Supplementary Information for:

A Big Bang model of human colorectal tumor growth

Andrea Sottoriva[1,6], Haeyoun Kang[2,3], Zhicheng Ma[1,6], Trevor A. Graham[4,5], Matthew P. Salomon[1], Junsong Zhao[1], Paul Marjoram[1], Kimberly Siegmund[1], Michael F. Press[2], Darryl Shibata[2], Christina Curtis[1, 6]

Affiliations

[1] Department of Preventive Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, CA USA

[2] Department of Pathology, Keck School of Medicine of the University of Southern California, Los Angeles, CA USA

[3] Department of Pathology, CHA University, Seongnam-si, Gyeonggi-do, South Korea

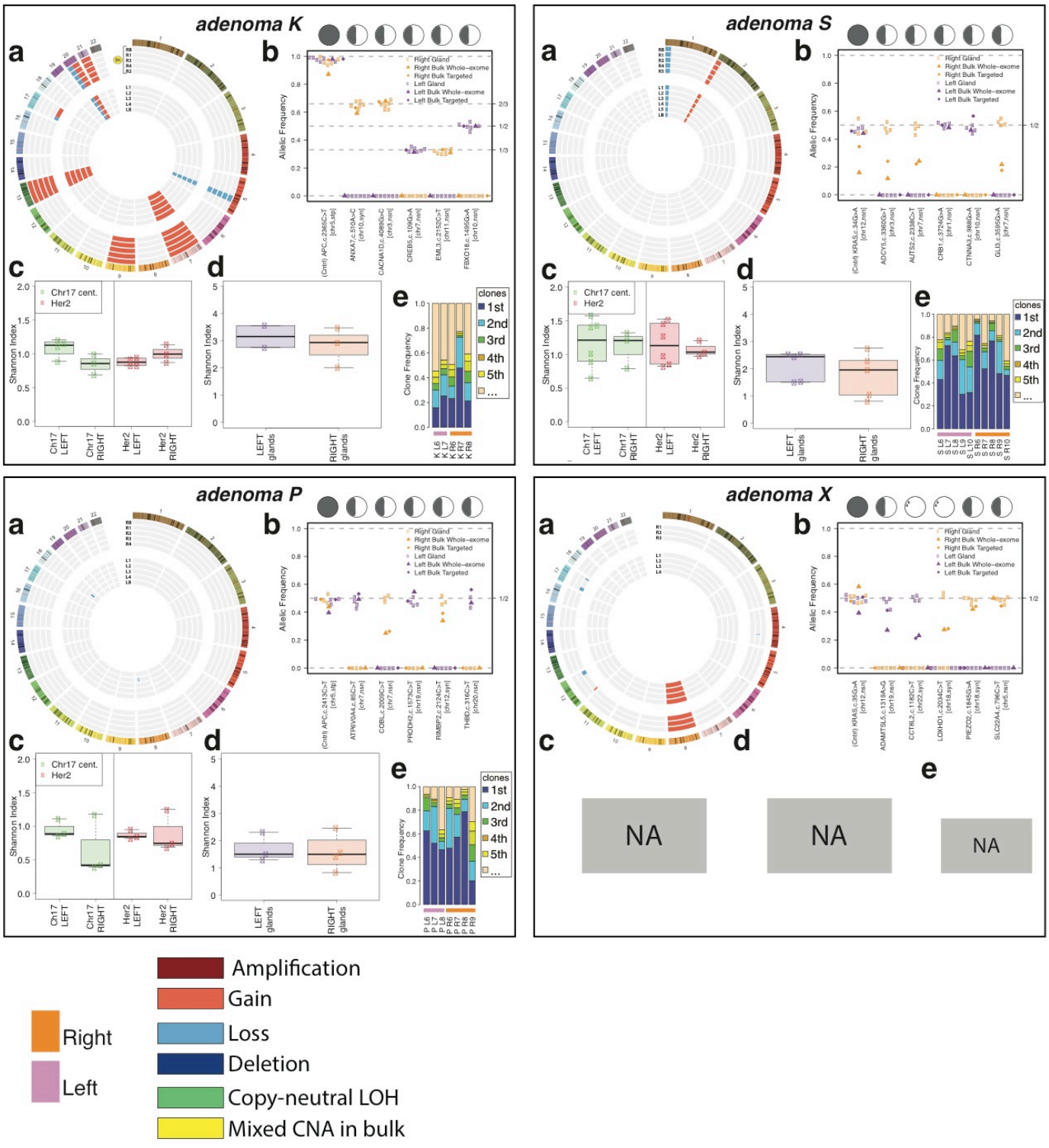
[4] Center for Evolution and Cancer, University of California, San Francisco, San Francisco, CA USA

[5] Centre for Tumor Biology, Barts Cancer Institute, Queen Mary University of London, London, UK

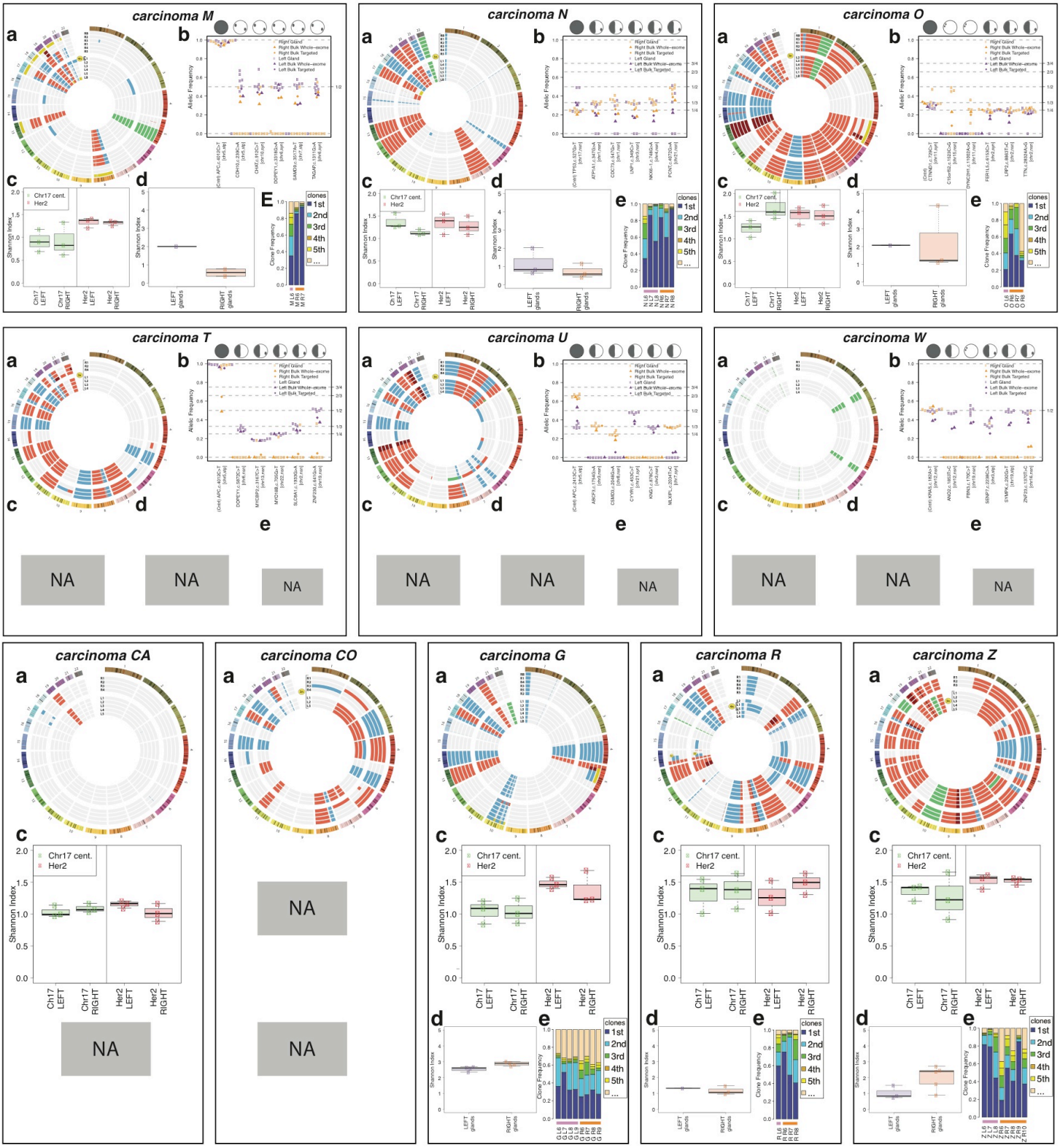
[6] Present addresses: Division of Molecular Pathology, Institute of Cancer Research, London, UK (A.S); Department of Medicine, Stanford University, Stanford, CA, USA (Z.M, C.C); Department of Genetics, Stanford University, Stanford, CA, USA (Z.M, C.C)

Correspondence should be addressed to: Darryl Shibata (dshibata@usc.edu) or Christina Curtis (cncurtis@stanford.edu)

Supplementary Figures

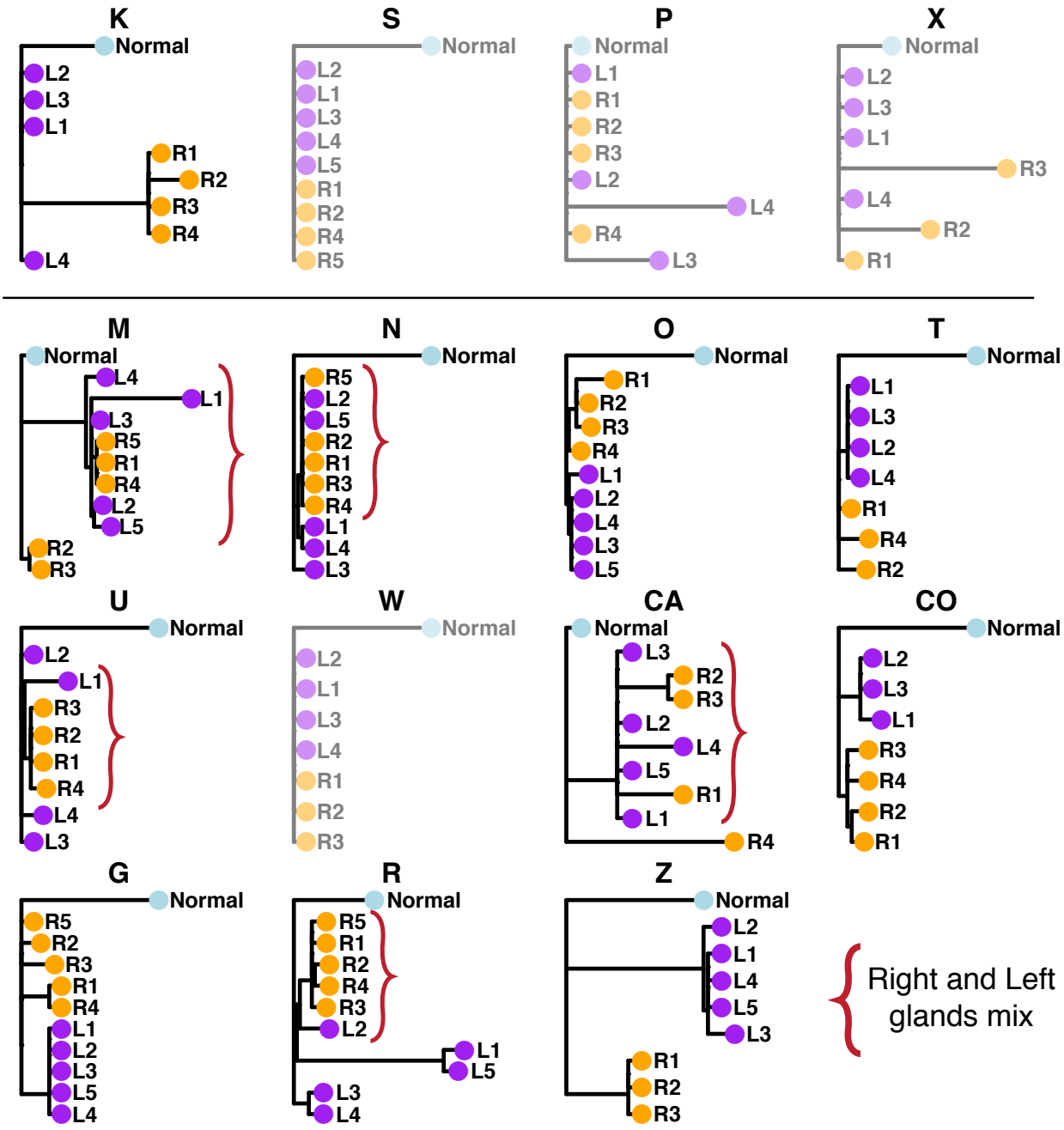


Supplementary Figure 1 (continued on next page).

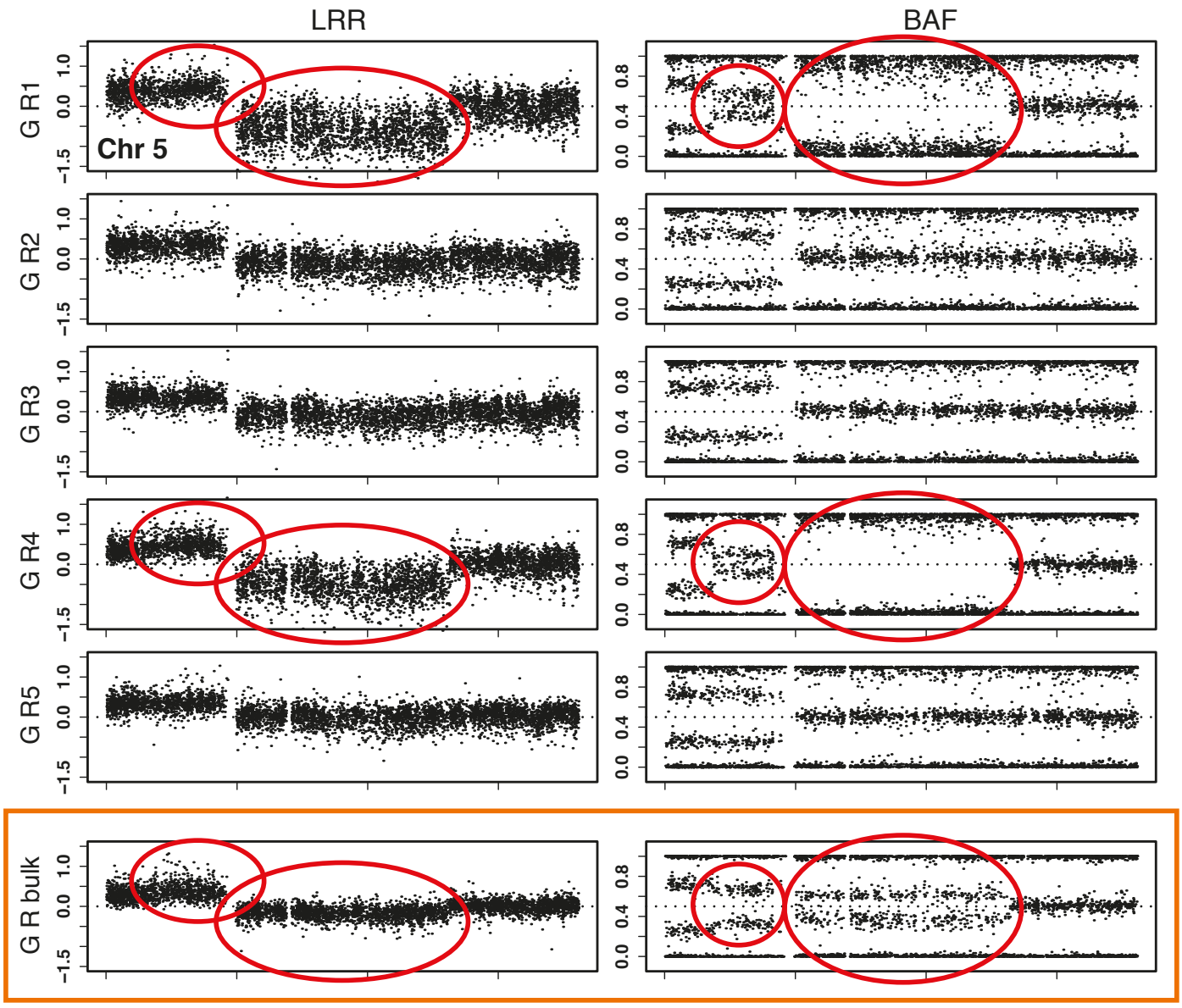


Supplementary Figure 1. Integrated genomic analysis of ITH across multiple spatial scales. (a) Circos plot representation of copy number aberration (CNA) profiles (where concentric circles correspond to individual glands and bulk samples) reveal extensive heterogeneity at the copy number level in the carcinomas, but not

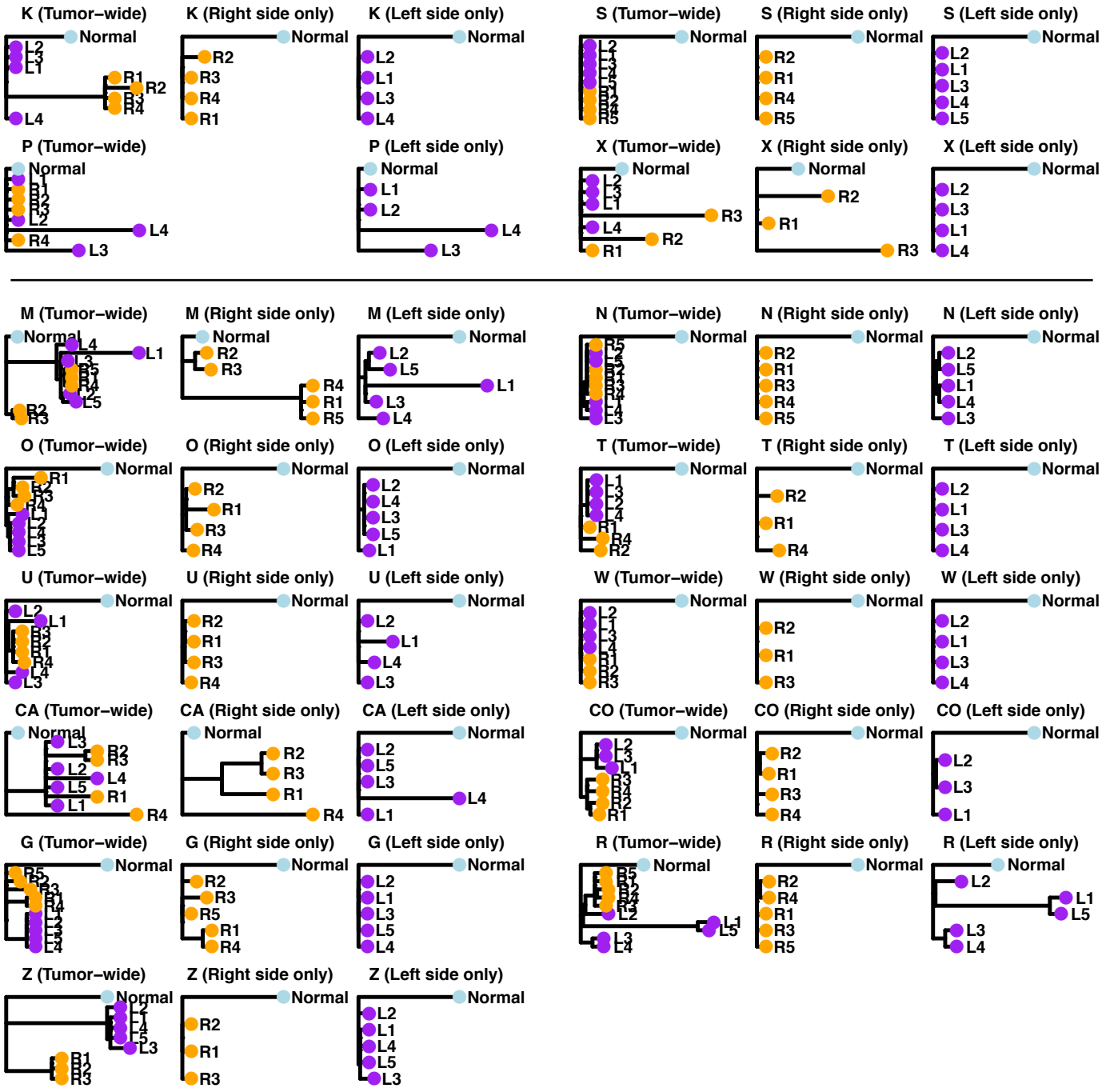
adenomas. Variegated and side-variegated CNAs were found in the majority of carcinomas (7/11), indicative of early mixing, followed by the scattering of sub-clones to distant tumor regions. (b) For all adenomas and a subset of carcinomas, WES was performed on bulk tumor samples, followed by targeted deep sequencing of a mutational panel in single glands. The allelic frequency of a subset of mutations is shown and color-coded according to side (right, left). Canonical drivers (*APC*, *KRAS*, or *TP53*) were commonly mutated and when present, were clonal within the tumor and thus used as positive controls. For tumor O, the canonical drivers were not mutated, and a clonal mutation in *CTNND1* was used as a positive control. For diploid samples, the expected allelic frequencies are 0.5 or 1 (when loss of heterozygosity has occurred). For triploid samples, the frequencies are 0.33, 0.66 or 1 (N, O, T, U and the right side of K) and for quadraploid samples: 0.25, 0.5, 0.75 or 1. The mutational profiles strongly corroborate the CNA data and reveal variegation in all carcinomas, with glands from opposite sides harboring the same private mutation. (c) FISH analysis of the *HER2* gene and corresponding chr17 centromere was performed on individual glands (n=3-6 per tumor side) and reveals extensive variability in copy number between physically adjacent cells, as summarized by Shannon Index of diversity. For each group, boxplots show the median, limited by the 25th (Q1) and 75th (Q3) percentiles, where whiskers represent the most extreme of the maximum or Q3+1.5(Q3−Q1) and the minimum or Q1−1.5(Q3−Q1), respectively. Importantly, ITH was uniformly high throughout the tumor, indicating the absence of recent clonal expansions, as postulated by the Big Bang model. The maximum possible heterogeneity corresponds to an index of 1.79 (99% of counts within [0,5], where max SI is ln(N) for N the number of possible count states, which is 6). (d) Molecular clock analysis based on neutral methylation tag sequencing corroborates the FISH data, with glands from different tumor regions exhibiting similar high levels of ITH throughout the tumor. Boxplots are formatted as in panel d. (e) The molecular clock analysis further highlights a complex hierarchy of mitotic sub-clones within glands, as demonstrated by the frequency distribution of intra-gland mitotic clones, where the five most common clones are shown.



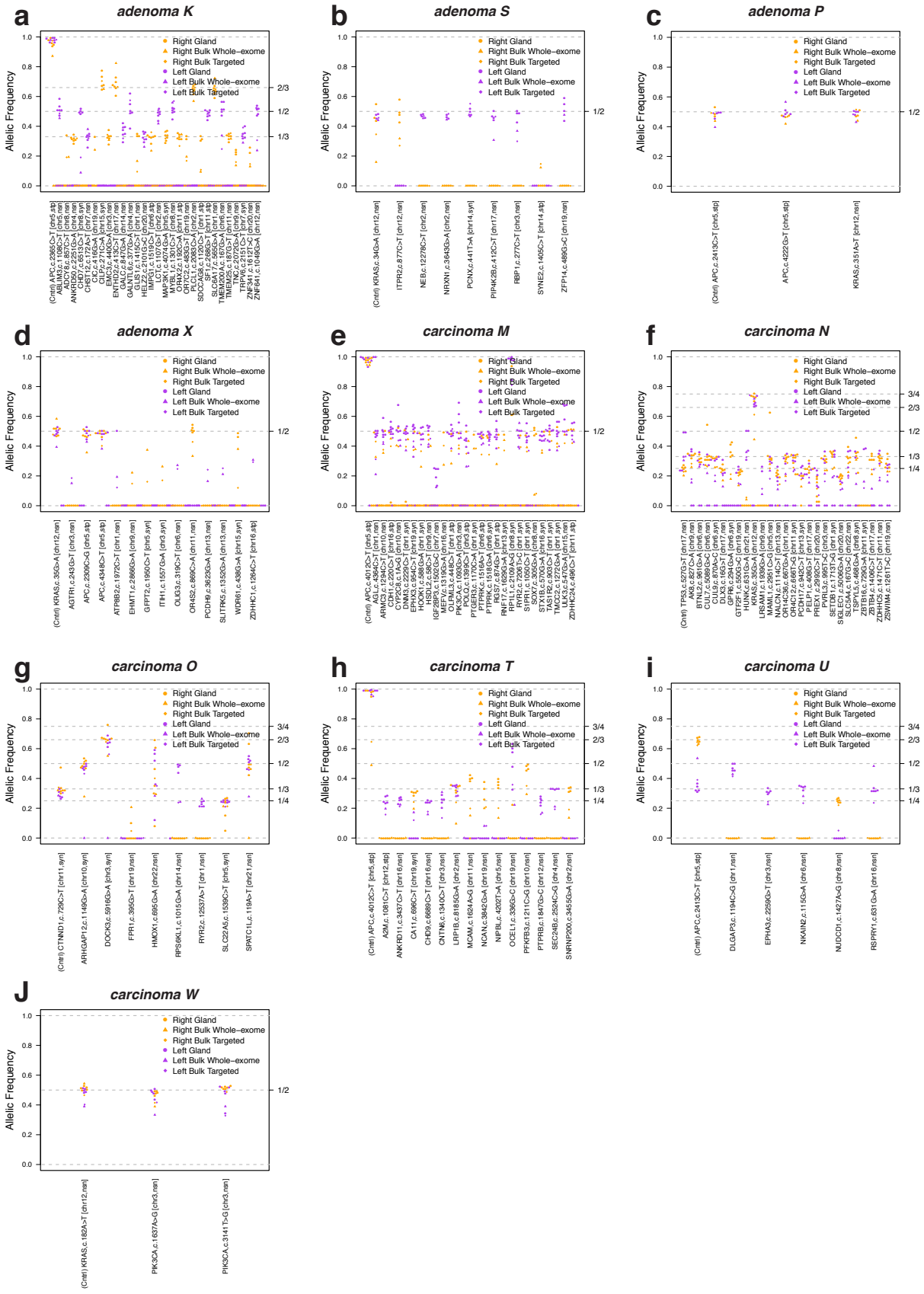
Supplementary Figure 2. Reconstruction of tumor phylogenetic trees. Tumor phylogenetic trees were reconstructed based on individual gland genome-wide CNA profiles. Sub-clone mixing between right and left glands is apparent in some of the carcinomas, where glands from one side were more similar to other glands from the opposite side than to neighboring glands (indicated by red parentheses). The accurate reconstruction of tumor phylogenies for adenomas S, P and X and carcinoma W (which was microsatellite instable, MSI) was not possible due to minimal CNAs (these are included for completeness, but indicated by transparent shading).



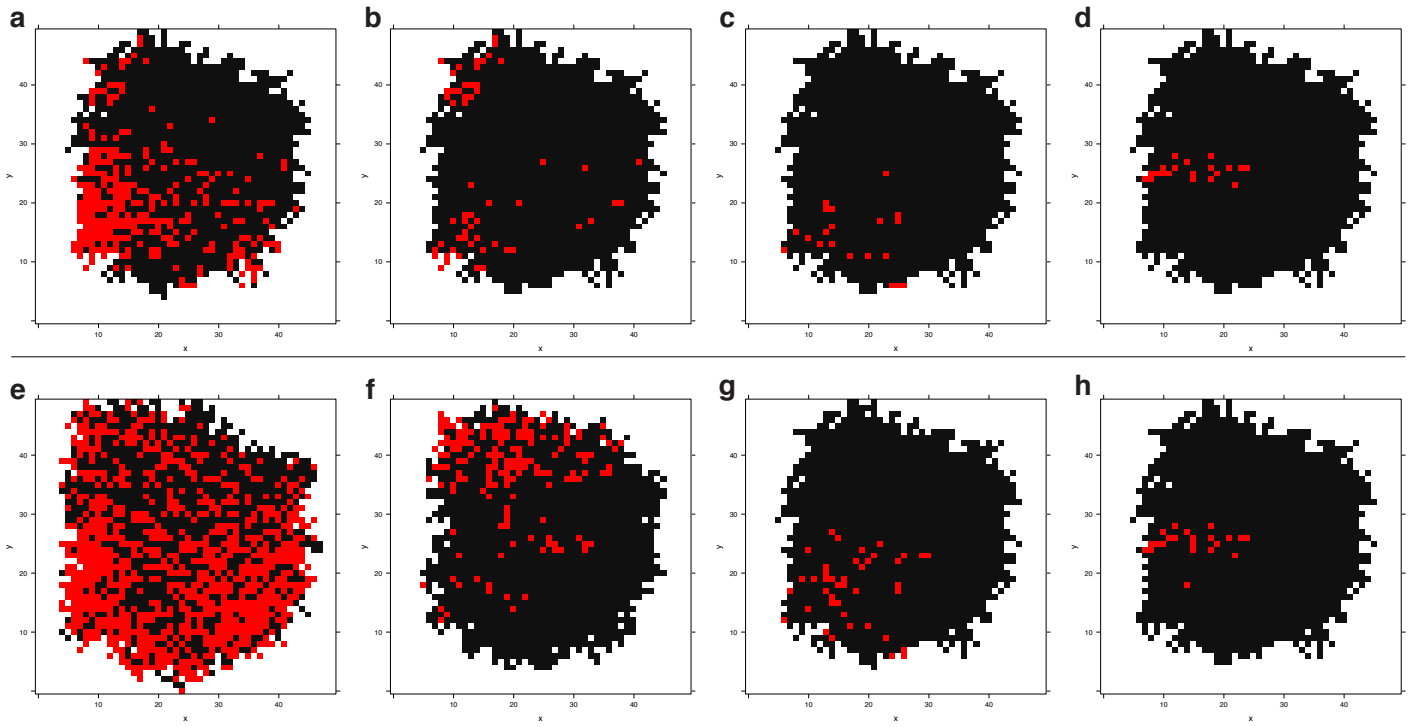
Supplementary Figure 3. Bulk tumor copy number profiles recapitulate single gland ITH. Individual tumor glands may exhibit distinct CNAs. When present, such heterogeneity was also noted in the bulk fragments collected from either the left or right side of the tumor, as illustrated here for representative carcinoma G.



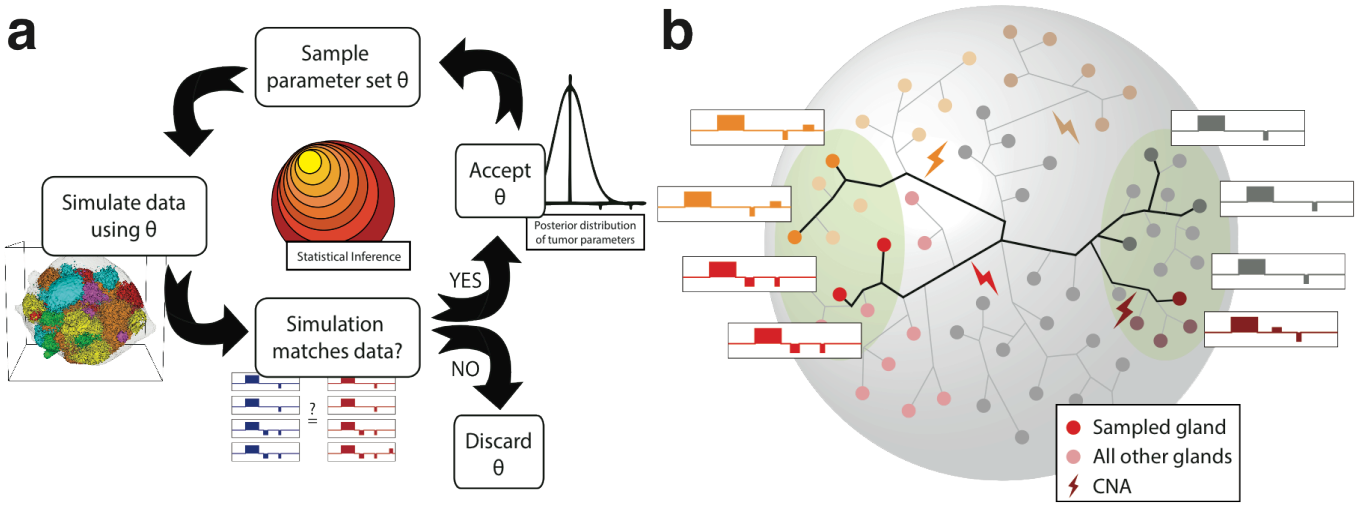
Supplementary Figure 4. The impact of sampling bias on tumor ancestral tree reconstruction. CNA profiles from individual glands sampled from either the right (R#) or left (L#) side of the tumor were used to generate phylogenetic trees and were compared with trees generated using tumor-wide data taken from both the right and left side. As a result of extensive ITH, the reconstruction of phylogenies based on a single tumor region does not yield a sub-tree that is representative of the combined tumor-wide, but rather an erroneous phylogeny.


a
adenoma K
b
adenoma S
c
adenoma P
d
adenoma X
e
carcinoma M
f
carcinoma N
g
carcinoma O
h
carcinoma T
i
carcinoma U
J
carcinoma W
Allelic Frequency
Right Gland
Right Bulk Whole−exome
Right Bulk Targeted
Left Gland
Left Bulk Whole−exome
Left Bulk Targeted

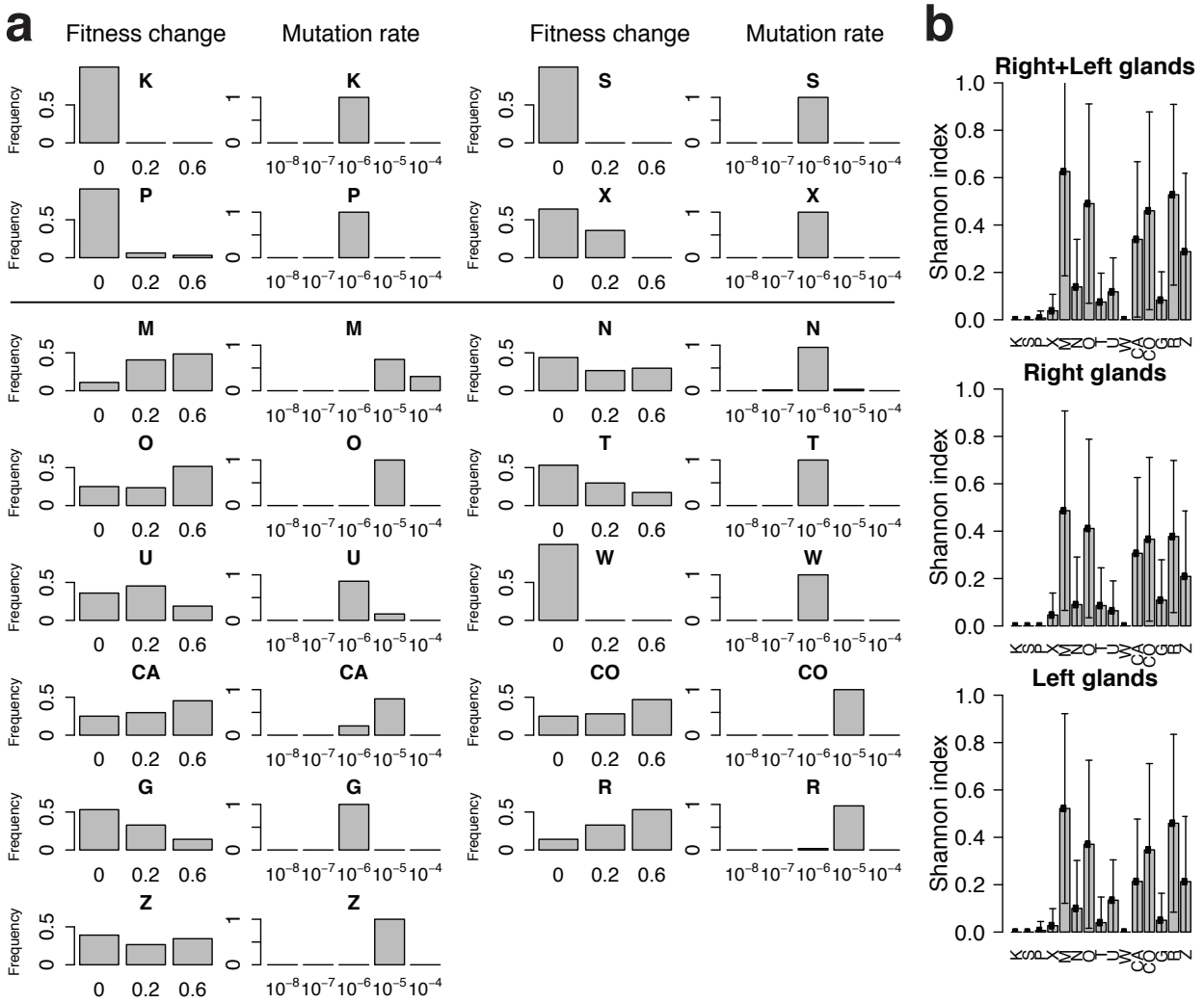
Supplementary Figure 5. Mutational profiling of single glands. The allelic frequencies of additional mutations based on single gland targeted deep sequencing for tumors K, S, X, M, N, O, T and U are presented here, and corroborate the findings reported in Supplementary Figure 1B. Note that the whole right side of adenoma K is triploid, hence the allelic frequencies are shifted to either 0.33 or 0.66. Sub-clone segregation is apparent in the adenomas, whereas variegation is typical of the carcinomas. Putative driver mutations found in all cells within a tumor (public) were used as a control. The mean depth of coverage per mutation was 626.58±20.2 (95% CI).



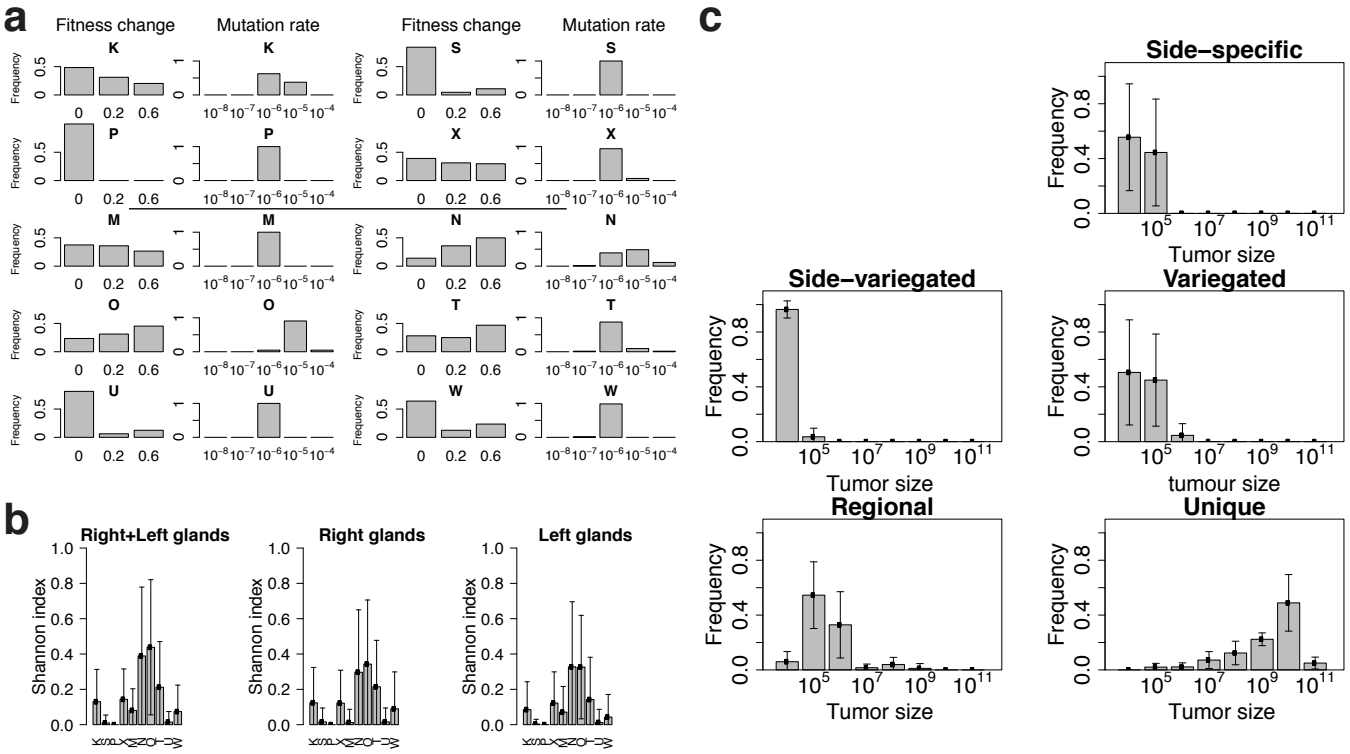
Supplementary Figure 6. Illustrative simulations show how variegation arises from early mixing in the primordial tumor followed by scattering due to growth, independent of a clone's fitness. In order to illustrate how genomic variegation can arise from simple growth dynamics, we simulated a small 2D (50x50 gland lattice) tumor that recapitulates the patterns of the larger 3D model used for statistical inference. Tumor glands are tubular in shape, but their average diameter can be approximated to 200 μm, hence a 50x50 gland lattice represents ~1cm^2 tumor. The simulation begins with a single gland surrounded by normal tissue (white, not modeled). We model the growth of the gland in a manner analogous to the 3D model where new space is allocated by pushing glands outwards, but we do not account for the within-gland structure. We then simulate the emergence of a new mutant gland (red) with no survival advantage (identical to all other glands) when the tumor is composed of 4 (a), 8 (b), 64 (c) and 1024 (d) glands. The new mutation is passed on to the gland progeny, defining the sub-clone shown in red. The results presented are representative of a large number of simulations, which show that a mutant clone introduced early can result in variegated patterns, whereas a mutant clone that arises at later time points tends to remain segregated. This simple illustration reflects our inference results, which demonstrate that variegated/side-variegated alterations occur early due to random mixing during growth in the primordial neoplasm in the absence of normal cell adhesion, followed by subsequent scattering during expansion. The same phenomenon occurs when the mutant clone harbors a considerable fitness advantage (20%) for a tumor composed of 4 (e), 8 (f), 64 (g), and 1024 (h) glands, indicating that the timing of a mutation is the critical factor in generating genetic variegation, not selection. These results suggest that in some tumors, abnormal cell mixing required for later invasion is expressed very early.



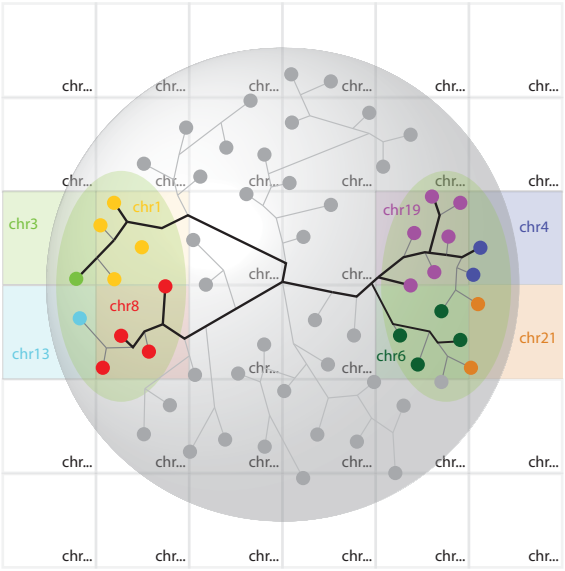
Supplementary Figure 7. Schematic representation of tumor evolutionary dynamic statistical inference framework. (a) In order to infer the evolutionary dynamics of neoplastic growth, including the mutational timeline, sub-clone fitness changes (σ), and the mutation rate (μ) from single gland genomic data, we employed a 3-dimensional mathematical model of tumor growth and statistical inference framework based on Approximate Bayesian Computation (ABC). The output is a set of posterior probability estimates for the tumor characteristics of interest given the data and the model of reference. (b) Schematic representation of the 3-dimensional spatial model of tumor growth. The model assumes gland fission, and accounts for the acquisition of CNAs or point mutations (indicated by a lightening symbol) that result in clones with a different relative fitness. Each simulation is run with a parameter set θ=(σ,μ) and generates *virtual* gland profiles such that the simulated data can be compared with the observed data within the inference method.



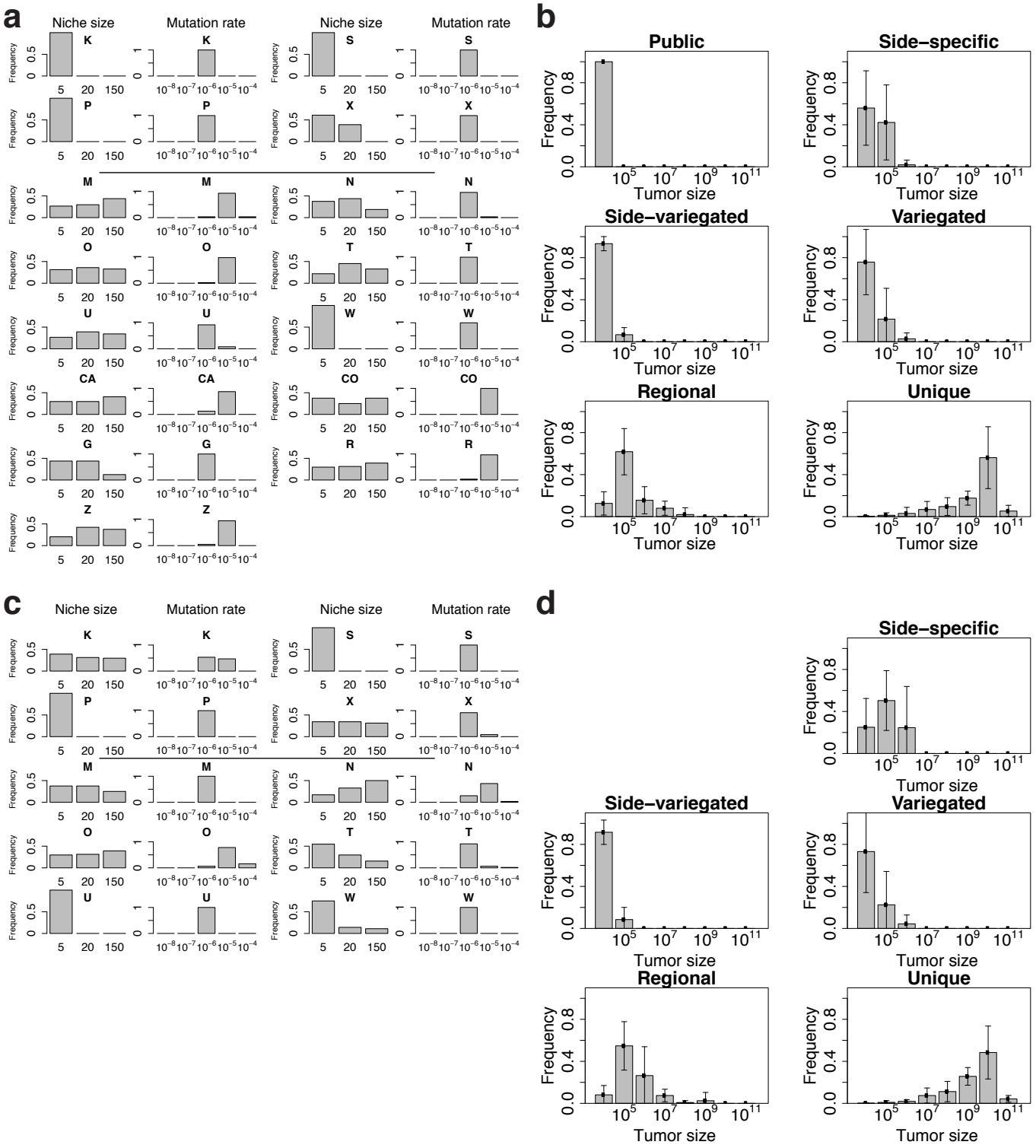
Supplementary Figure 8. Inferred tumor fitness parameters and mutation rates based on the CNA data. The inferred posterior probability distributions (a) for changes in sub-clone fitness and mutation (CNA) rate are illustrated for all samples. The inference results indicate that the adenomas do not exhibit sub-clone fitness changes and have a mutation rate of 10^{-6}. In contrast, the carcinomas were more variable, as N, T, U, W and G have mutation rates on the order 10^{-6}, while the remainder exhibit an elevated mutation rate (10^{-5}), and all apart from W (MSI+) show moderate to large fitness changes. (b) Glands within the same tumor and even the same side exhibit variation in fitness based on the Shannon index. This suggests that despite evidence for selection, selective sweeps that would homogenize the population are rare and as a result, sub-clones with different fitness levels coexist within the same region. Error bars correspond to the standard deviation.



Supplementary Figure 9. Statistical inference on single gland mutational data corroborates findings at the CNA level. (a) The inferred posterior probability distributions for changes in sub-clone fitness and mutation rate based on mutational data are in agreement with the CNA-based inference. The variation in fitness between adjacent glands within the tumor (b) and the timeline during which different classes of alterations arise (c) are also congruent with the results based on the CNA data. Note that public mutations that occur *after* the transition to an established neoplasm were not found in the simulations selected by the inference framework. Error bars represent the standard deviation.

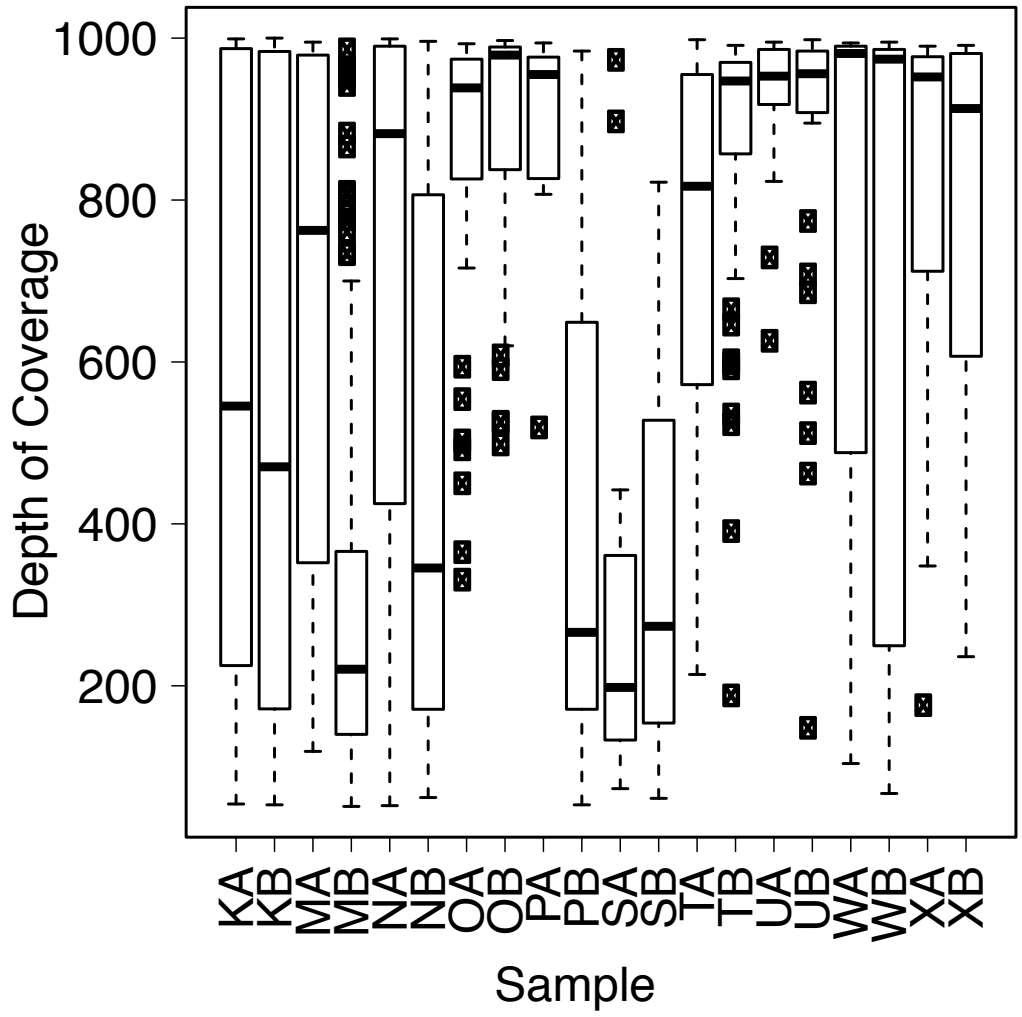


Supplementary Figure 10. Schematic of microenvironment-aware simulation of tumor growth. In order to model the contribution of local microenvironment to selection, microenvironmental niches that select for different genotypic characteristics were simulated. To achieve this, the lattice in which the tumor grows was divided into a grid of separate microenvironmental niches, each of which selects for a specific alteration in a random chromosome (indicated by colored circles) by inducing a higher survival probability in glands that harbor that aberration.

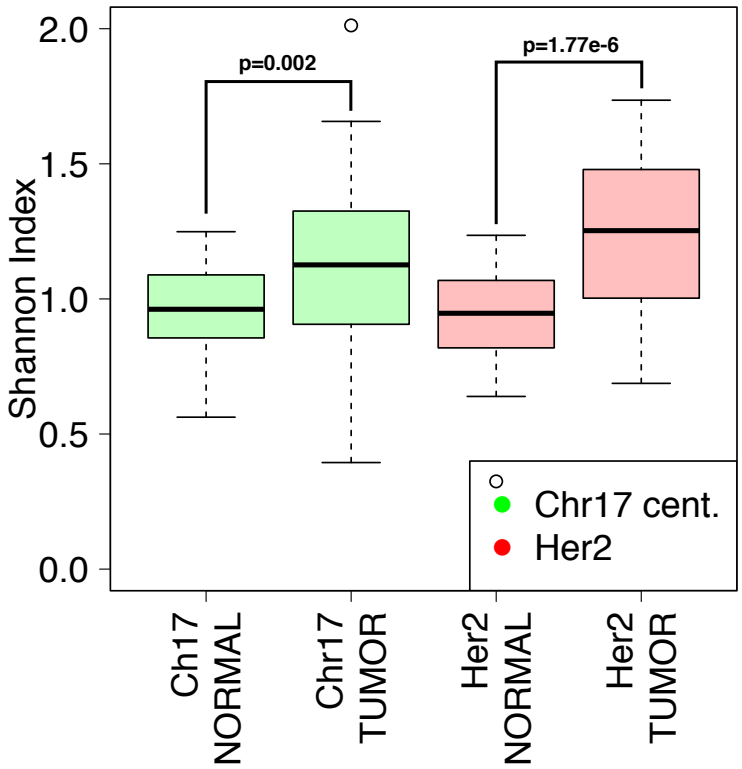


Supplementary Figure 11. The mutation rate and mutational timeline do not change under the assumption of microenvironmental niche-driven growth. (a) The inferred posterior probability distributions for niche size

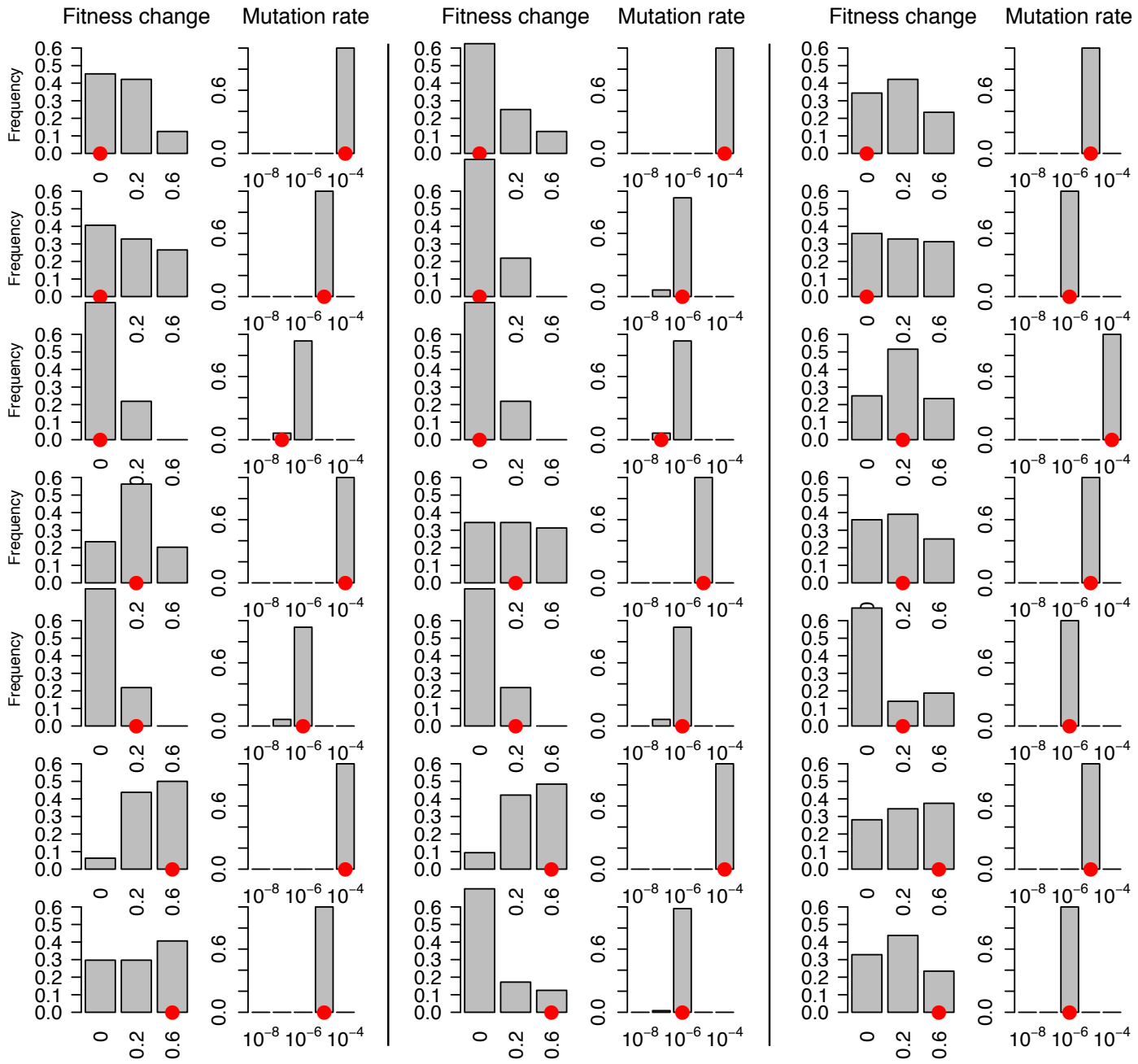
indicate variable sizes for adenomas versus carcinomas. Despite the fact that microenvironments may differ within different regions of a tumor, the inferred CNA rates are equivalent to the microenvironment-free model. (b) Similarly, the inferred mutational timeline, shown here summarized across all samples, is in agreement with the microenvironment-free model. In particular, the parameter estimates are consistent across cases (as illustrated by the error bars), and indicate that public, as well as the majority of private alterations (side-variegated, variegated, and side-specific) occur early after the transition to an advanced tumor, whereas unique mutations occur late. (c) As in (a), similar results for the niche size and mutation rates were obtained independently using single gland targeted mutational profiles as input to the inference framework. (d) As in (b), the inferred mutational timeline is equivalent using single gland targeted mutational profiles. Public mutations occurring *after* the transition to an established neoplasm were not found in the simulations selected by the inference framework. Error bars represent the standard deviation.



Supplementary Figure 12. Depth of coverage for targeted deep sequencing. Single glands were sequenced to an average depth of coverage of 626.58±20.2 (95% CI), where boxplots are presented for individual samples (A, right; B, left) from each tumor. Boxplots show the median, limited by the 25th (Q1) and 75th (Q3) percentiles, where whiskers represent the most extreme of the maximum or Q3+1.5(Q3−Q1) and the minimum or Q1−1.5(Q3−Q1), respectively. The filled squares represent outliers.



Supplementary Figure 13. Summary of variability in FISH copy number estimates between tumor and adjacent normal glands based on the Shannon Index. In order to determine whether the estimates of heterogeneity from the FISH data were biased as a result of using 5 μm thick sections (the diameter of colorectal tumor cells may be as large as 10 μm), we compared the number of aberrant events in tumor glands with the counts for adjacent normal colon tissue. As shown in the boxplots, estimates of ITH based on the Shannon Index differ significantly (t-test, p-value<0.01) between the tumor and adjacent normal. For each group, boxplots show the median, limited by the 25th (Q1) and 75th (Q3) percentiles, where whiskers represent the most extreme of the maximum or Q3+1.5(Q3−Q1) and the minimum or Q1−1.5(Q3−Q1), respectively. The open circles represent outliers.



Supplementary Figure 14. Validation of the statistical inference framework using synthetic data. To verify the accuracy of our inference framework we generated a synthetic dataset for 21 tumors with different parameters (σ, μ) using our spatial tumor growth model. The results illustrate that the correct parameter values (red circles) are recovered for the majority of cases, indicating the robustness of this approach.

Supplementary Tables

Supplementary Table 1. Summary of patient clinical information for colorectal adenoma and carcinoma samples. Here, MSI refers to microsatellite instability, whereas MSS indicates microsatellite stability.

Patient	Age	Gender	Stage	Size (cm)	Type	MSI status
K	50	M	1	6	Adenoma	MSS
P	71	M	0	3.5	Adenoma	MSS
S	66	M	0	6	Adenoma	MSS
X	69	F	0	2.5	Adenoma	MSS
M	76	F	2	3	Carcinoma	MSS
N	71	M	1	2.3	Carcinoma	MSS
O	41	M	3	9.5	Carcinoma	MSS
T	85	M	3	5.7	Carcinoma	MSS
U	78	F	2	3.9	Carcinoma	MSS
W	51	M	1	3.4	Carcinoma	MSI
CO	83	M	2	8	Carcinoma	MSS
CA	57	M	2	8.5	Carcinoma	MSS
G	61	M	3	5.3	Carcinoma	MSS
R	51	M	3	3	Carcinoma	MSS
Z	83	M	3	7.5	Carcinoma	MSS

Supplementary Table 2. Summary of the number of CNAs detected within each class per tumor.

Patient	Public	Side-specific	Side-variegated	Variegated	Regional	Unique
Adenomas						
K	5	3	0	0	0	2
S	2	0	0	0	0	0
P	0	0	0	0	0	3
X	1	0	0	0	0	6
Carcinomas						
M	1	0	23	2	4	10
N	14	0	2	0	1	2
O	25	2	2	1	2	7
T	25	3	0	0	0	9
U	20	2	0	1	0	11
W	6	0	0	0	0	0
CA	0	0	1	0	1	5
CO	12	6	2	0	1	5
G	17	2	0	0	3	3
R	18	4	3	0	14	5
Z	25	16	0	0	2	2