Supplementary Note

Evolutionary Model

We considered all 92 patients for whom mutations were sequenced in both the initial and recurrent tumor samples. To exclude false positives, only variants with an allele frequency of at least 5% were used. Variants occurring at a cellular fraction of at least 95% were classified as clonal in a sample, and others were considered subclonal. Below we use the abbreviations “C” for clonal, “S” for subclonal, and “X” for absent. Clonality status provides evidence for the timing of the mutation: Clonal mutations occur prior to the most recent common ancestor of the sample (defined as occurring within that sample’s *branch*), while subclonal mutations occur after that common ancestor’s lineage has bifurcated (defined as occurring within that sample’s *diversification*). Considering both the initial and recurrent samples, there are five mutational patterns that can be explained by a single mutation event in the model depicted shown in Figure 3B:

CC: The variant occurs clonally in both the initial and recurrent samples. (Mutation event in the shared branch.)

CX: The variant occurs clonally in the initial sample and is absent from the recurrence. (Mutation event in the initial branch.)

XC: The variant occurs clonally in the recurrent sample and is absent from the initial. (Mutation event in the recurrent branch.)

SX: The variant occurs subclonally in the initial sample and is absent from the recurrence. (Mutation event in the initial diversification.)

XS: The variant occurs subclonally in the recurrent sample and is absent from the initial. (Mutation event in the recurrent diversification.)

The other three mutational patterns can be explained by positing at least two mutation events:

CS: The variant occurs clonally in the initial sample and subclonally in the recurrence. (Case 1: Mutation in the shared branch, back-mutation in the recurrent diversification. Case 2: Mutation in the initial branch, same mutation in the recurrent diversification.)

SC: The variant occurs subclonally in the initial sample and clonally in the recurrence. (Case 1: Mutation in the shared branch, back-mutation in the initial diversification. Case 2: Mutation in the initial diversification, same mutation in the recurrent branch.)

SS: The variant occurs subclonally in both the initial and recurrent samples (Mutation in the initial diversification, same mutation in the recurrent diversification.)

Notably, this model assumes that the initial and recurrent samples are monophyletic – that is, each forms a distinct evolutionary clade, having diverged from a common ancestor sometime in the past. In reality, they may exhibit more complex evolutionary patterns. For example, the recurrent sample may be nested within the initial clade (as in Supplementary Figure 7B), or the two samples may be evolutionarily intertwined. We determined that, of the 92 patients studied, 45 fit the monophyletic model well (see

"Assessing Model Fit," below). Of the remaining 47 patients, many had mutational patterns that could not be explained by our monophyletic model, unless unrealistically many mutations occurred twice, in two separate lineages. As an extreme case, Patient R009 had 58 of 100 mutations appear subclonally in both samples. This mutational pattern suggests that the two samples are evolutionarily interwined. The model does, however, allow for less extreme levels of recurrent- or back-mutation: there are well-fitting patients with up to 23 mutations clonal in one sample and subclonal in the other.

The expected number of mutations of each pattern was computed using a second-order approximation, i.e., the probability that the same mutation occurs (or back-mutates) three or more times is zero. The following parameters are needed for the computation:

- Substitution rates:
    - Per-site, per-year substitution rates u_1 (pre-treatment) and u_2 (post-treatment);
    - Per-site, per-year back-substitution rates v_1 (pre-treatment) and v_2 (post-treatment);
- Times:
    - Branch lengths, in years: t_S (shared branch), t_I (initial sample branch), t_{R1} (recurrence branch prior to treatment), and t_{R2} (recurrence branch after treatment)
    - Time between the most recent common ancestor of a sample and the collection of that sample (t_{MRCA}), assumed to be the same for both samples.

    - The times are constrained so that the age at diagnosis equals $t_S + t_I + t_{MRCA}$, the age at recurrence equals $t_S + t_{R1} + t_{R2} + t_{MRCA}$, and $t_I + t_{MRCA} = t_{R1}$.
- The effective genome length, *L*. If substitutions are equally probable at all sites, then this parameter simply equals the length of the entire sequenced genome (exome length, 3×10^7 bp). Since mutation may not be possible at many sites, and since not every mutation is equally probable, the fitted value is typically smaller. Given the same genomic subsitution rate (product of *L* and a per-site rate), a larger value of *L* decreases the probability that the same mutation occurs twice or is reversed by back-mutation.
- The effective sample sizes s_I (initial) and s_R (recurrence). A larger sample increases the probability that subclonal variants may be found. Given the average exome coverage for most patients (~150x), we capped the effective samples sizes at 200.

As a first approximation, the initial sample branch length, pre-treatment forward substitution rate, post-treatment substitution rate, and effective genome length are related to the number of clonal mutations by $n_{CC} \approx u_1 L t_S$, $n_{CX} \approx u_1 L t_I$, and $n_{XC} \approx u_1 L t_{R1} + u_2 L t_{R2}$, where each subscripted *n* is the number of mutations of a particular pattern and the time t_{MRCA} is ignored. Solving these three equations produces reasonable point estimates for the time t_I and the genomic substitution rates $u_2 L$, $u_1 L$. Our actual parameter estimates accounted for the number of mutations of all eight patterns and used a negative binomial likelihood for the number of mutations, where the overdispersion of the negative binomial distribution was also fitted. (This negative binomial fit is related mathematically to the notion of gamma rate variation commonly

used in phylogenetics[59,60].) We considered each patient separately and used a Bayesian MCMC approach to obtain posterior distributions for each parameter. The model was implemented using PyStan v2.8.0.2, an interface for the Bayesian inference programming language Stan[61]. A total of 250,000 Hamiltonian Monte Carlo iterations (burn-in of 125,000) was sufficient for convergence in nearly all patients (effective sample sizes >200, $\hat{\rho}$<1.001). Commented Stan code is provided in Supplementary Code, fully describing the model and priors.

Assessing Model Fit

Patients were deemed to fit the model well if they passed all the following criteria:

- MCMC convergence. The effective sample size for all fitted parameters must be at least 200, and the $\hat{\rho}$ for all fitted parameters must be at most 1.001.
- Genome length. The 95th percentile of L must be at least 10^6. Some patients had surprisingly many subclonal variants shared between the untreated and recurrence samples. These patients were fitted with a short genome length L, as a small “target size” could explain the occurrence of the same exact mutation twice.
- Overdispersion. The median overdispersion for the negative binomial distribution must be at most 15.
- Similarity of forward- and back-substitution rates. The two rates must not differ by more than a factor of 10, or, if they do, then the p-value corresponding to this difference must exceed 0.05.

- No outlier mutation data. All eight mutation pattern counts must lie between the 1st and 99th percentiles of the fitted negative binomial distributions.
- Two-hit approximation valid. The Stan code (Supplementary Code) computes the probability that a mutation occurs a second time (or back-mutates) along any branch or within either diversification. The 97.5th percentile of each of these probabilities must not exceed 0.14, and the sum of all of these 97.5th percentile values must not exceed 0.5. Larger values would indicate that the same mutation is likely to occur three or more times.

Sensitivity analysis of cellular frequency clonality cutoff

In the branching model of tumor evolution described in the Main Text, mutations occurring with a cellular frequency of at least 0.95 are deemed clonal, while those with a lower frequency are deemed subclonal. Here we apply different definitions of clonality to the same model. For each definition, we estimated pre- and post-treatment substitution rates and time between lineage divergence and diagnosis, as done previously in Figure 3C,D, Supplementary Figure 8, and Supplementary Figure 9. Generally, varying the cellular frequency clonality cutoff between 0.92 and 0.98 only slightly altered model fits (Additional Figures M1-M2), with per-patient comparison R^2 values between 0.74 and 0.81 for each parameter (Additional Figures M3A, M4A, M5A). Increasing the cutoff reduced the number of patients for whom the model fit well (49, 45, and 27 out of 92 patients fit well, for cutoffs of 0.92, 0.95, and 0.98 respectively), as a higher cutoff treats more patients as having subclonal mutations shared between both tumor samples.