Supplementary Methods

Evolutionary pressures shape soft tissue sarcoma development and radiotherapy response

Erik S. Blomain[1], Shaghayegh Soudi[2], Anish Somani[2], Ajay Subramanian[2], Eniola Oladipo[2], Christin New[3], Deborah E. Kenney[3], Neda Nemat-Gorgani[2], Raffi S. Avedian[3], Robert J. Steffner[3], David G. Mohler[3], Susan M. Hiniker[2], Alex Chin[2], Anusha Kalbasi[2], Michael S. Binkley[2], Matt van de Rijn[4], Everett J. Moding[2,5,*]

[1]Department of Radiation Oncology, Thomas Jefferson University, Philadelphia, Pennsylvania, USA
[2]Department of Radiation Oncology, Stanford University, Stanford, California, USA
[3]Department of Orthopedic Surgery, Stanford University, Stanford, California, USA
[4]Department of Pathology, Stanford University, Stanford, California, USA
[5]Stanford Cancer Institute, Stanford University, Stanford, California, USA

*Corresponding author: Everett J. Moding, M.D., Ph.D., Assistant Professor, Department of Radiation Oncology, Stanford Cancer Institute, 875 Blake Wilbur Drive, Stanford, CA 94305-5847. Email: emoding@stanford.edu. Tel: 650-498-1625.

Estimation of purity and quality control

Initial copy number estimates and tumor purity were determined from whole exome sequencing using TitanCNA[1], and the best solution was selected based on the S_dbw validity index. Although uncertainty in variant allele fraction (*vaf*) measured by next-generation sequencing will lead to over or underestimation of cancer cell fraction (*ccf*) in a subset of variants, we utilized the fact that clonal mutations should have a *ccf* of 1 to confirm the TitanCNA purity estimate. The initial copy number estimates from TitanCNA were utilized in combination with the *vaf* measured by ultra-deep targeted resequencing to estimate the *ccf* of each single nucleotide variant (SNV). Samples with >40% of their SNVs with an estimated *ccf* > 1.5 or zero SNVs with a *ccf* > 0.9 were flagged as having potentially erroneous purity estimations. For these samples (<10% of all regions), alternative TitanCNA purity solutions were explored based on the lowest S_dbw value for a different ploidy or by analyzing the 2nd and 3rd best solutions for the same ploidy. If alternative solutions yielded *ccf* estimates closer to 1 for clonal SNVs, they were utilized for further analysis. Regions with purity < 0.1 or an average depth < 100 were excluded from the analysis.

Estimating cancer cell fraction

A modified method based upon previous work[2,3] was utilized to calculate the cancer cell fraction (*ccf*) of each SNV depending on the copy number status of the location.

Mutations in diploid regions of the genome without CNAs

For mutations in diploid regions of the genome without a CNA, the *ccf* was calculated using the observed variant allele frequency (*vaf*) and tumor purity (*pu*).

$$ccf = \frac{vaf \times 2}{pu}$$

Mutations in regions with subclonal CNAs

In regions with both a copy number alteration (CNA) and a SNV, we calculated the probability that the CNA occurred before the SNV and vice versa utilizing a beta distribution to determine the most likely sequencing of mutations. In total, 4 distinct scenarios could be observed.

1) Early Major: a SNV on the major allele proceeded the CNA. Three cell populations were assumed to exist in such a case: (a) normal diploid cells, (b) diploid cells with the SNV, and (c) copy number altered cells which contained the SNV.

2) Early Minor: a SNV on the minor allele proceeded the CNA. Again, three populations were assumed to exist in such a case: (a) normal diploid cells, (b) diploid cells with the SNV, and (c) copy number altered cells which contained the SNV.

3) Late: the CNA proceeded the SNV on that allele. Three populations were assumed to exist in such a case: (a) normal diploid cells, (b) copy number altered cells without the SNV, and (c) copy number altered cells which contained the SNV.

4) Independent: the CNA and the SNV occurred in different lineages. In such a case, three populations were assumed to exist: (a) normal diploid cells, (b) copy number altered cells without the SNV, and (c) diploid cells with the SNV.

In cases with subclonal CNAs, the ordering of the CNA relative to the SNV needed to be determined to calculate *ccf*. We utilized a beta distribution to determine the most likely timing of the mutations. We first calculated the proportion of sequenced cells with the CNA, or segmental aneuploid genome proportion (*sAGP*) based on the *pu* and the prevalence of the CNA (*pa*).

$$sAGP = pu \times pa$$

The effective copy number (*nc2*) was then determined based on the *sAGP* and total copy number (*nt*).

$$nc2 = nt \times sAGP + 2 \times (1 - sAGP)$$

Upper and lower bounds for the somatic allele frequency (*SAF*) in each scenario were calculated using the *sAGP*, *nc2*, *pu, pa,* major copy number (*na*), and minor copy number (*nb*).

$$SAF_{upper}(Early\ Major) = \frac{na \times sAGP + (1 - pa) \times pu}{nc2}$$

$$SAF_{lower}(Early\ Major) = \frac{na \times sAGP}{nc2}$$

$$SAF_{upper}(Early\ Minor) = \frac{nb \times sAGP + (1 - pa) \times pu}{nc2}$$

$$SAF_{lower}(Early\ Minor) = \frac{nb \times sAGP}{nc2}$$

$$SAF_{upper}(Late) = \frac{sAGP}{nc2}$$

$$SAF_{lower}(Late) = 0$$

$$SAF_{upper}(Independent) = \frac{(1 - pa) \times pu}{nc2}$$

$$SAF_{lower}(Independent) = 0$$

Using the beta distribution for the observed *vaf*, we calculated the probability that the true *SAF* was between the upper and lower bounds for each scenario (*Prob*(*scenario*)).

$$Prob(scenario) = Pbeta\left(SAF_{upper}\right) - Pbeta(SAF_{lower})$$

The evolutionary scenario with the highest probability was chosen for further analysis.

Mutations in regions with clonal CNAs

For SNVs in regions with clonal CNAs, there would be a strong prior that the SNV occurred late. However, we considered the possibility that a clonal SNV could have occurred prior to the CNA. The probability of the SNV occurring late was calculated analogously to SNVs in regions with a subclonal CNA.

$$P_{Late} = Pbeta(\frac{sAGP}{nc2})$$

To calculate the probability of an early SNV in regions of a clonal CNA ($P_{Early\ Clonal}$), the lower bound of the *SAF* was set to 99% of the *SAF* upper bound. To account for the possibility of the SNV occurring at an intermediate copy level state, we calculated the probability of an early SNV across every potential copy number state from 2 to the copy number of the major allele.

$$P_{Early\ Clonal} = max\big(Pbeta(p_a) - Pbeta(.99 * p_a)\big)$$
$$\text{where } p_a = \frac{n_x \times sAGP}{nc2} \text{ and } n_x \in Z = \{2, \dots, na\}$$

The values for P_{Late} and $P_{Early\ Clonal}$ were compared to determine the timing of the SNV relative to the clonal CNA.

Estimation of ccf in regions with CNAs

Following determination of the evolutionary scenario for each SNV, the *ccf* could be calculated.

$$ccf_{Early\ Major} = max(\frac{vaf \times nc2 - na \times sAGP}{pu} + pa, \quad nc2 \times vaf\ /\ (pu \times na))$$

$$ccf_{Early\ Minor} = max(\frac{vaf \times nc2 - nb \times sAGP}{pu} + pa, \quad nc2 \times vaf\ /\ (pu \times nb))$$

$$ccf_{Late\ or\ Independent} = \frac{vaf \times nc2}{pu}$$

Intratumoral heterogeneity metrics

Prior to calculation of the metrics, SNVs were categorized as public (clonal) or private (subclonal) using thresholds adapted from prior work[2] adjusted based on our simulated sarcomas. SNVs needed to meet the following 3 criteria to be classified as private: 1) at least one region with an estimated *ccf* below 0.78, 2) the total binomial probability < 0.05 of seeing the observed number of reads across regions if the SNV was public, and 3) at least one region with an estimated *ccf* +/- 95% CI < 0.95. Although the maximum possible *ccf* is 1, the estimated *ccf* in a region can exceed 1 due to sampling errors leading to uncertainty in the measured *vaf*. For calculation of the metrics, SNVs with an estimated *ccf* > 1.5 were collapsed to 1.5. We utilized the estimated *ccf* for the private SNVs to calculate 5 metrics of intratumoral heterogeneity: 1) the fraction of high-frequency subclonal SNVs (fHsub), 2) the fraction of high frequency region-specific subclonal SNVs out of all region-specific subclonal SNVs (fHRS), 3) dissimilarity of the site frequency spectrum between regions measured by Kolmogorov-Smirnov distance (KSD), 4) Wright's fixation index (FST), 5) and the area of under the curve for the

cumulative site frequency spectrum divided by the area under the curve for a theoretical neutral site frequency spectrum (rAUC). All metrics were calculated as described previously[2].

Estimating plasma cancer cell fraction

For calculation of plasma *ccf*, median copy number estimates were taken from the tumor regions at the same timepoint. Tumor purity was estimated for each SNV assuming Early Major timing and clonal CNAs at that location.

$$pu = 2 \times vaf/\big(na + vaf(2 - nt)\big)$$

The median of the top 3 deciles for estimated purity was used for calculation of *ccf* as described above for tumor samples. Plasma samples with fewer than 30% of variants with at least one sequencing read were not included in the prevalence analysis.

References

1. Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* **24**, 1881–1893 (2014).
2. Sun, R. *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat Genet* **49**, 1015–1024 (2017).
3. Li, B. & Li, J. Z. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol* **15**, 473 (2014).