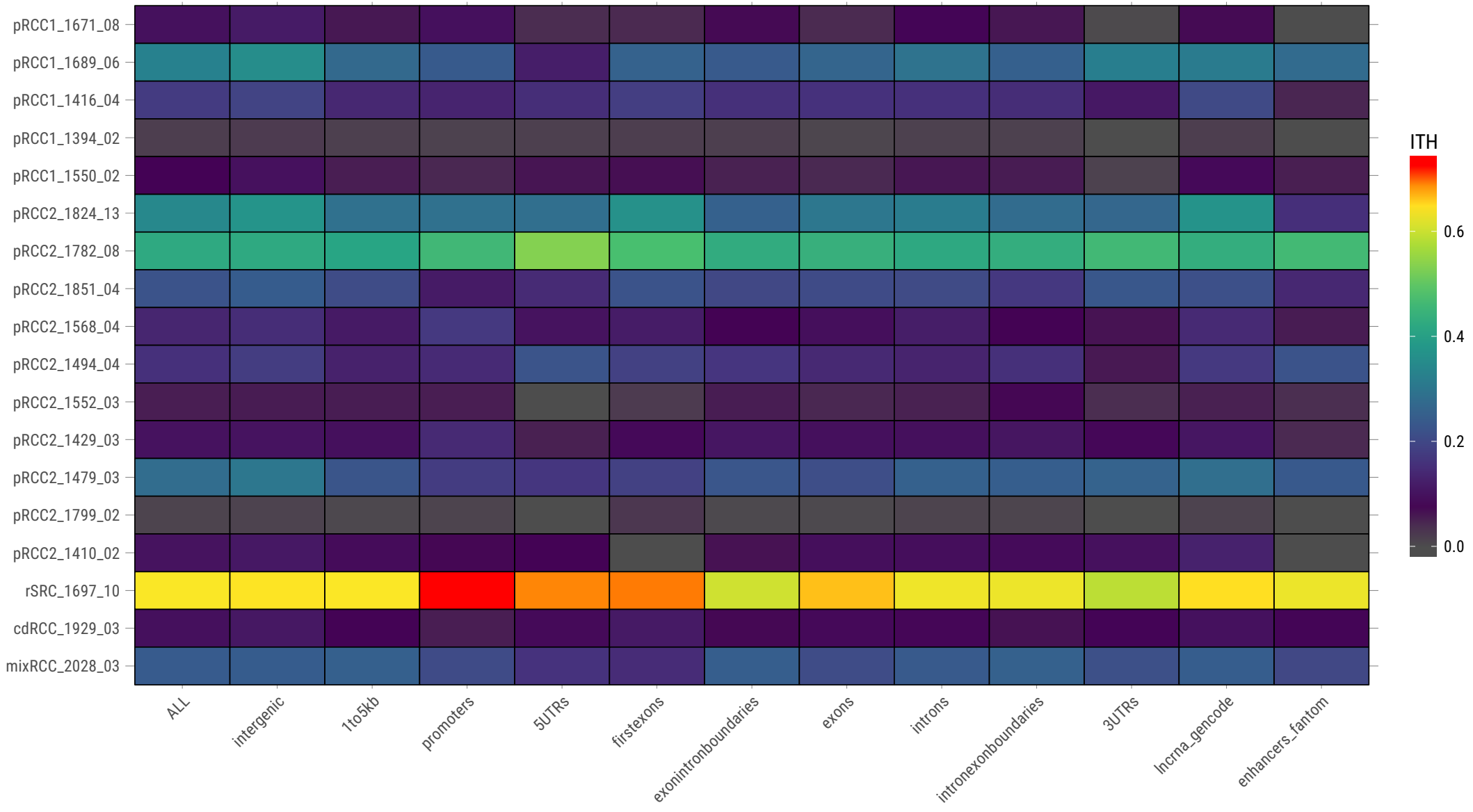
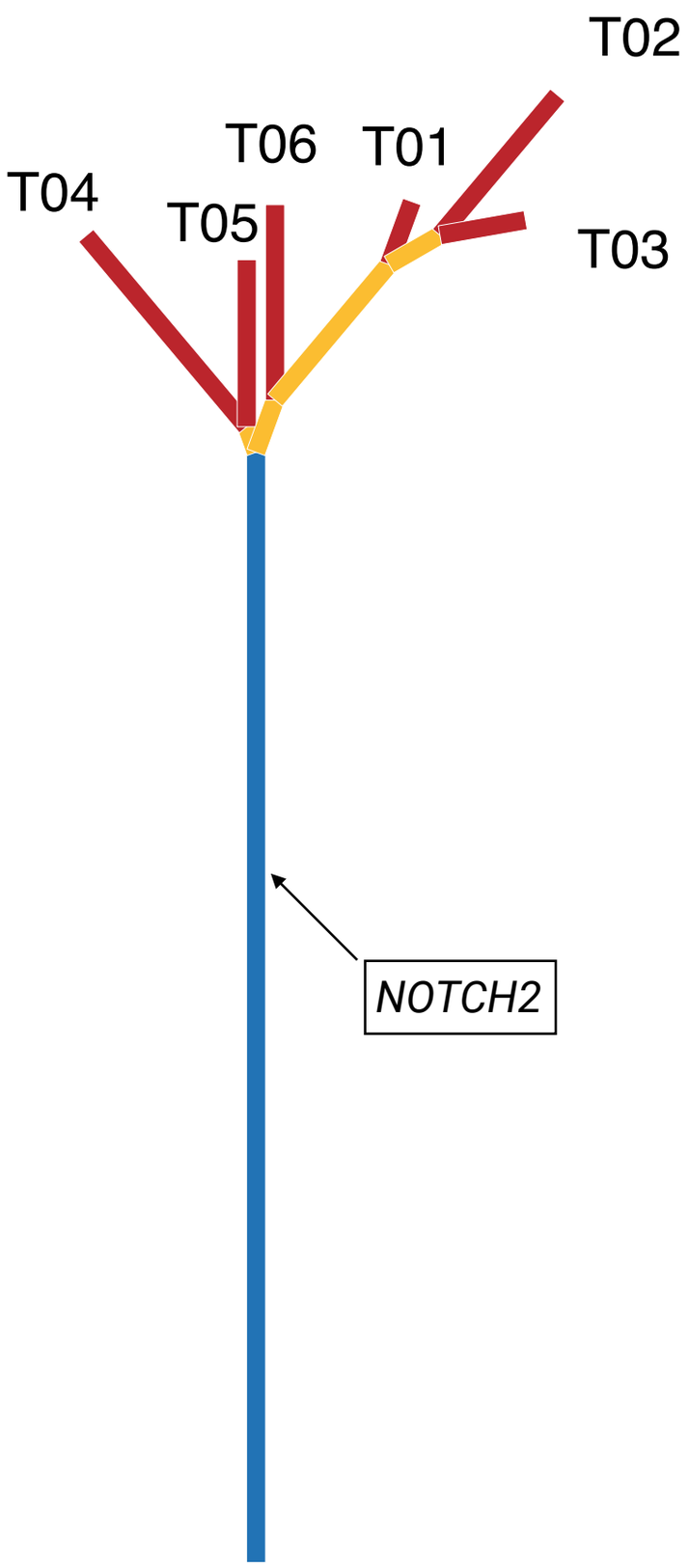
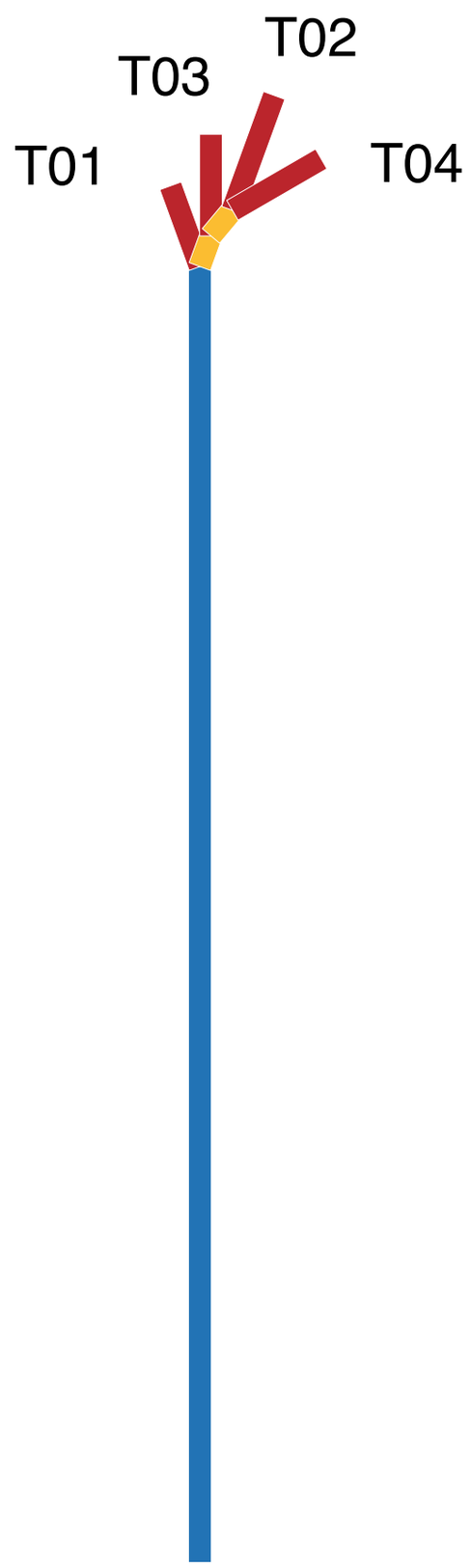
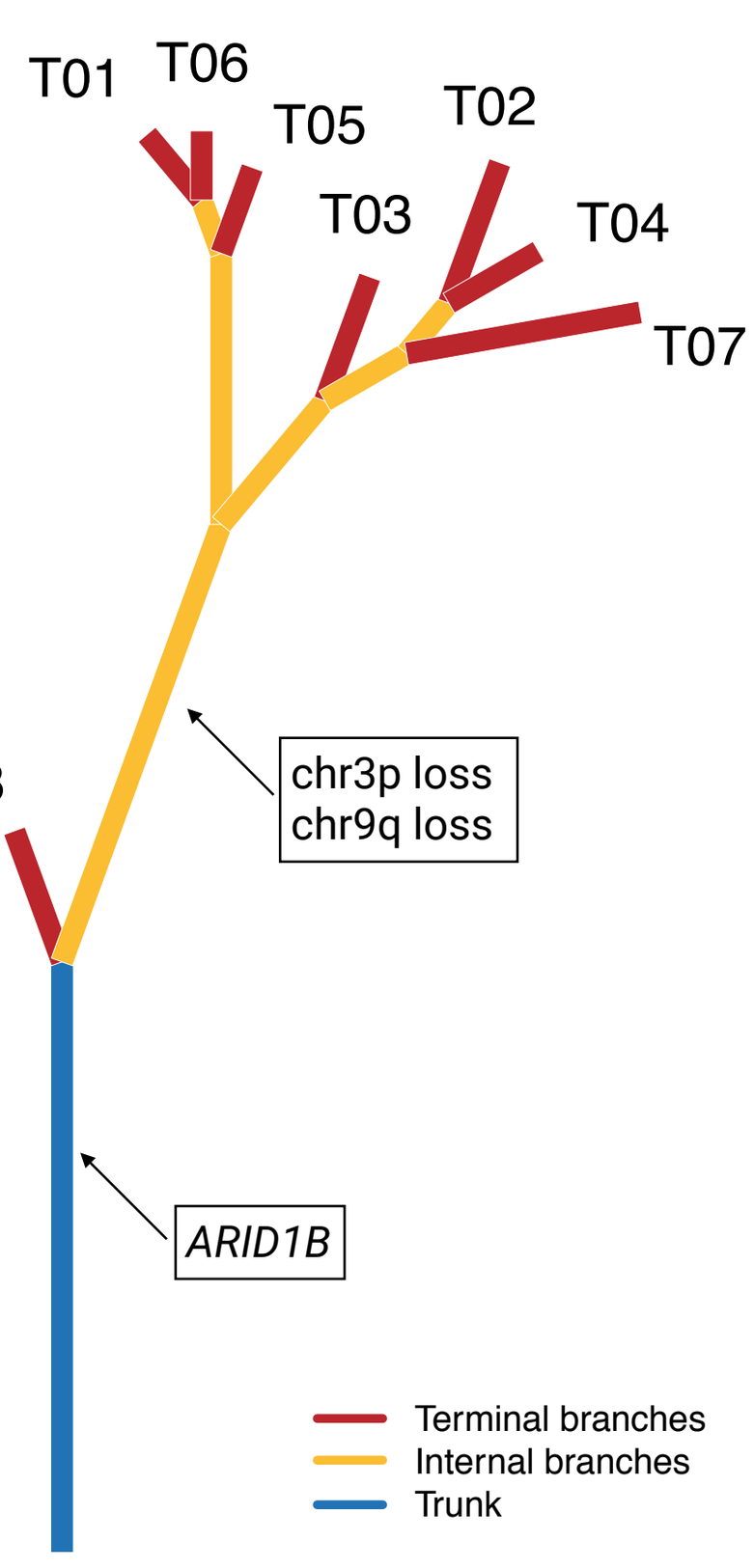

pRCC1_1671_08
pRCC1_1689_06
pRCC1_1416_04
pRCC1_1394_02
pRCC1_1550_02
pRCC2_1824_13
pRCC2_1782_08
pRCC2_1851_04
pRCC2_1568_04
pRCC2_1494_04
pRCC2_1552_03
pRCC2_1429_03
pRCC2_1479_03
pRCC2_1799_02
pRCC2_1410_02
rSRC_1697_10
cdRCC_1929_03
mixRCC_2028_03
ALL
intergenic
1to5kb
promoters
5UTRs
firstexons
exonintronboundaries
exons
introns
intronexonboundaries
3UTRs
lncrna_gencode
enhancers_fantom
ITH
0.6
0.4
0.2
0.0

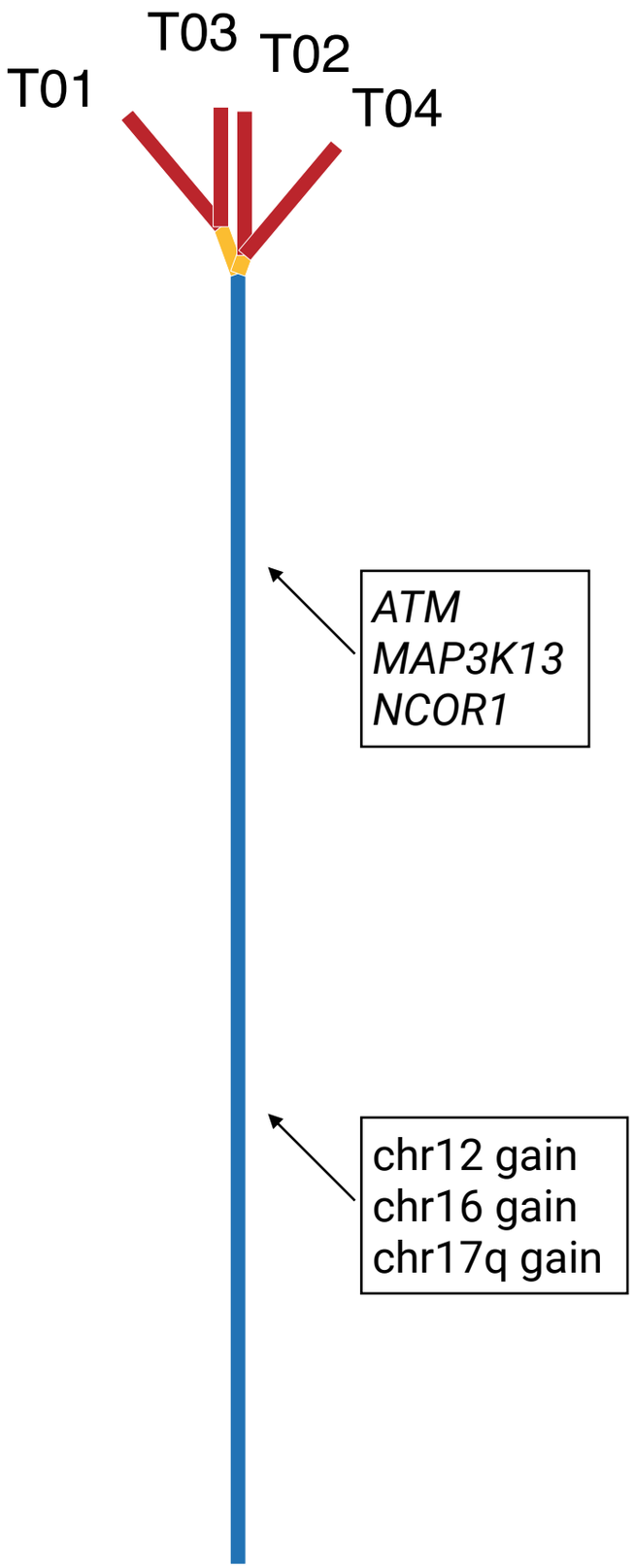
Figure S5: Multi-regional trees (MRT) for each tumor. The branches represent the tumor regions and the lengths of branch or trunk are proportional to the number of single nucleotide variants (SNVs) shared by the tumor regions. The root of the trees represent normal cells without somatic SNVs. SNVs and SVs in driver genes and recurrent somatic copy number alterations are noted on the trees.



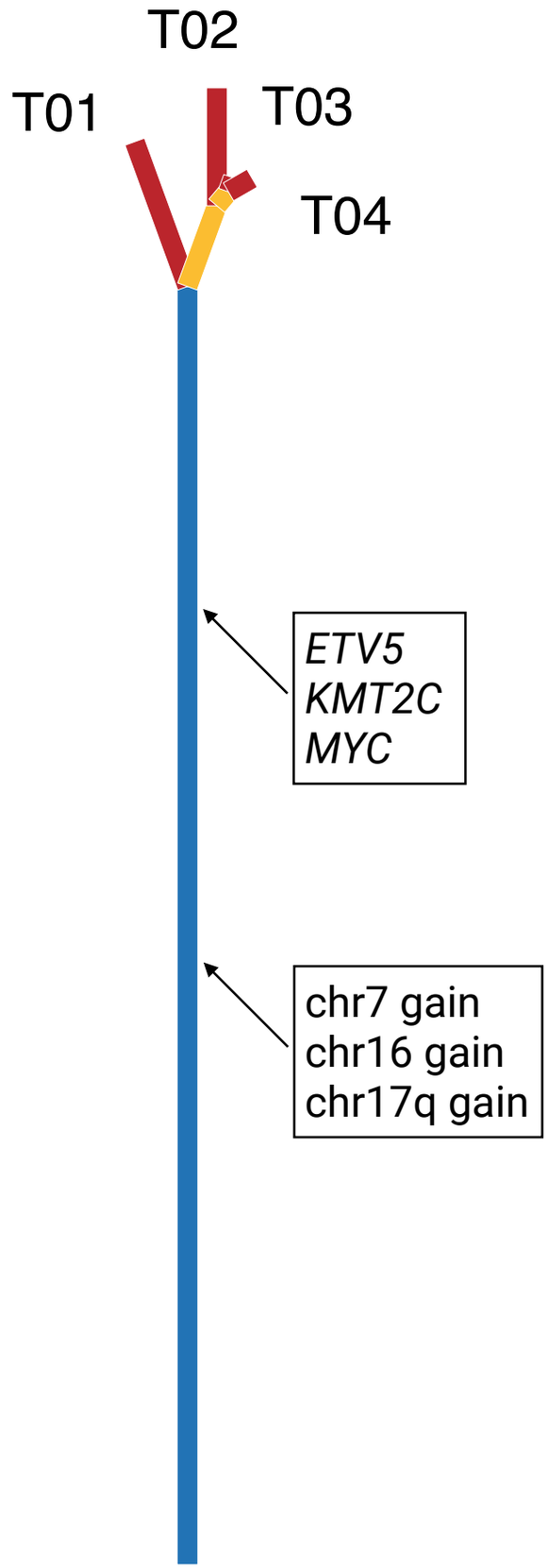




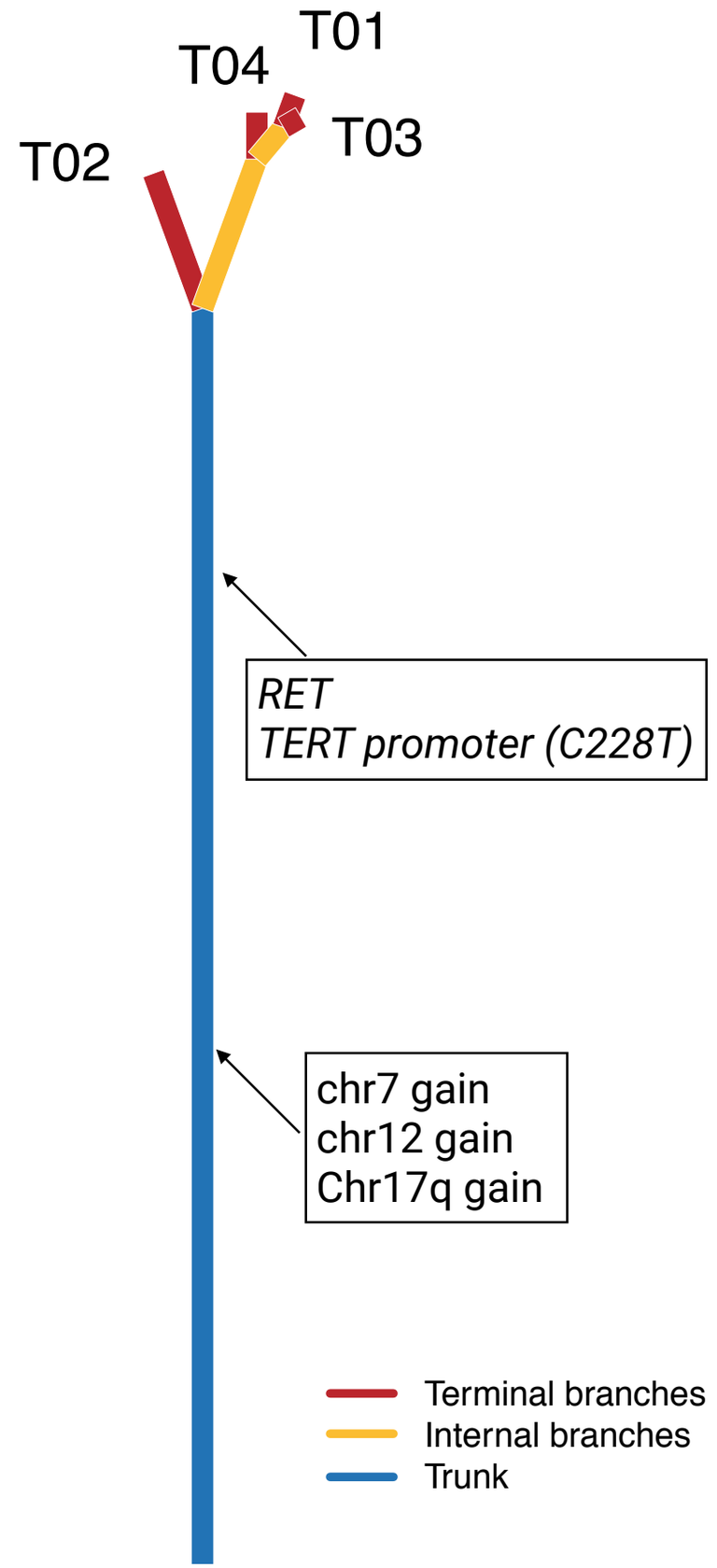
(d) pRCC2_1851_04



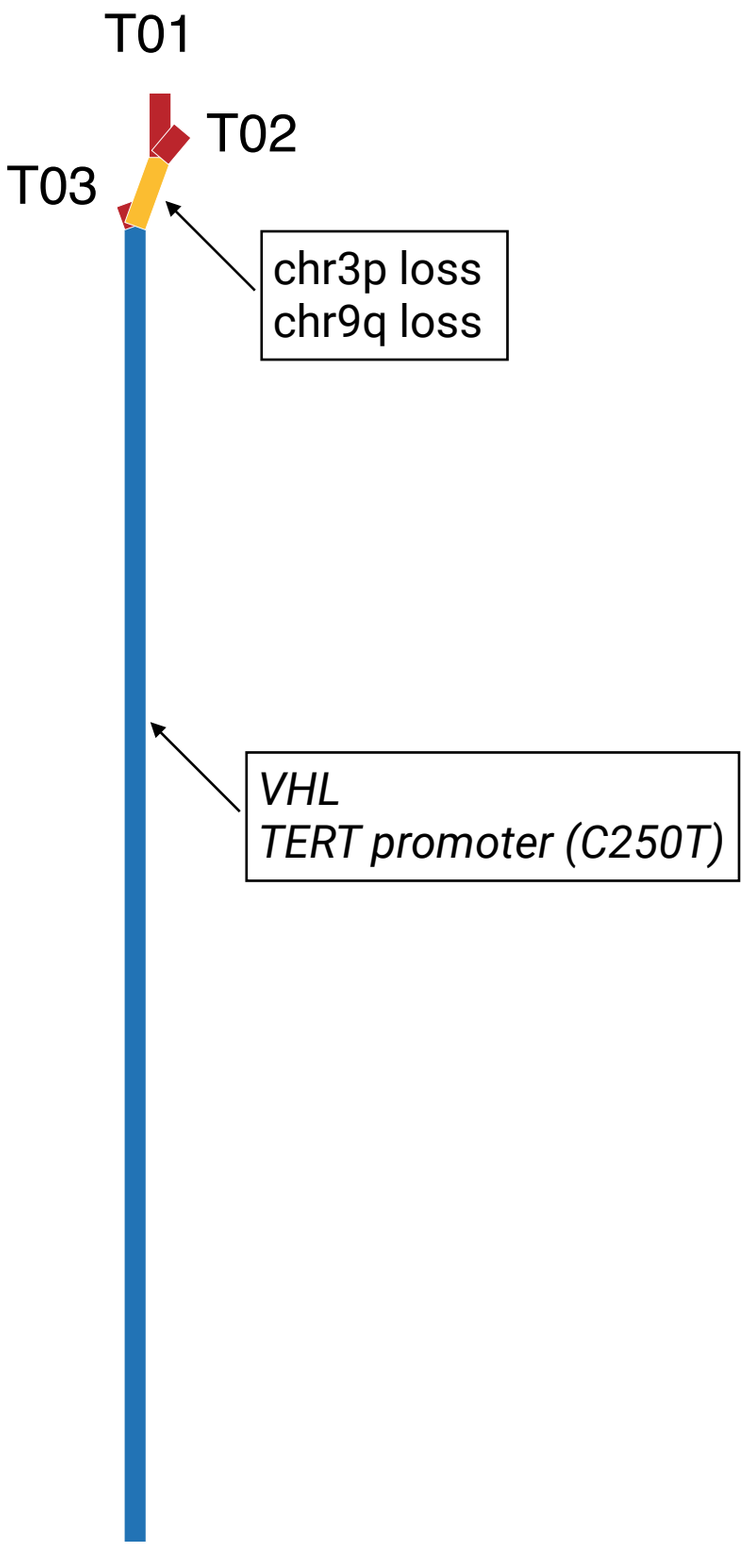
(e) pRCC2_1494_04



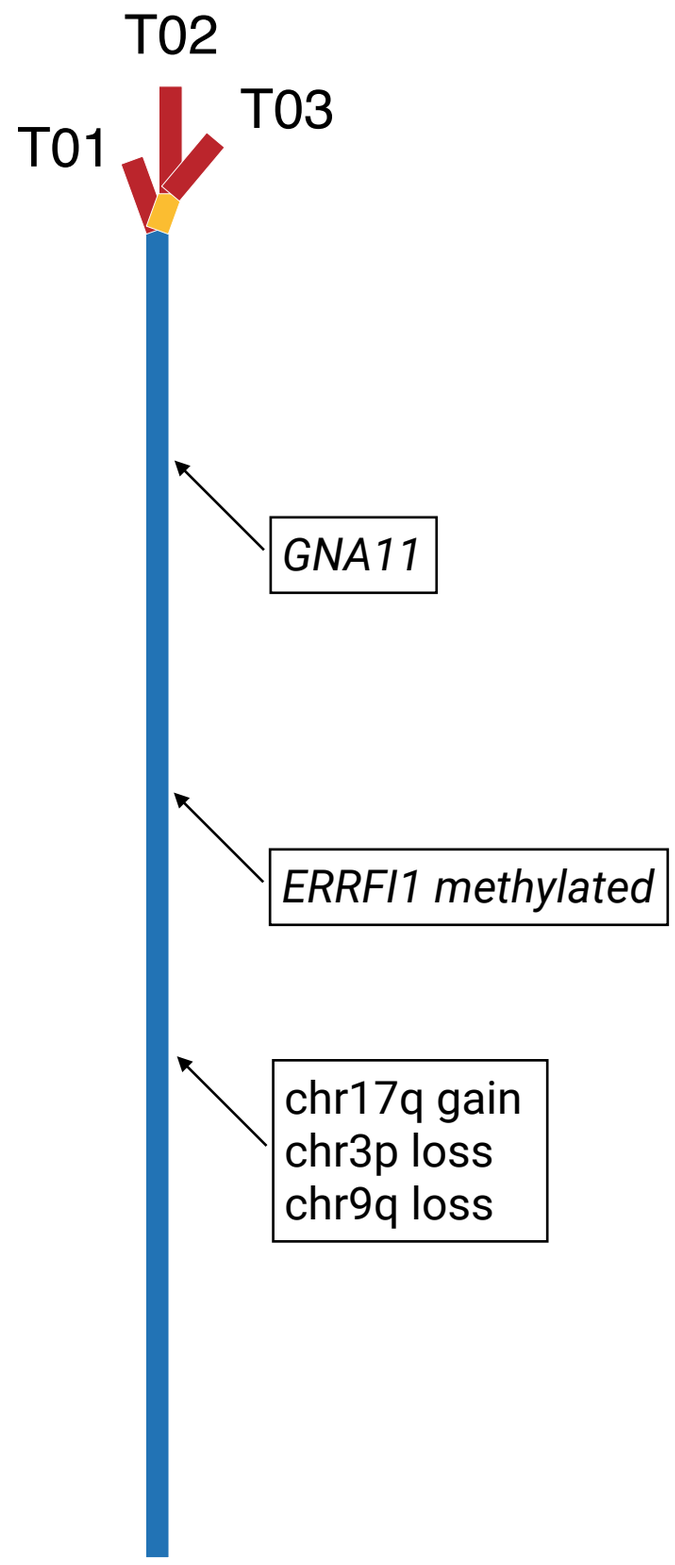
(f) pRCC2_1568_04



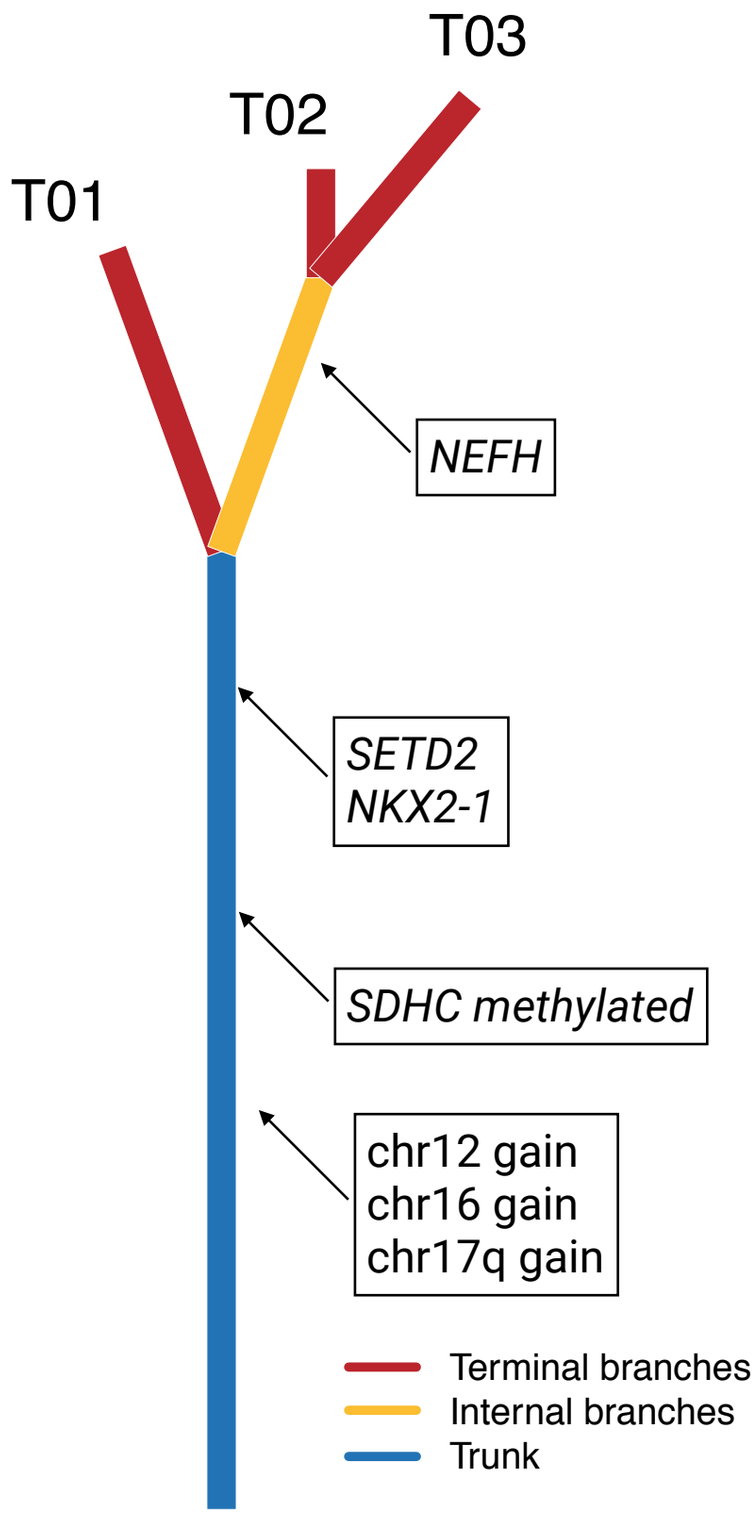
(g) pRCC2_1552_03



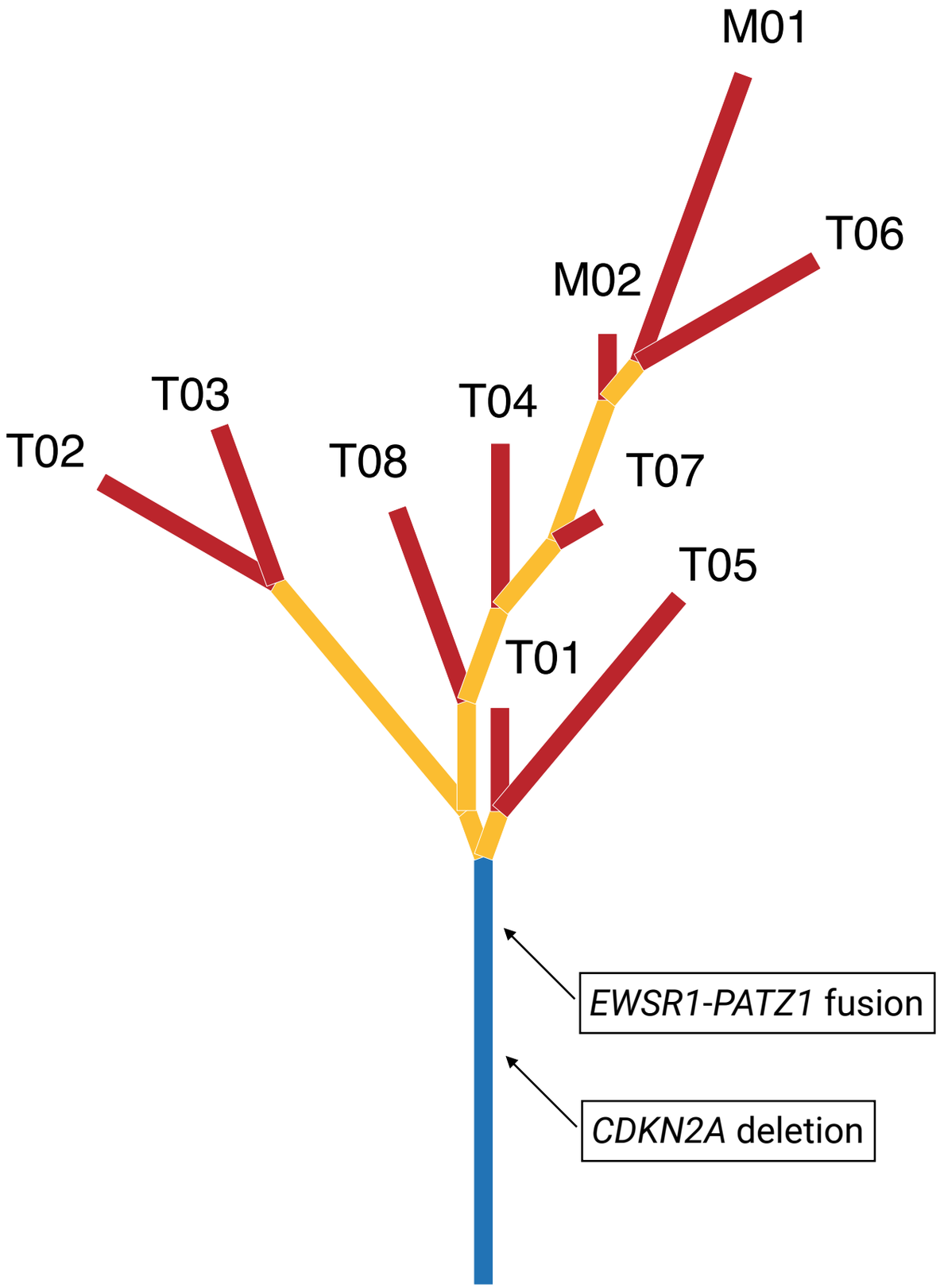
(h) pRCC2_1429_03



(i) pRCC2_1479_03



(j) rSRC_1697_10



(k) cdRCC_1929_03

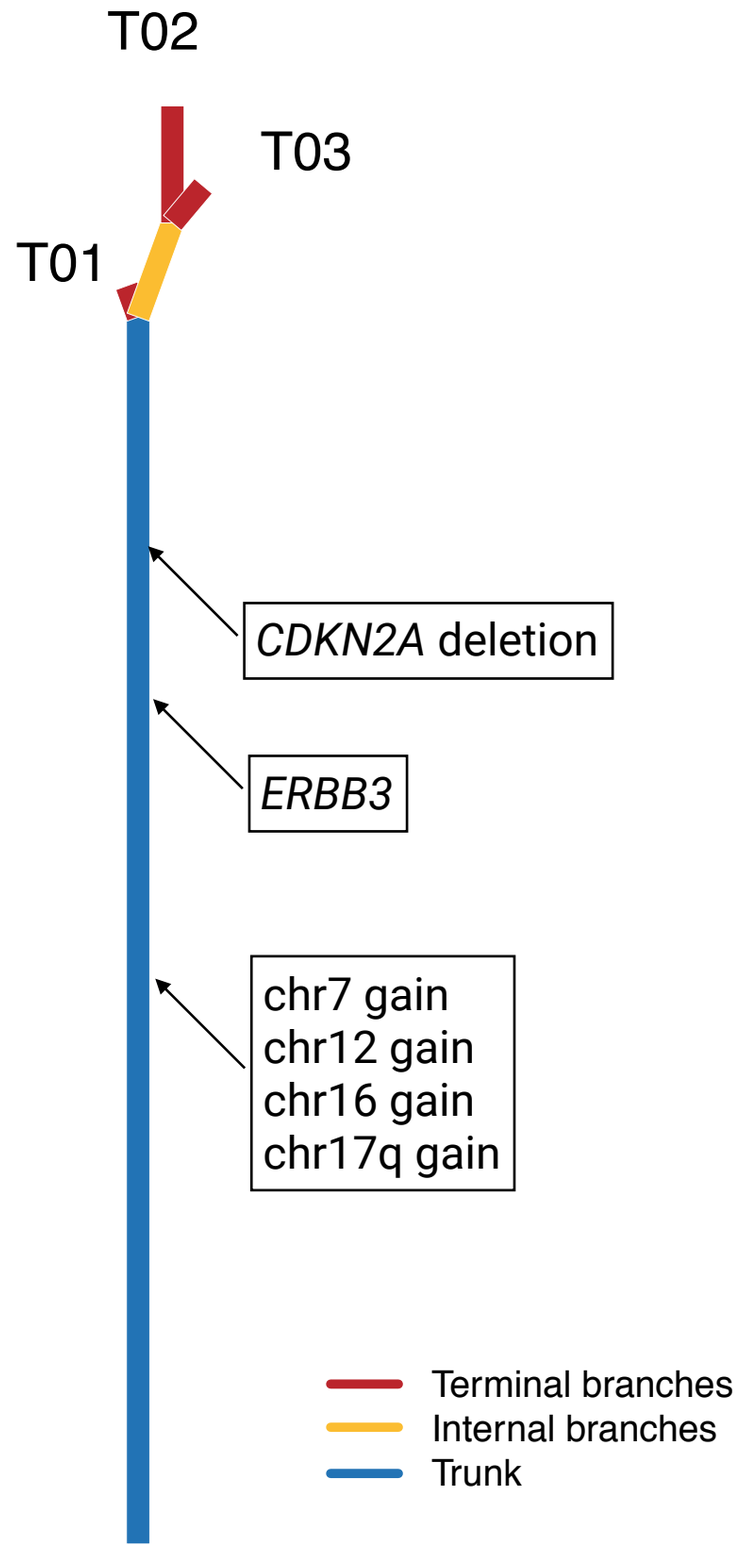
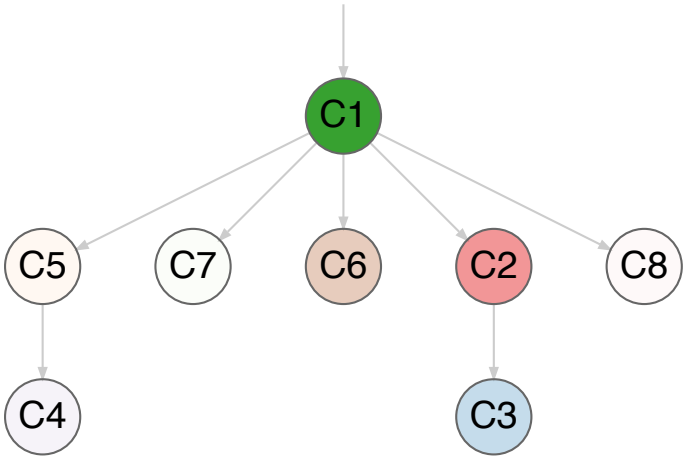


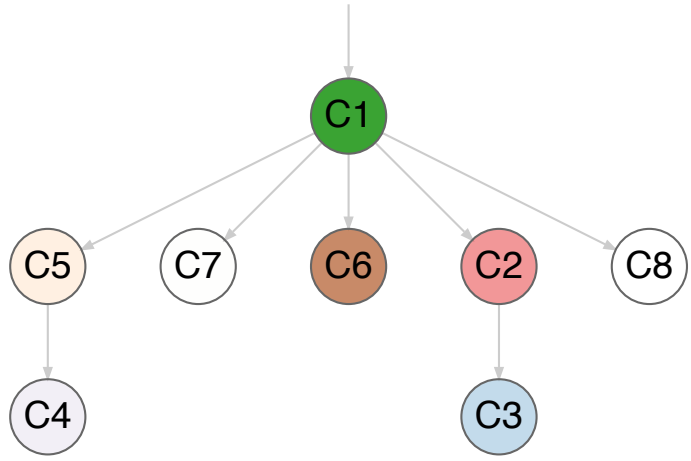
Figure S6: Lineage phylogenetic trees (LPT) for each tumor. The evolutionary history is described for each tumor region, with circles representing subclones (colored: subclone present; blank: subclone absent. Subclone numbers and graduation of circle colors are based on SNV CCFs ranking, from the highest to the lowest). Arrows link the parent and descendant subclones. The subclones were numbered based on cancer-cell fraction (CCF) ranking, from highest to lowest.

(a) pRCC1_1689_06

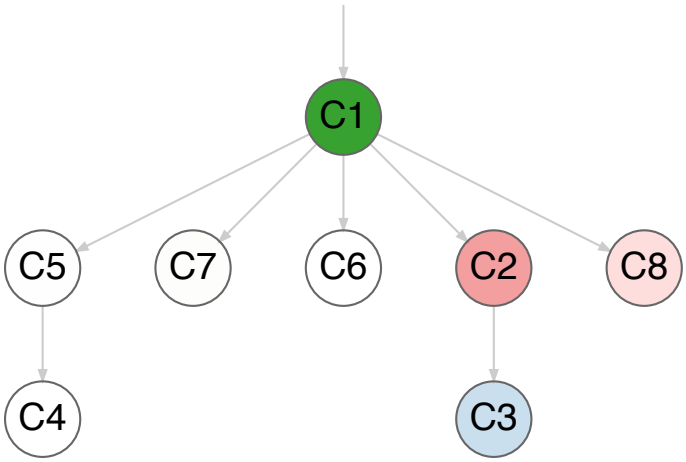
T01



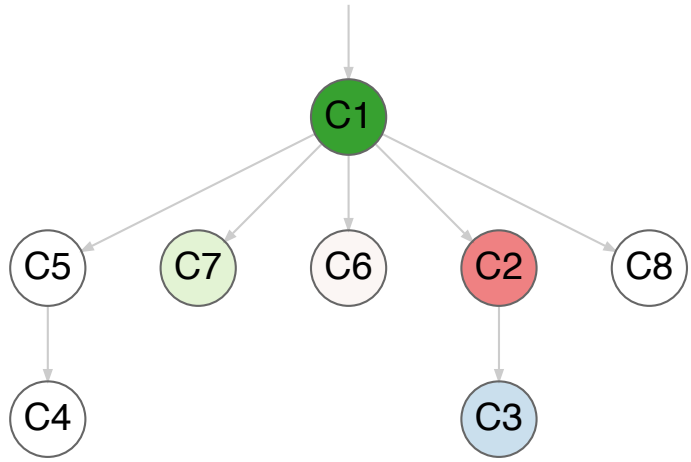
T03



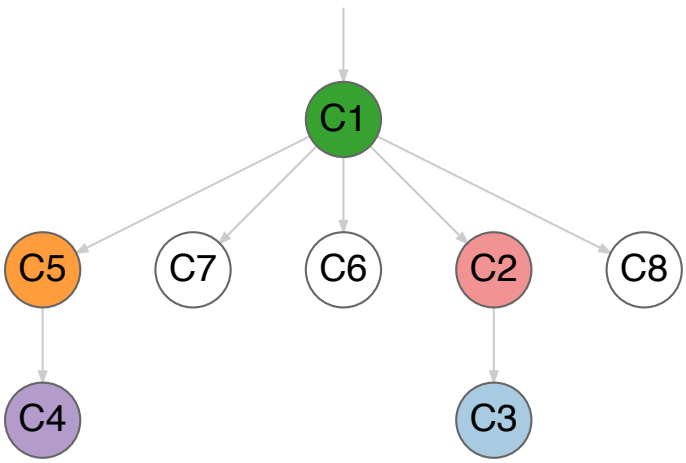
T04



T05

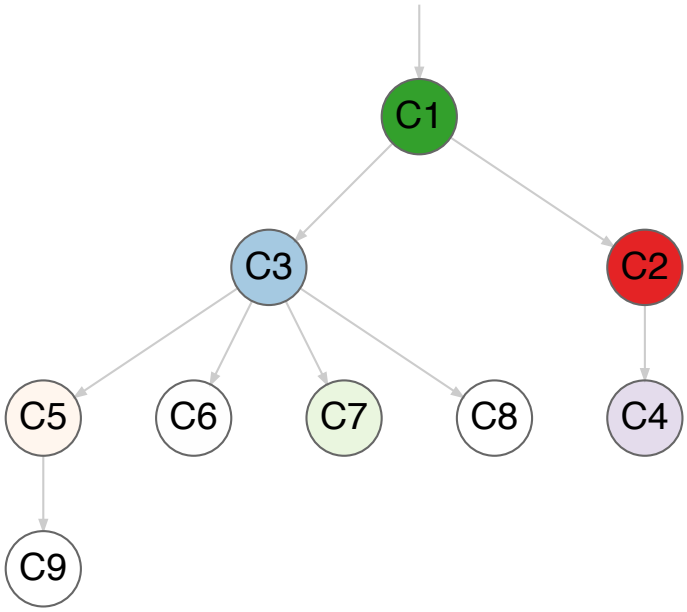


T06

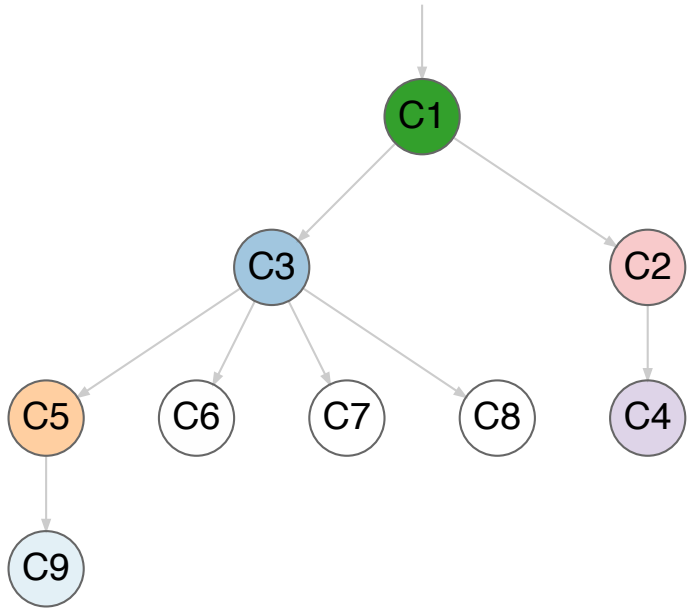


(b) pRCC1_1416_04

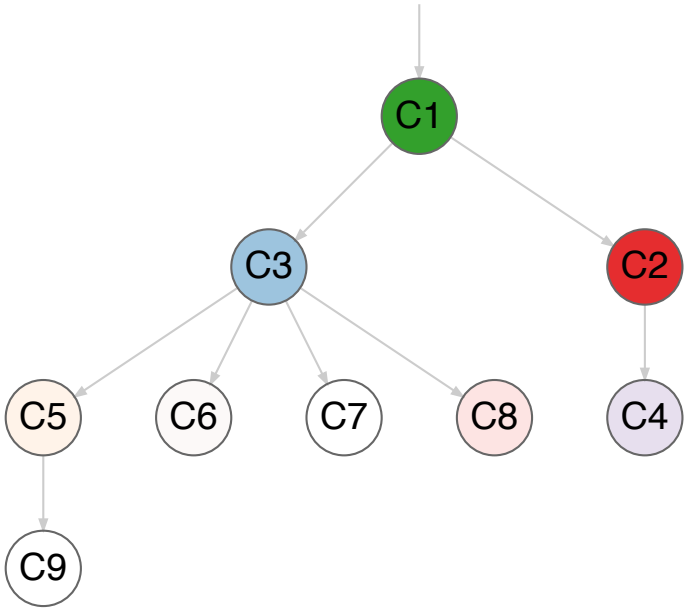
T01



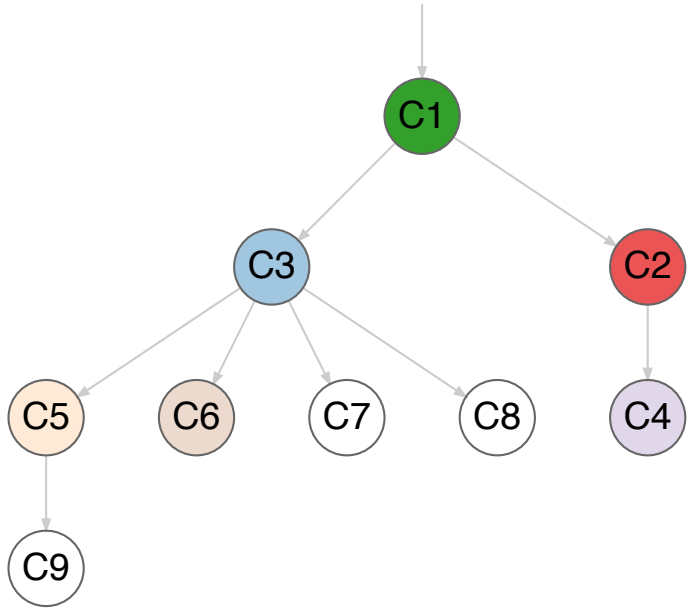
T02



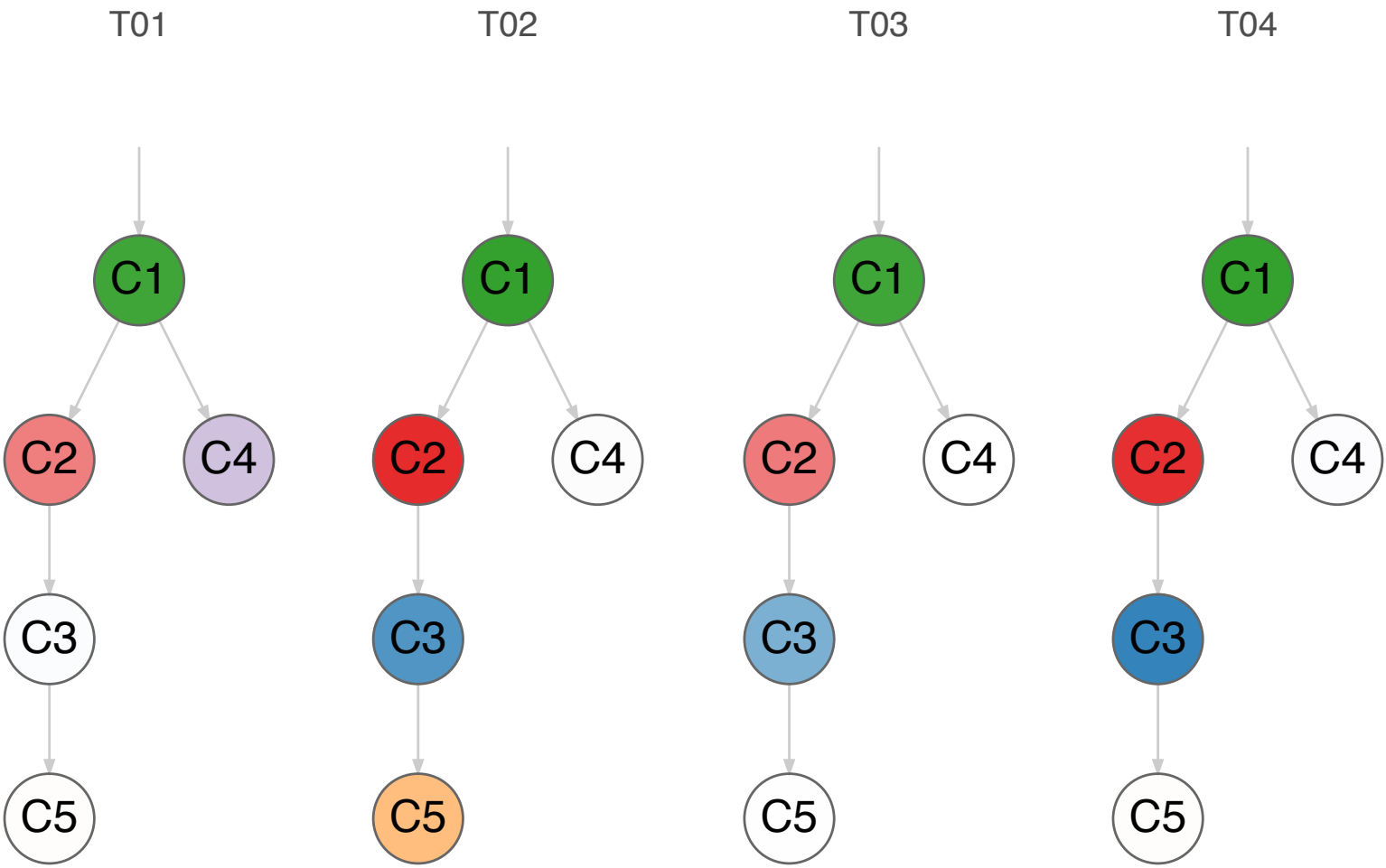
T03



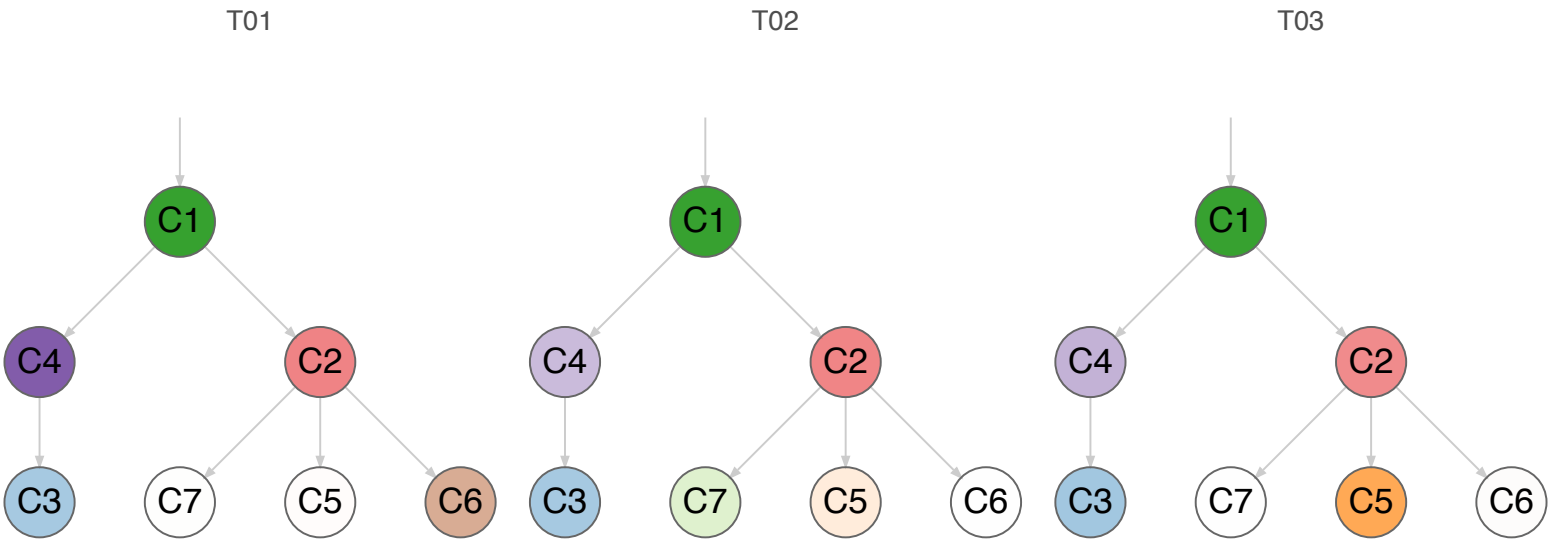
T04



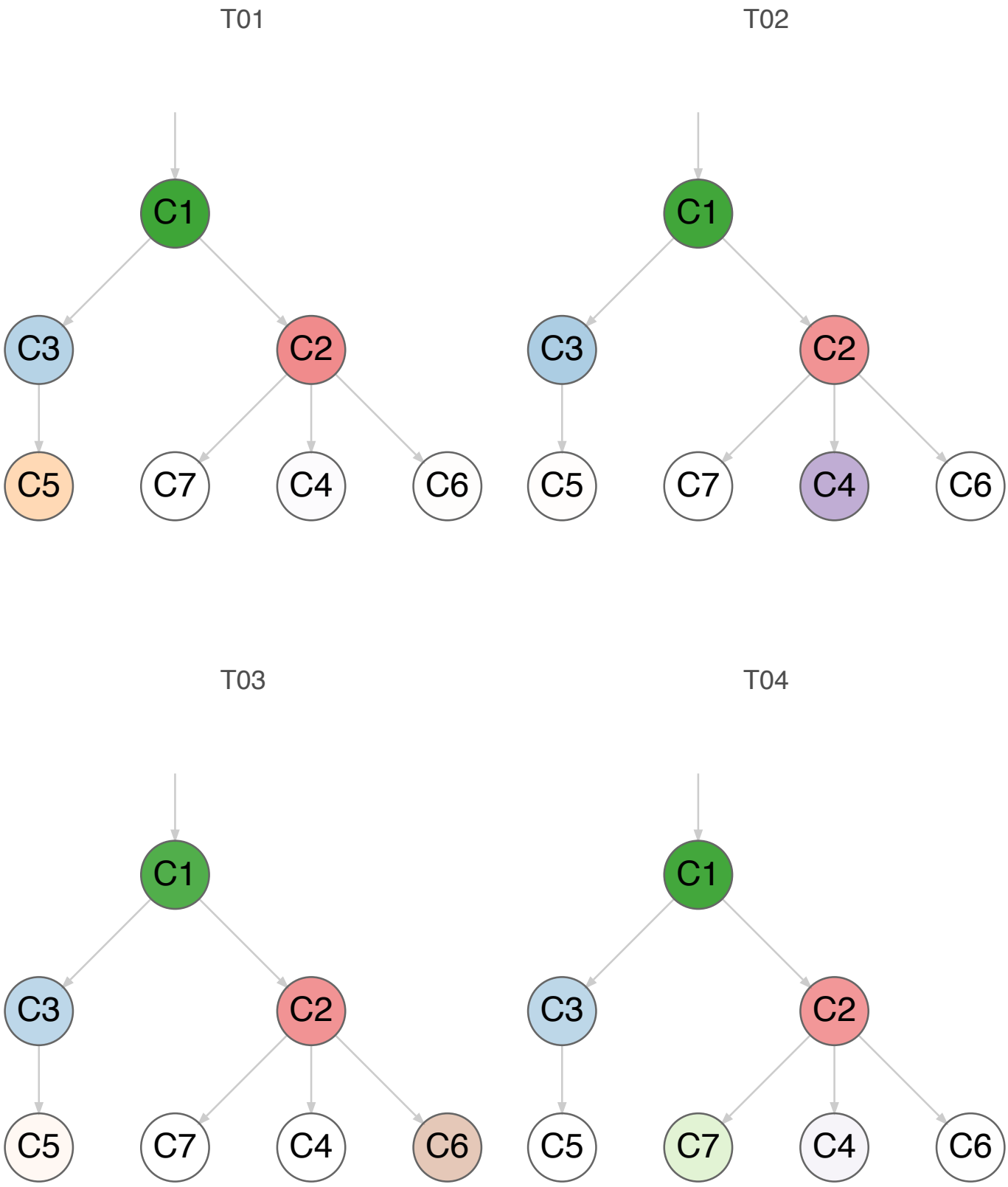
(c) pRCC2_1494_04



(d) pRCC2_1429_03

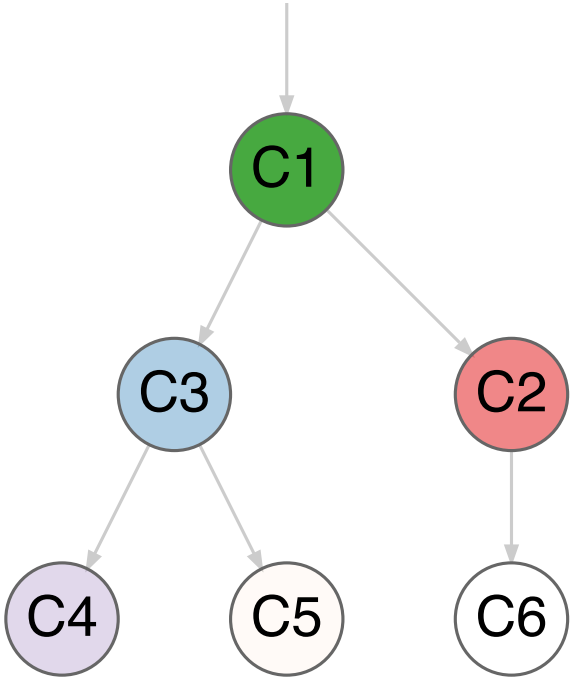


(e) pRCC2_1851_04

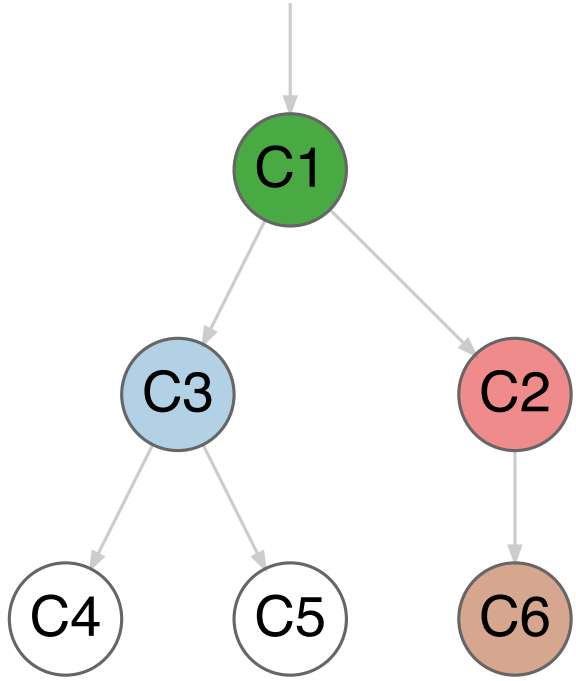


(f) pRCC2_1568_04

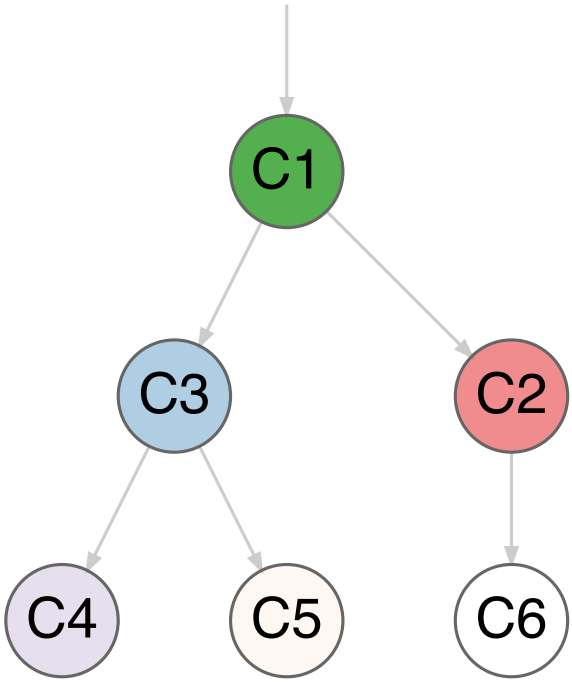
T01



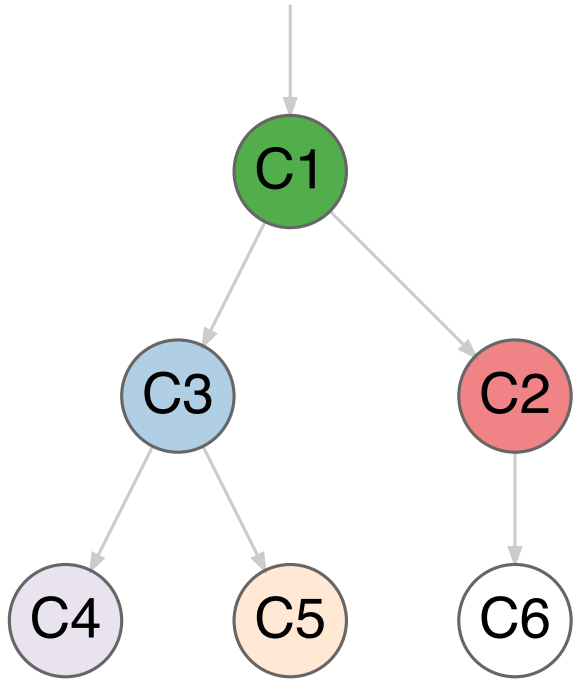
T02



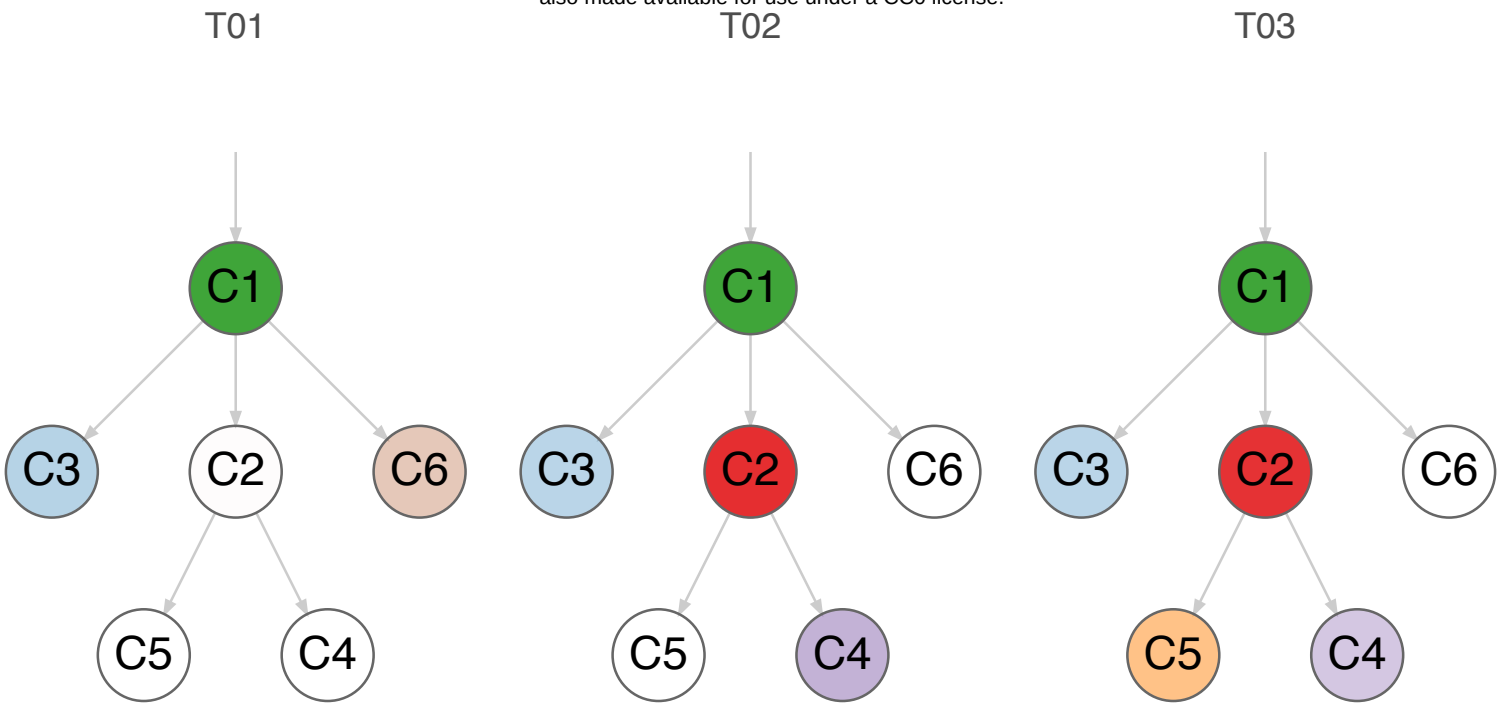
T03



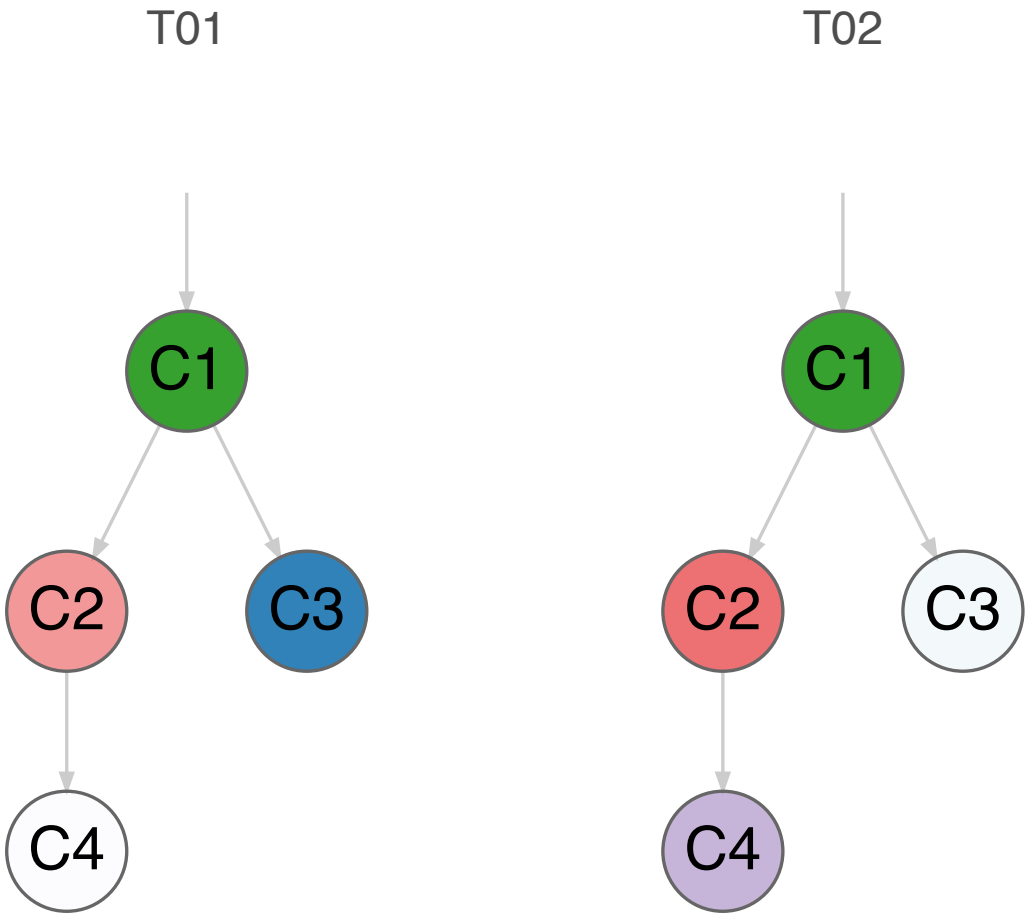
T04



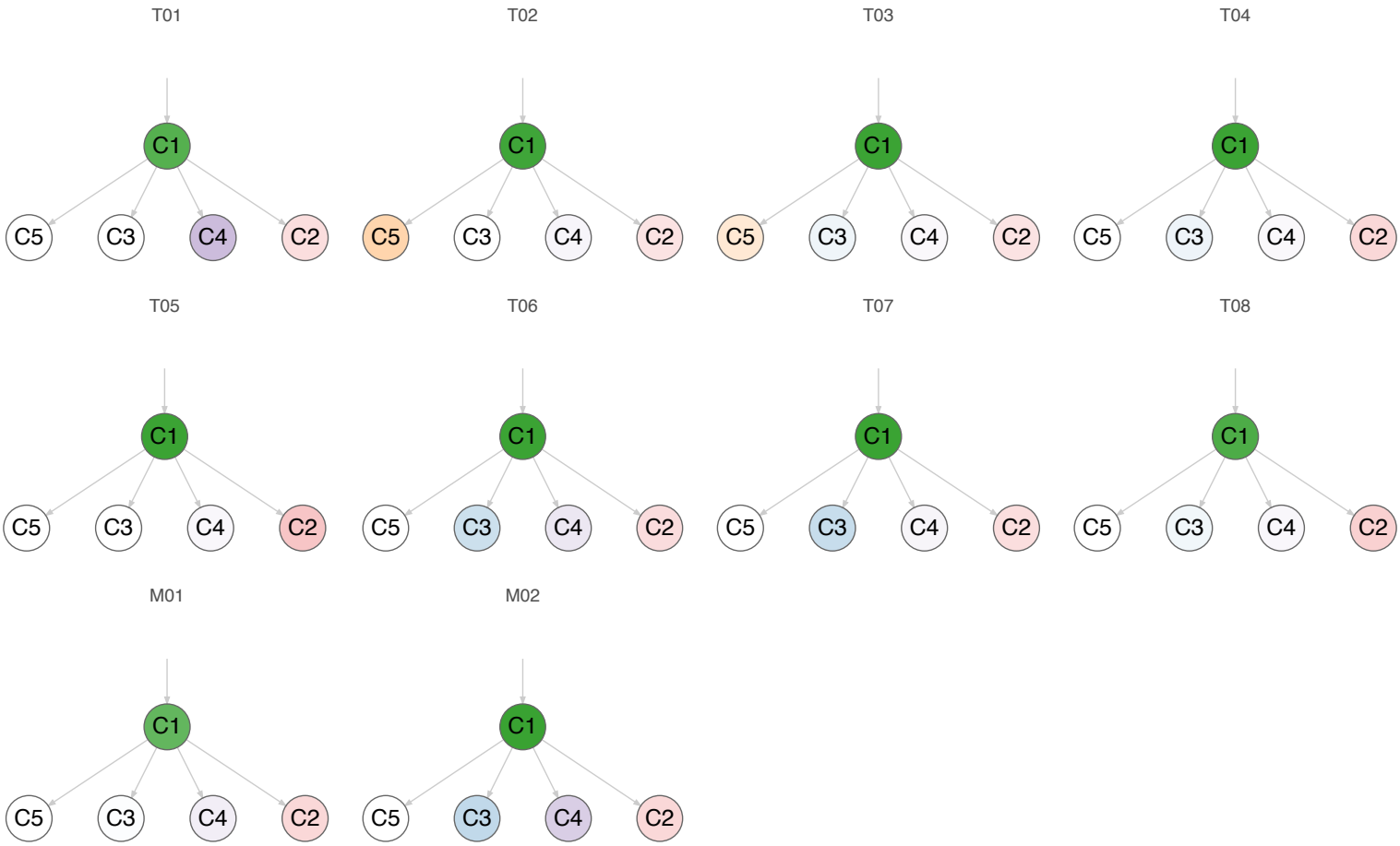
(g) pRCC2_1479_03



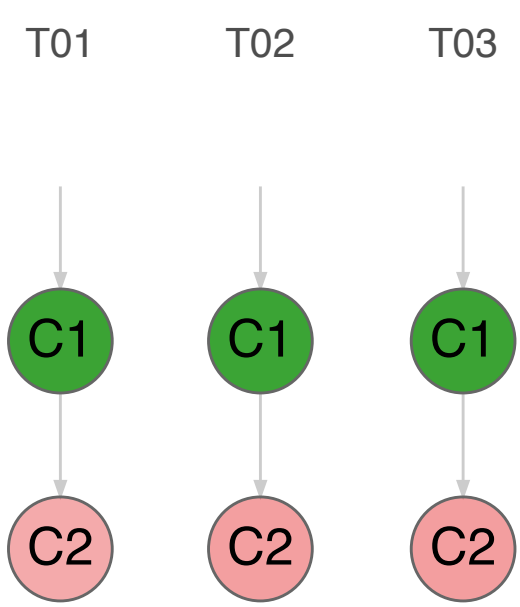
(h) pRCC2_1552_03



(i) rSRC_1697_10



(j) cdRCC_1929_03



(k) mixRCC_2028_03

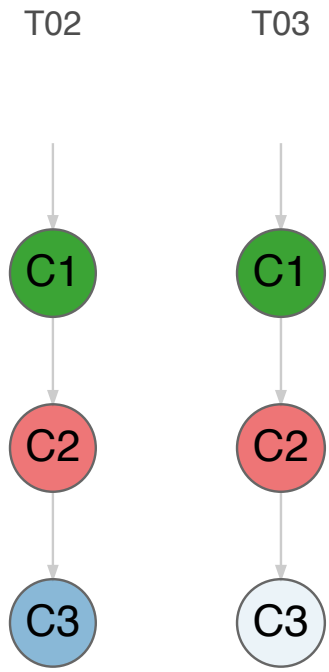
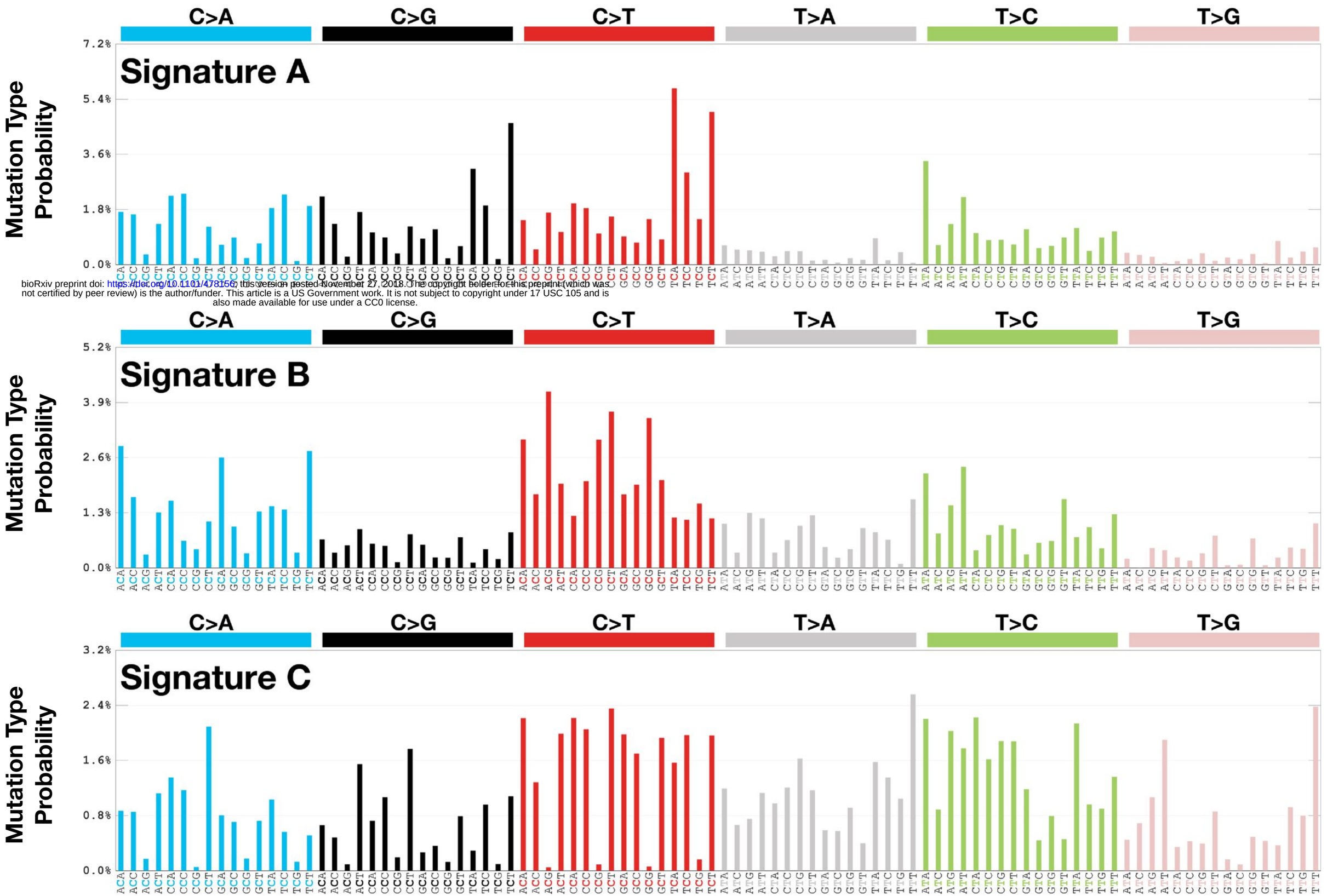


Figure S7: The four SNV mutational signatures identified by *de novo* extraction. The mutational signature is displayed by 96 mutation types, defined by the mutated base and its sequence context immediately 3′ and 5′ on the horizontal axes. The vertical axes depict the percentage of mutations attributed to each mutation type



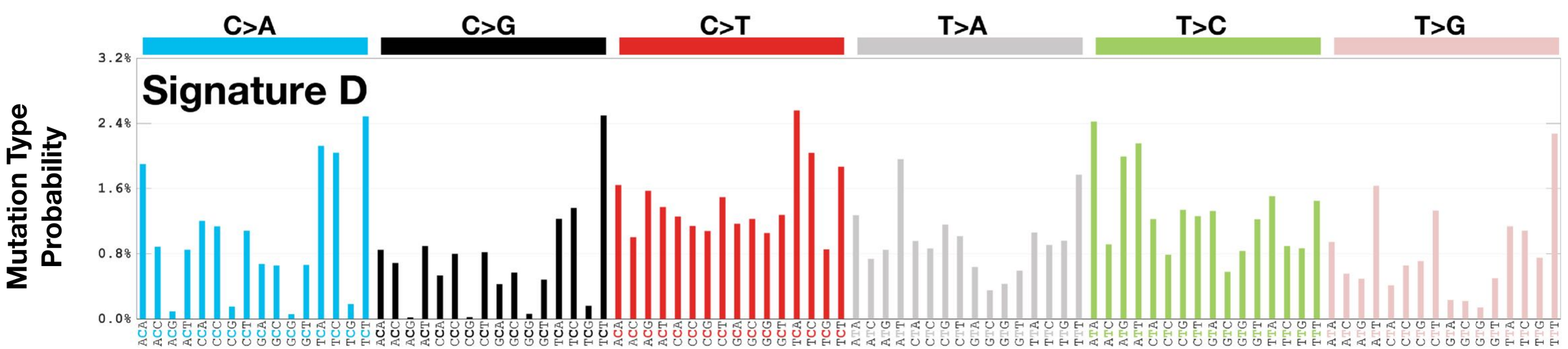


Figure S8: Proportions of mutational signatures. The contribution of known mutational signatures in each sample.

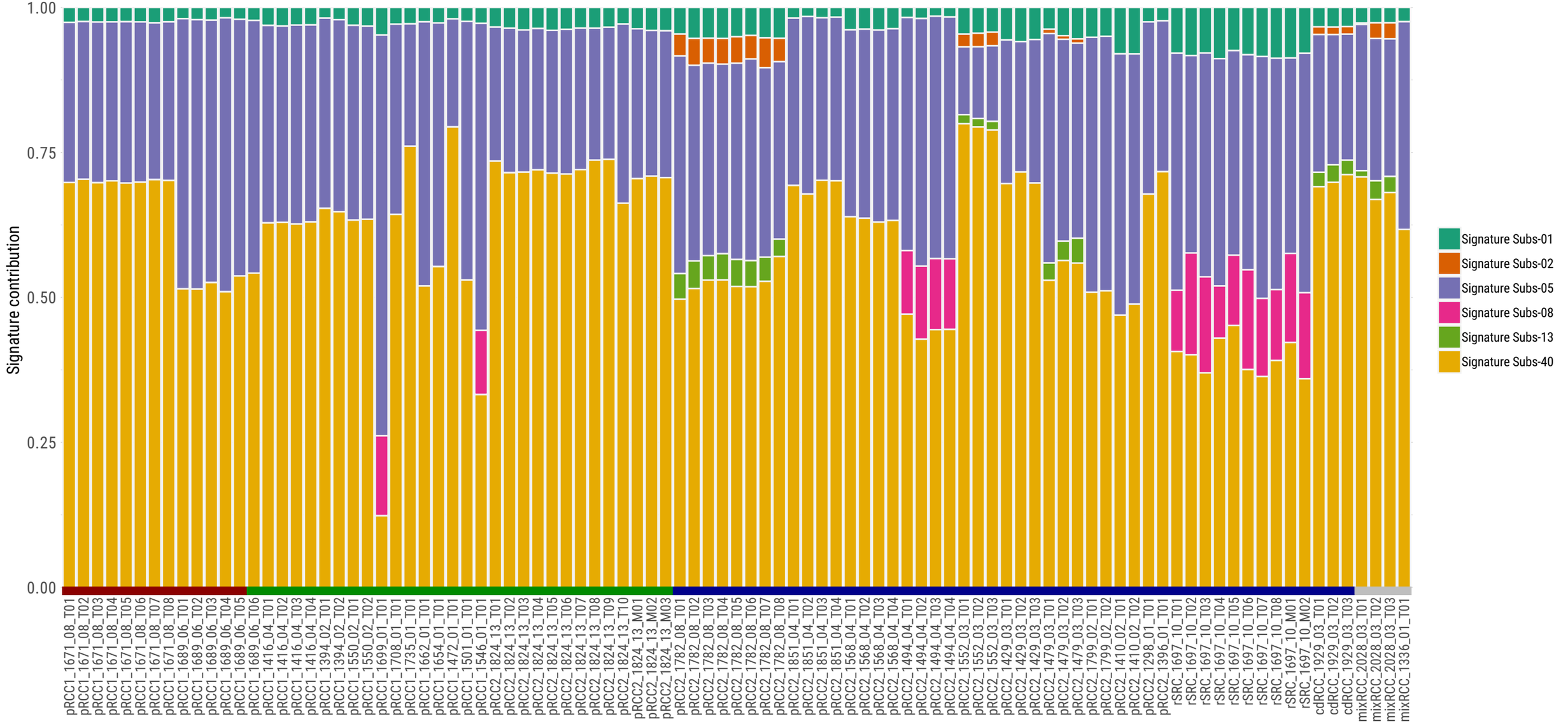

1.00
0.75
0.50
0.25
0.00
Signature contribution
Signature Subs-01
Signature Subs-02
Signature Subs-05
Signature Subs-08
Signature Subs-13
Signature Subs-40
pRCC1_1671_08_T01
pRCC1_1671_08_T02
pRCC1_1671_08_T03
pRCC1_1671_08_T04
pRCC1_1671_08_T05
pRCC1_1671_08_T06
pRCC1_1671_08_T07
pRCC1_1671_08_T08
pRCC1_1689_06_T01
pRCC1_1689_06_T02
pRCC1_1689_06_T03
pRCC1_1689_06_T04
pRCC1_1689_06_T05
pRCC1_1689_06_T06
pRCC1_1416_04_T01
pRCC1_1416_04_T02
pRCC1_1416_04_T03
pRCC1_1416_04_T04
pRCC1_1394_02_T01
pRCC1_1394_02_T02
pRCC1_1550_02_T01
pRCC1_1550_02_T02
pRCC1_1699_01_T01
pRCC1_1708_01_T01
pRCC1_1735_01_T01
pRCC1_1662_01_T01
pRCC1_1654_01_T01
pRCC1_1472_01_T01
pRCC1_1501_01_T01
pRCC1_1546_01_T01
pRCC2_1824_13_T01
pRCC2_1824_13_T02
pRCC2_1824_13_T03
pRCC2_1824_13_T04
pRCC2_1824_13_T05
pRCC2_1824_13_T06
pRCC2_1824_13_T07
pRCC2_1824_13_T08
pRCC2_1824_13_T09
pRCC2_1824_13_T10
pRCC2_1824_13_M01
pRCC2_1824_13_M02
pRCC2_1824_13_M03
pRCC2_1782_08_T01
pRCC2_1782_08_T02
pRCC2_1782_08_T03
pRCC2_1782_08_T04
pRCC2_1782_08_T05
pRCC2_1782_08_T06
pRCC2_1782_08_T07
pRCC2_1782_08_T08
pRCC2_1851_04_T01
pRCC2_1851_04_T02
pRCC2_1851_04_T03
pRCC2_1851_04_T04
pRCC2_1568_04_T01
pRCC2_1568_04_T02
pRCC2_1568_04_T03
pRCC2_1568_04_T04
pRCC2_1494_04_T01
pRCC2_1494_04_T02
pRCC2_1494_04_T03
pRCC2_1494_04_T04
pRCC2_1552_03_T01
pRCC2_1552_03_T02
pRCC2_1552_03_T03
pRCC2_1429_03_T01
pRCC2_1429_03_T02
pRCC2_1429_03_T03
pRCC2_1479_03_T01
pRCC2_1479_03_T02
pRCC2_1479_03_T03
pRCC2_1799_02_T01
pRCC2_1799_02_T02
pRCC2_1410_02_T01
pRCC2_1410_02_T02
pRCC2_1298_01_T01
pRCC2_1396_01_T01
rSRC_1697_10_T01
rSRC_1697_10_T02
rSRC_1697_10_T03
rSRC_1697_10_T04
rSRC_1697_10_T05
rSRC_1697_10_T06
rSRC_1697_10_T07
rSRC_1697_10_T08
rSRC_1697_10_M01
rSRC_1697_10_M02
cdRCC_1929_03_T01
cdRCC_1929_03_T02
cdRCC_1929_03_T03
mixRCC_2028_03_T01
mixRCC_2028_03_T02
mixRCC_2028_03_T03
mixRCC_1336_01_T01

Figure S9: Clonality of known mutational signatures for all samples and by subtypes. The proportion of mutational signatures are reported for trunks and branches.

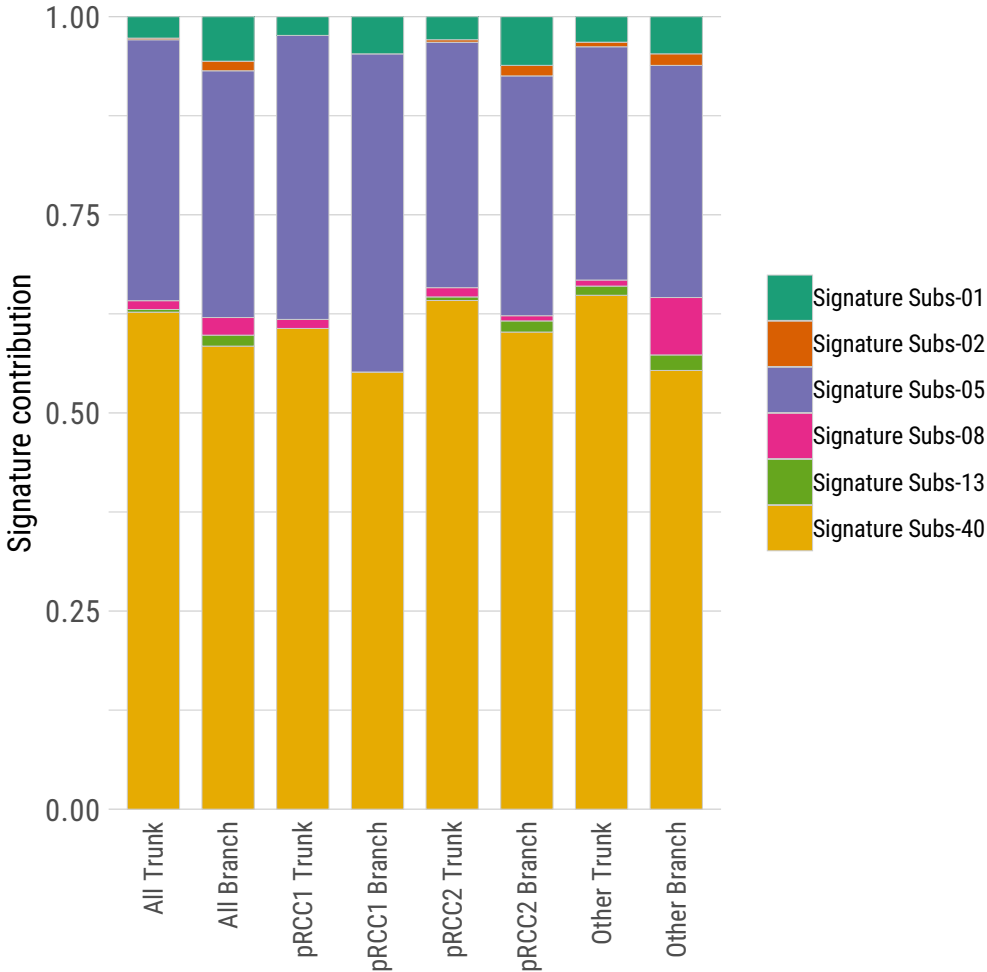
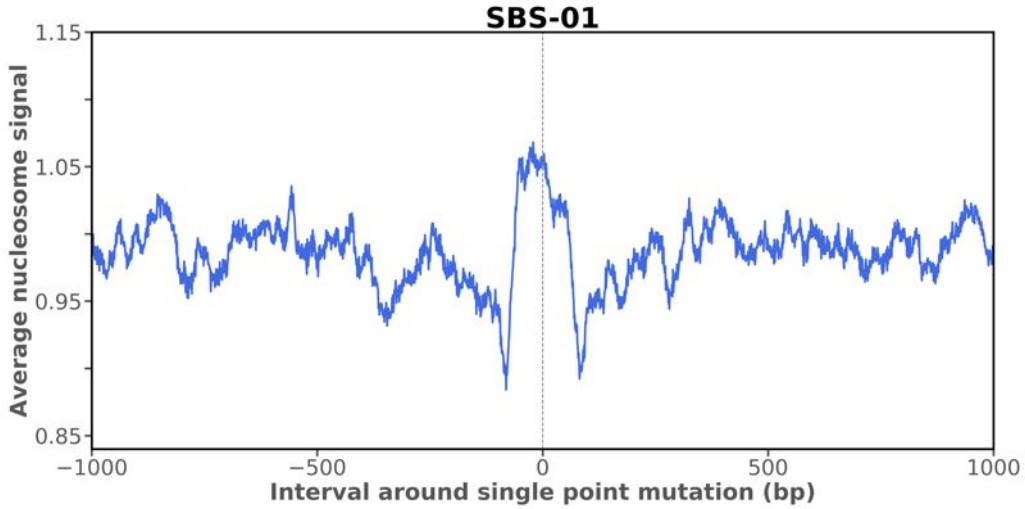

1.00
0.75
0.50
0.25
0.00
Signature contribution
All Trunk
All Branch
pRCC1 Trunk
pRCC1 Branch
pRCC2 Trunk
pRCC2 Branch
Other Trunk
Other Branch
Signature Subs-01
Signature Subs-02
Signature Subs-05
Signature Subs-08
Signature Subs-13
Signature Subs-40

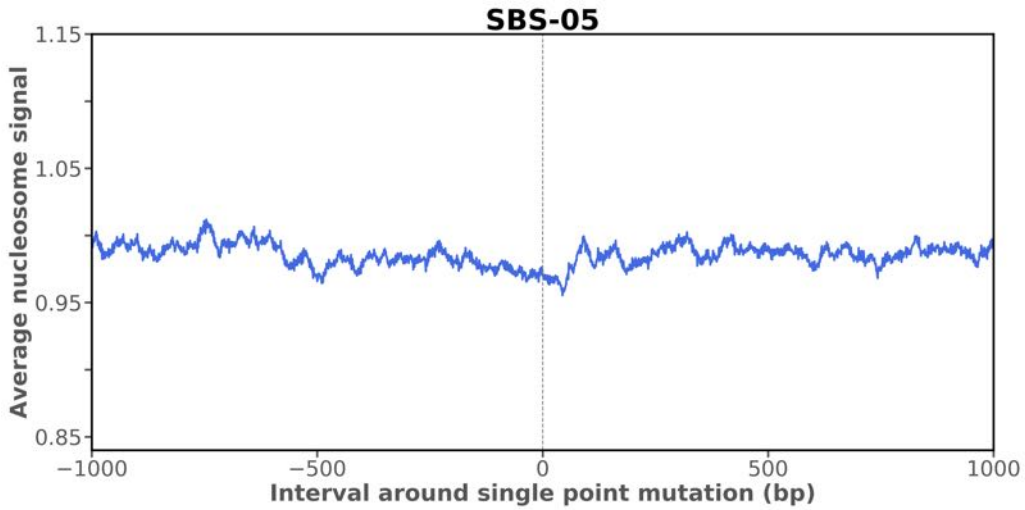
Figure S10: Mutational signatures for primary and metastatic tumors. The contribution of known mutational signatures in tumor pRCC2_1824_13 and rSRC_1697_10.

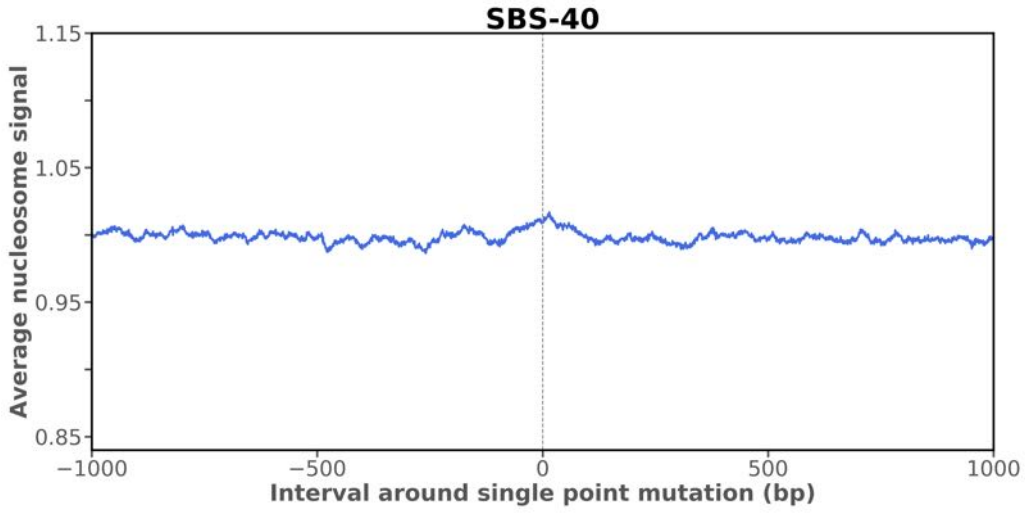
1.00
0.75
0.50
0.25
0.00
Signature contribution
pRCC2_1824_13, Metastasis
pRCC2_1824_13, Primary
rSRC_1697_10, Metastasis
rSRC_1697_10, Primary
Signature Subs-01
Signature Subs-02
Signature Subs-05
Signature Subs-08
Signature Subs-13
Signature Subs-40

Figure S11: Topography of mutational signatures. (a) The distributions of nucleosome density signals (y-axes) are shown in a 2 kb window centered on each mutation (position 0 on the x-axes), for each signature. The averaged signal was calculated as the total amount of signal observed at each point divided by total number of mutations contributing to that signal. **(b)** Distribution of the base substitution signatures across the cell cycle. Replication domains were identified by using conservatively defined transition zones in DNA replication time data. Data were separated into deciles, with each segment containing exactly 10% of the observed replication time signal. Normalized mutation density per decile is presented for early (left) to late (right) replication domains. **(c)** Transcription and replication strand-bias for mutational signatures and types of single-point somatic mutations.

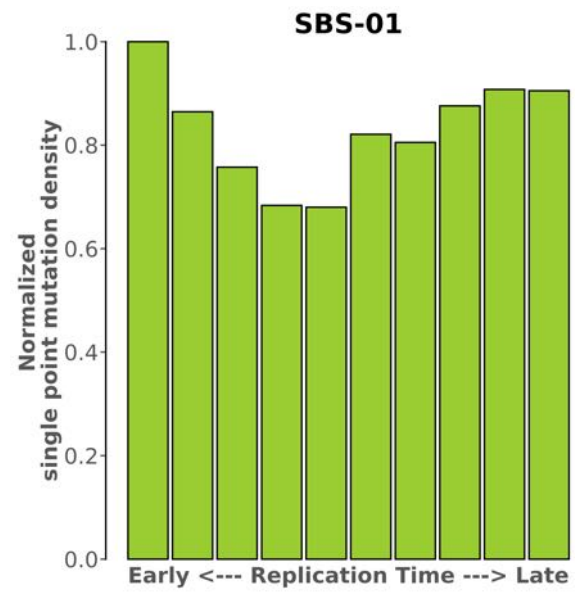
a

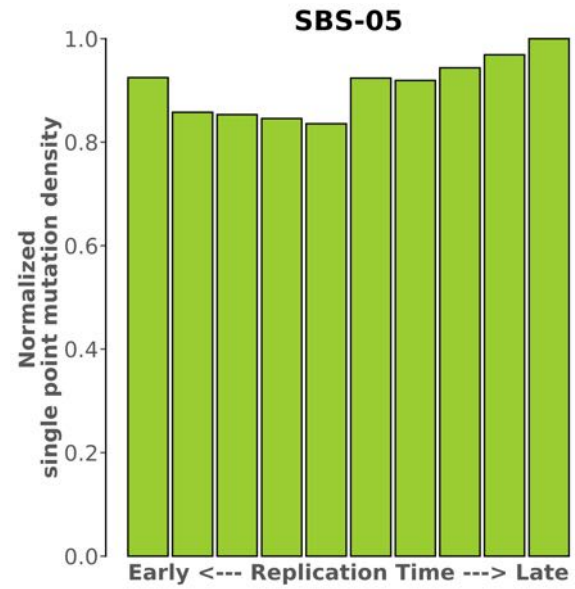


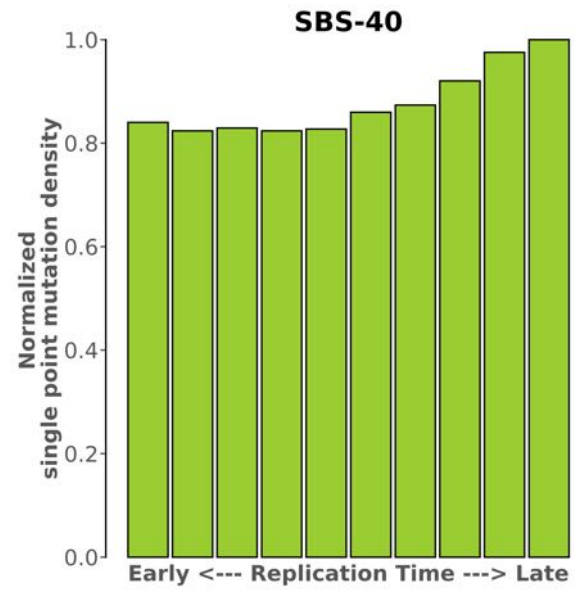




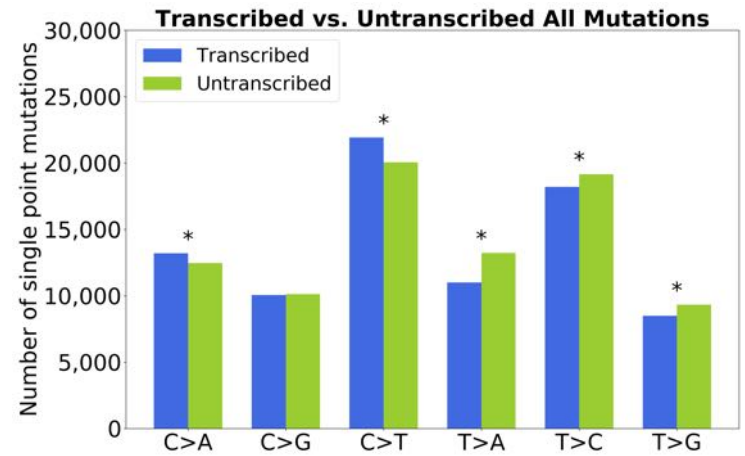
b

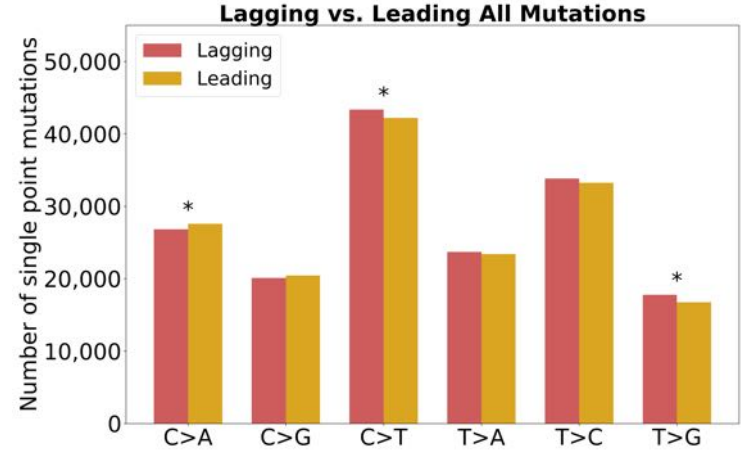


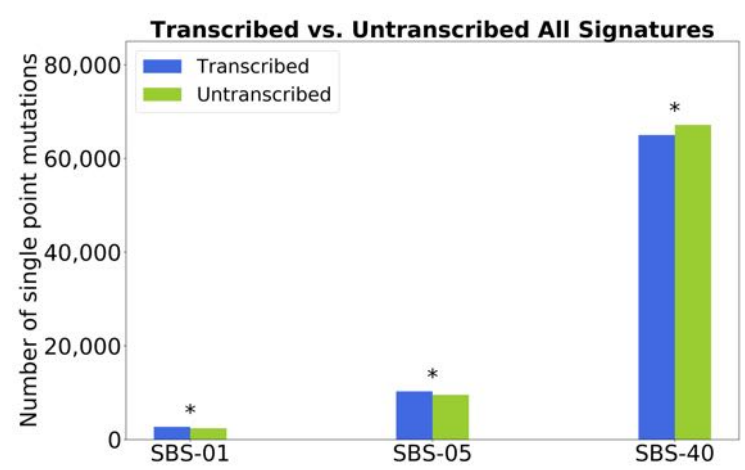


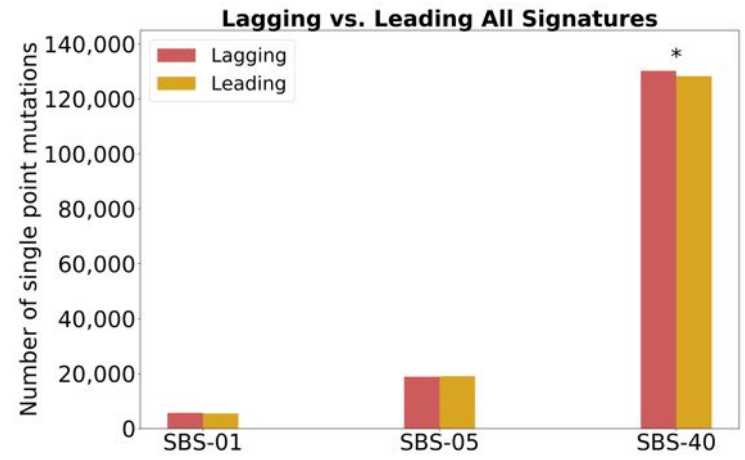


c



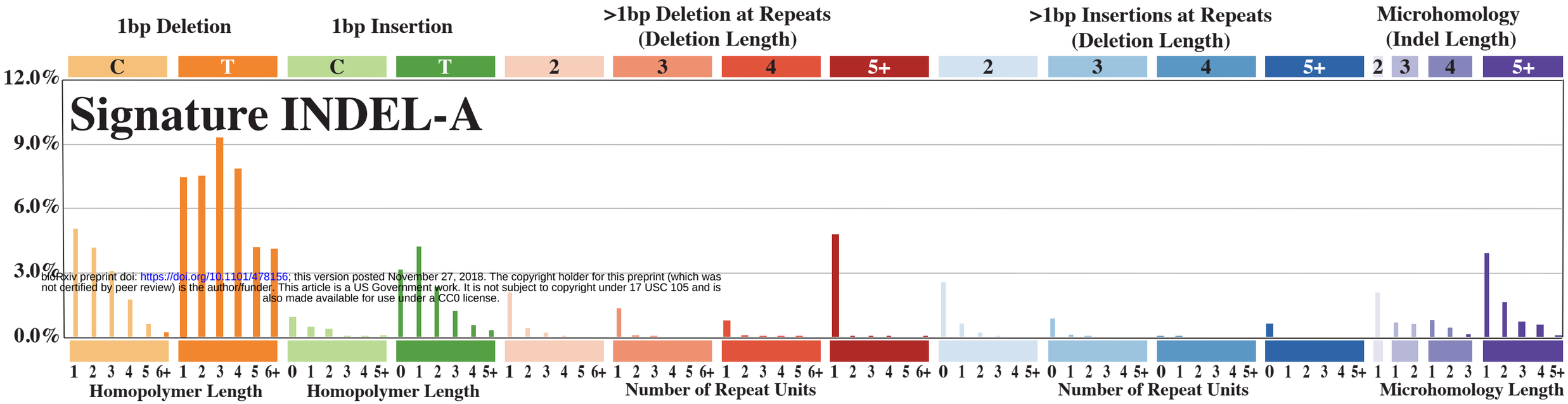


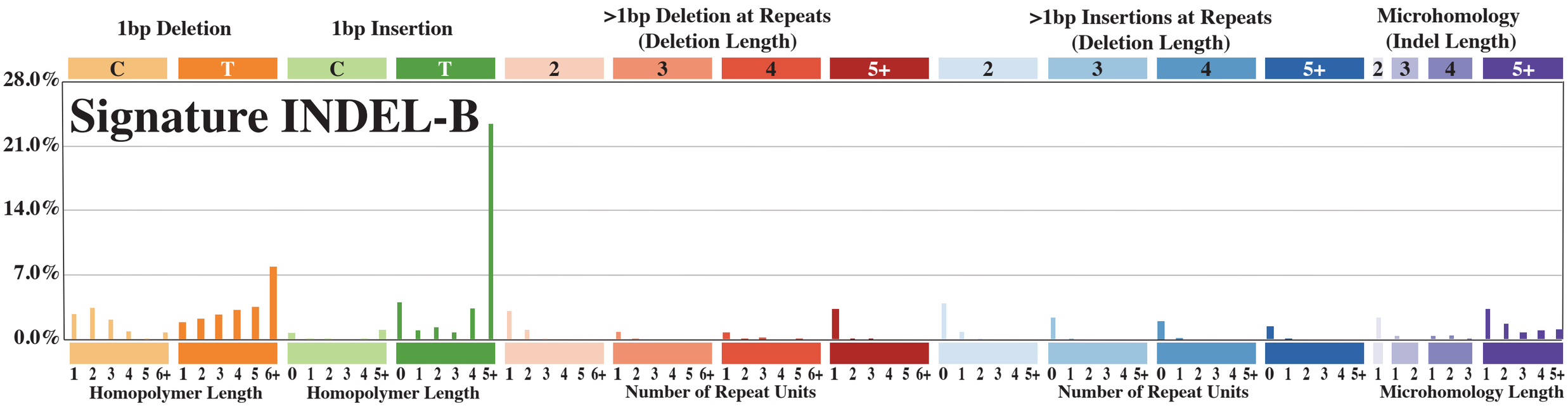




* shows FDR BH adjusted p-value <= 0.05

Figure S12: De novo extraction of indel mutational signatures. Profile of the three *de novo* extracted signatures of small insertions and deletions (indels) is provided. Indels were classified as deletions or insertions and, when of a single base, as C or T and according to the length of the mononucleotide repeat tract in which they occurred. Longer indels were classified as occurring at repeats or with overlapping microhomology at deletion boundaries, and according to the size of indel, repeat, and microhomology.





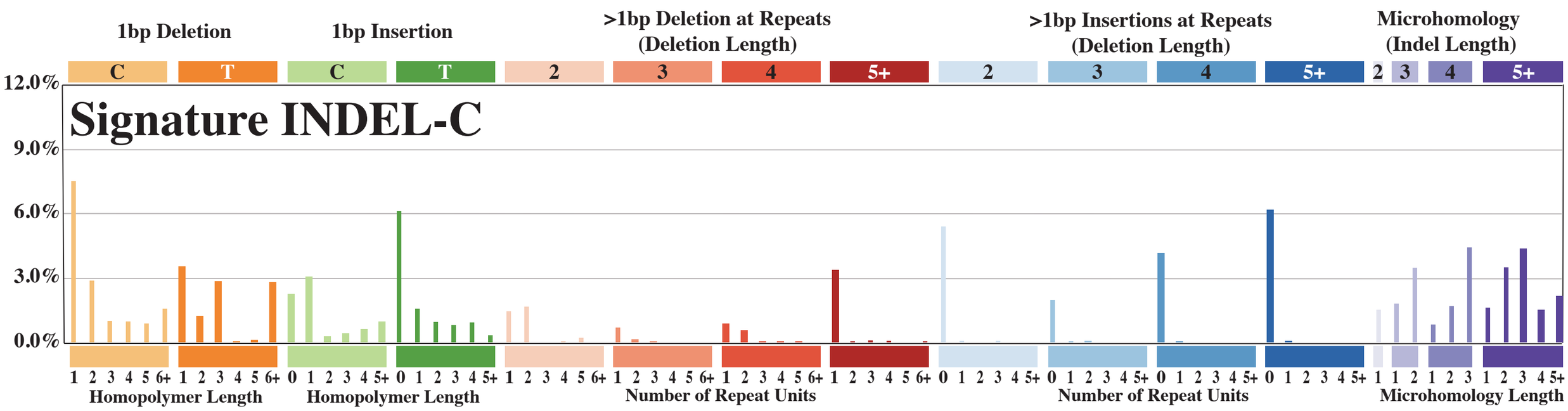


Figure S13: Telomere length (TL) for each sample. TL is based on the abundance of telomere motif sequence $(TTAGGG/CCCTAA)_4$.

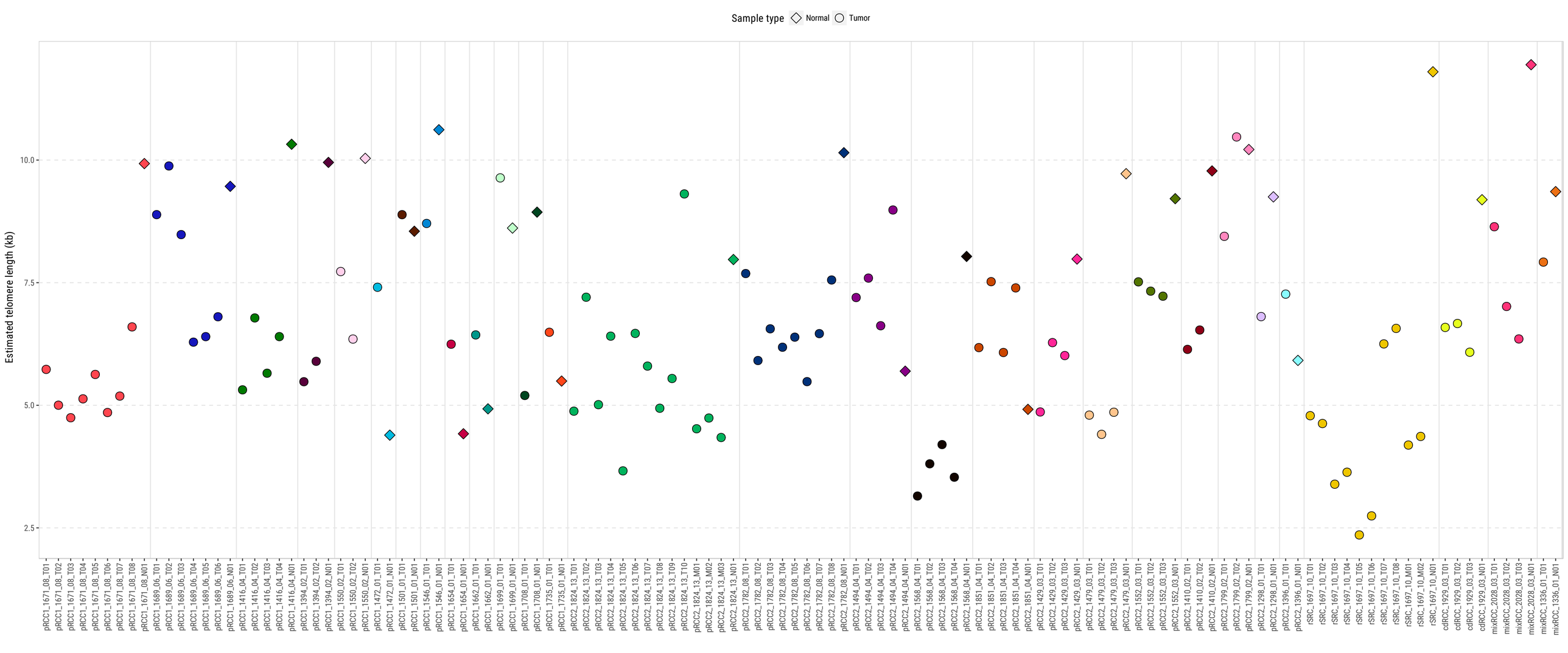

Sample type
Normal
Tumor
Estimated telomere length (kb)
10.0
7.5
5.0
2.5
pRCC1_1671_08_T01
pRCC1_1671_08_T02
pRCC1_1671_08_T03
pRCC1_1671_08_T04
pRCC1_1671_08_T05
pRCC1_1671_08_T06
pRCC1_1671_08_T07
pRCC1_1671_08_T08
pRCC1_1671_08_N01
pRCC1_1689_06_T01
pRCC1_1689_06_T02
pRCC1_1689_06_T03
pRCC1_1689_06_T04
pRCC1_1689_06_T05
pRCC1_1689_06_T06
pRCC1_1689_06_N01
pRCC1_1416_04_T01
pRCC1_1416_04_T02
pRCC1_1416_04_T03
pRCC1_1416_04_T04
pRCC1_1416_04_N01
pRCC1_1394_02_T01
pRCC1_1394_02_T02
pRCC1_1394_02_N01
pRCC1_1550_02_T01
pRCC1_1550_02_T02
pRCC1_1550_02_N01
pRCC1_1472_01_T01
pRCC1_1472_01_N01
pRCC1_1501_01_T01
pRCC1_1501_01_N01
pRCC1_1546_01_T01
pRCC1_1546_01_N01
pRCC1_1654_01_T01
pRCC1_1654_01_N01
pRCC1_1662_01_T01
pRCC1_1662_01_N01
pRCC1_1699_01_T01
pRCC1_1699_01_N01
pRCC1_1708_01_T01
pRCC1_1708_01_N01
pRCC1_1735_01_T01
pRCC1_1735_01_N01
pRCC2_1824_13_T01
pRCC2_1824_13_T02
pRCC2_1824_13_T03
pRCC2_1824_13_T04
pRCC2_1824_13_T05
pRCC2_1824_13_T06
pRCC2_1824_13_T07
pRCC2_1824_13_T08
pRCC2_1824_13_T09
pRCC2_1824_13_T10
pRCC2_1824_13_M01
pRCC2_1824_13_M02
pRCC2_1824_13_M03
pRCC2_1824_13_N01
pRCC2_1782_08_T01
pRCC2_1782_08_T02
pRCC2_1782_08_T03
pRCC2_1782_08_T04
pRCC2_1782_08_T05
pRCC2_1782_08_T06
pRCC2_1782_08_T07
pRCC2_1782_08_T08
pRCC2_1782_08_N01
pRCC2_1494_04_T01
pRCC2_1494_04_T02
pRCC2_1494_04_T03
pRCC2_1494_04_T04
pRCC2_1494_04_N01
pRCC2_1568_04_T01
pRCC2_1568_04_T02
pRCC2_1568_04_T03
pRCC2_1568_04_T04
pRCC2_1568_04_N01
pRCC2_1851_04_T01
pRCC2_1851_04_T02
pRCC2_1851_04_T03
pRCC2_1851_04_T04
pRCC2_1851_04_N01
pRCC2_1429_03_T01
pRCC2_1429_03_T02
pRCC2_1429_03_T03
pRCC2_1429_03_N01
pRCC2_1479_03_T01
pRCC2_1479_03_T02
pRCC2_1479_03_T03
pRCC2_1479_03_N01
pRCC2_1552_03_T01
pRCC2_1552_03_T02
pRCC2_1552_03_T03
pRCC2_1552_03_N01
pRCC2_1410_02_T01
pRCC2_1410_02_T02
pRCC2_1410_02_N01
pRCC2_1799_02_T01
pRCC2_1799_02_T02
pRCC2_1799_02_N01
pRCC2_1298_01_T01
pRCC2_1298_01_N01
pRCC2_1396_01_T01
pRCC2_1396_01_N01
rSRC_1697_10_T01
rSRC_1697_10_T02
rSRC_1697_10_T03
rSRC_1697_10_T04
rSRC_1697_10_T05
rSRC_1697_10_T06
rSRC_1697_10_T07
rSRC_1697_10_T08
rSRC_1697_10_M01
rSRC_1697_10_M02
rSRC_1697_10_N01
cdRCC_1929_03_T01
cdRCC_1929_03_T02
cdRCC_1929_03_T03
cdRCC_1929_03_N01
mixRCC_2028_03_T01
mixRCC_2028_03_T02
mixRCC_2028_03_T03
mixRCC_2028_03_N01
mixRCC_1336_01_T01
mixRCC_1336_01_N01

Figure S14: Percentage of somatic copy number alterations (SCNA) across cytoband regions. HOMD: homozygous deletion; DLOH: hemizygous deletion loss of heterozygosity; NLOH: copy neutral loss of heterozygosity; ALOH: amplified loss of heterozygosity; ASCNA: allele-specific copy number amplification; BCNA: balanced copy number amplification.

Figure S15: Genome-wide profiles of SCNA events of each sample. Samples are labeled by copy number type, histology subgroup, tissue type and purity. HOMD: homozygous deletion; DLOH: hemizygous deletion loss of heterozygosity; HET: diploid heterozygous; NLOH: copy neutral loss of heterozygosity; ALOH: amplified loss of heterozygosity; ASCNA: allele-specific copy number amplification; BCNA: balanced copy number amplification.

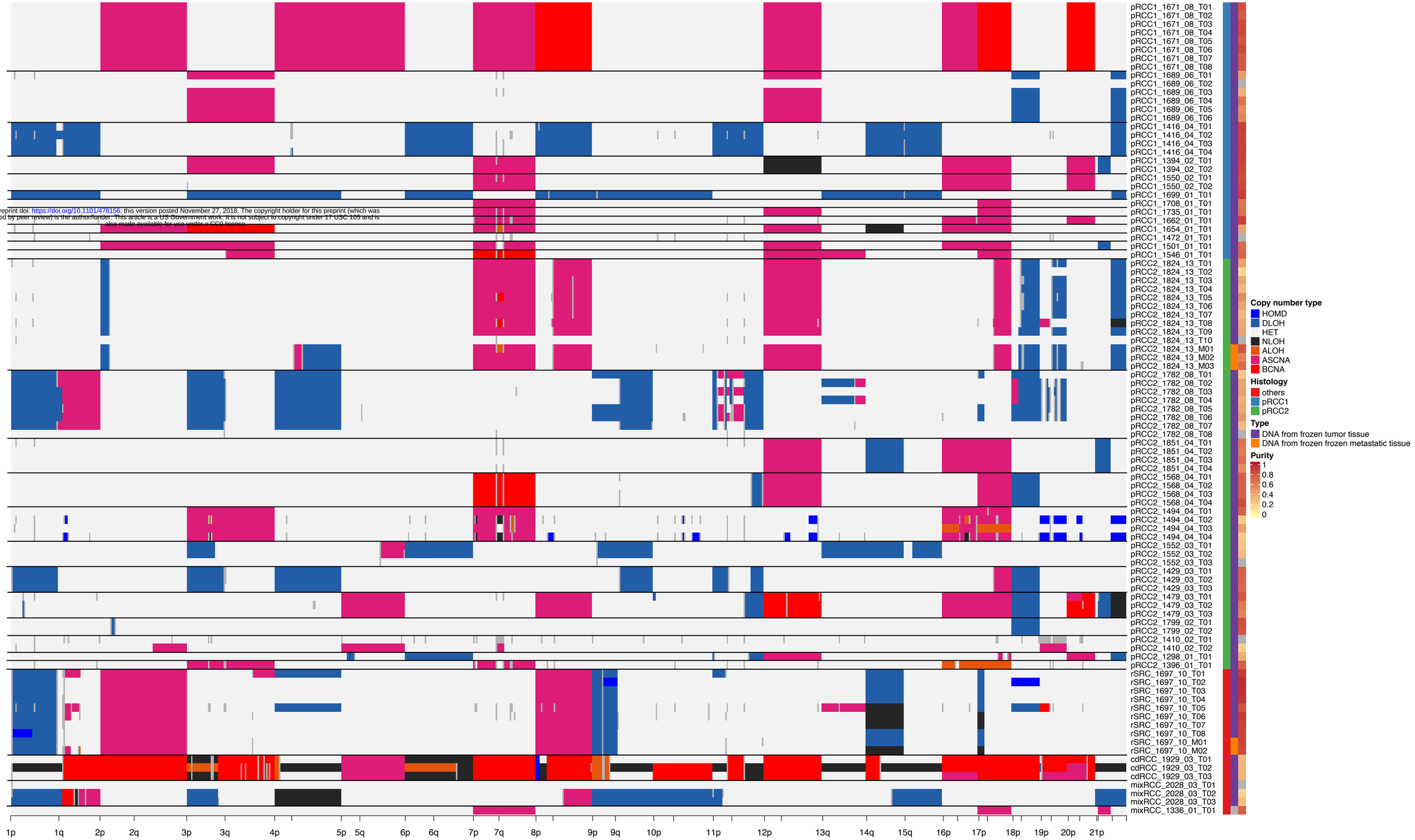


Figure S16: Cluster Dendrogram of SCNA profiles. The similar SCNA profiles were clustered hierarchically.

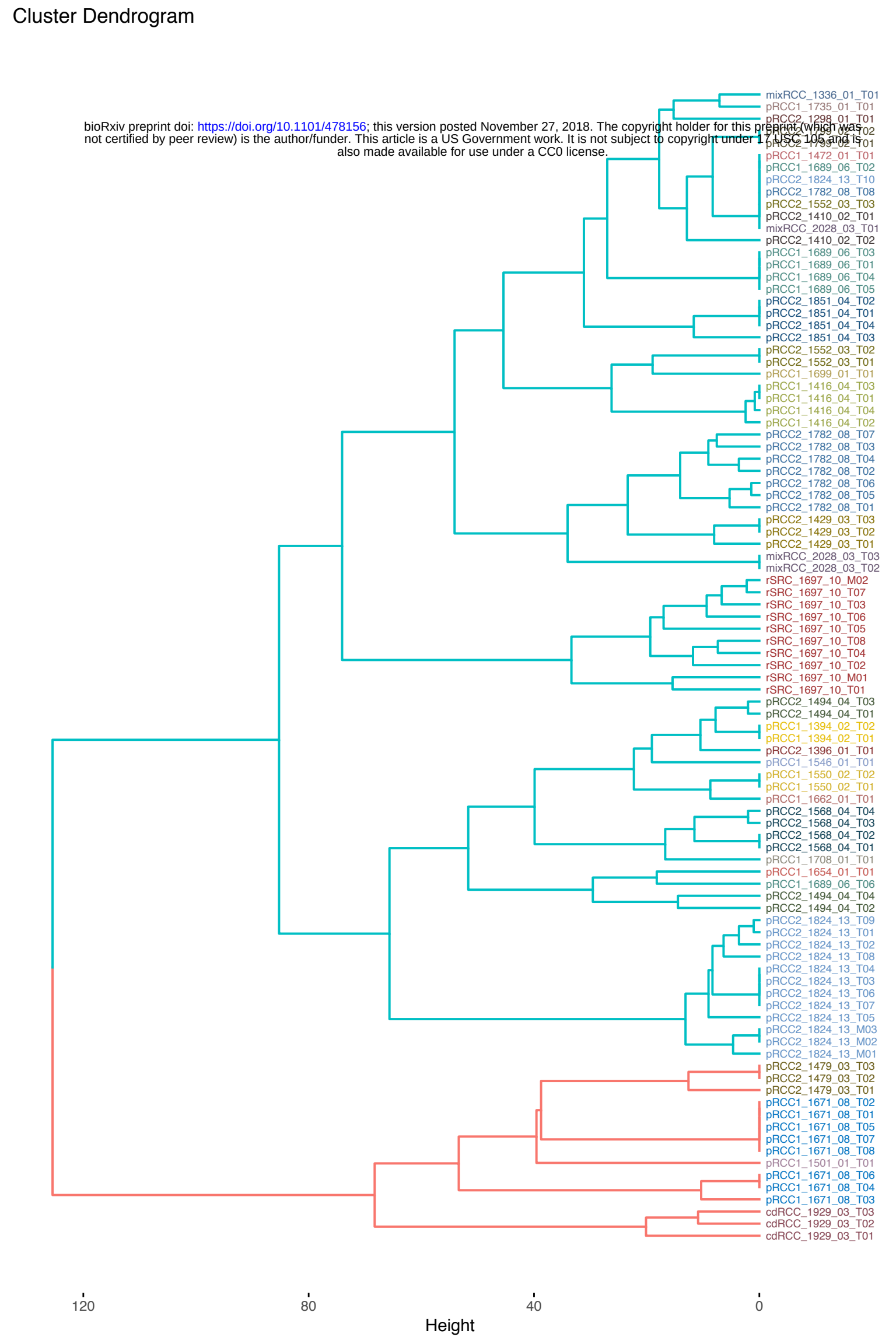


Figure S17: Focal SCNA events for each sample. DLOH: hemizygous deletion loss of heterozygosity; HET: diploid heterozygous; NLOH: copy neutral loss of heterozygosity; ALOH: amplified loss of heterozygosity; ASCNA: allele-specific copy number amplification; BCNA: balanced copy number amplification.

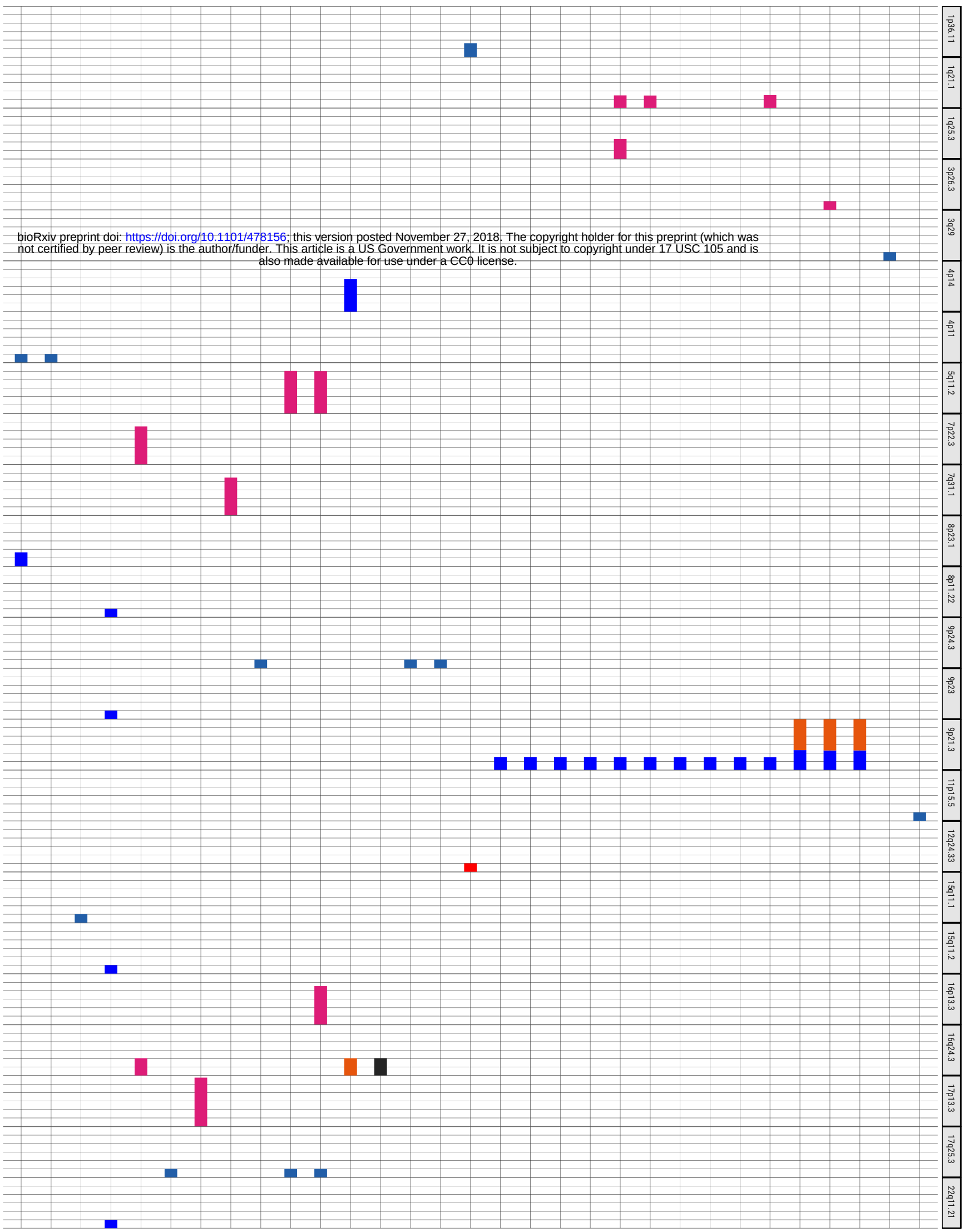


Figure S18: rSRC copy number profile for chromosome 9. The clonal focal homozygous deletion of *CDKN2A* is located at 9p21.3. For each sample, the blue dots are observed values and red lined are estimated ones; The first two panels show the profiles of logR and logOR over chromosomes; The last panel indicates estimated total copy numbers and minor copy number over chromosomes with cancer cell fraction (cf-em) in bottom, estimated by the expectation–maximization (em) algorithm.

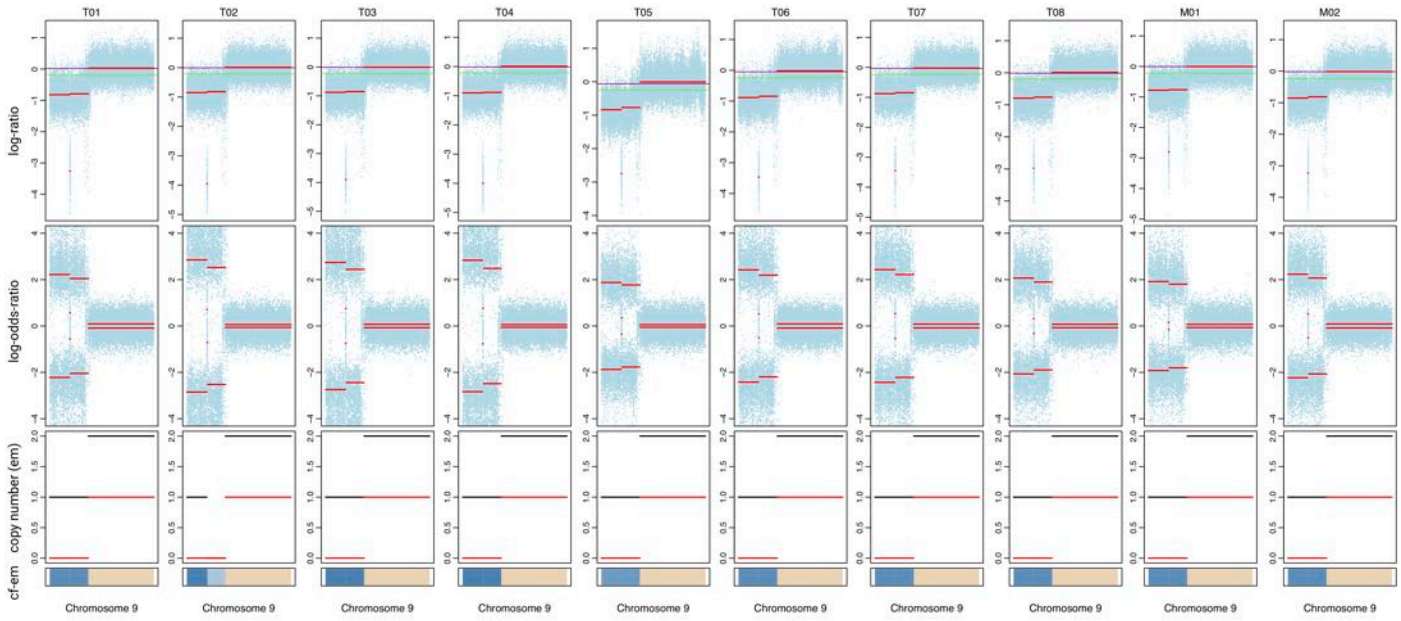
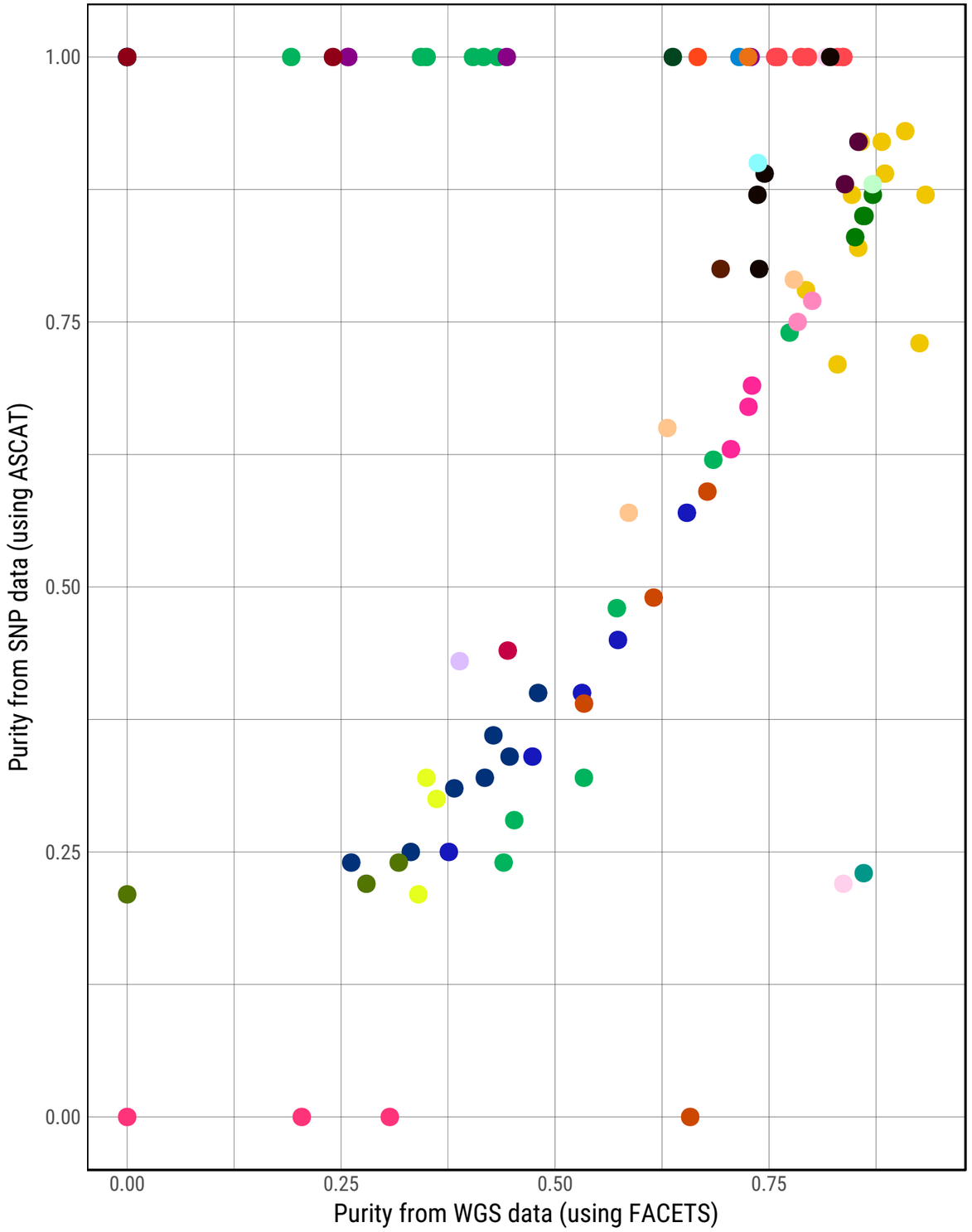


Figure S19: Concordance of purity and ploidy across platforms. The scatterplots of purity (left panel) and ploidy (right panel) are estimated based on WGS and SNP genotyping.


1.00
0.75
0.50
0.25
0.00
Purity from SNP data (using ASCAT)
0.00
0.25
0.50
0.75
Purity from WGS data (using FACETS)

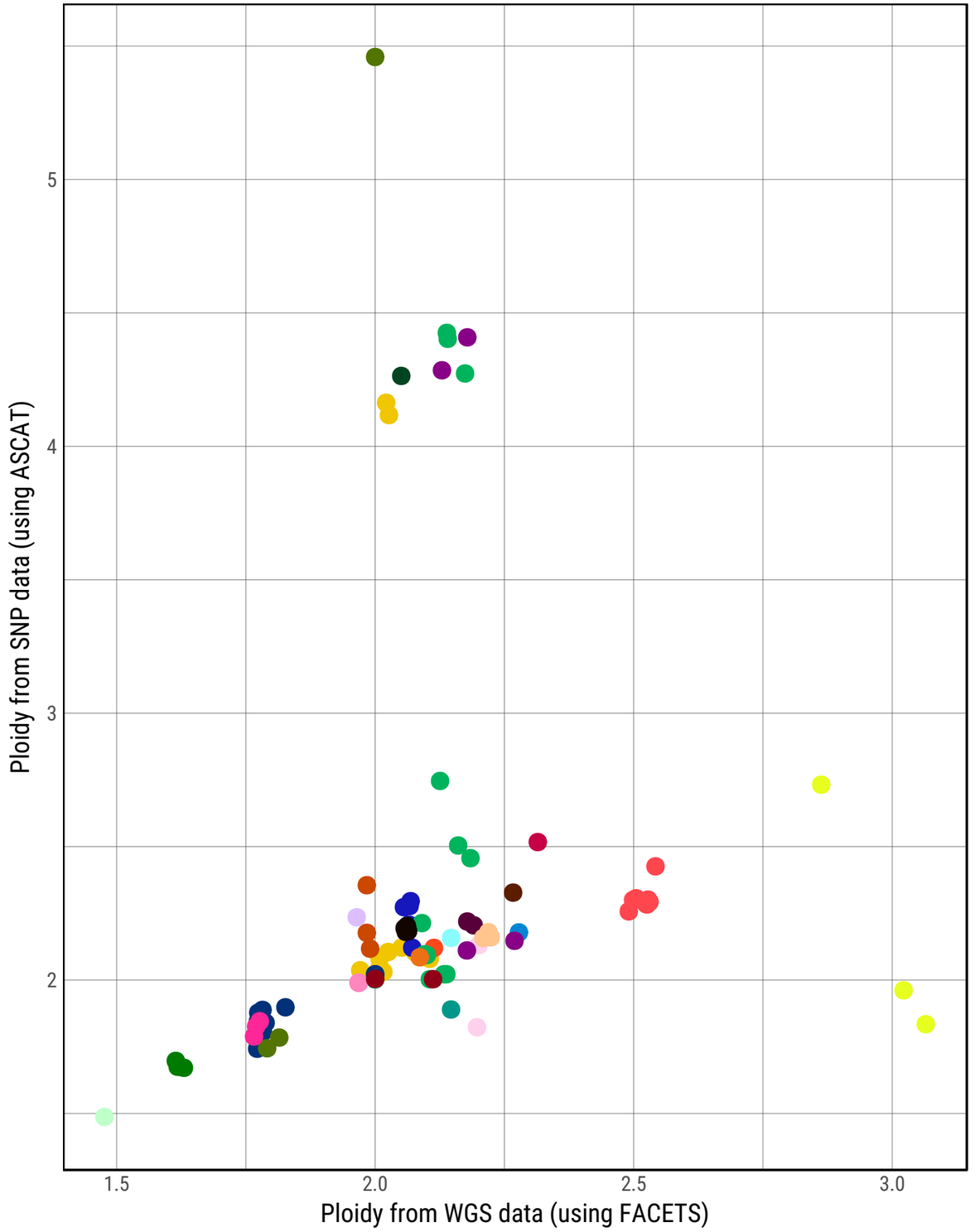

5
4
3
2
Ploidy from SNP data (using ASCAT)
1.5
2.0
2.5
3.0
Ploidy from WGS data (using FACETS)

Figure S20: Validation of gene fusions. PCR to detect putative fusions were performed using indicated primer sets and samples (Panel a). Individual amplicons indicated by arrows in Panel b were gel purified and sequenced by Sanger sequencing. Numbers at arrows indicate approximate molecular weights. Fusions and breakpoints detected by WGS between EWSR/PATZ1 and MALAT1/TFEB were validated by this method. The MET-MET product amplified by primer set B is the normally annotated transcript without the putative indel.

(a)

Forward Primer		Reverse Primer			Reference
METex1F	GTTCCTGGGCACCGAAAG	METex4R	CACATTTGTCGTGGCACCAG		novel
METex2F-2	TGCCATGTGTGCATTCCCTA	METex4R			novel
STRN-ALKF	CCACAAGTTGAAATACGGGACAGAA	STRN-ALKR	TCAGCTTGTACTCAGGGCTCT		http://www.pnas.org/content/111/11/4233.short
ALKex19F	TGATCCTCTCTGTGGTGACCT	STRNex4R	GTGGCTGCACTTCTGTTTCA		http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0087170#pone.0087170.s002
EWS-720F	ATGGTCAACAAAGCAGCTATGGG	ZSG-S5R	GTCAGGAACCGAATGGGACGA		Oncogene 2000 19 3799-3804 Mastrangelo et al
EWS-720F		ZSG-A1R	GCAGGGCACCTTGTCTTCATG		Oncogene 2000 19 3799-3804 Mastrangelo et al
EWS-720F		PATZ1ex1R	GAGCTGGAGATGCACACTATCAG		novel
AlphaRT-1f	TAA CGC ATT TAC TAA ACG CAG ACG	TFEBexon2r	AAC CCT ATG CGT GAC GCC ATG GTG G		Ian J. Davis, 6051–6056, doi: 10.1073/pnas.0931430100

(b)

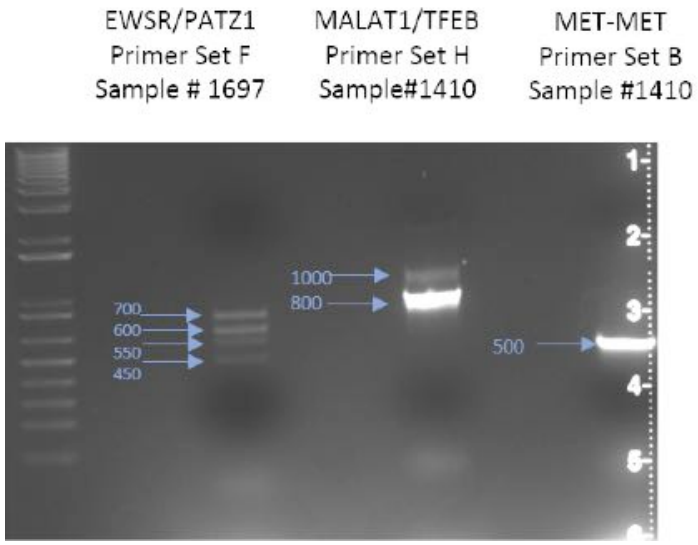


Figure S21: SV hotspots in rSRC_1697_10. Each arch links the breakpoints of SV fragments with colors denoting different SV types.

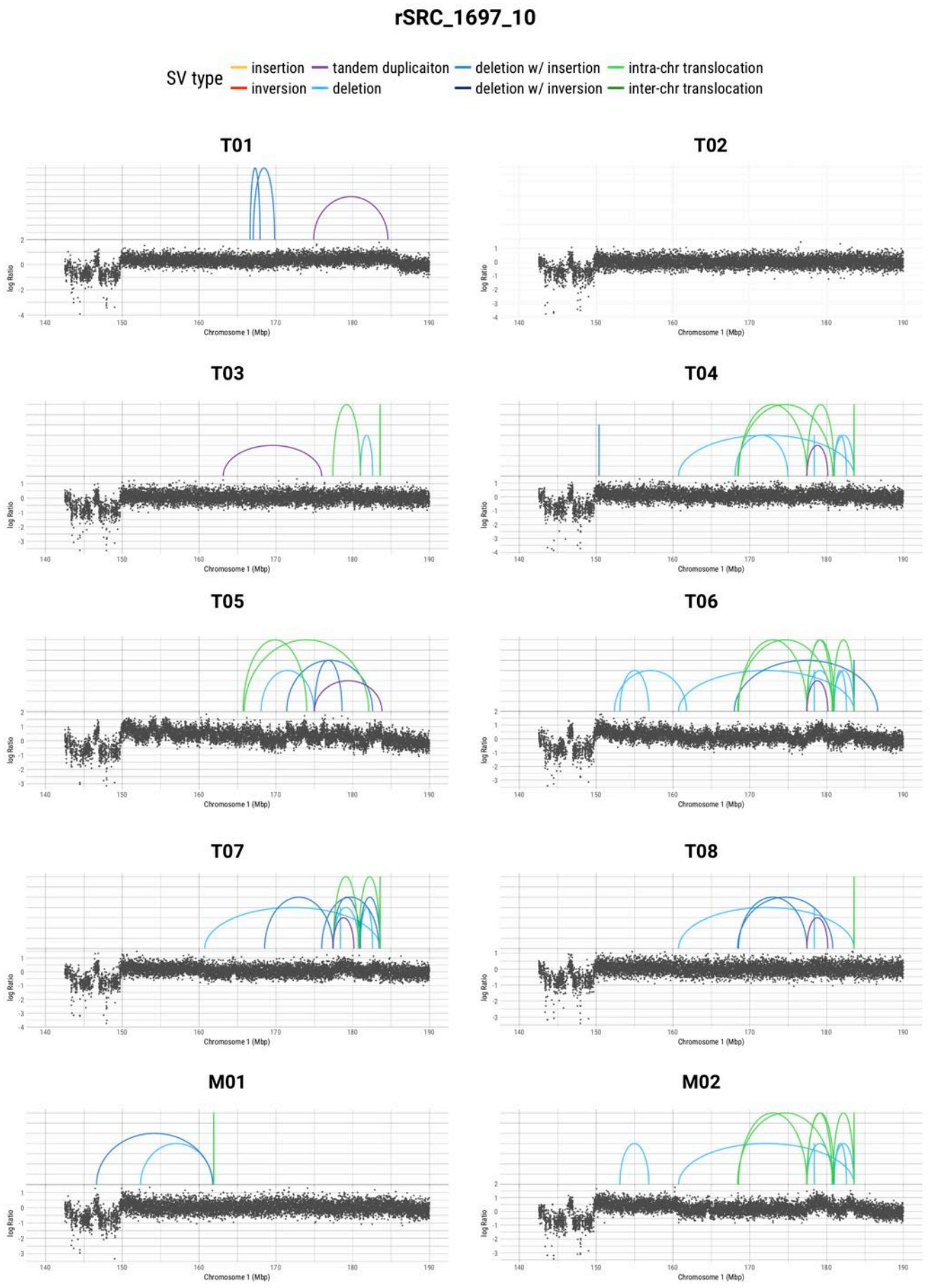


Figure S22: LINE-1 retrotransposition events in pRCC2_1782_08. Genes involved in the retrotransposons insertions are indicated.

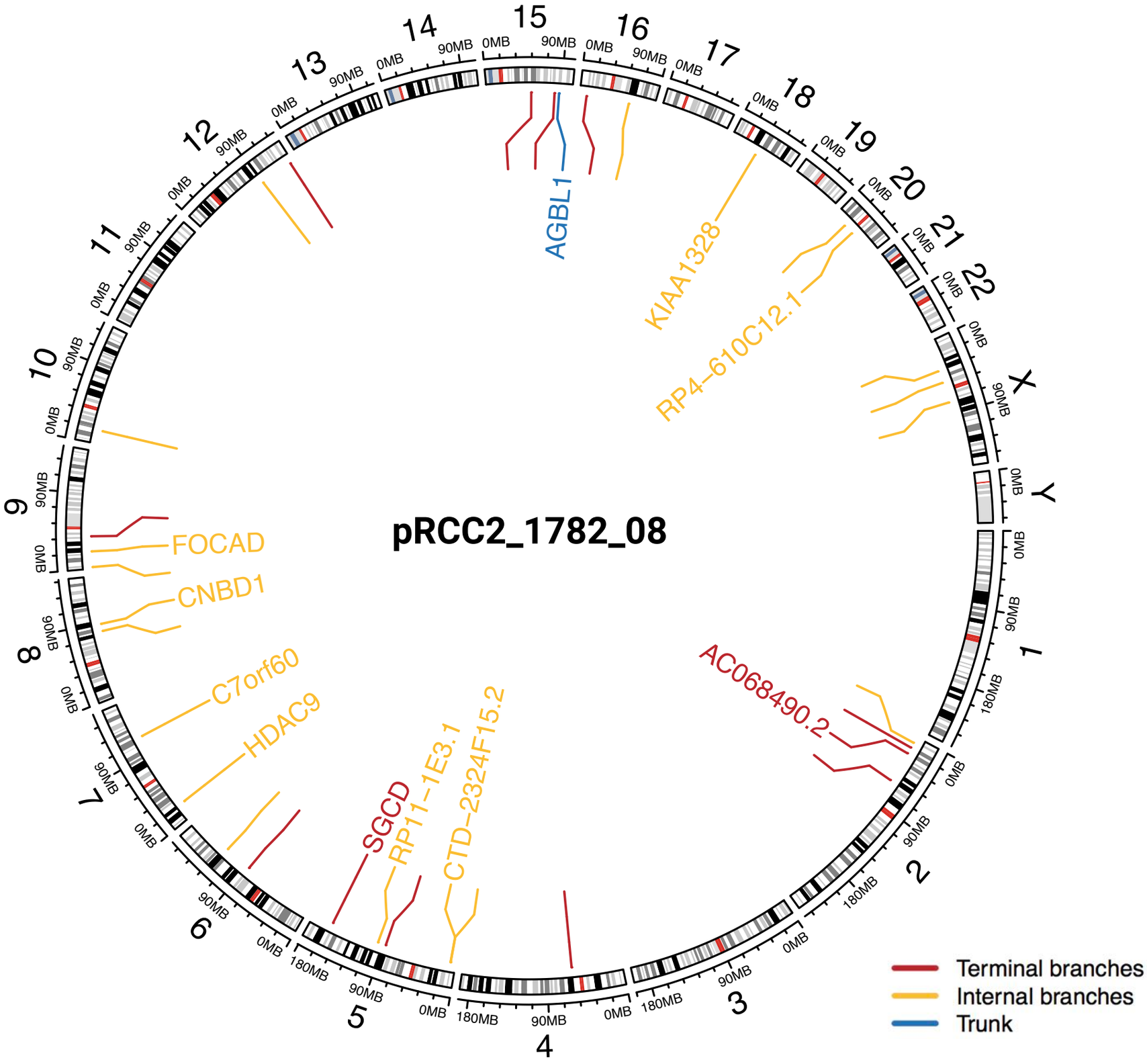
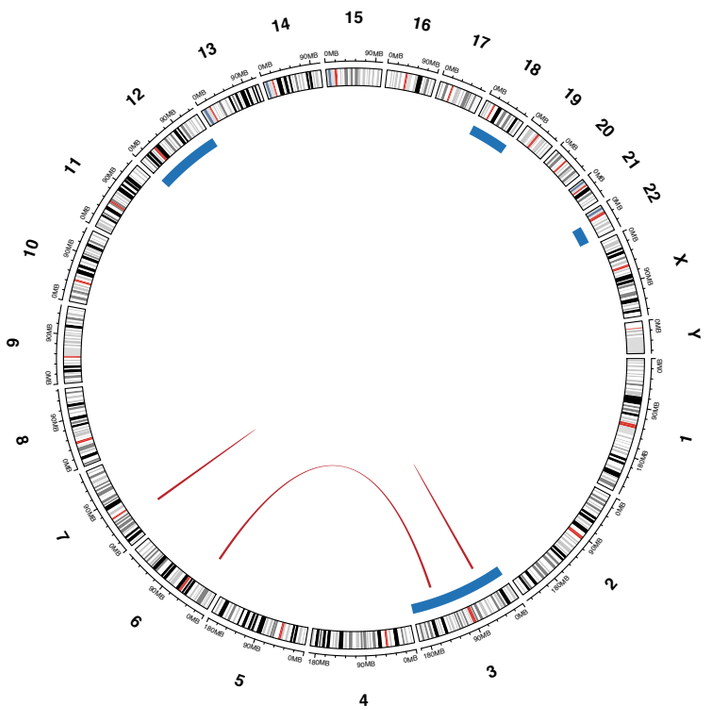
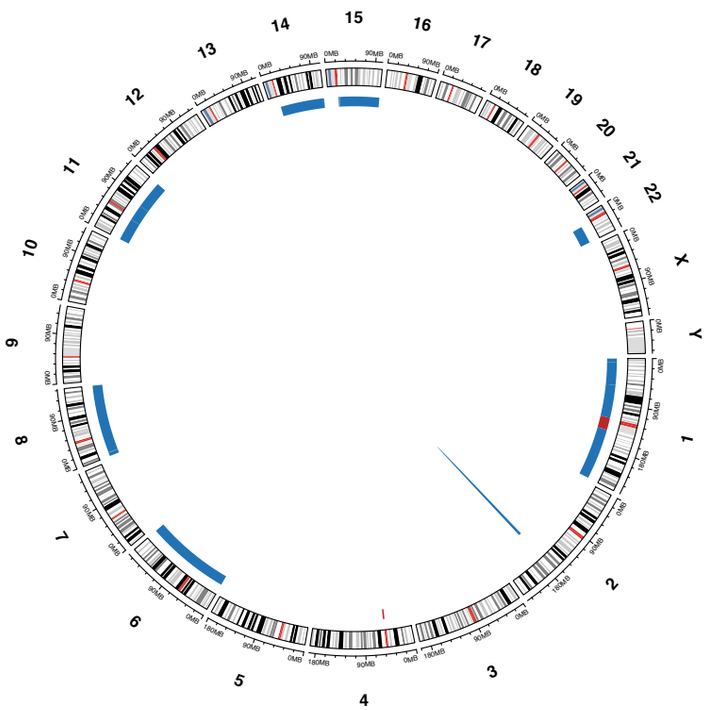


Figure S23: Circos plots of SV (linked by arches) and SCNA (in the inner circle) for each tumor. Branch and trunk events are notated by different colors.

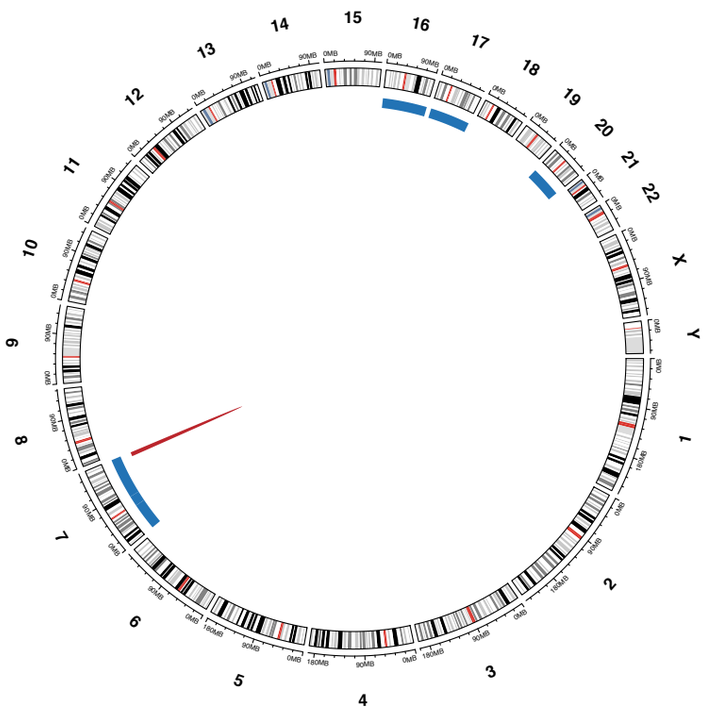
(a) pRCC1_1689_06



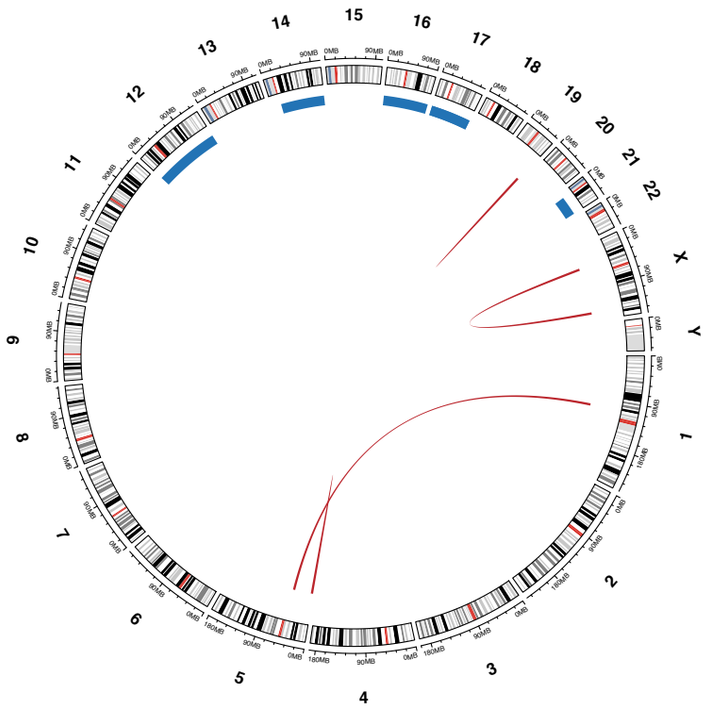
(b) pRCC1_1416_04



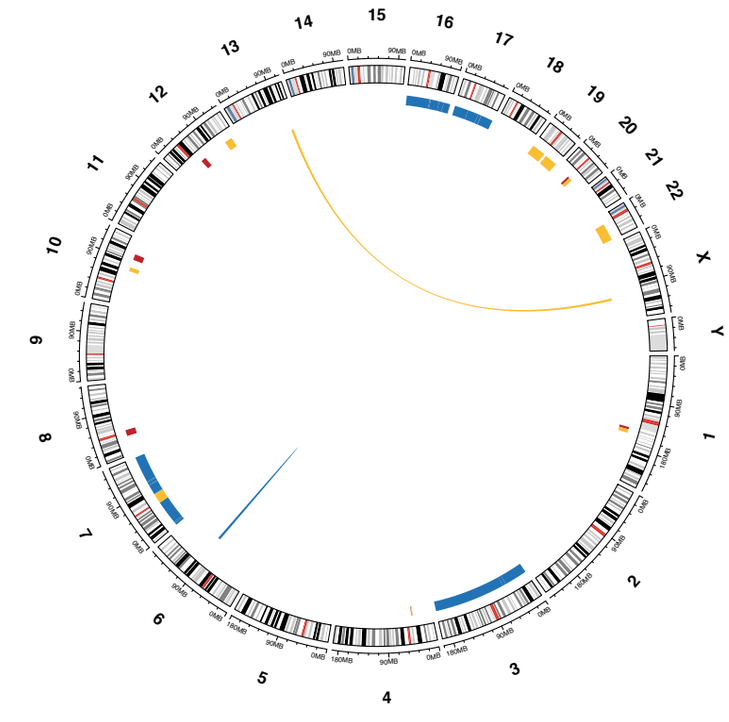
(c) pRCC1_1550_02



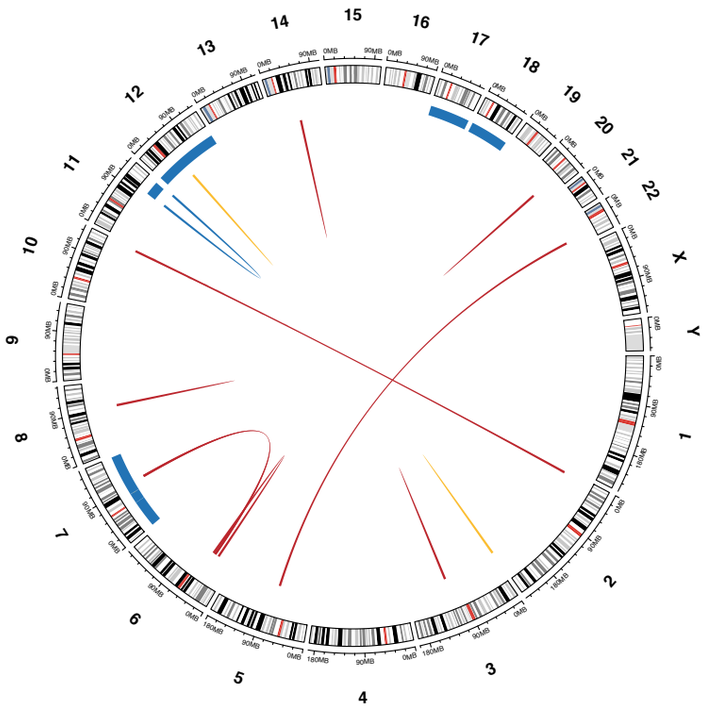
(d) pRCC2_1851_04



(e) pRCC2_1494_04

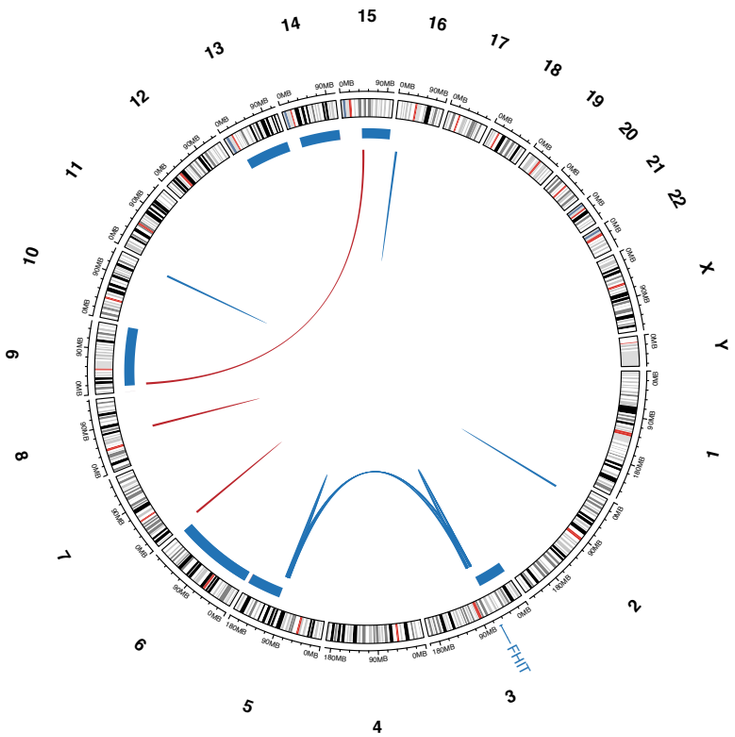


(f) pRCC2_1568_04

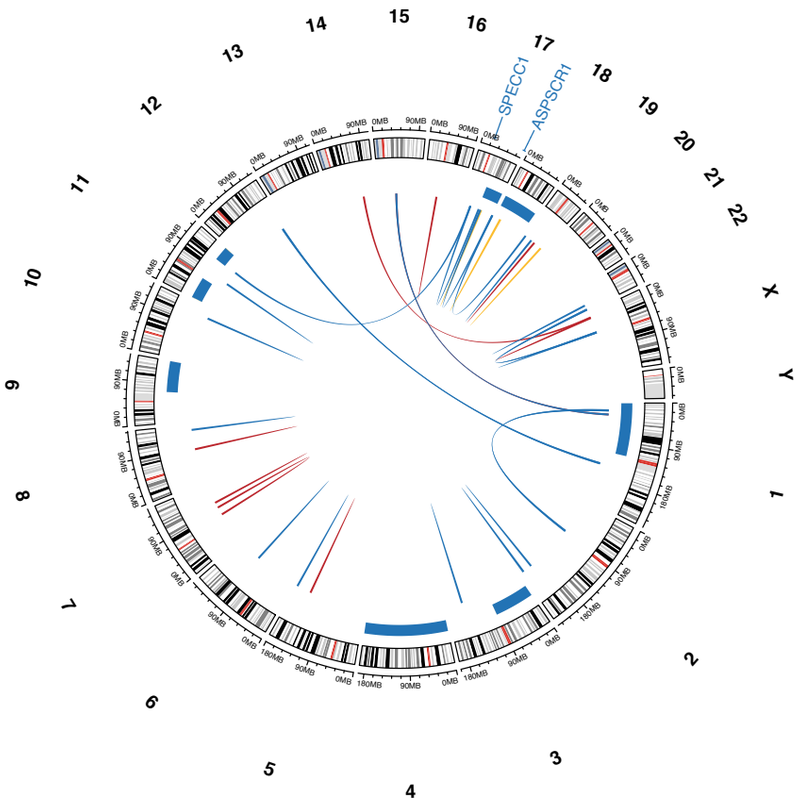


Terminal branches
Internal branches
Trunk

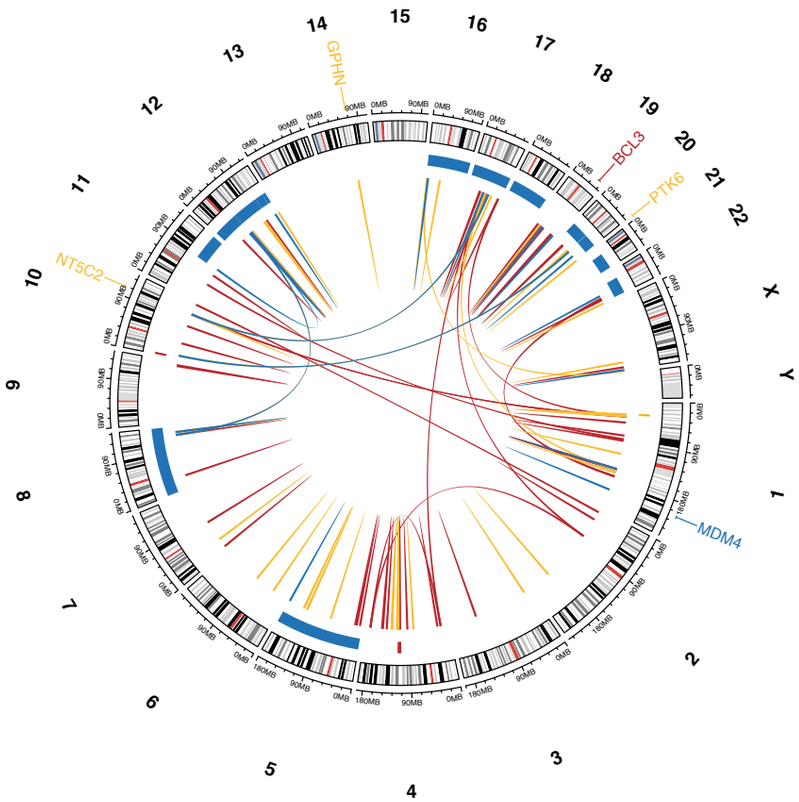
(g) pRCC2_1552_03



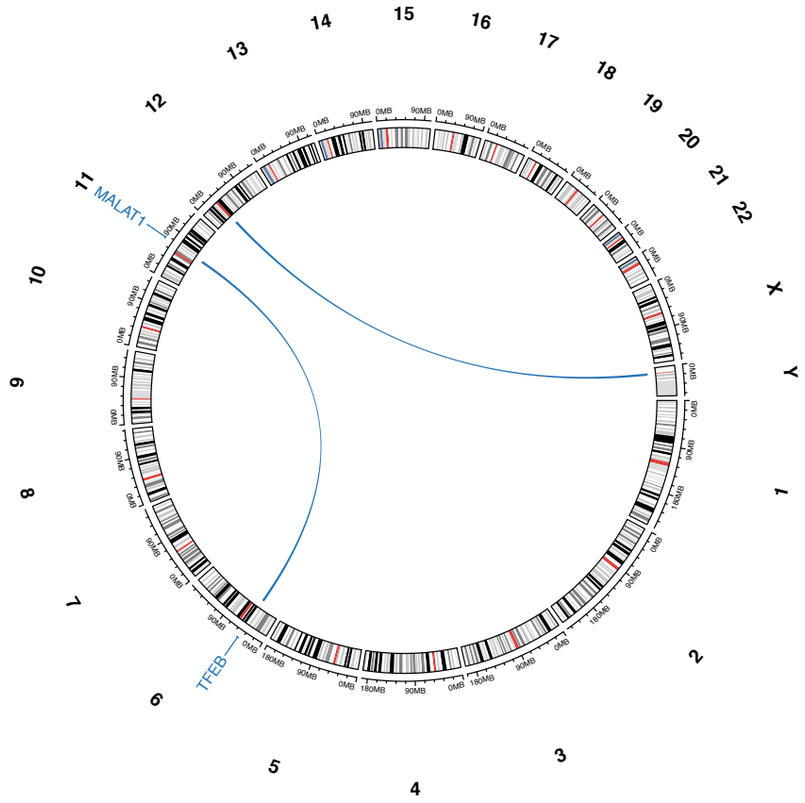
(h) pRCC2_1429_03



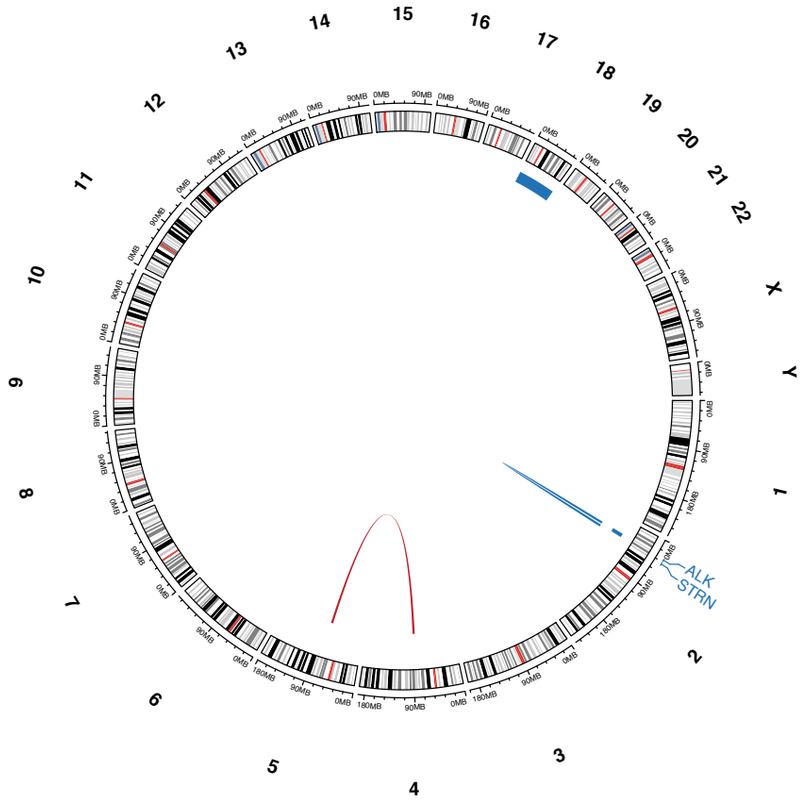
(i) pRCC2_1479_03

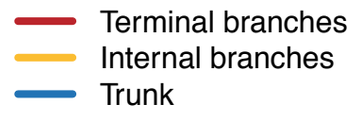


(j) pRCC2_1410_02

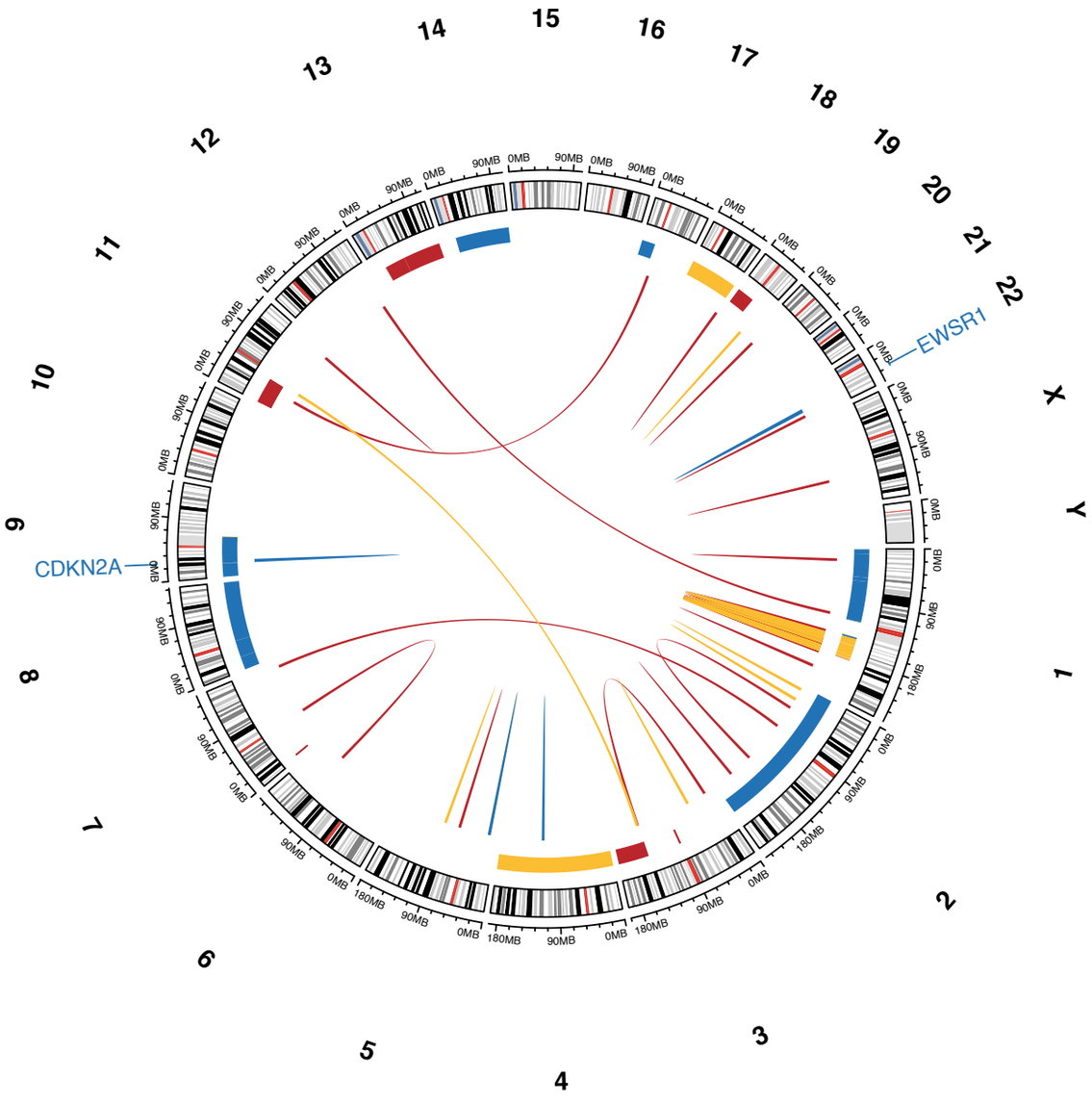


(k) pRCC2_1799_02

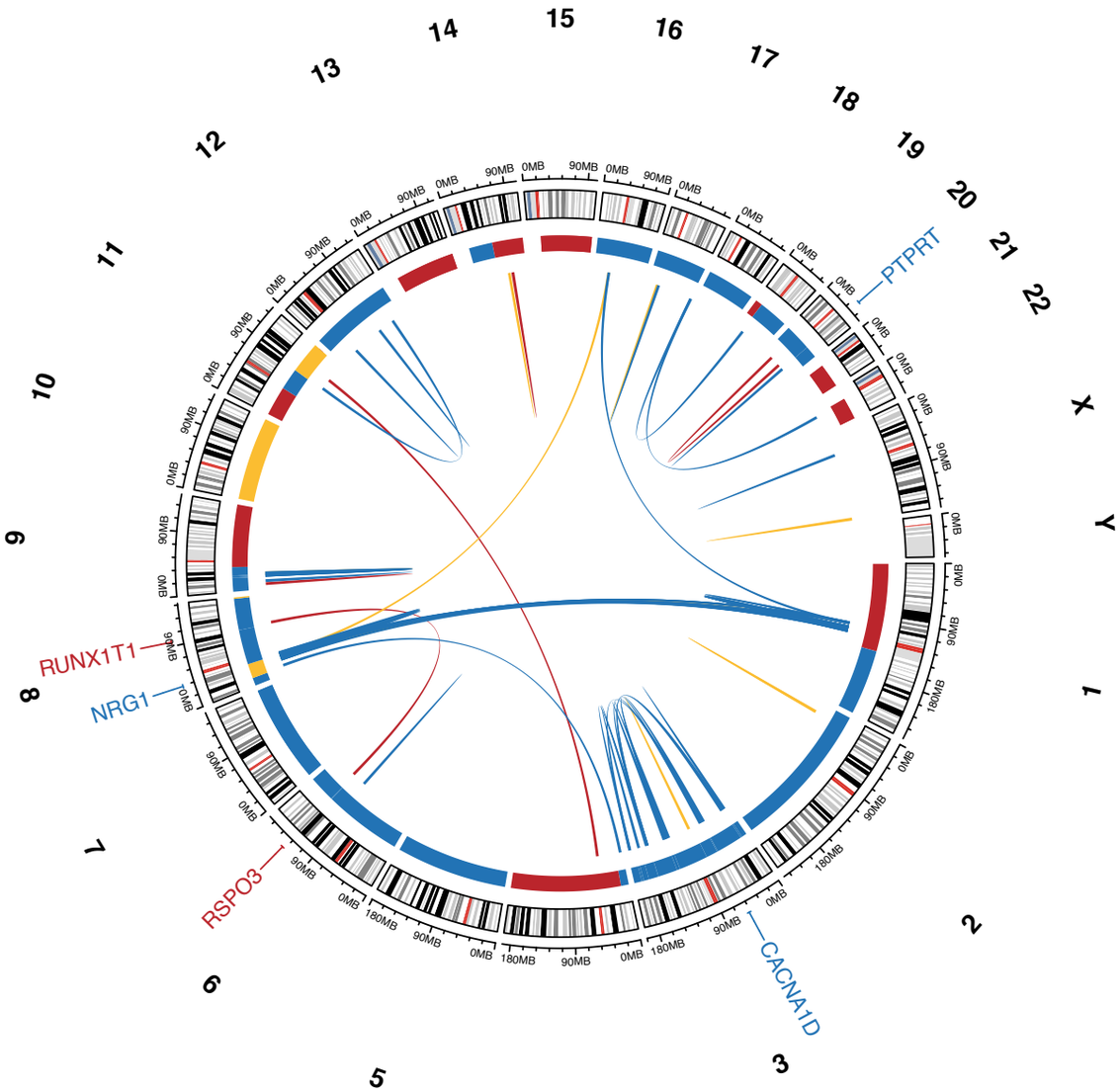




(l) rSRC_1697_10



(m) cdRCC_1929_03



Terminal branches
Internal branches
Trunk

Figure S24: Validation of the WGS-detected SV events. A PCR-based sequencing methodology was used (AmpliSeq) to validate the whole-genome sequencing based SV events. Read counts were colored based on their magnitude. Bottom blue bar indicates SVs shared by all regions; orange bar SVs shared by part of regions; and red bar SVs found in one region only.


PAP1_1671_08_T01
PAP1_1671_08_T02
PAP1_1671_08_T03
PAP1_1671_08_T04
PAP1_1671_08_T05
PAP1_1671_08_T06
PAP1_1671_08_T07
PAP1_1671_08_T08
PAP1_1689_06_T01
PAP1_1689_06_T03
PAP1_1689_06_T04
PAP1_1689_06_T05
PAP1_1689_06_T06
PAP1_1416_04_T01
PAP1_1416_04_T02
PAP1_1416_04_T03
PAP1_1416_04_T04
PAP2_1824_13_M01
PAP2_1824_13_M02
PAP2_1824_13_M03
PAP2_1824_13_T01
PAP2_1824_13_T02
PAP2_1824_13_T03
PAP2_1824_13_T04
PAP2_1824_13_T05
PAP2_1824_13_T06
PAP2_1824_13_T07
PAP2_1824_13_T08
PAP2_1824_13_T09
PAP2_1782_08_T01
PAP2_1782_08_T02
PAP2_1782_08_T03
PAP2_1782_08_T04
PAP2_1782_08_T05
PAP2_1782_08_T06
PAP2_1782_08_T07
PAP2_1494_04_T01
PAP2_1494_04_T02
PAP2_1494_04_T03
PAP2_1494_04_T04
PAP2_1568_04_T01
PAP2_1568_04_T02
PAP2_1568_04_T03
PAP2_1568_04_T04
PAP2_1851_04_T02
PAP2_1851_04_T03
PAP2_1851_04_T04
PAP2_1429_03_T01
PAP2_1429_03_T02
PAP2_1429_03_T03
PAP2_1479_03_T01
PAP2_1479_03_T02
PAP2_1479_03_T03
PAP2_1552_03_T01
PAP2_1552_03_T02
NA12878
Trunk
Internal branches
Terminal branches
log2(Read count):
3 4 5 6 7 8 9 10 11

Figure S25: Clonality of methylation status. The proportion of methylation levels based on the multi-regional trees (MRT) in trunks, internal branches, and terminal branches for each tumor.

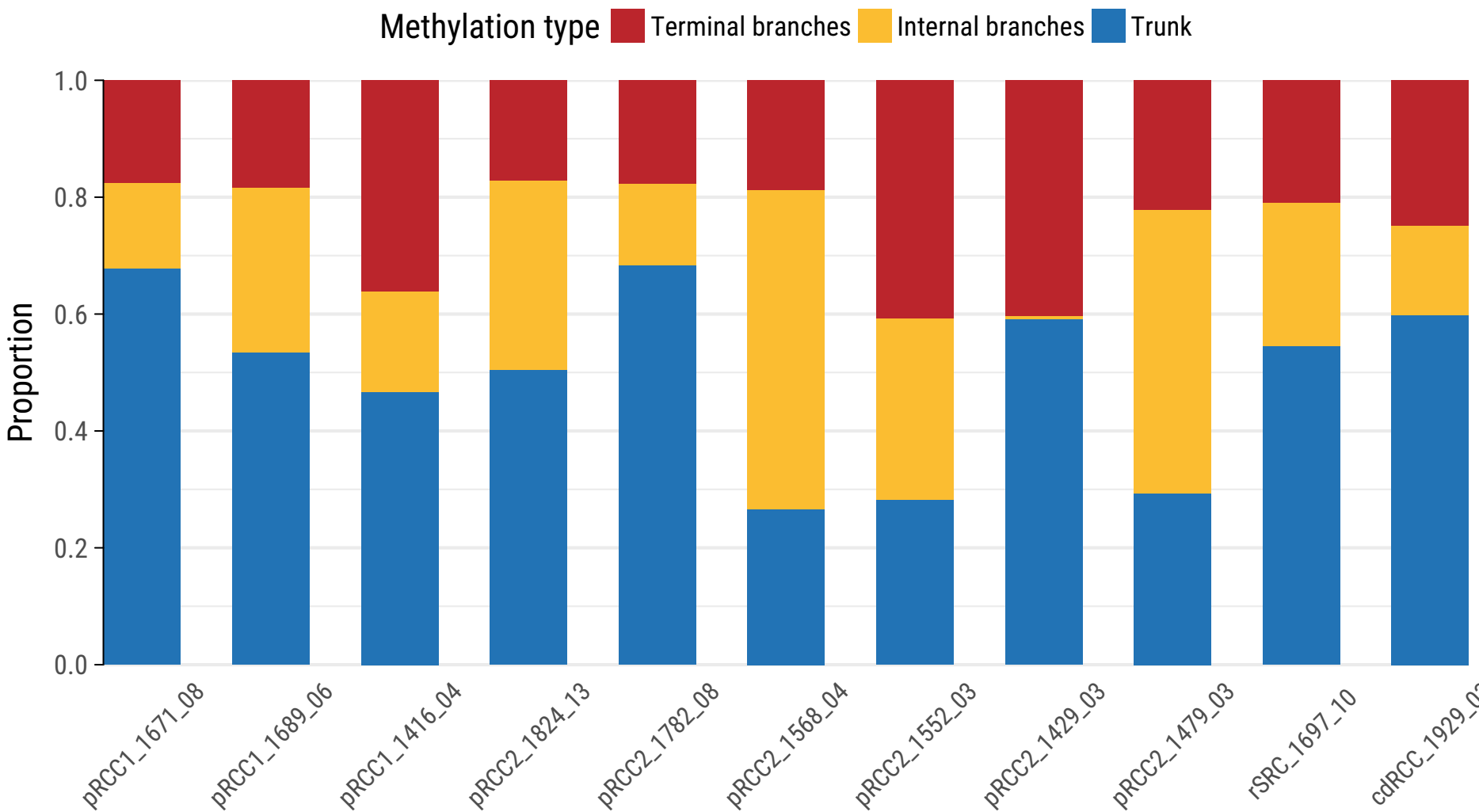

Methylation type
Terminal branches
Internal branches
Trunk
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
pRCC1_1671_08
pRCC1_1689_06
pRCC1_1416_04
pRCC2_1824_13
pRCC2_1782_08
pRCC2_1568_04
pRCC2_1552_03
pRCC2_1429_03
pRCC2_1479_03
rSRC_1697_10
cdRCC_1929_03

Figure S26: Copy number profile of the T10 sample of pRCC2_1824_13. The blue and gray dots are observed values and red lined are estimated ones; The first two panels show the profiles of logR and logOR over chromosomes; The last panel indicates estimated total copy numbers and minor copy number over chromosomes with cancer cell fraction (cf-em) in bottom, estimated by the expectation–maximization (em) algorithm.

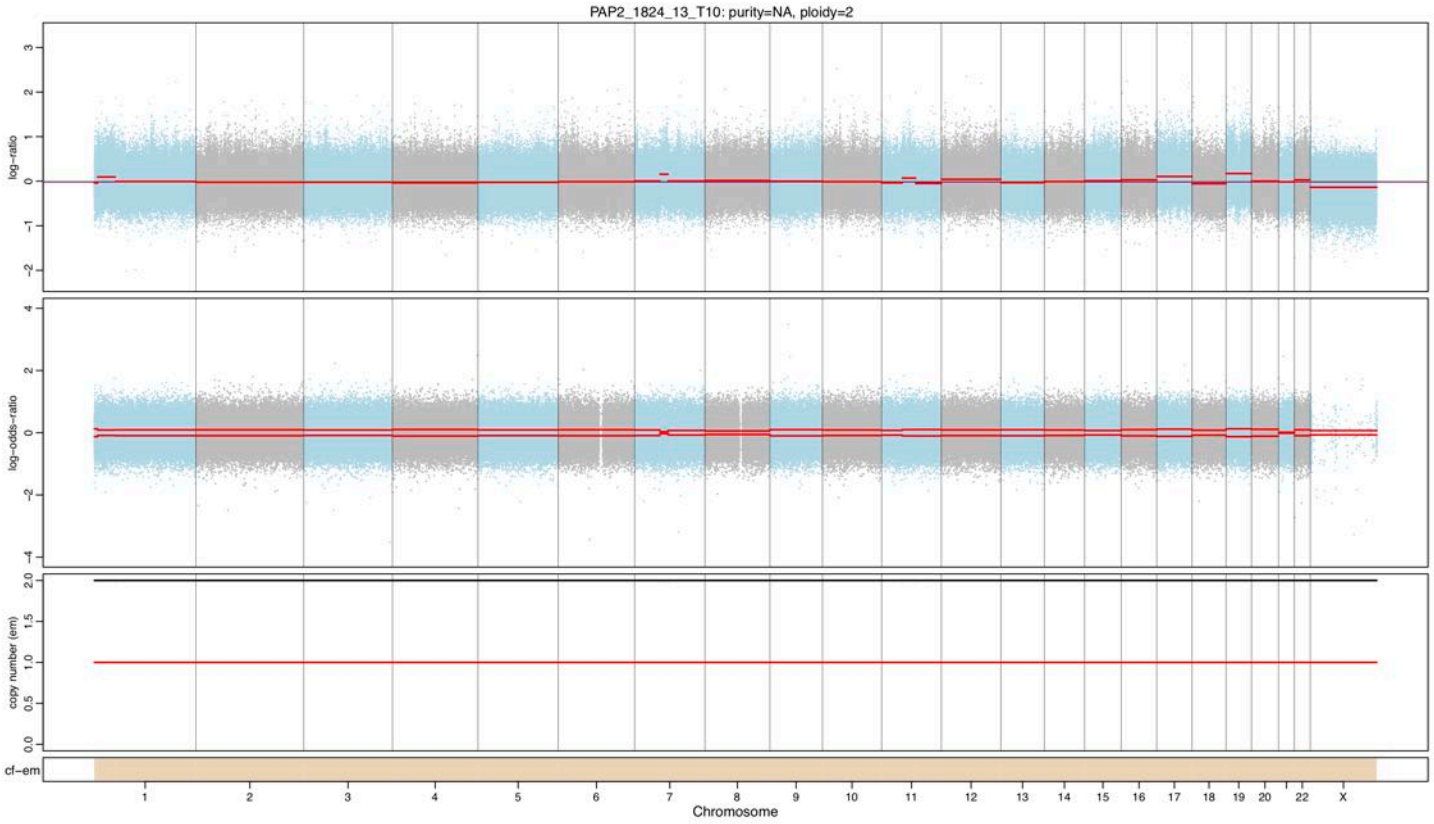
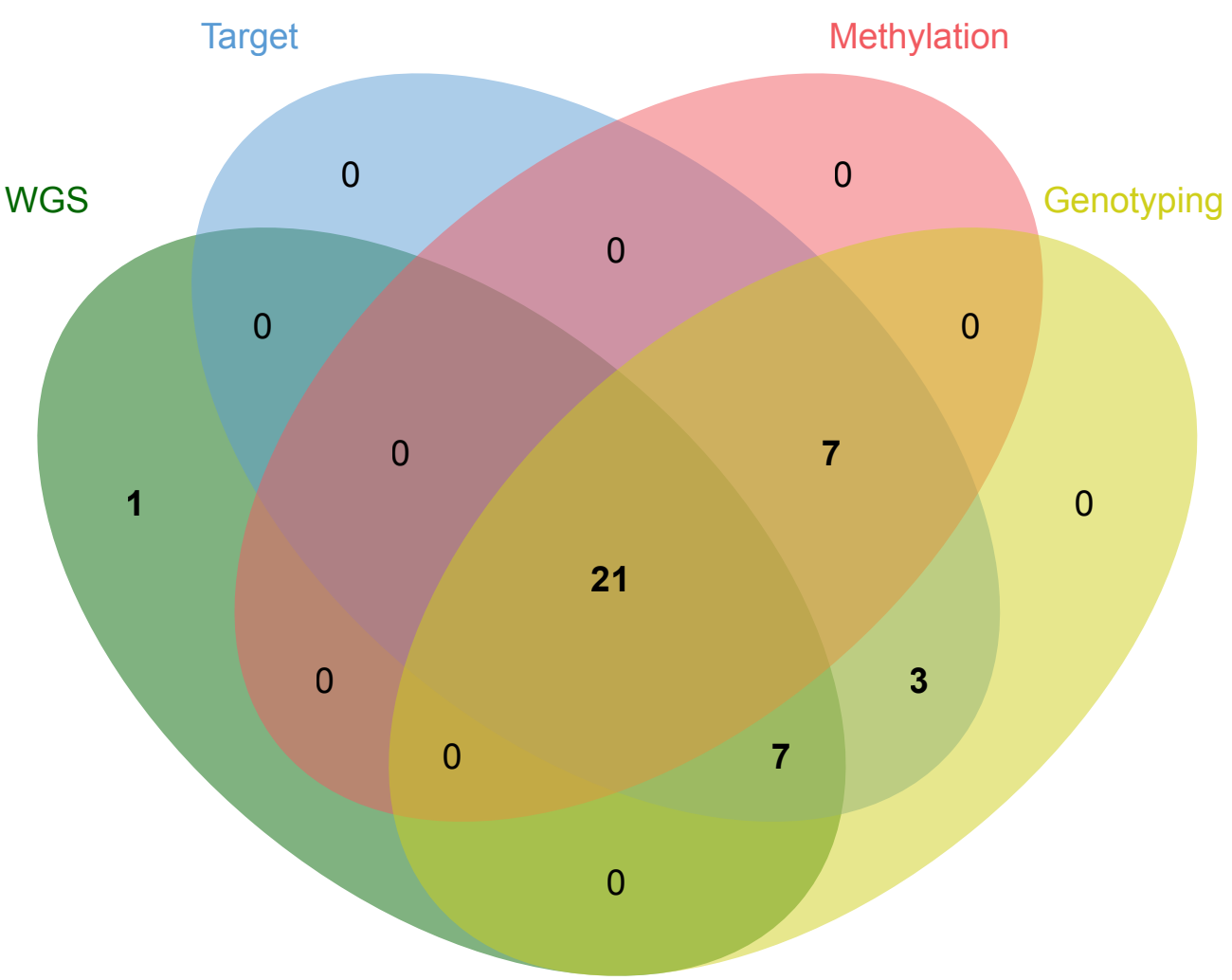
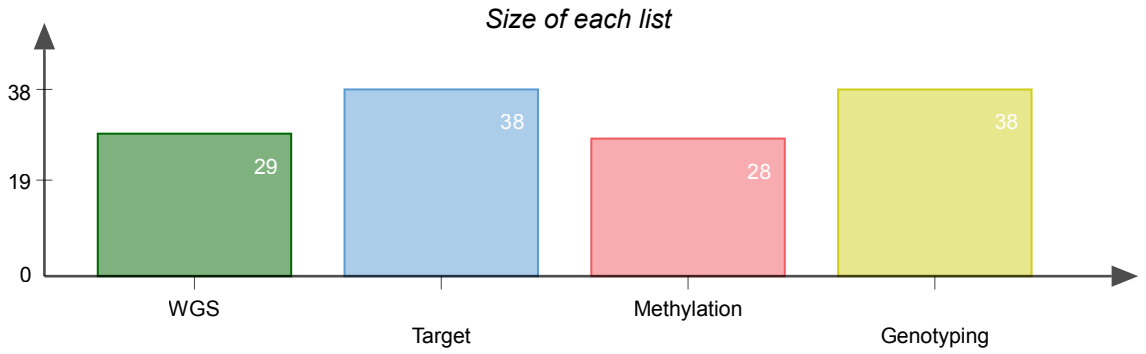
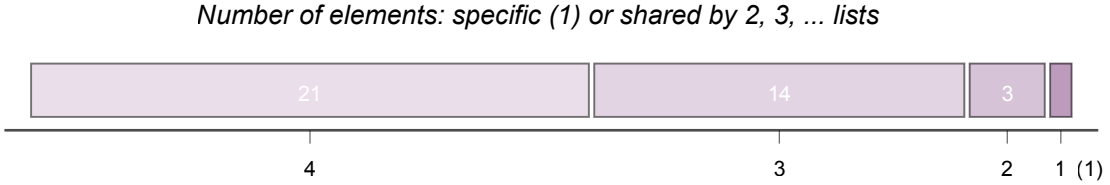


Figure S27: Venn Diagram of genomic data available across subjects. The figure describes the intercept of whole genome sequencing, deep targeted sequencing, methylation array and genotyping array data across all 29 subjects.


Target
Methylation
WGS
Genotyping
0
0
0
0
0
0
1
0
7
0
21
0
3
0
7
0


Size of each list
38
19
0
29
38
28
38
WGS
Target
Methylation
Genotyping


Number of elements: specific (1) or shared by 2, 3, ... lists
21
14
3
4
3
2
1 (1)

Click here to access/download

Supplemental Videos and Spreadsheets

Supplemental Tables_FINAL.xlsx