SUPPLEMENTARY METHODS

Patient material selection criteria

The search term was "metastatic adenocarcinoma" and the years included were 2000 through 2011 (*N* =3823). To be included for further analysis, the patients had to be deceased, have detailed clinical data on primary cancer, axillary metastasis as well as distant metastasis available, and enough paraffin embedded material to enable exome sequencing, gene expression and immunohistochemical stains from each site. Core and fine needle biopsies were not eligible for inclusion. In total twenty patients met the criteria. Formalin fixed paraffin embedded (FFPE) tissue sections were retrieved from all lesions. From the majority of primary cancers and metastases, multiple tumor areas of different topography were isolated (>5 mm distance from each other) resulting in 104 samples. Five metastatic samples (two samples of bone relapse in patient 6, one region of local recurrence sample in patient 13 and two samples of bone relapse in patient 12) failed during exome sequencing due to insufficient DNA, resulting in a total of 99 samples.

Tissue microarray (TMA) and IHC staining

FFPE sections were conditioned in CC1 solution (Ventana Medical Systems, Tucson, AZ, USA) for 36 min (Ki67) to 64 min (PR) and incubated with mouse monoclonal antibodies for Ki67 (clone Mib-1) (Dako A/S, Glostrup, Denmark) and rabbit monoclonal primary antibodies (Ventana) for ER (clone SP1), PR (clone 1E2), and HER2 (clone 4B5) at 35 °C (HER2) or 37 °C (others) for 16 min (Ki67) to 44 min (ER) according to the manufacturer's instructions, and finally counterstained with hematoxylin. Two independent pathologists (NFM and GS) at Karolinska Institutet performed scoring of ER, PR, HER2 and Ki67 and the consensus values were used to determine IHC-based surrogate subtype for each cancer

sample. The assessments of ER, PR, HER2 and Ki67 IHC were combined into surrogate subtypes using definitions recommended by expert recommendations (1-3). For a laboratory specific threshold for Ki67 in Tissue Microarray (TMA) specimens, we incorporated digital image analysis of a previously published cohort (*n*=130) of consecutive cancer specimens collected at the Department of Pathology, Uppsala University Hospital, Uppsala, Sweden from January 1 1987 through December 31 1989 (4-7). Surrogate subtype classification based on IHC is illustrated below:

Luminal A-like: ER ≥ 1 % and PR ≥ 20 %. HER2 "negative" and Ki67 < 4.1 %.

Luminal B-like: ER ≥ 1 % or PR ≥ 1 % and HER2 "negative" and Ki67 ≥ 4.1 %, or

ER ≥ 1 % or PR ≥ 1 % and HER2 "positive" Any Ki67or ER ≥ 1 % and PR < 20 % and HER2 "negative". Any Ki67.

HER2-enriched-like: ER < 1 % and PR < 1 %. HER2 "positive". Any Ki67.

Basal-like: ER < 1 % and PR < 1 %. HER2 "negative". Any Ki67.

PAM50 molecular subtyping after subgroup-specific gene-centering

PAM50 molecular subtyping (8) of each tumour sample was performed after subgroup-specific gene-centering (9). The population- based Stockholm cohort with primary breast cancer patients (10) (GEO:GSE1456) was used as training cohort. The subgroup of patients with breast cancer relapse within the first five years was used to mimic the tumour progression cohort. All molecular subtype analysis was done in R/Bioconductor.

The PAM50 centroids and Entrez Gene IDs in the pam50 data object in the package genefu was used. The hgu113a.db and hgu133b.db annotation packages were used for the Stockholm data and 49/50 PAM50 genes had mapped probesets on the Affymetrix HG-U133A and HG-U133B arrays. For probesets that were present on both arrays, the average value was used. For probesets that were mapped to the same Entrez Gene ID, the one with highest

interquartile range was selected. For each PAM50 gene, the subgroup-specific percentile of the global median in the training cohort was identified. The value 50 (i.e. the median) was imputed for the one gene (KRT17) where gene-expression data was missing.

In the tumour progression cohort, all PAM50 genes have mapped probesets on the Affymetrix Human Transcriptome Array (HTA) 2.0 platform (GEO:GPL17586) as given by the manufacturer's annotations. Again, for probesets that were mapped to the same Entrez Gene ID, the one with highest interquartile range was selected. The baseline expression of each gene was assigned at the subgroup-specific percentile of the breast samples in the tumour progression cohort (median aggregated by patient). Thereafter expression data for each sample was gene-centered by subtracting the baseline expression.

For each sample in the tumour progression cohort data, the Spearman's rank correlation between the sample after subgroup-specific gene-centering and each of the five PAM50 subtype centroids was calculated and the class of the most highly correlated centroid was assigned to the sample. Finally, a stringent criterion of nearest centroid correlation coefficient, larger than 0.25, was applied to assign a final subtype classification.

Main assumptions in Dollo parsimony

We used a variant of parsimony-based phylogenetic reconstruction method named Dollo parsimony to reconstruct phylogenetic tree for each patient. We used *Rdollop()* from R package Rphylip, which uses the implementation "dollop" given in PHYLIP version 3.696. Following are the main assumptions in Dollo parsimony:

1. We know the state of each ancestral site (in germline) to be 0.
2. The sites (mutations) evolve independently.
3. Each lineage in the phylogenetic tree evolve independently of each other.
4. Probability of acquiring a mutation, i.e., changing from state 0 to 1 is small.

5. Probability of a losing a mutation (a deletion), i.e., changing from state 1 to 0 is also small, but still far greater than the probability of acquiring a mutation.

Validation of phylogenetic trees

We validated the phylogenetic trees produced by Dollo parsimony using two approaches. First, we performed phylogenetic reconstruction by an orthogonal method "LICHeE" v1.0 (60), We used the following parameters: -minVAFPresent 0.05 -minClusterSize 10 -maxClusterDist 0.25 -maxVAFAbsent 0.

Second, we validated that the phylogenetic trees are not affected by variable coverage and/or different tumor purity between samples. We adapted a modified approach from Yates et al (16) to identify and remove mutations whose presence or absence in any sample from a patient is indeterminate due to either read coverage or lower tumor purity, i.e., they can be missed by chance. Then, phylogenetic trees were reconstructed using Dollo parsimony after removing all indeterminate mutations. Supplementary Figure 11 contains a side-by-side comparison of the trees in each patient.

To identify indeterminate mutations, we computed the upper 95% confidence interval (CI) of VAF for each absent mutation in each sample according to the binomial distribution. If the upper 95% CI exceeded a threshold VAF_{thr}, the mutation was marked as indeterminate. VAF_{thr} is defined for each mutation as the maximum observed VAF for that mutation in other samples from the same patient multiplied by the ratio between tumor purity in the sample having the maximum VAF and tumor purity of the considered sample. Although this approach does not take into account copy number information and assumes similar underlying cancer cell fraction, we believe that it removes majority of mutations that have ambiguous placement in phylogetentic trees. Binomial confidence intervals were computed according to the "bayes" method using *binom.confint()* function in binom package in R.

Subset analysis to validate the robustness of phylogenetic inference

Intratumor heterogeneity in the primary cancer (11) can complicate the inference of seeding origin of metastases. In order to ameliorate this effect, we sequenced multiple primary blocks in some patients which demonstrated, for instance in patient 4, how different primary regions seeded different metastases (Fig. 3b). However, on the other hand, this also raises the question whether the number of primary samples sequenced affects the inference of progression model. This is termed as incomplete taxon sampling problem in phylogenetic inference. In order to show that Dollo parsimony is robust to this problem, we performed subset analysis for the following two cases.

i. In case of patient 4, a parallel progression case, where we have 6 primary samples, taking all 62 possible subsets of primary samples with three metastases and estimating the probability of linear progression. A case where we observe lower probability of linear progression in each subset will ultimately support a higher probability for the existing inference of parallel progression.

ii. In case of patient 5, a linear progression case where we have 2 primary samples, taking the 2 possible subsets of primary samples with two metastases and estimating the probability of linear progression. A case where we observe higher probability of linear progression in each subset will ultimately support the existing hypothesis reported in the manuscript.

We used the following method to infer the probability of linear progression. We reconstructed 1000 bootstrap trees from available subset of samples as described in the Methods. Then, for each of the bootstrap tree, we used the separating property to test whether any of the primary samples is blocking the path among the metastases. If blocking, we have a NO result for

linear progression; if not blocking, we have a YES result for linear progression. Finally we combined the results across all the 1000 trees to estimate the probability of linear progression.

Subset analysis for patient 4

In patient 4, we have 62 possible subsets. This includes 6 possible subsets containing 1 primary sample, 15 possible subsets containing 2 primary samples, 20 possible subsets containing 3 primary samples, 15 possible subsets containing 4 primary samples, and 6 possible subsets containing 5 primary samples. The results are given in Supplementary Table 8. We observe from the results that, across all possible 62 subsets, we obtain either zero or almost zero probability that all three metastases are seeded in a linear fashion. This confirms that the primary tumor has seeded at least two or all three metastases in parallel, which is in line with the results for all samples taken together (Fig. 3).

Next, we take into account the paired metastases cases (Uterus to Brain, Uterus to Colon, and Brain to Colon) where, for a metastases pair, the earlier metastasis has seeded the latter metastases in a linear fashion. For Uterus to Brain pair, we see only 1/62 case with more than 50% probability meaning that 98% of the subsets support Uterus did not seed the Brain metastases. For the rest of two possible cases (Uterus-Colon, and Brain-Colon), there is not a single case with a probability of 50% or higher of linear progression in any pair. Overall, the subset analysis supports the parallel progression model for patient 4.

Subset analysis for patient 5

In patient 5, we have 2 possible subsets where we take one primary sample each with the two bone metastases. The results are given in Supplementary Table 9. We observe from the results that the probability of linear progression in both subsets is almost 100% which supports the results when full data is used for inferring the progression model.

In summary, we see that Dollo parsimony is robust to the number and combination of primary samples taken for inferring the phylogenetic tree. This, in turns, means that progression model

inference performed using separating property also does not change when different subsets are considered.

Parameter values used in PyClone

Out of the available three models in PyClone, we used the authors' recommended genotype-aware PyClone-beta-binomial model with all model's parameter values set to recommended values (the rest of the two models are genotype-naive infinite binomial mixture model and infinite beta-binomial mixture model). We tested the robustness of cellular prevalence (CP) cut-off of 0.05 as follows. We set the cellular prevalence cut-off to 0.04 and 0.02 and compared it to CP cut-off of 0.05 to check if the seeding patterns are altered. We observed that, overall, the progression and lymph node seeding results were not changed for CP threshold of 0.04 and they were quite similar for the CP threshold of 0.02 (Supplementary Table 10). Regarding the number of iterations in MCMC, the following criterion was used. If the number of samples in patient were less than 5, 10000 iterations were used; if the number of samples in a patient were between 5 and 7, 15000 iterations were used; if the number of samples in a patient were between 7 and 10, 20000 iterations were used; and if the number of samples were more than 10, 50000 iterations were used. The first 25% percent iterations were thrown as burnin, thereafter every 10th sample was considered, i.e., a thinning value of 10 was used. To test convergence, we ran two independent PyClone analyses for each patient and compared the results. For patient 11, we found that using 15000 iterations for MCMC sampling were not enough for convergence. Subsequently, we used 30000 iterations and observed convergence.

Mutational Signatures

We extracted a number of signatures ranging between 2-10 with five repetitions, and computed the residuals sum of squares (RSS) and the explained variance between the

observed profile and fitted spectrum for different number of signatures. The final number of signatures (four) was decided based on the first inflection point when plotting RSS and explained variance change with number of signatures (Supplementary Fig. 7a). The accuracy of the fitted signatures is dependent on the number of samples used for extraction. To allow higher accuracy of fitting, we merged our cohort with an external in-house cohort of primary breast cancers from 129 patients with exome sequencing. The external cohort analysis was performed in a similar pipeline, which excludes potential batch effects.

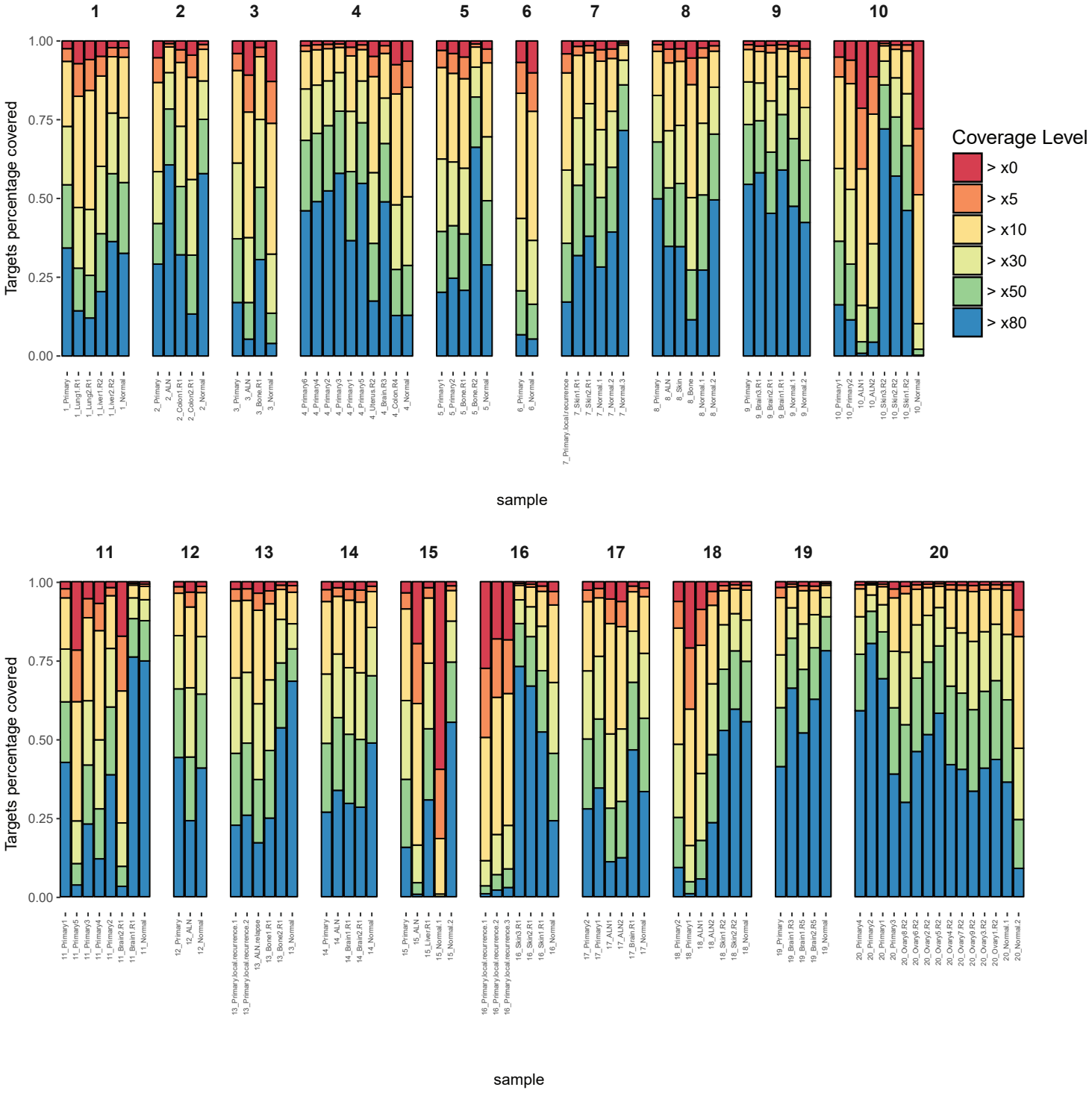
To identify the biological processes underlying each signature, the Euclidean distance was computed between the frequencies of different mutation classes in our four signatures and those in the validated signatures published by Alexandrov et al (12). Based on the shortest Euclidean distance, we were able to reliably map signatures S1 and S2 to the age-associated signature 1 and APOBEC-associated signature 2 from Alexandrov et al respectively (Supplementary Fig. 7b). Signature S3 had a similar distance to several published signatures. We believe that the best candidate for S3 is signature 8 which has an unknown etiology since they share the characteristic of weak strand bias in C>A substitutions (Supplementary Fig. 7c) and since signature 8 was also found in breast cancer. We found that elevated contribution of signature S4 is significantly associated with BRCA1/2 deleterious germline mutations in the external cohort (p-value = 0.0009, Mann-Whitney). Consequently, S4 was mapped to signature 3 in Alexandrov et al (12) which is associated with homologous recombination deficiency.

References

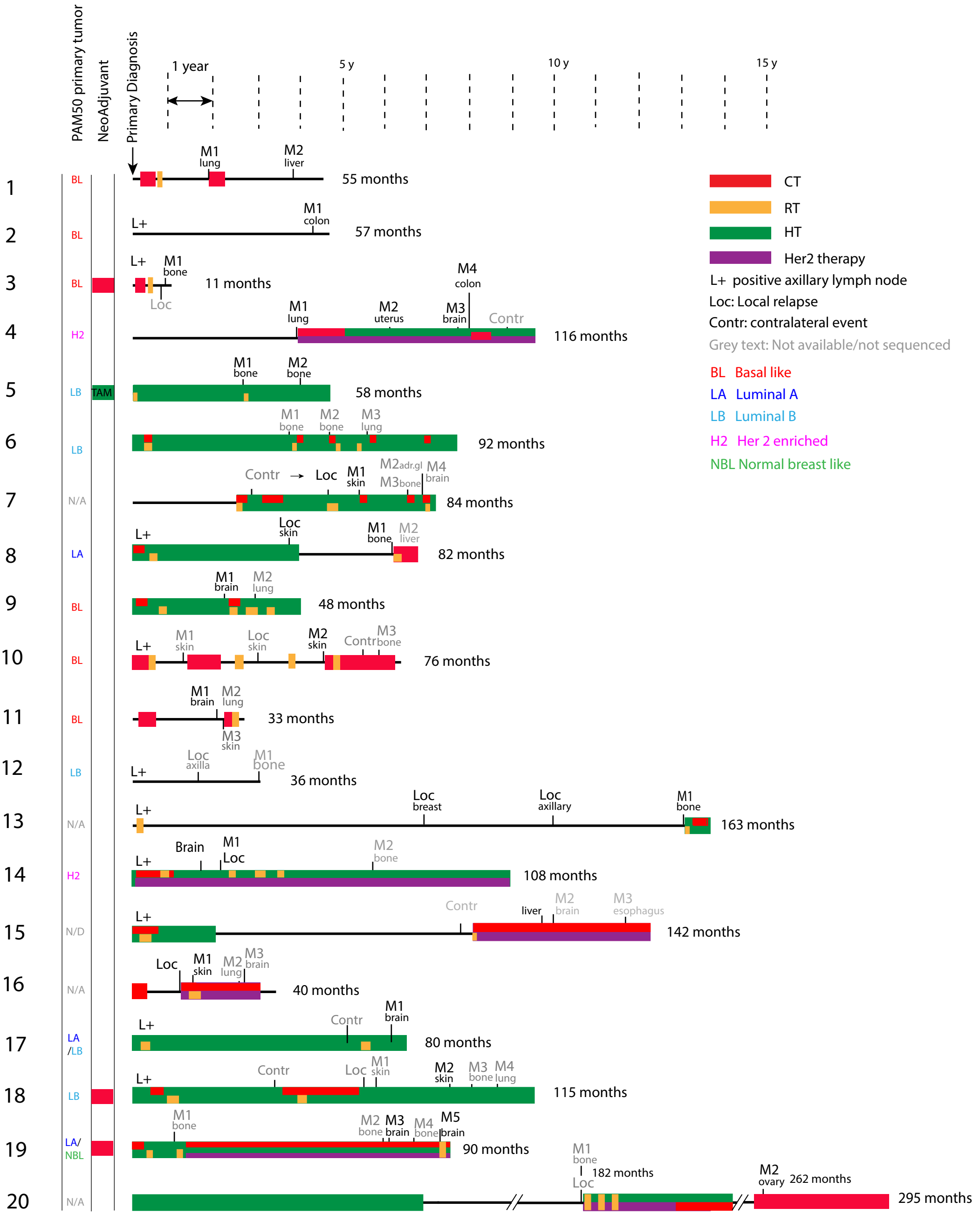
1. Guiu S, Michiels S, Andre F, Cortes J, Denkert C, Di Leo A, Hennessy BT, Sorlie T, Sotiriou C, Turner N, et al. Molecular subclasses of breast cancer: how do we define them? The IMPAKT 2012 Working Group Statement. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO.* 2012;23(12):2997-3006.

2. Coates AS, Winer EP, Goldhirsch A, Gelber RD, Gnant M, Piccart-Gebhart M, Thurlimann B, Senn HJ, and Panel M. Tailoring therapies--improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO.* 2015;26(8):1533-46.
3. Prat A, Cheang MC, Martin M, Parker JS, Carrasco E, Caballero R, Tyldesley S, Gelmon K, Bernard PS, Nielsen TO, et al. Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal A breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* 2013;31(2):203-9.
4. Lindahl T, Landberg G, Ahlgren J, Nordgren H, Norberg T, Klaar S, Holmberg L, and Bergh J. Overexpression of cyclin E protein is associated with specific mutation types in the p53 gene and poor survival in human breast cancer. *Carcinogenesis.* 2004;25(3):375-80.
5. Sjogren S, Inganas M, Norberg T, Lindgren A, Nordgren H, Holmberg L, and Bergh J. The p53 gene in breast cancer: prognostic value of complementary DNA sequencing versus immunohistochemistry. *J Natl Cancer Inst.* 1996;88(3-4):173-82.
6. Linderholm B, Karlsson E, Klaar S, Lindahl T, Borg AL, Elmberger G, and Bergh J. Thrombospondin-1 expression in relation to p53 status and VEGF expression in human breast cancers. *European journal of cancer.* 2004;40(16):2417-23.
7. Stalhammar G, Fuentes Martinez N, Lippert M, Tobin NP, Molholm I, Kis L, Rosin G, Rantalainen M, Pedersen L, Bergh J, et al. Digital image analysis outperforms manual biomarker assessment in breast cancer. *Mod Pathol.* 2016;29(4):318-29.
8. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* 2009;27(8):1160-7.
9. Zhao X, Rodland EA, Tibshirani R, and Plevritis S. Molecular subtyping for clinically defined breast cancer subgroups. *Breast Cancer Res.* 2015;17(29.
10. Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.* 2005;7(6):R953-64.
11. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, Aas T, Alexandrov LB, Larsimont D, Davies H, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med.* 2015;21(7):751-9.
12. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. Signatures of mutational processes in human cancer. *Nature.* 2013;500(7463):415-21.

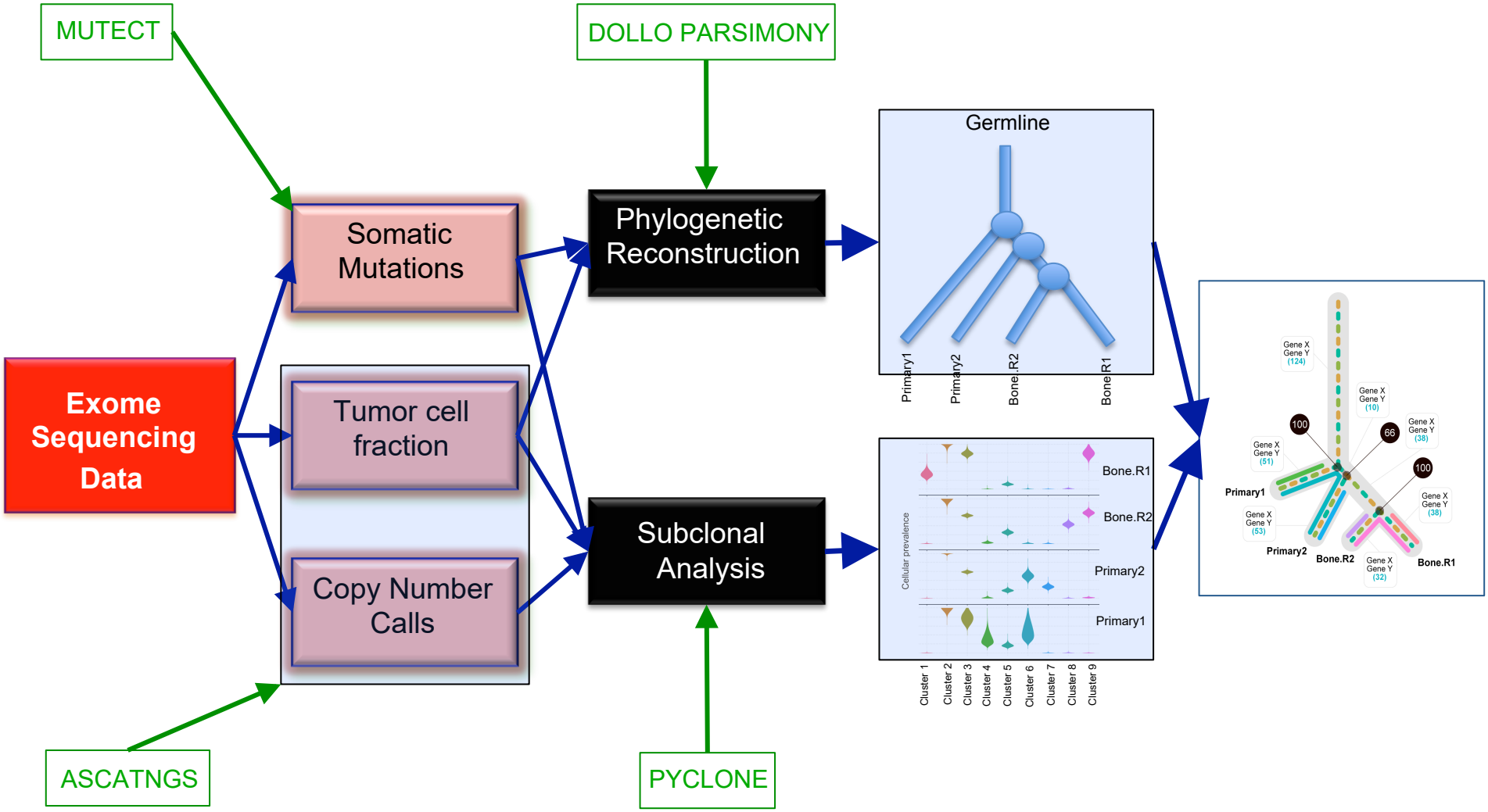
Supplementary Figure 1



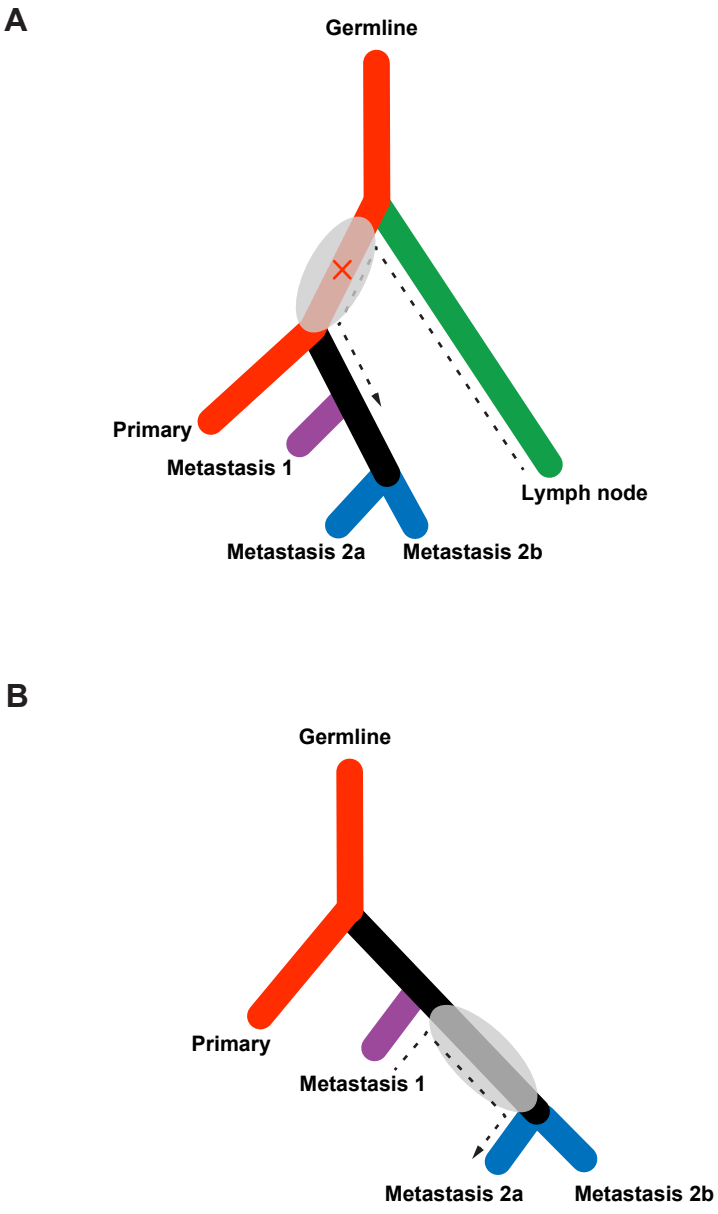
Supplementary Figure 2



Supplementary Figure 3

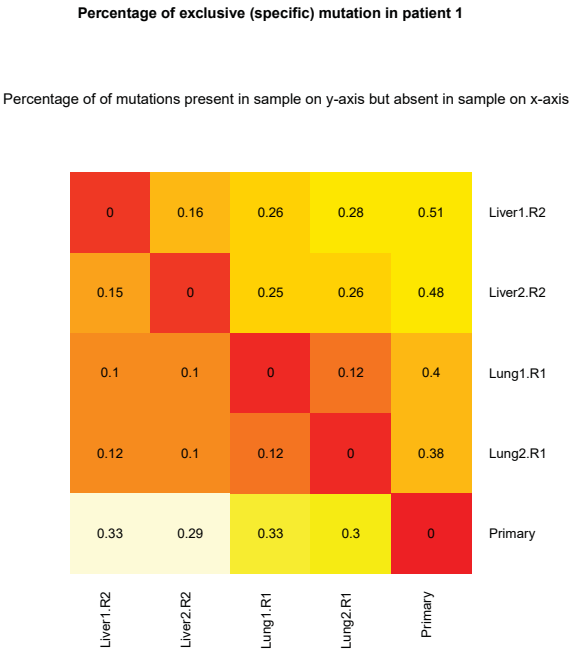


Supplementary Figure 4

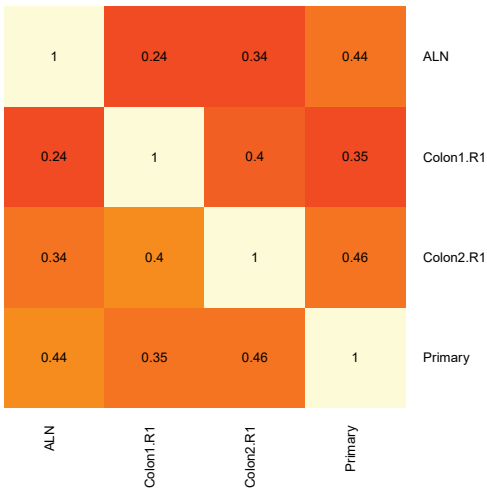


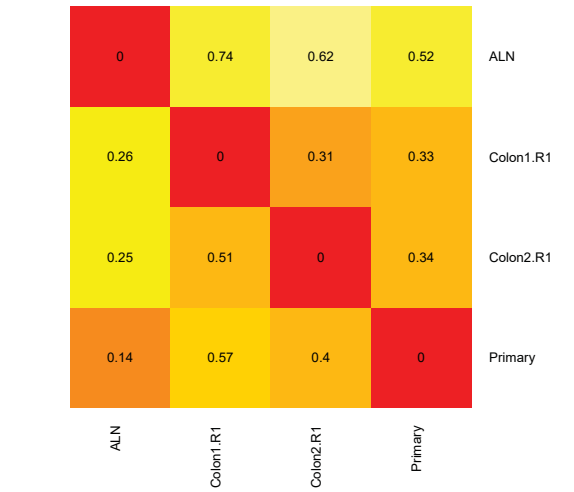
Supplementary Figure 5

A



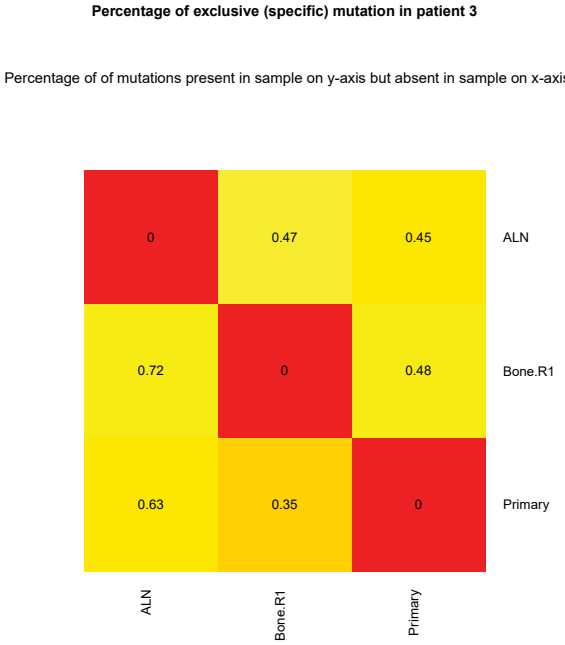
B



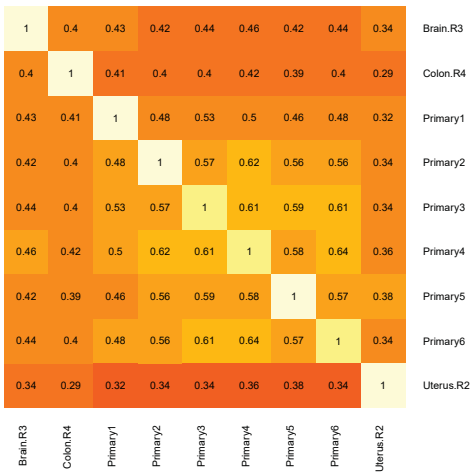


Supplementary Figure 5

C

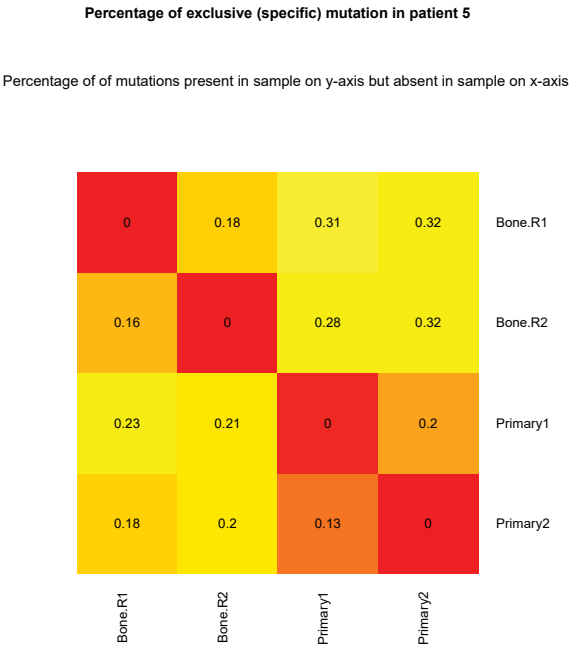


D

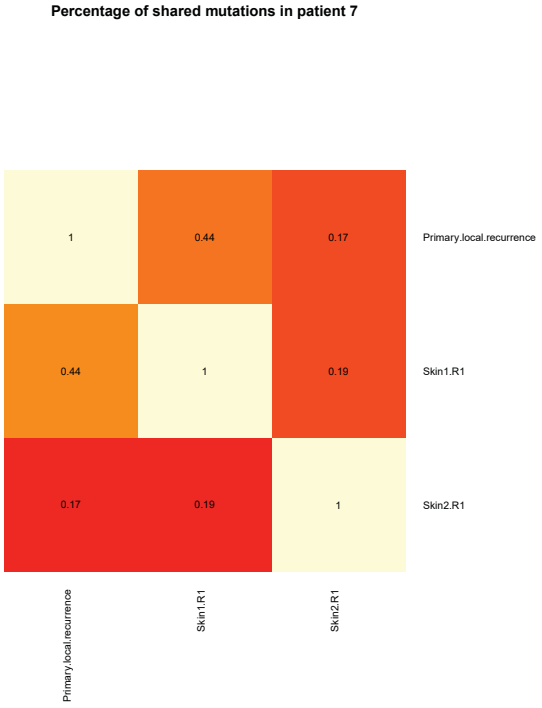


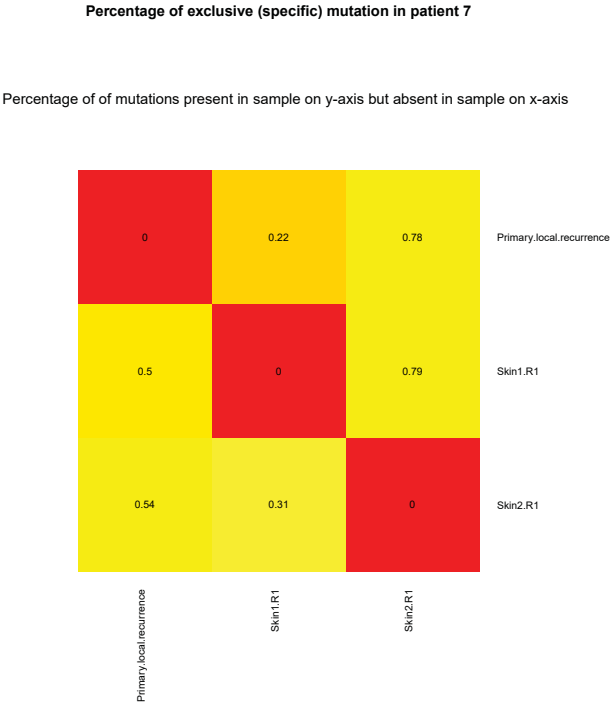
Supplementary Figure 5

E



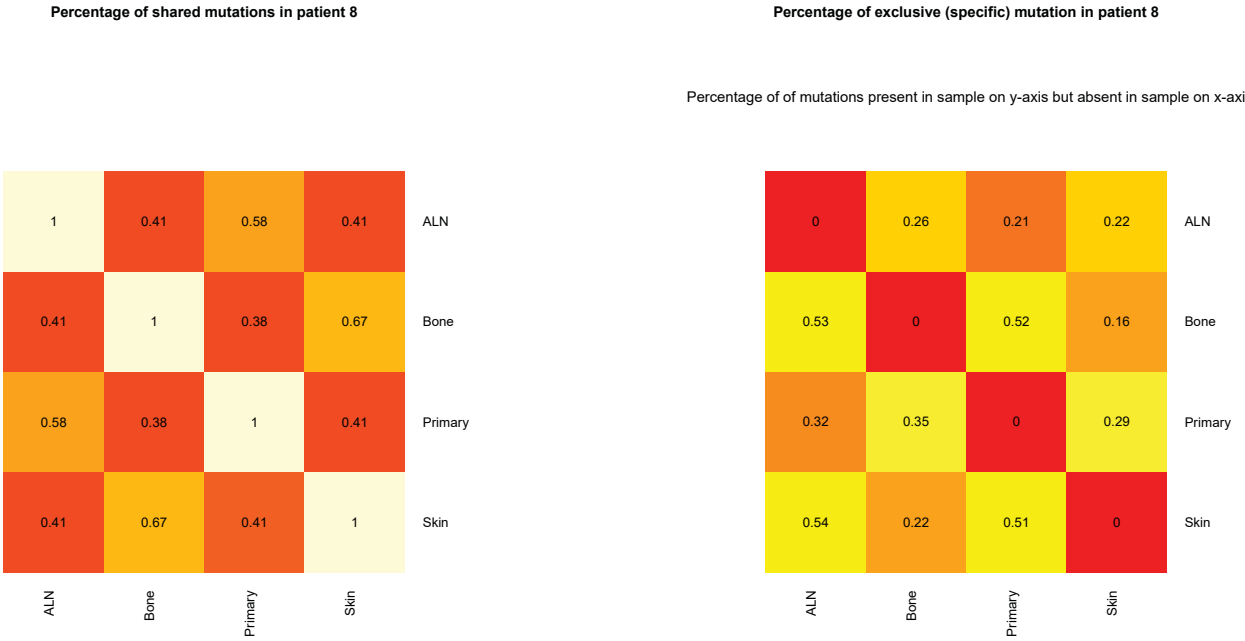
F



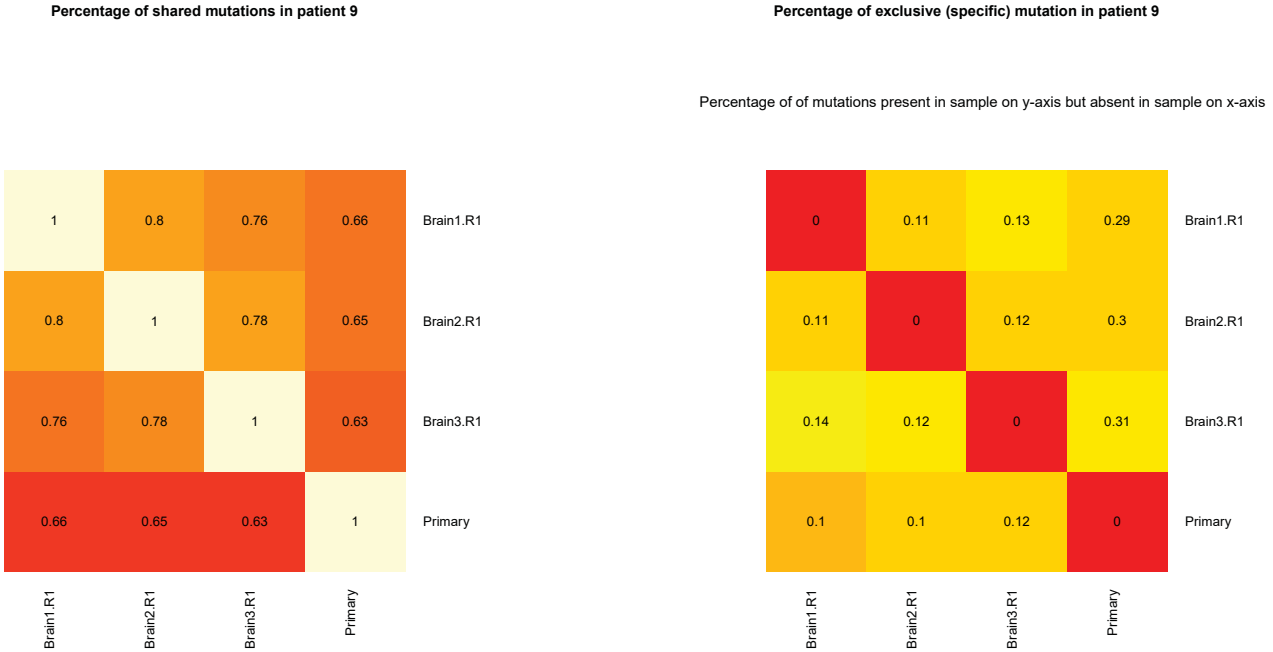


Supplementary Figure 5

G



H



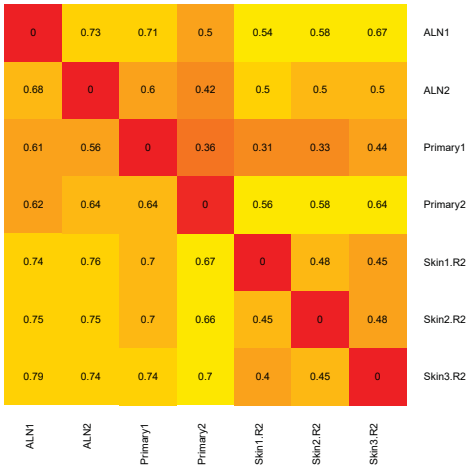
Supplementary Figure 5

I

Percentage of shared mutations in patient 10

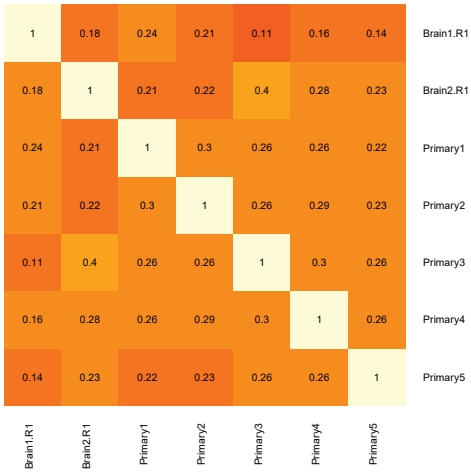
Percentage of exclusive (specific) mutation in patient 10

Percentage of of mutations present in sample on y-axis but absent in sample on x-axis



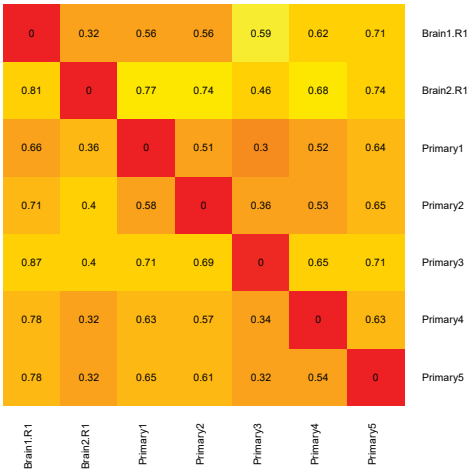
J

Percentage of shared mutations in patient 11



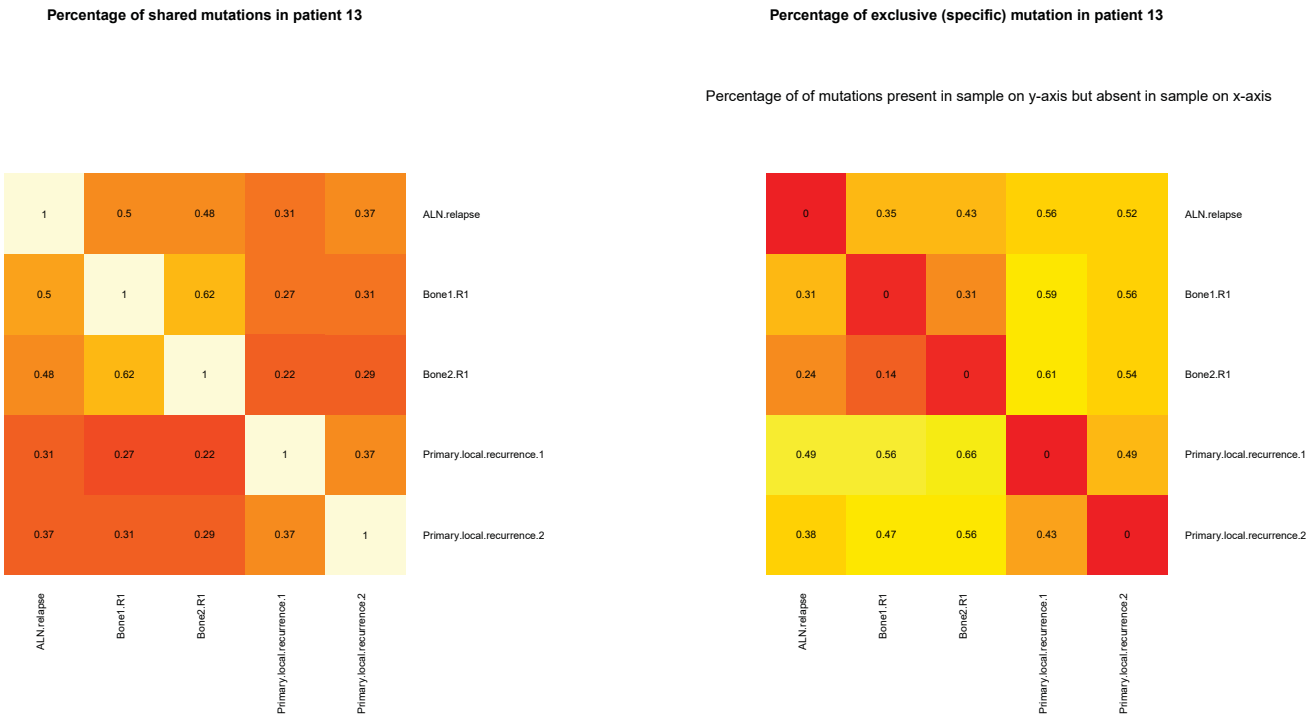
Percentage of exclusive (specific) mutation in patient 11

Percentage of of mutations present in sample on y-axis but absent in sample on x-axis

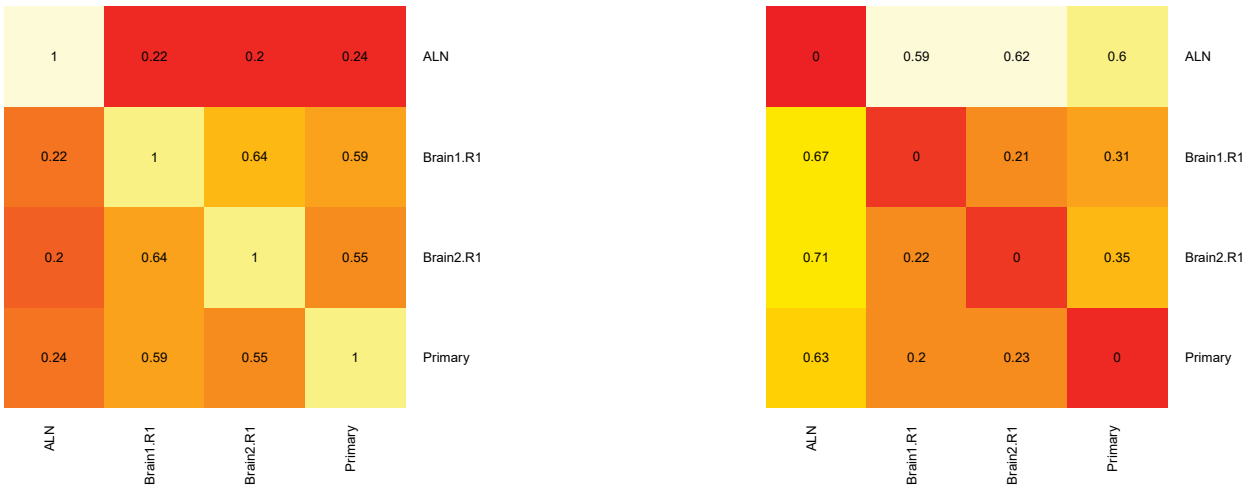


Supplementary Figure 5

K

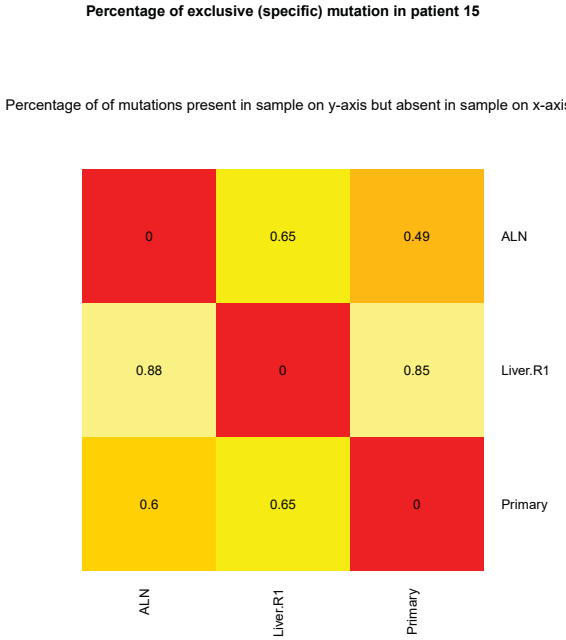


L

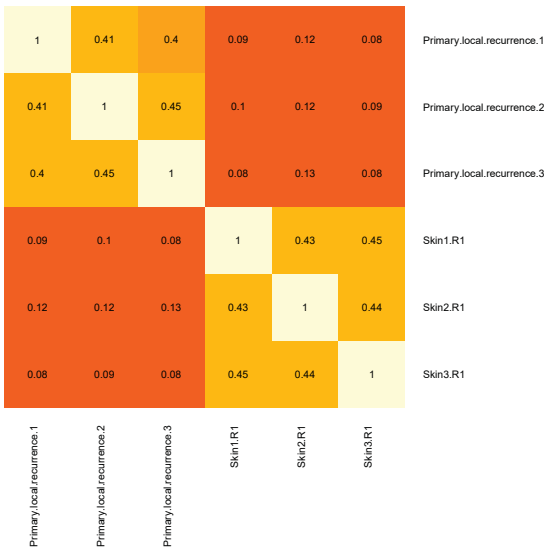


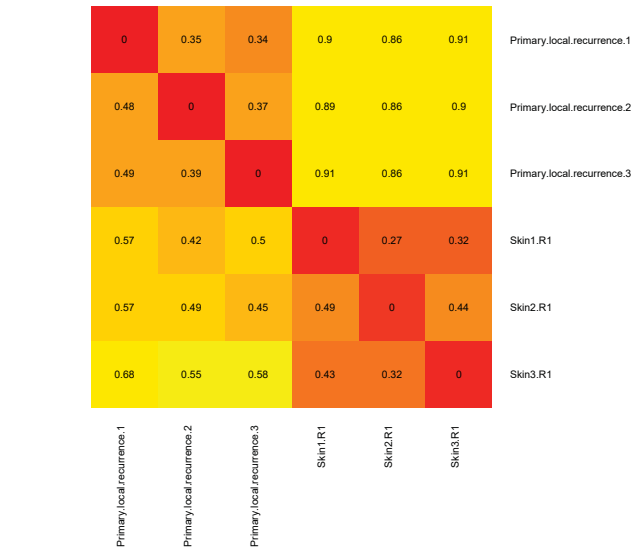
Supplementary Figure 5

M



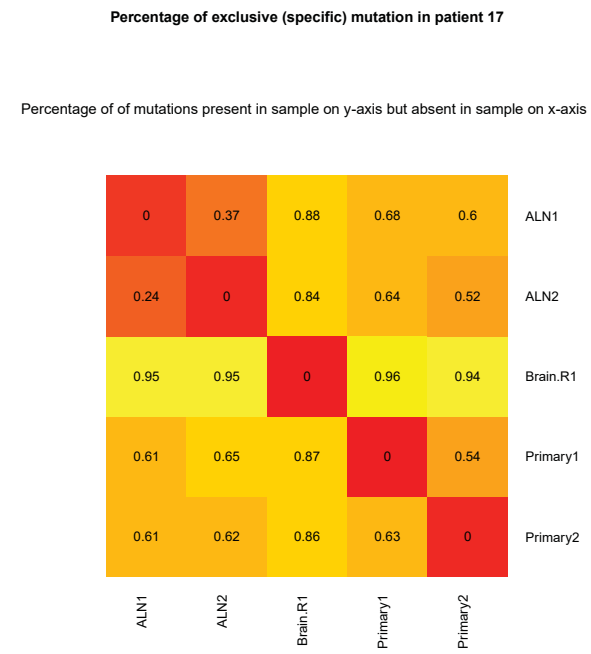
N



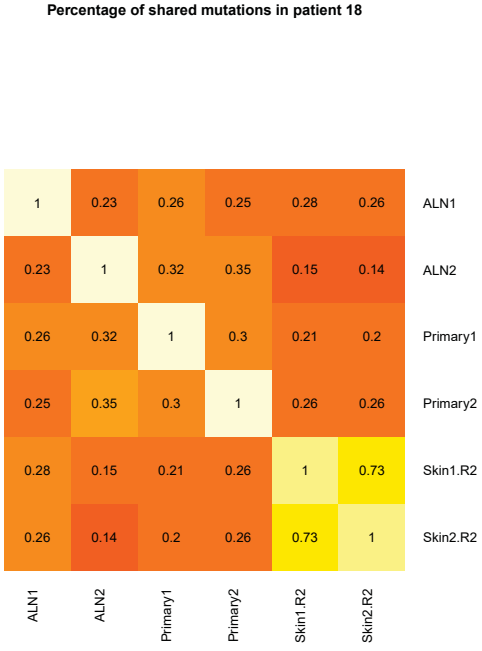


Supplementary Figure 5

O



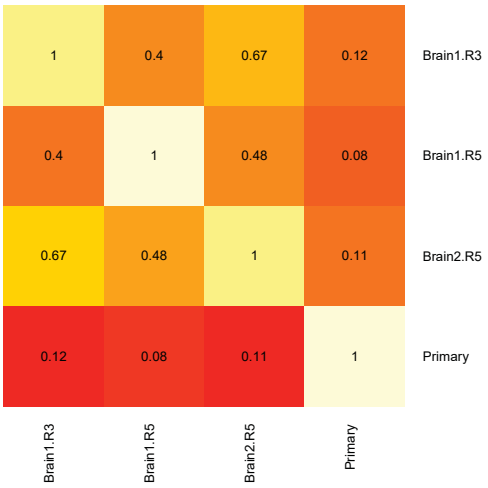
P



Supplementary Figure 5

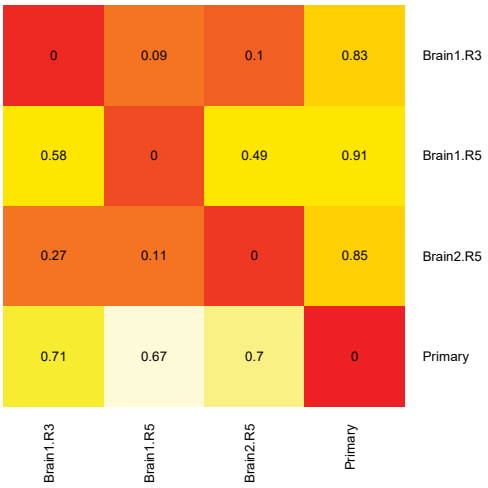
Q

Percentage of shared mutations in patient 19



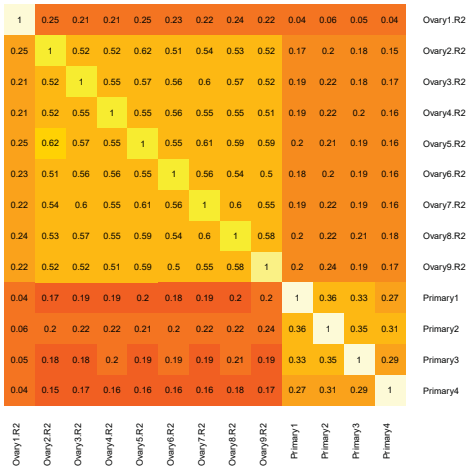
Percentage of exclusive (specific) mutation in patient 19

Percentage of of mutations present in sample on y-axis but absent in sample on x-axis



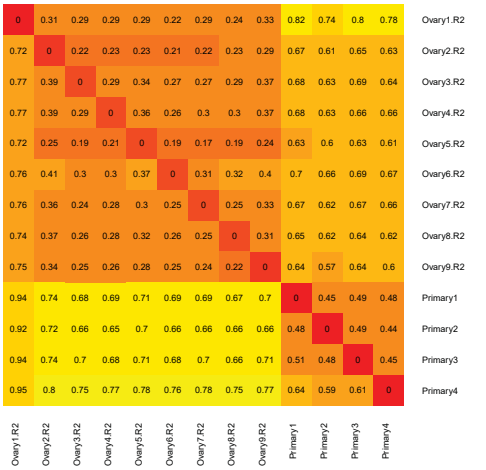
R

Percentage of shared mutations in patient 20



Percentage of exclusive (specific) mutation in patient 20

Percentage of of mutations present in sample on y-axis but absent in sample on x-axis

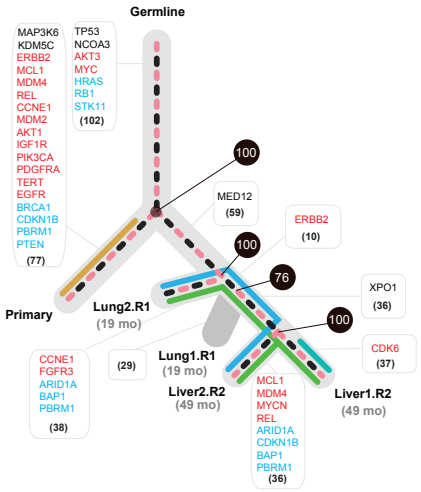


Supplementary Figure 6

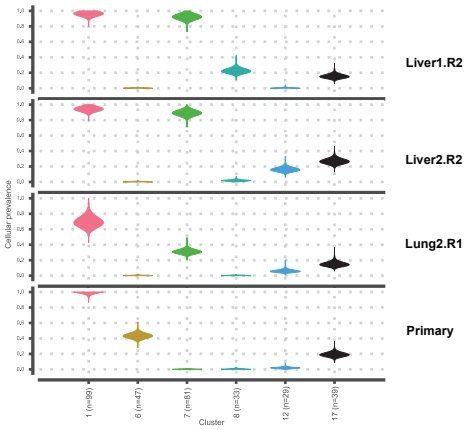
A

Patient 1: ER-/PR-/HER2-

Phylogenetic Tree



Density Plot



Cluster Table

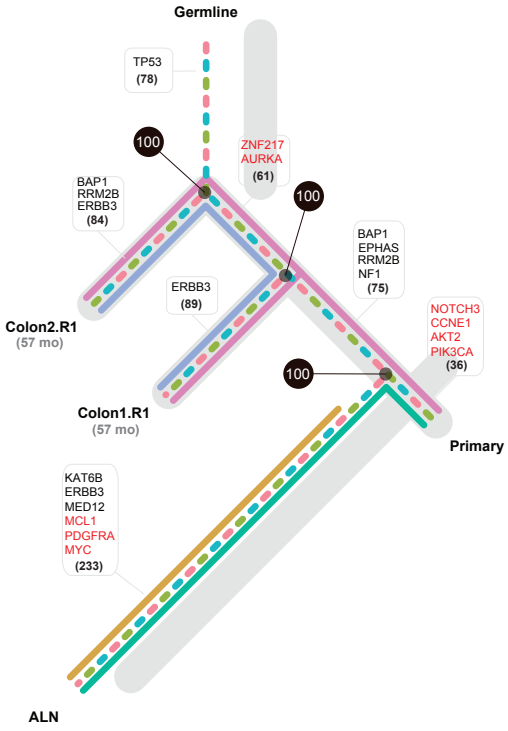
	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		1	99	TP53;NCOA3
		17	39	0
Primary		6	47	MAP3K6;KDM5C
Metastasis		7	81	XPO1;MED12
		12	29	0
		8	33	0

Supplementary Figure 6

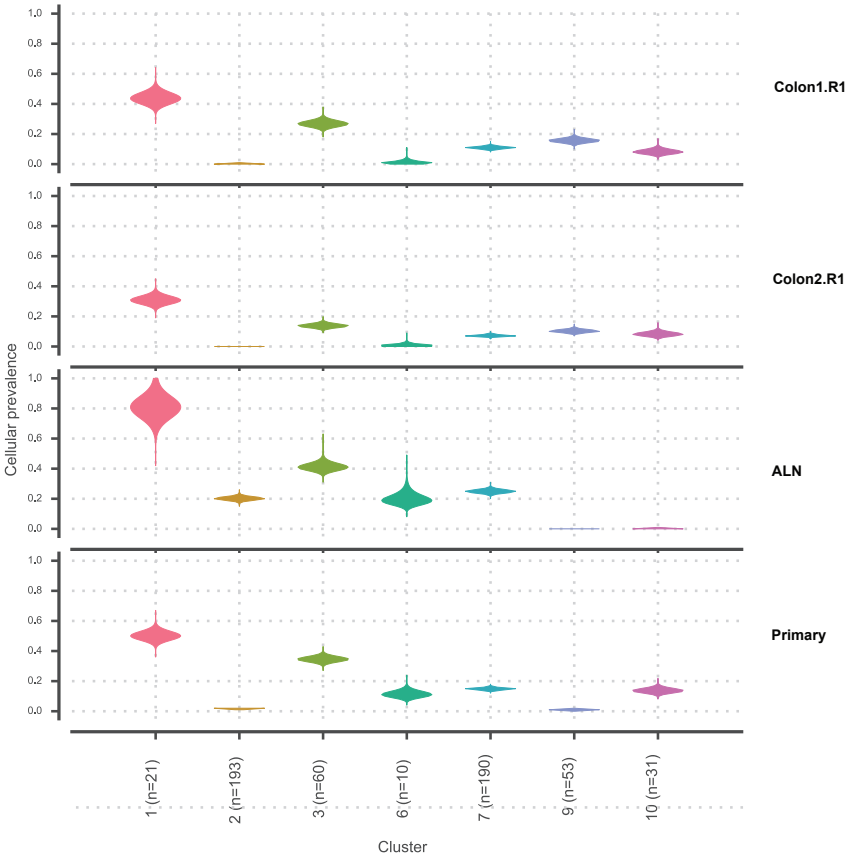
B

Patient 2: ER-/PR-/HER2-

Phylogenetic Tree



Density Plot



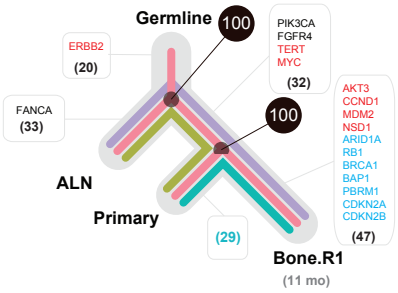
Cluster Table

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		1	21	TP53
		3	60	BAP1
		7	190	ERBB3;NF1;EPHA5; RRM2B
Primary & Lymph		10 (except lymph)	31	0
		6	10	0
Lymph Metastasis		2	193	KAT6B;MED12
		9	53	0

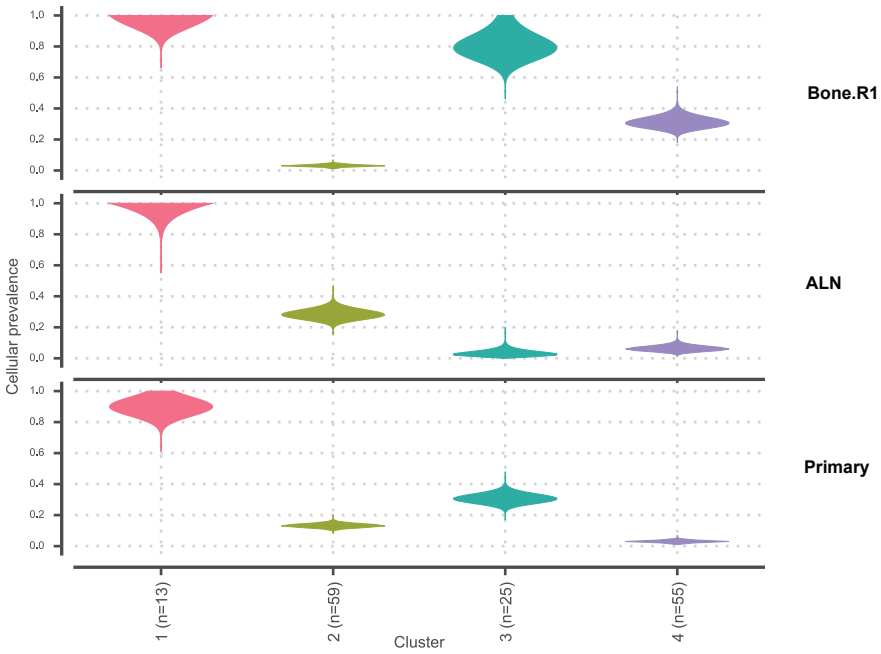
Supplementary Figure 6

C Patient 3: ER-/PR-/HER2+

Phylogenetic Tree



Density Plot



Cluster Table

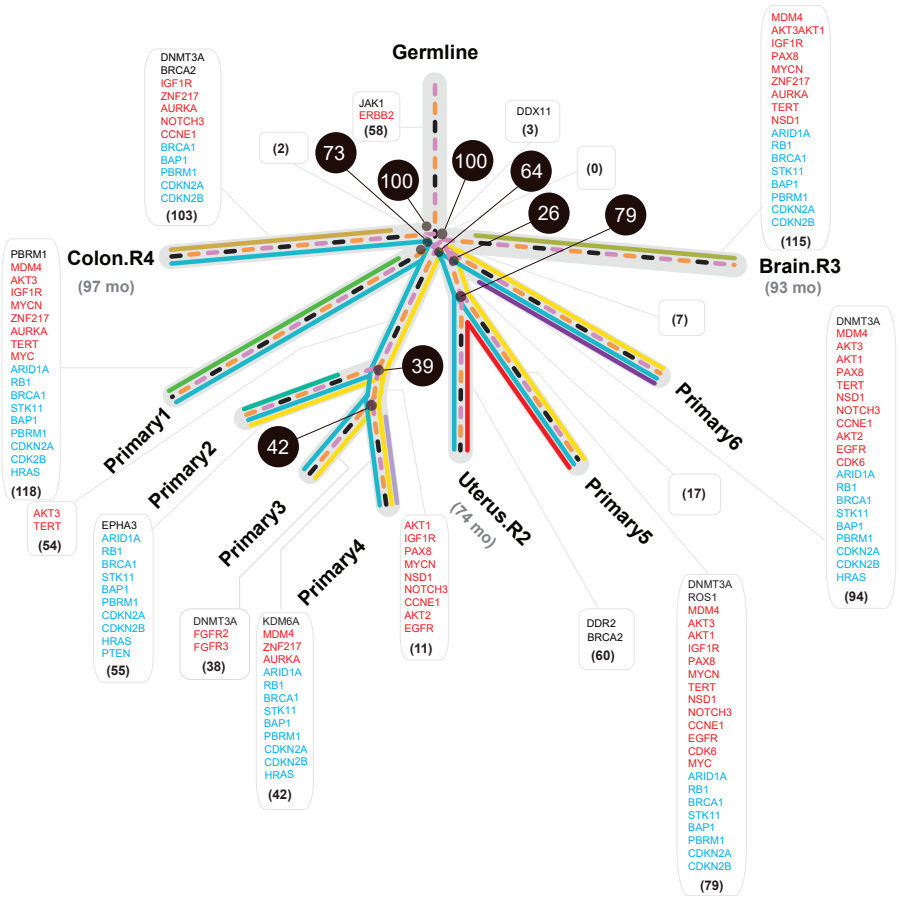
	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		1	13	0
Primary & Lymph		2	59	FANCA
Primary & Metastasis		3	25	PIK3CA;FGFR4
Lymph & Metastasis		4	55	0

Supplementary Figure 6

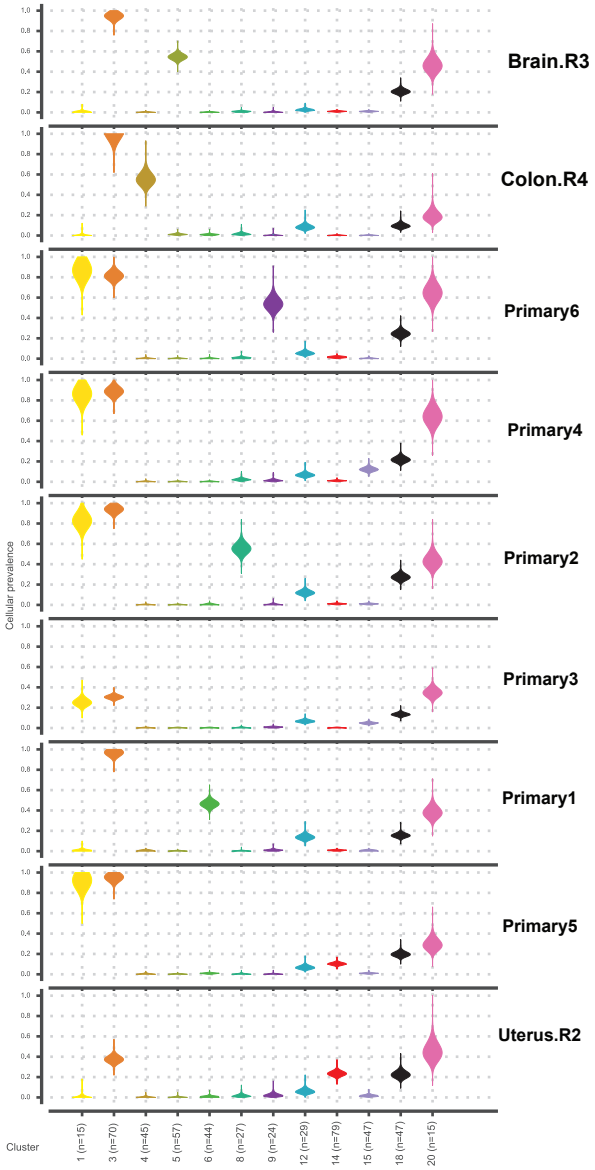
D

Patient 4: ER-/PR-/HER2+

Phylogenetic Tree



Density Plot



Cluster Table

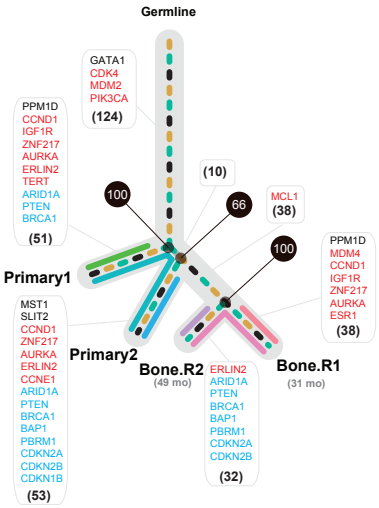
	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		20	15	0
		18	47	DNMT3A
		3	70	JAK1
		12 (except Brain.R3)	29	0
Primary		1 / P6,4,2,3,5	15	0
		6 / P1	44	PBRM1
		8 / P2	27	EPHA3
		15 / P4	47	PLCG1
		9 / P6	24	0
Primary 16 & Uterus.R2		14 / P5, Uterus.R2	79	BRCA2;DDR2;ROS1;KDM6A
Metastasis		4 / Colon	45	BRCA2
		5 / Brain	57	0

Supplementary Figure 6

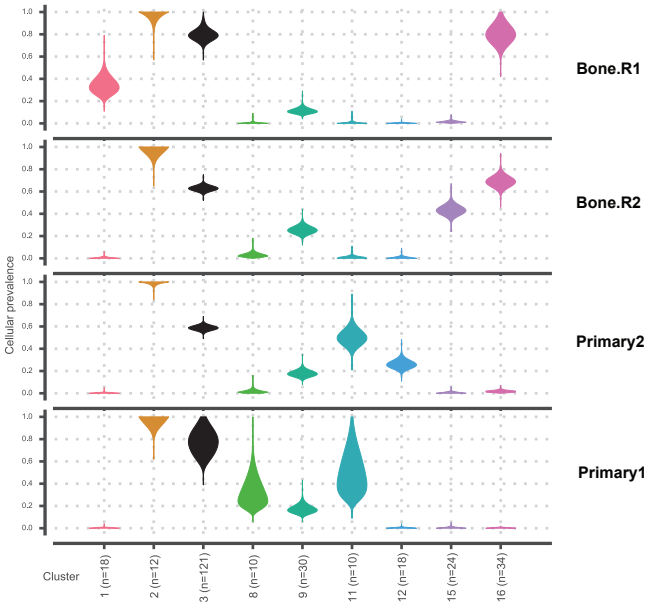
E

Patient 5: ER+/PR+/HER2-

Phylogenetic Tree



Density Plot



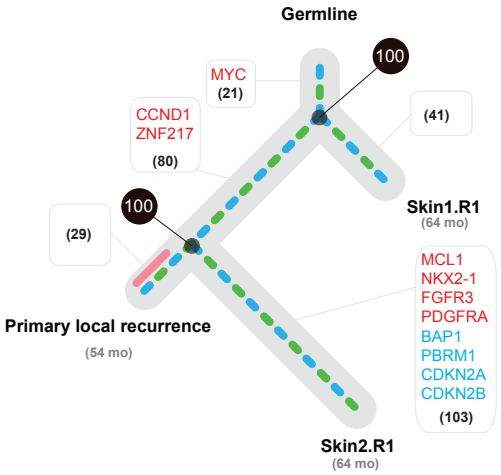
Cluster Table

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		2	12	0
		3	121	GATA1
		9	30	PPM1D
Primary		8	10	0
		11	10	0
		12	18	0
Metastasis		16	34	0
		1	18	0
		15	24	0

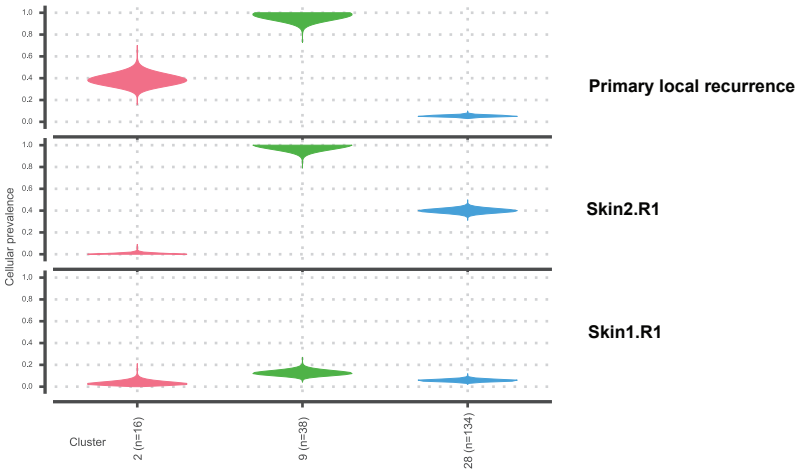
Supplementary Figure 6

F Patient 7: ER+/PR-/HER2- (Skin Metastasis IHC, Primary Tumor data NA)

Phylogenetic Tree



Density Plot



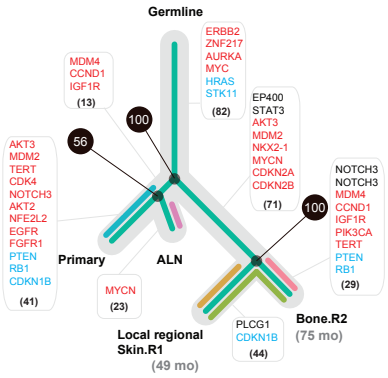
Cluster Table

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		9	38	TP53
		28	134	0
Primary		2	16	0

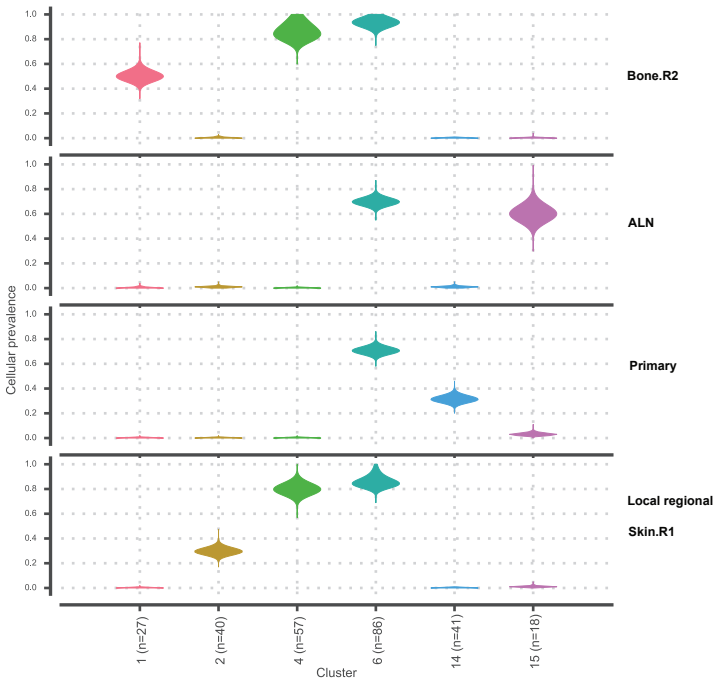
Supplementary Figure 6

G Patient 8: ER+/PR+/HER2- (Based on IHC on axillary Lymph and Bone metastasis

Phylogenetic Tree



Density Plot



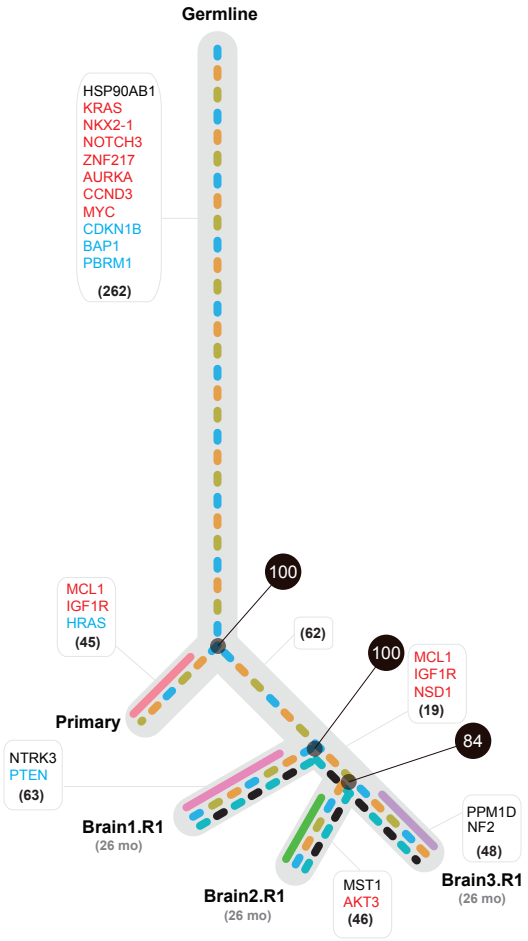
Cluster Table

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		6	86	0
Primary		14	41	0
Lymph		15	18	0
Metastasis		4	57	EP400;STAT3
		2	40	PLCG1
		1	27	NOTCH3

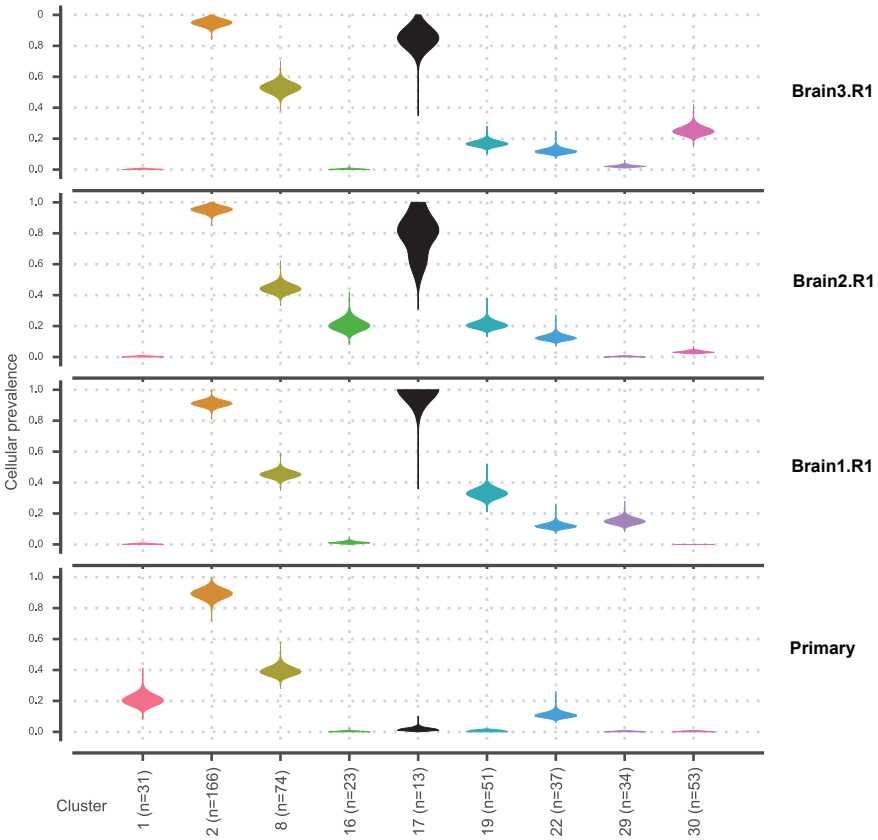
Supplementary Figure 6

H Patient 9: ER-/PR-/HER2-

Phylogenetic Tree



Density Plot



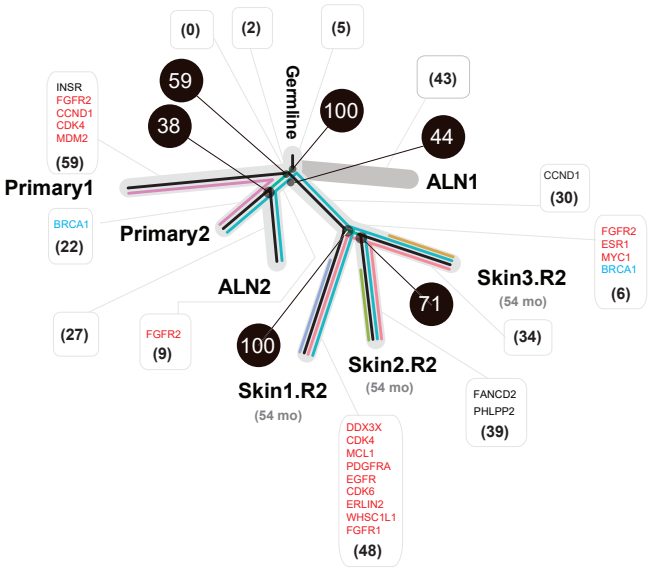
1. Cluster Table

	Cluster color	Cluster ID	Mutation count	
Truncal		2	166	0
		8	74	HSP90AB1
		22	37	0
Primary		1	31	0
Metastasis		30	53	PPM1D;NF2
		29	34	NTRK3
		16	23	MST1
		17	13	0
		19	51	0

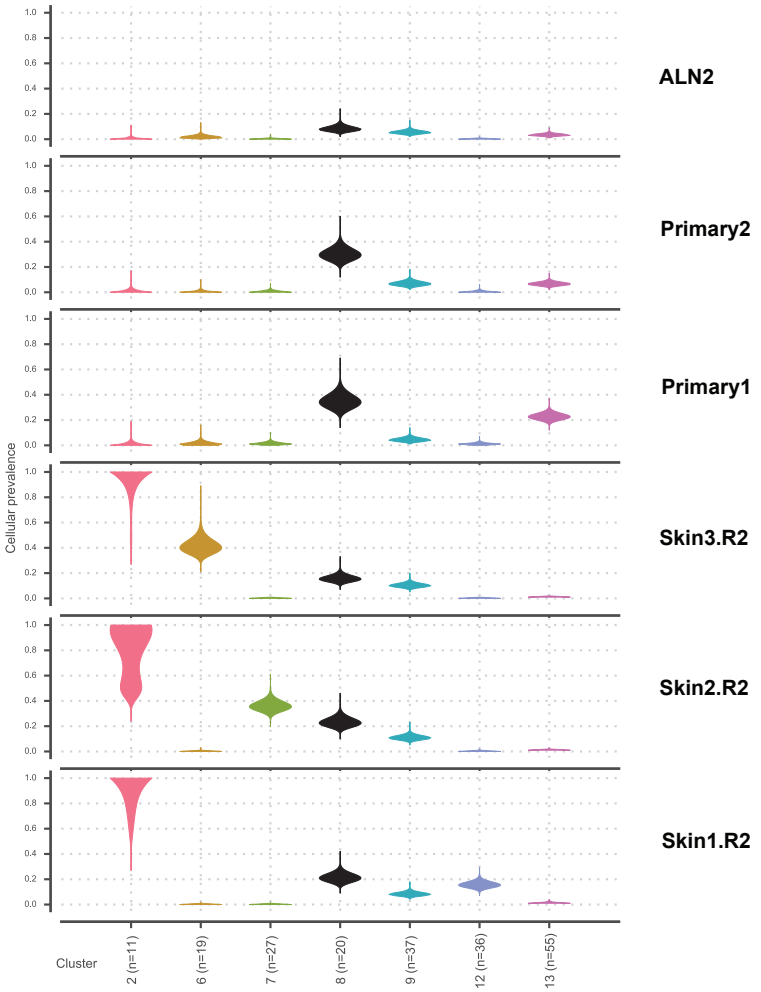
Supplementary Figure 6

I Patient 10: ER-/PR-/HER2-

Phylogenetic Tree



Cluster Table



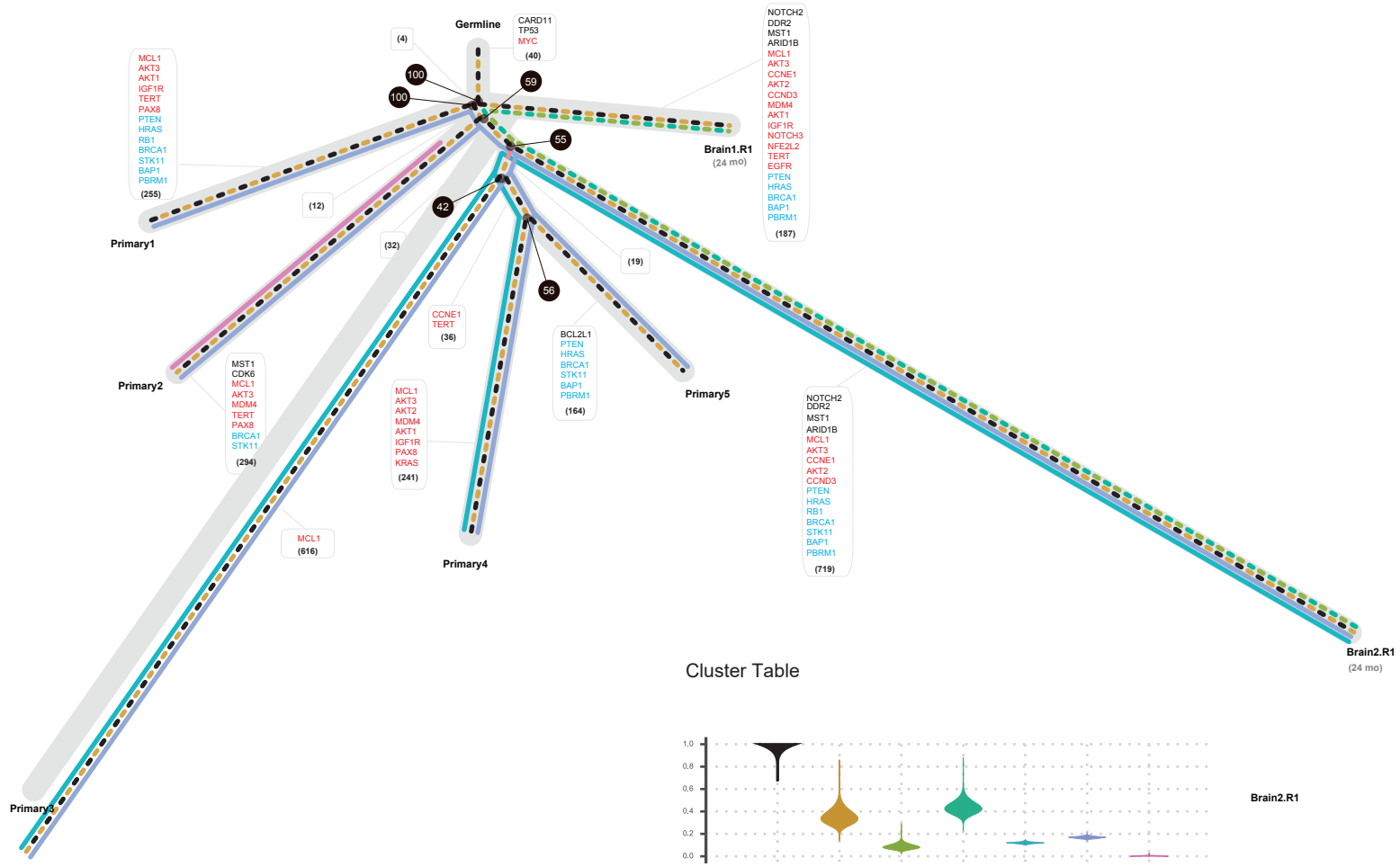
Density Plot

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		8	20	0
		9 (except primary1)	37	FANCD2
Primary		13	55	INSR
Metastasis		2	11	0
		12	36	DDX3X
		7	27	PHLPP2
		6	19	0

Supplementary Figure 6

J Patient 11: ER-/PR-/HER2-

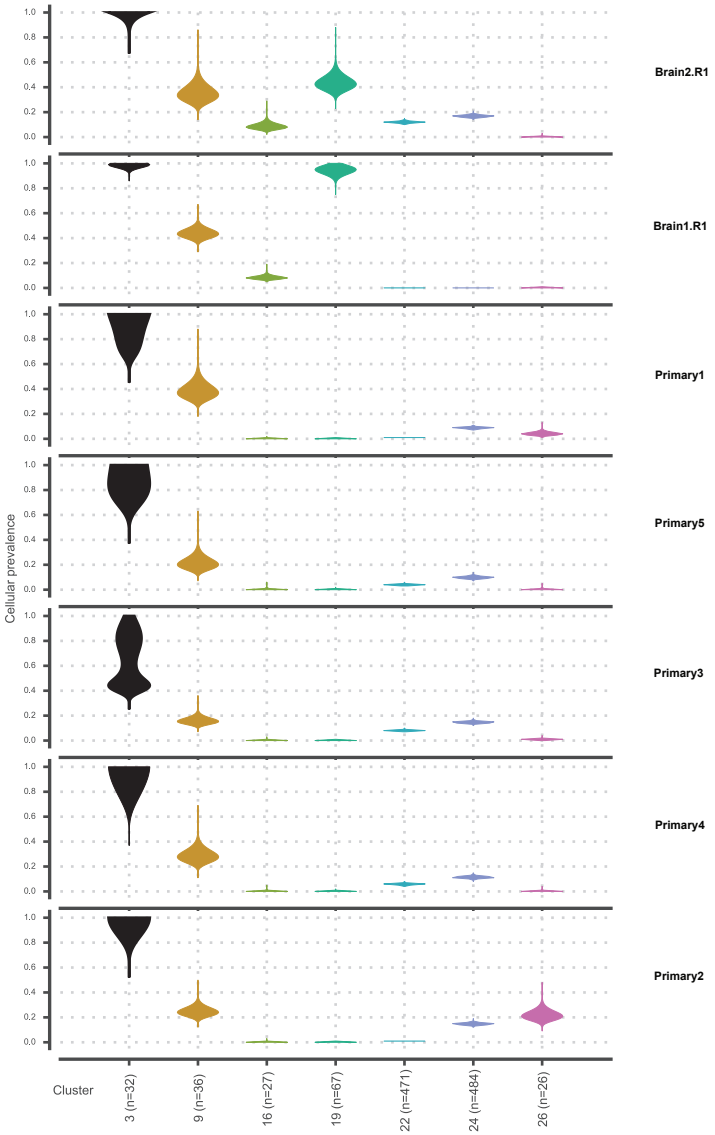
Phylogenetic Tree



Cluster Table

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		3	32	CARD11
		9	36	0
		24 (expect Brain1.R1)	484	0
Primary		26	26	CDK6
Primary & Metastasis		22	471	BCL2L1
Metastasis		19	67	NOTCH2;DDR2
		16	27	0

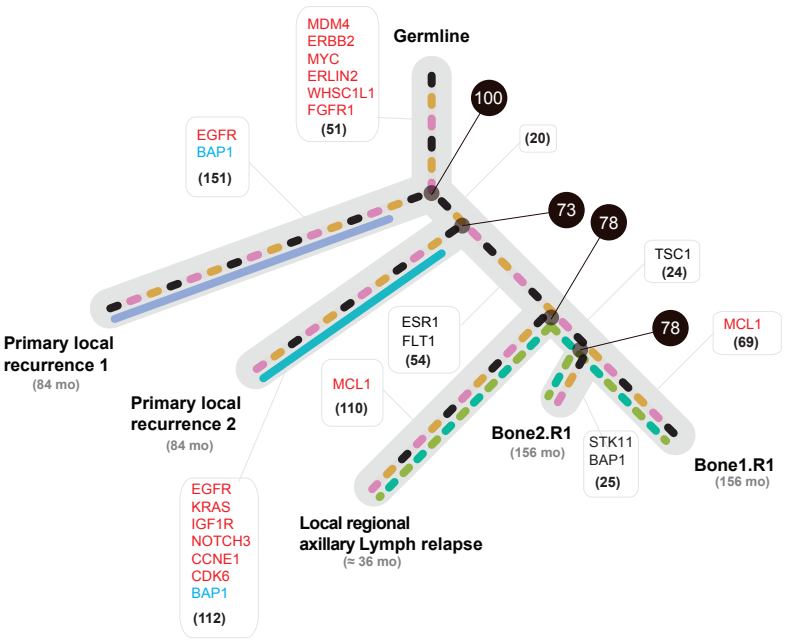
Density Plot



Supplementary Figure 6

K Patient 13: ER+/PR+/HER2- (Based on Bone Metastasis IHC, Primary Tumor data NA)

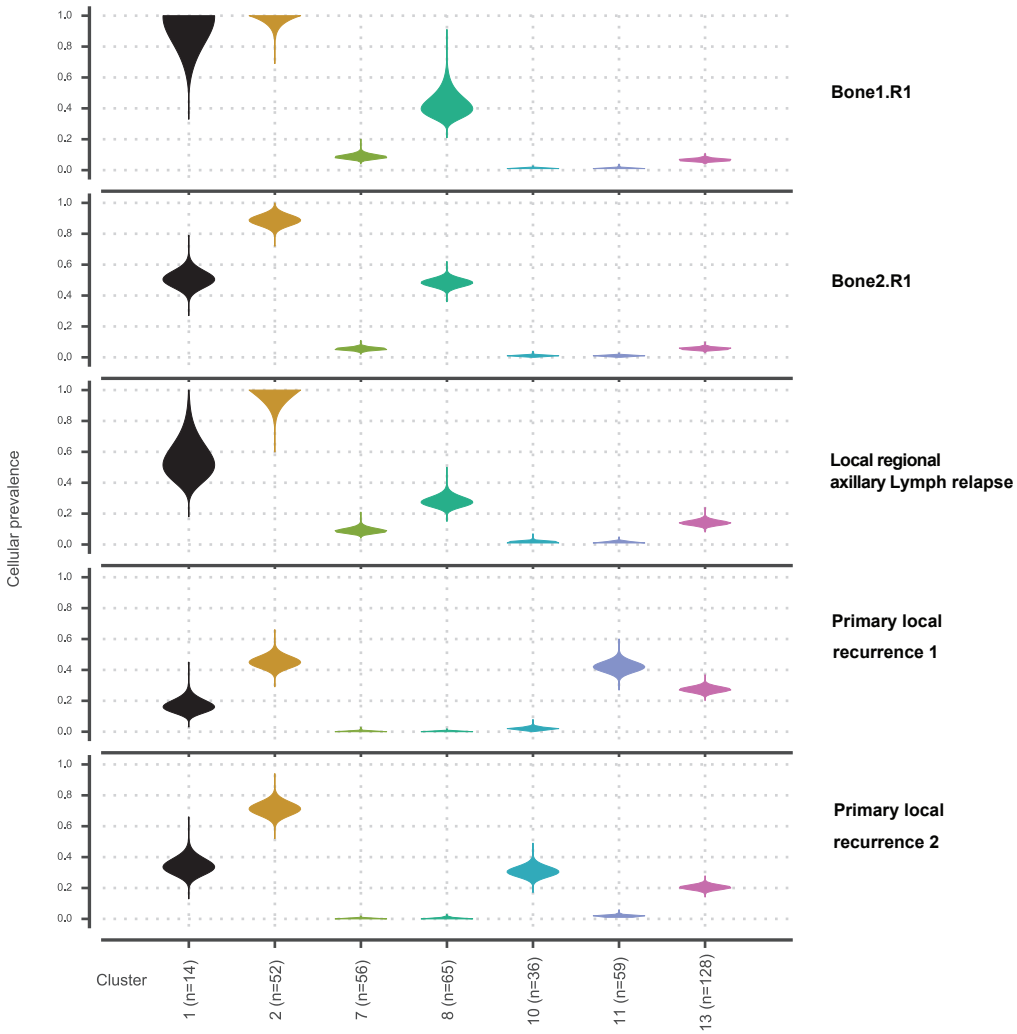
Phylogenetic Tree



Cluster Table

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		2	52	0
		1	14	0
		13	128	0
Primary local reccurence		11	59	NF1
		10	36	0
Lymph relapse & Metastasis		7	56	0
		8	65	FLT1;ESR1;TSC1

Density Plot

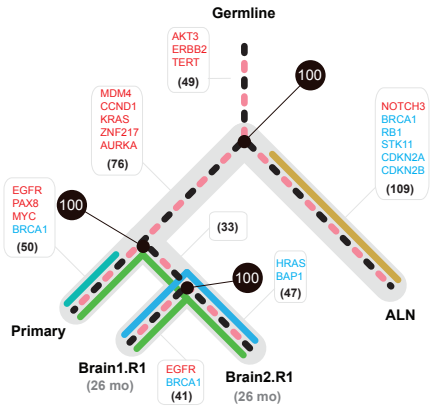


Supplementary Figure 6

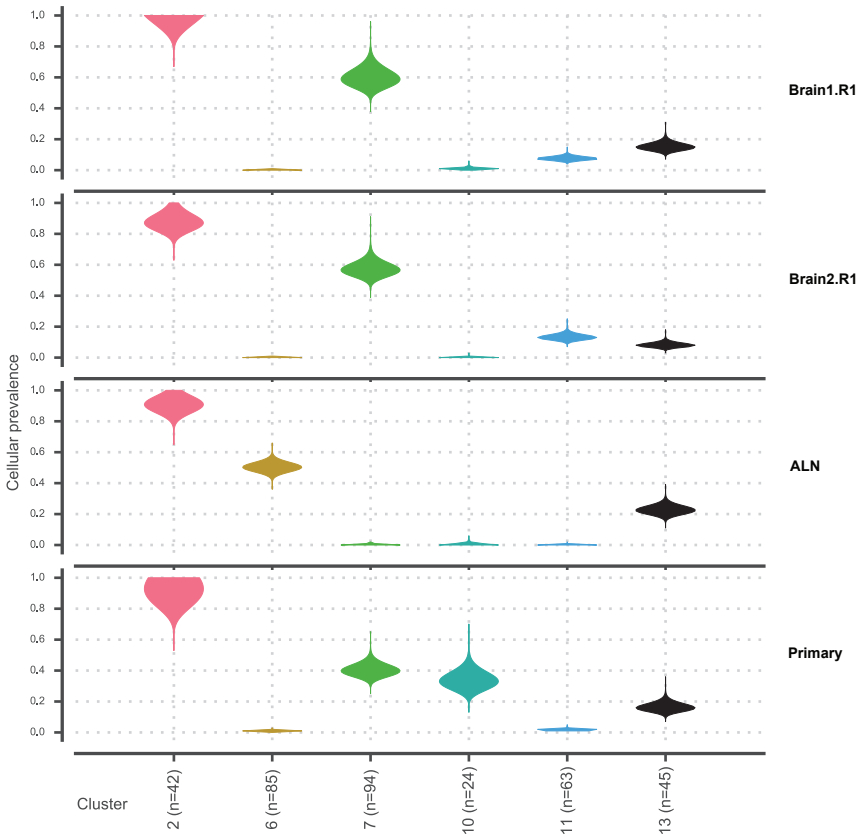
L

Patient 14: ER+/PR+/HER2+

Phylogenetic Tree



Density Plot



Cluster Table

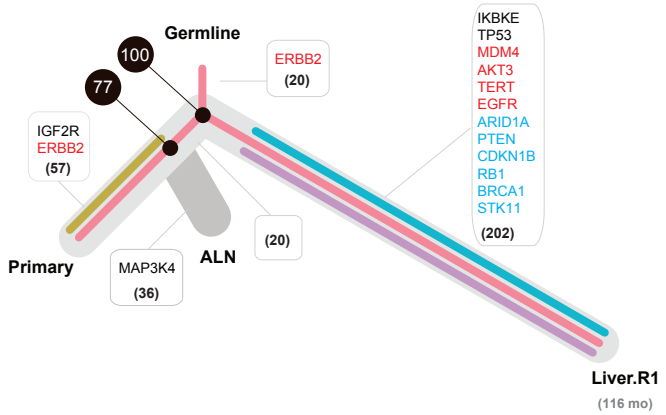
	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		2	42	0
		13	45	0
		7 (except lymph)	94	0
Primary		10	24	0
Lymph		6	85	CCNE1
Metastasis		11	63	0

Supplementary Figure 6

M

Patient 15: ER+/PR+/HER2-

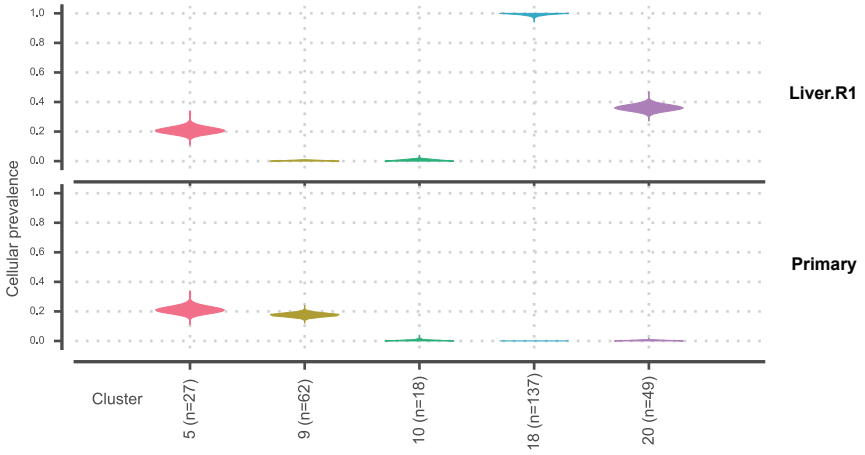
Phylogenetic Tree



Density Plot

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		5	27	0
Primary		9	62	IGF2R
Metastasis		18	137	TP53;IKBKE
		20	49	0

Cluster Table

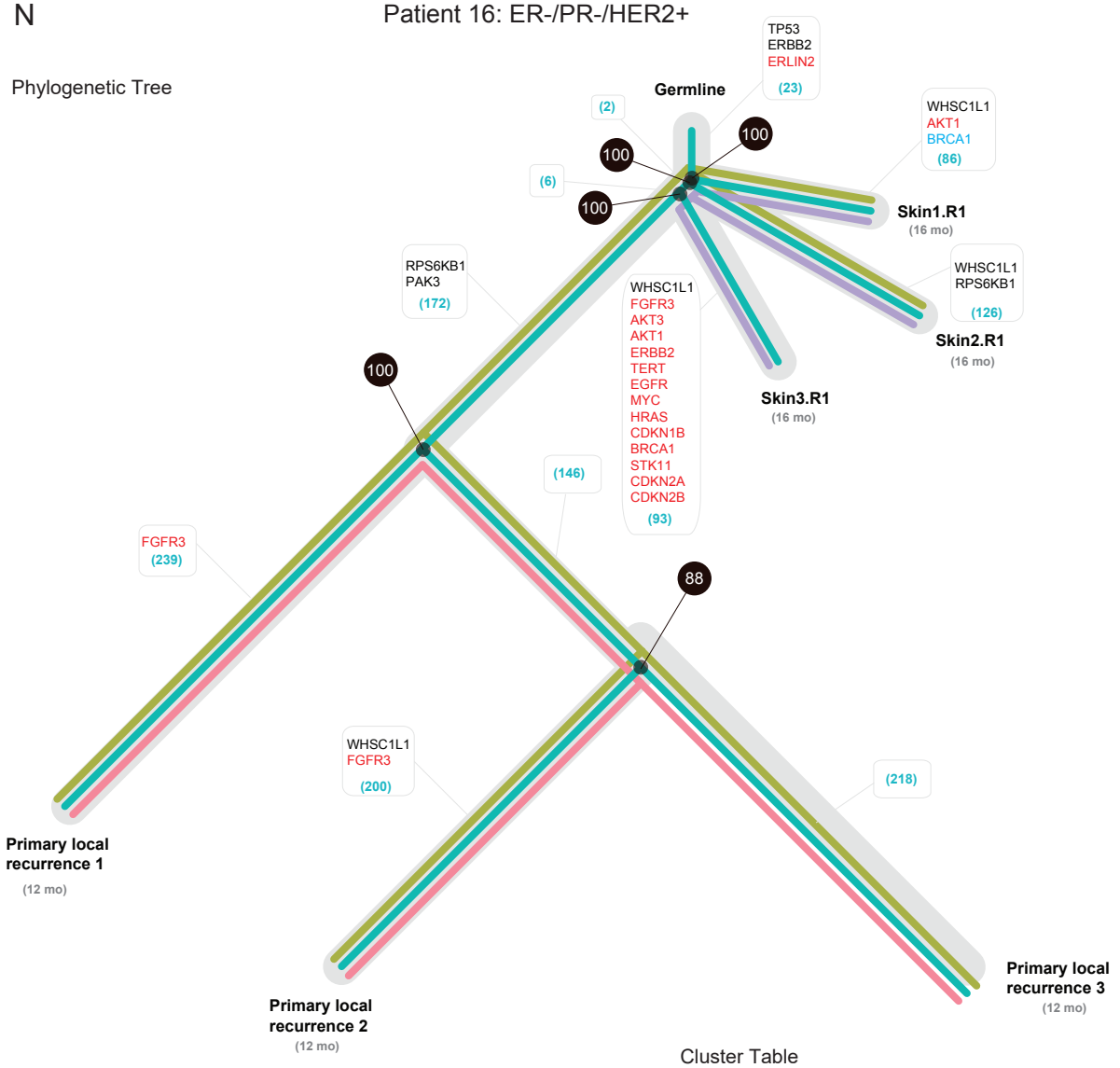


Supplementary Figure 6

N

Patient 16: ER-/PR-/HER2+

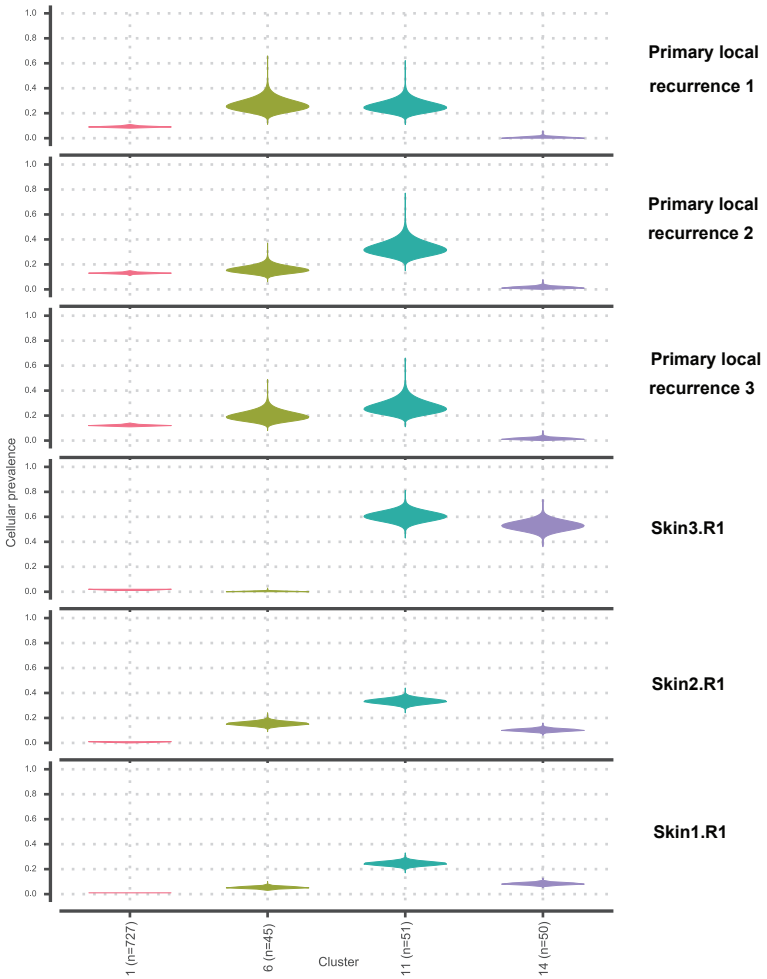
Phylogenetic Tree



Density Plot

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		11	51	TP53
		6 (except Skin3)	45	RPS6KB1
Primary local recurrence		1	727	PAK3
Metastasis		14	50	RPS6KB1

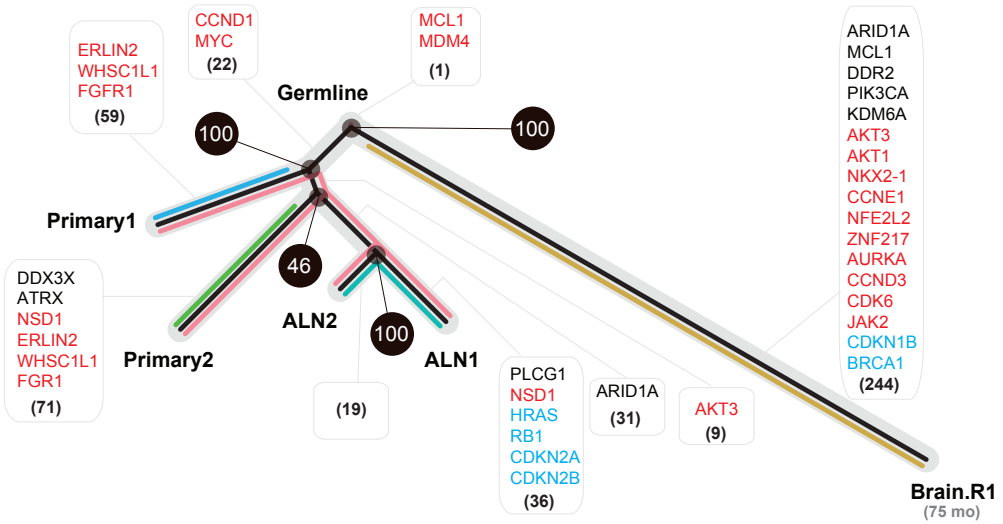
Cluster Table



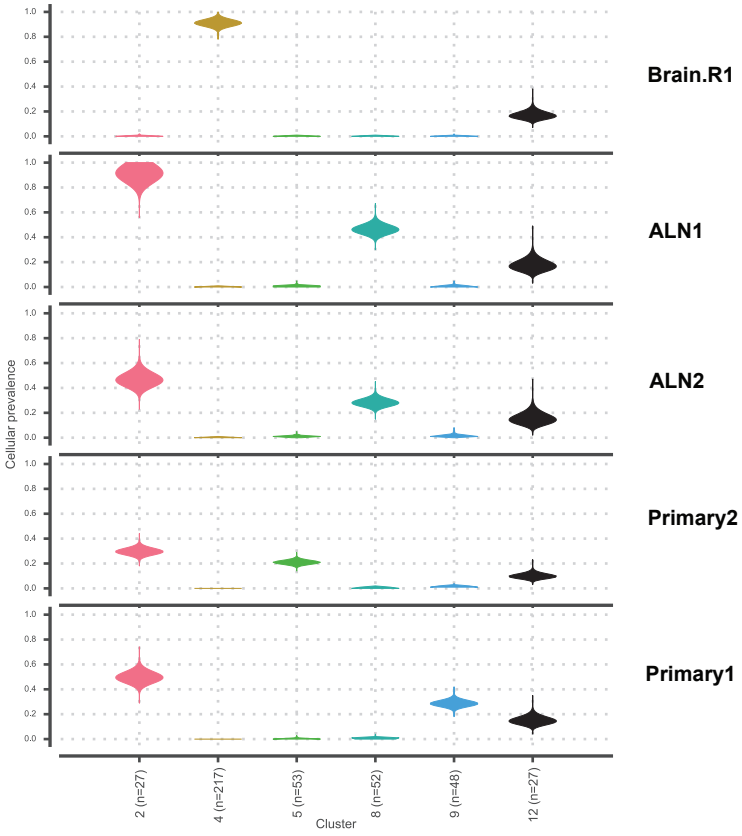
Supplementary Figure 6

O Patient 17: ER+/PR+/HER2-

Phylogenetic Tree



Density Plot



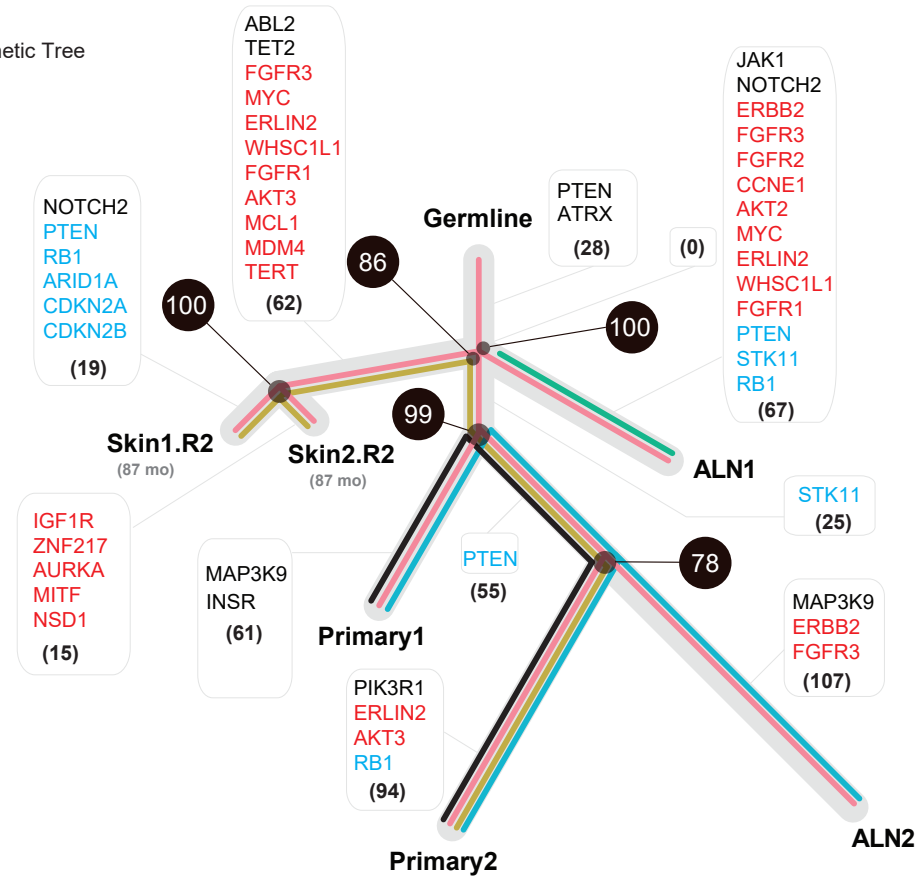
Cluster Table

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		12	27	0
		2 (except Brain)	27	0
Primary		9	48	0
		5	53	DDR2;DDX3X;ATRX
Lymph		8	52	ARID1A;PLCG1
Metastasis		4	217	MCL1;DDR2;ARID1A;PIK3CA;KDM6A

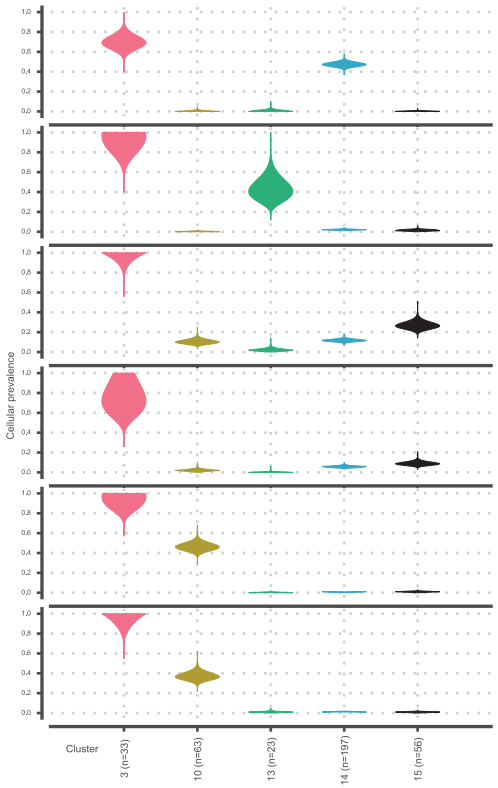
Supplementary Figure 6

P Patient 18: ER+/PR+/HER2-

Phylogenetic Tree



Density Plot



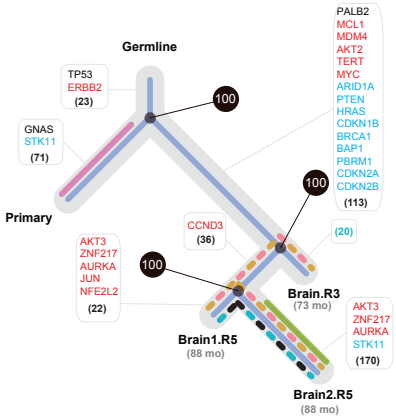
Cluster Table

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		3	33	PTEN;ATRX
Primary		15	56	INSR;PIK3R1
Primary & Lymph 2		14	197	MAP3K9
Lymph1		13	23	JAK1
Skin Metastasis & Primary 2		10	63	ABL2;TET2

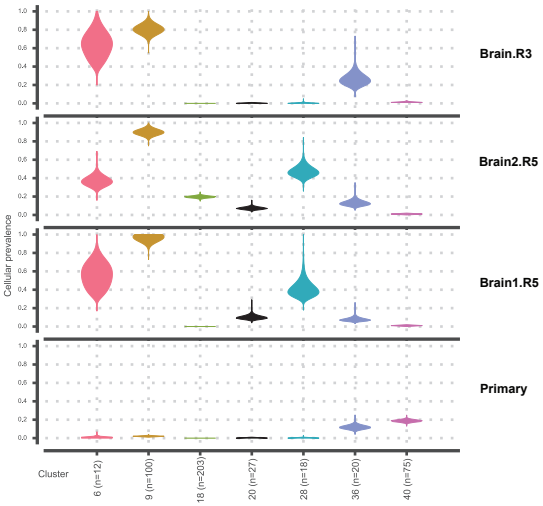
Supplementary Figure 6

Q Patient 19: ER+/PR+/HER2+

Phylogenetic Tree



Cluster Table



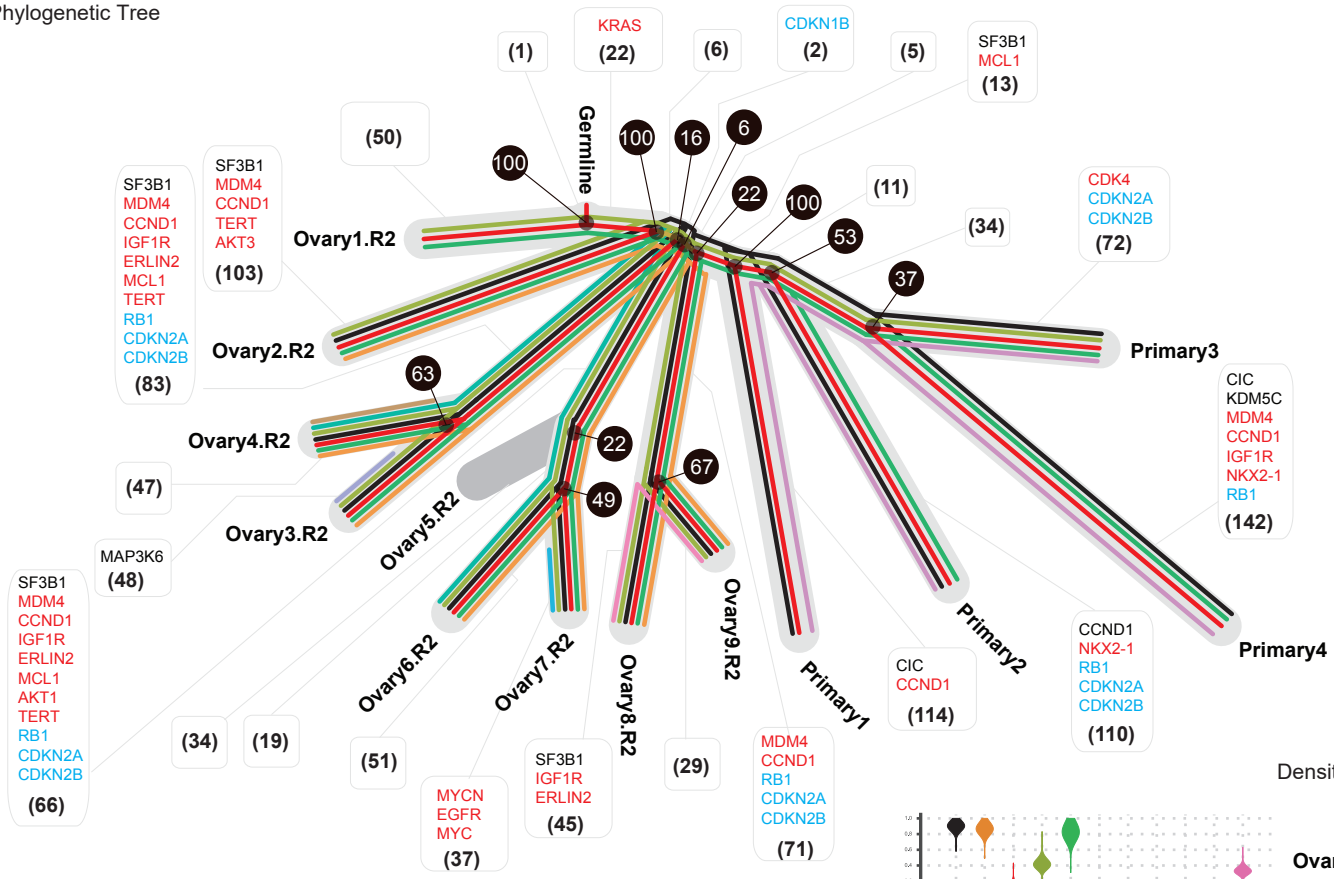
Density Plot

	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		36	20	0
Primary		40	75	GNAS
Metastasis		9	100	PALB2;TP53
		6	12	0
		28	18	0
		20	27	0
		18	203	0

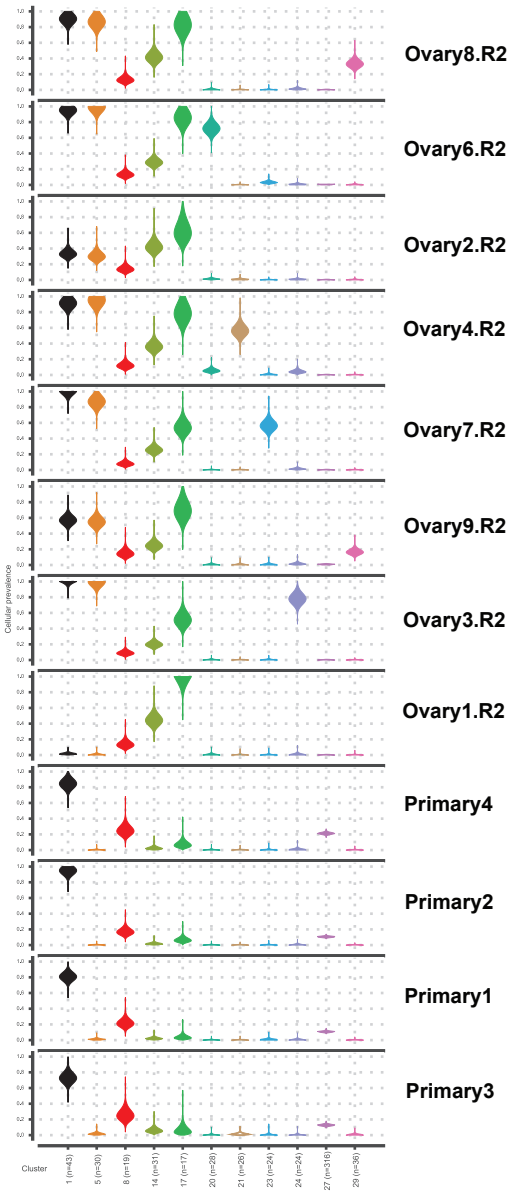
Supplementary Figure 6

R Patient 20: ER+/PR+/HER2-

Phylogenetic Tree



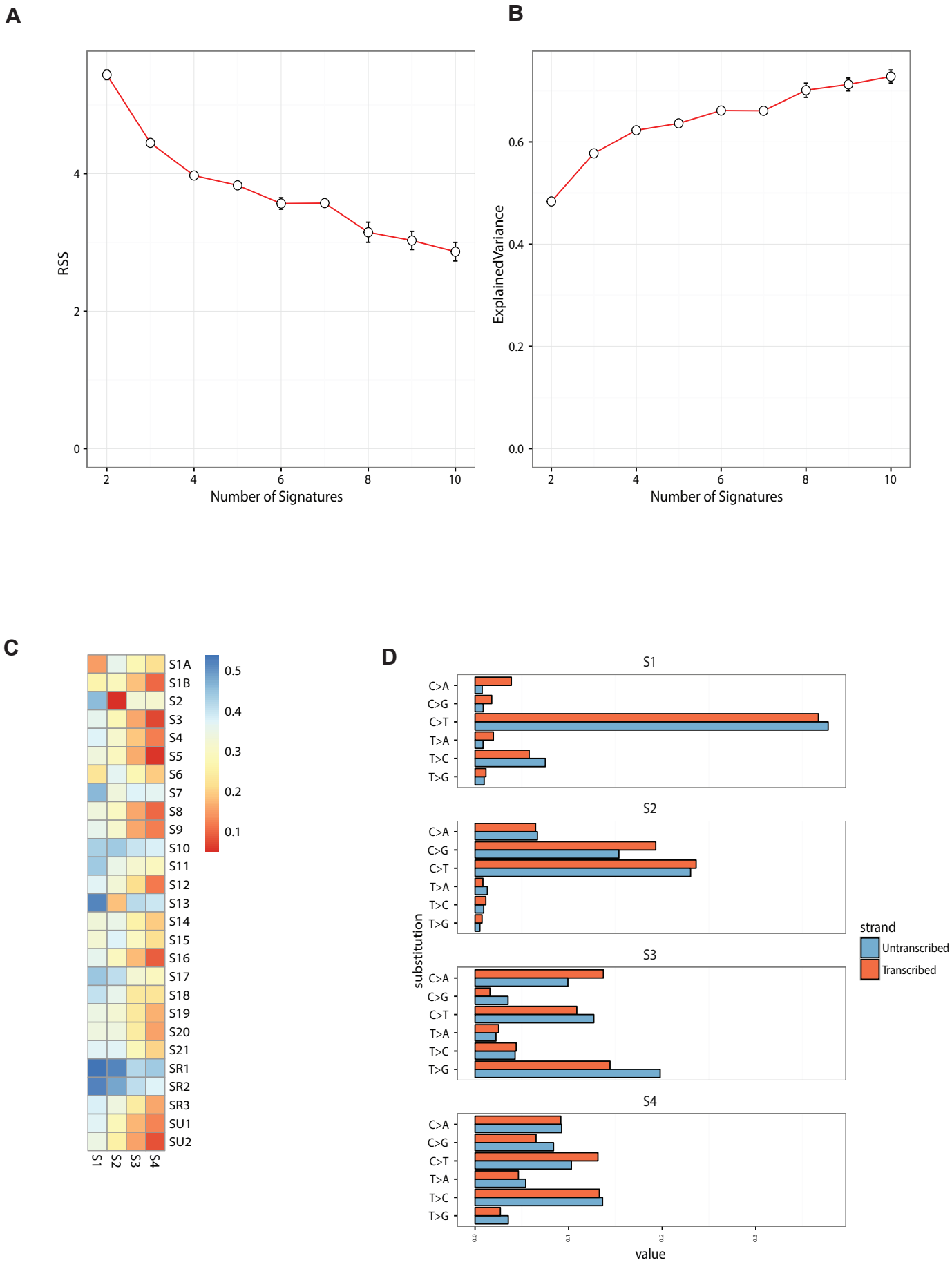
Density Plot



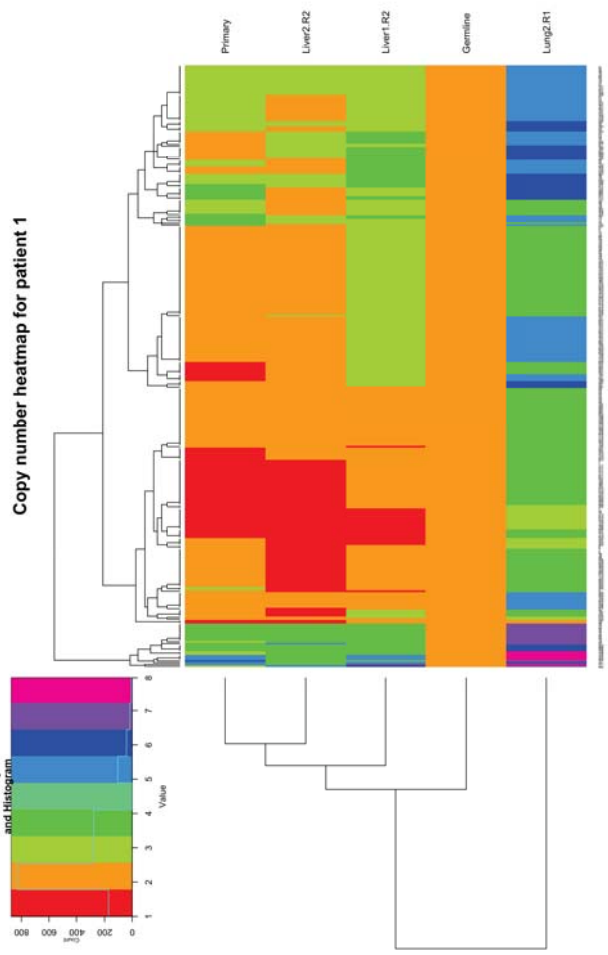
Cluster Table

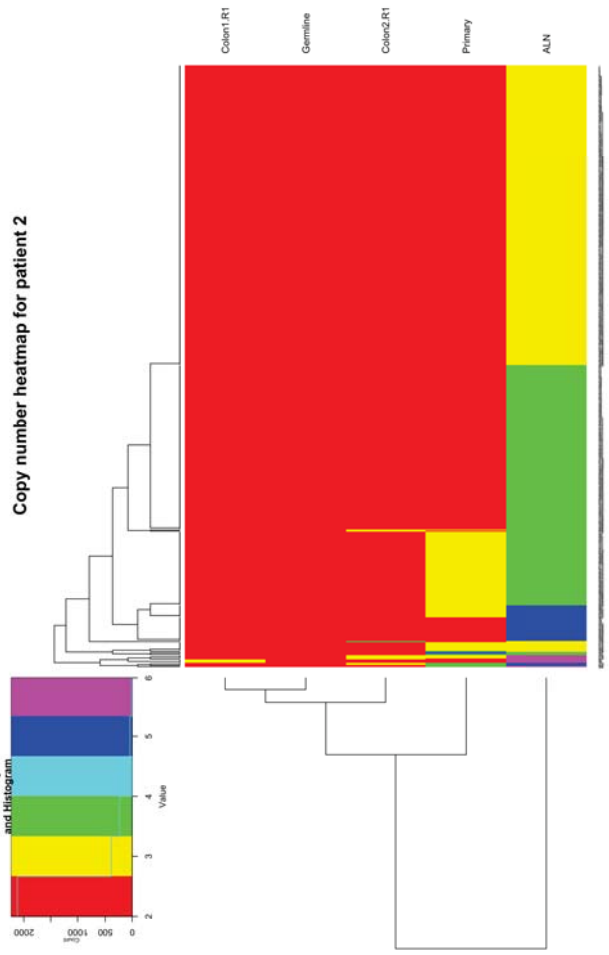
	Cluster color	Cluster ID	Mutation count	Driver genes
Truncal		8	19	0
		1 (except ovary1.R2)	43	SF3B1
		17 (except primary1)	17	0
Primary		27	316	CIC; KDM5C
Primary & Metastasis		14 (In primary3 & all ovary metastasis)	31	0
Metastasis		5 (except ovary1.R2)	30	0
		20	28	0
		21	26	0
		24	24	MAP3K6
		23	24	0
		29	36	0

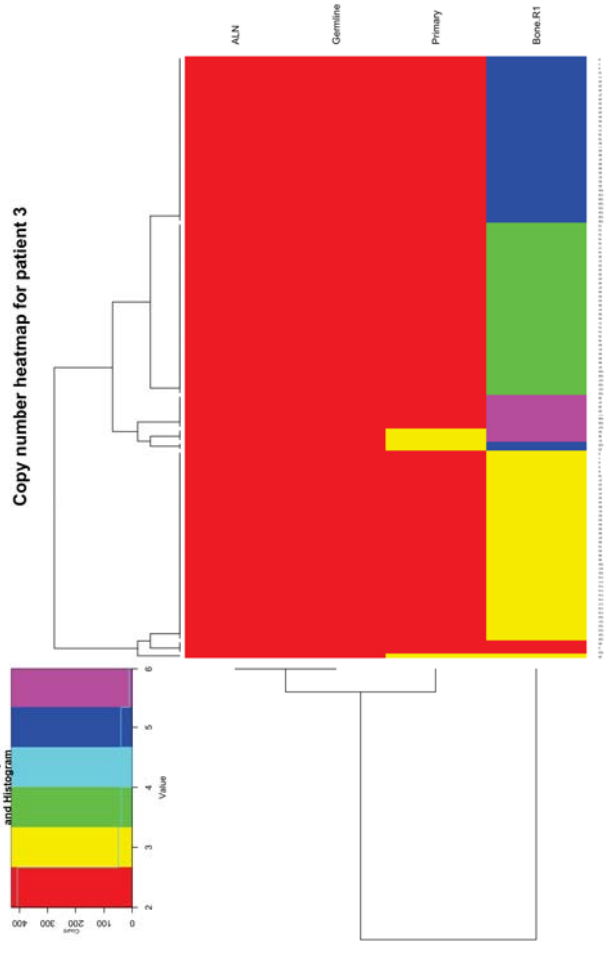
Supplementary Figure 7

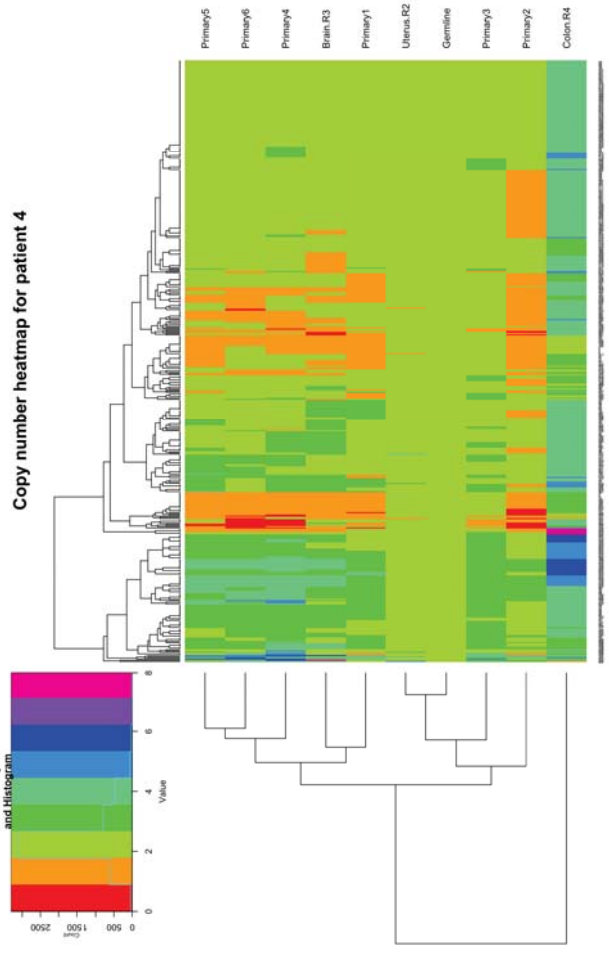


Supplementary Figure 8

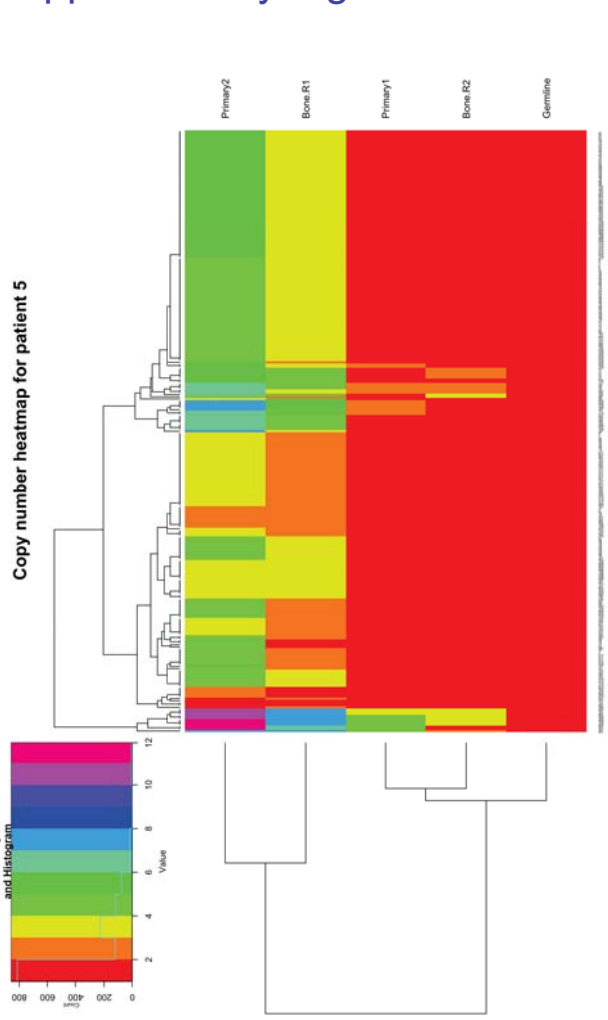


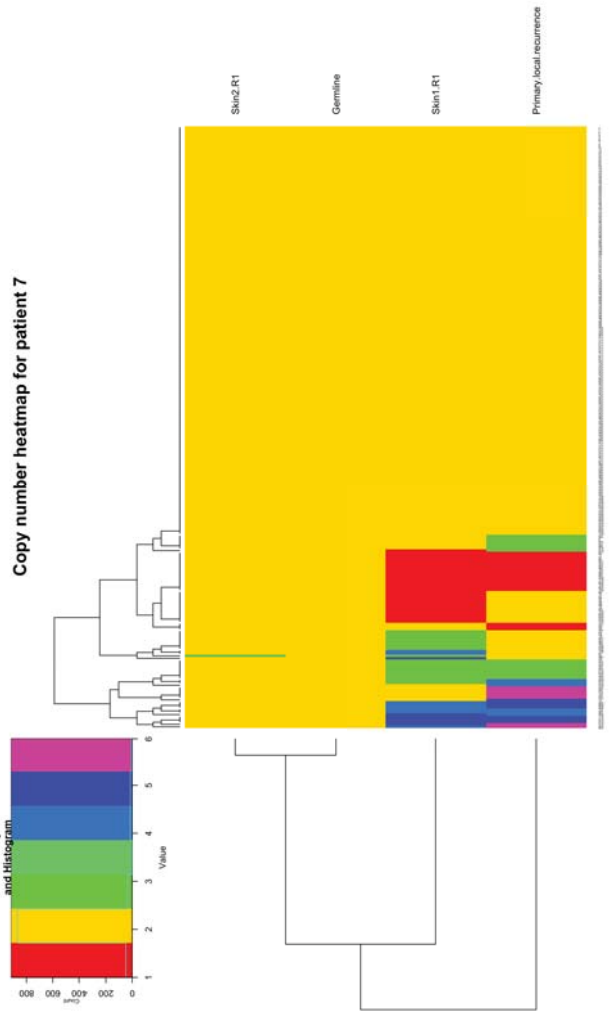


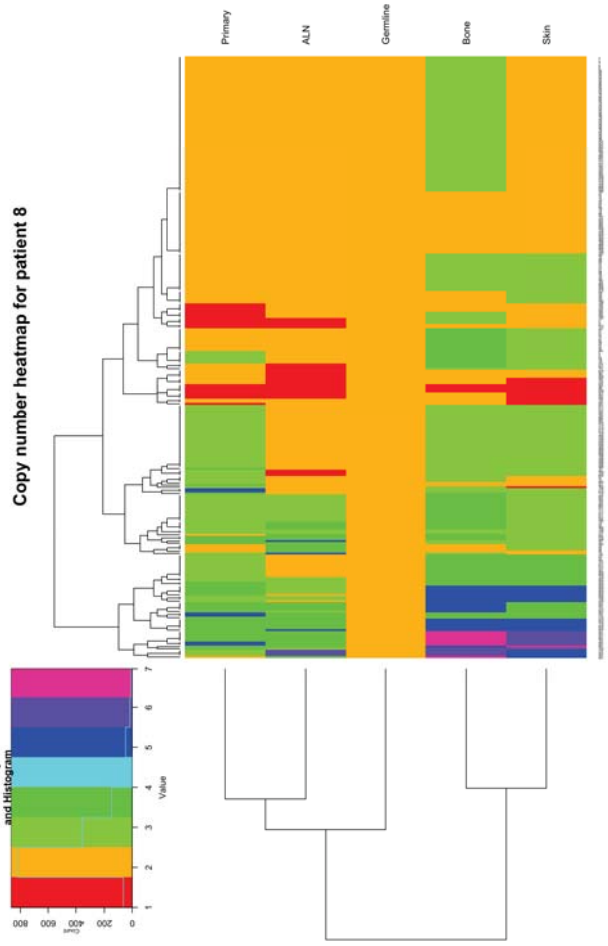


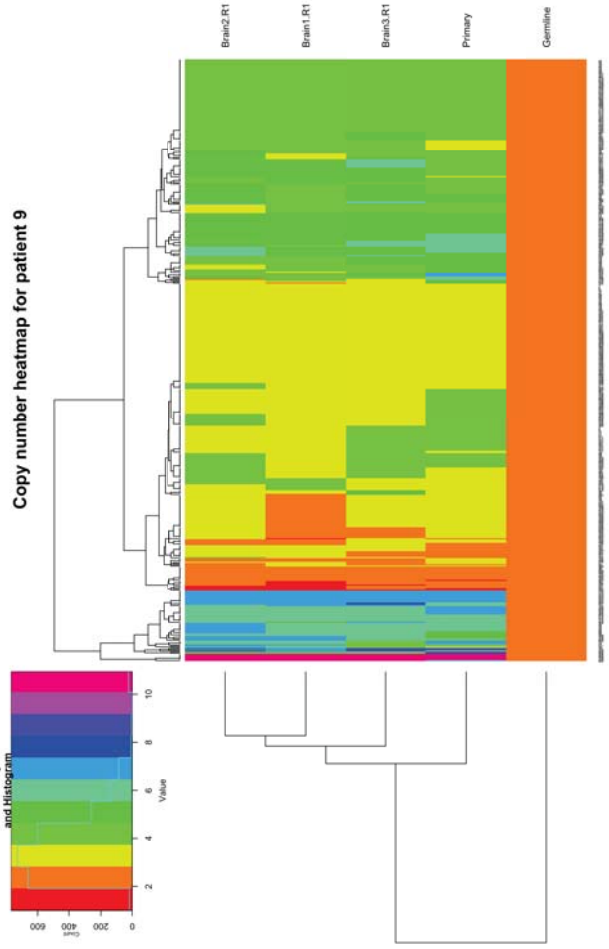


Supplementary Figure 8

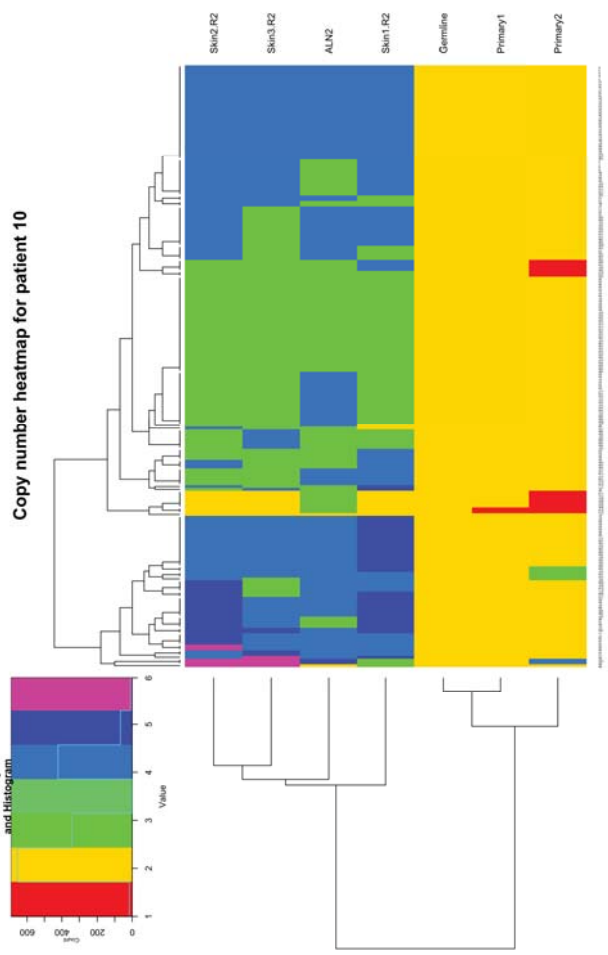


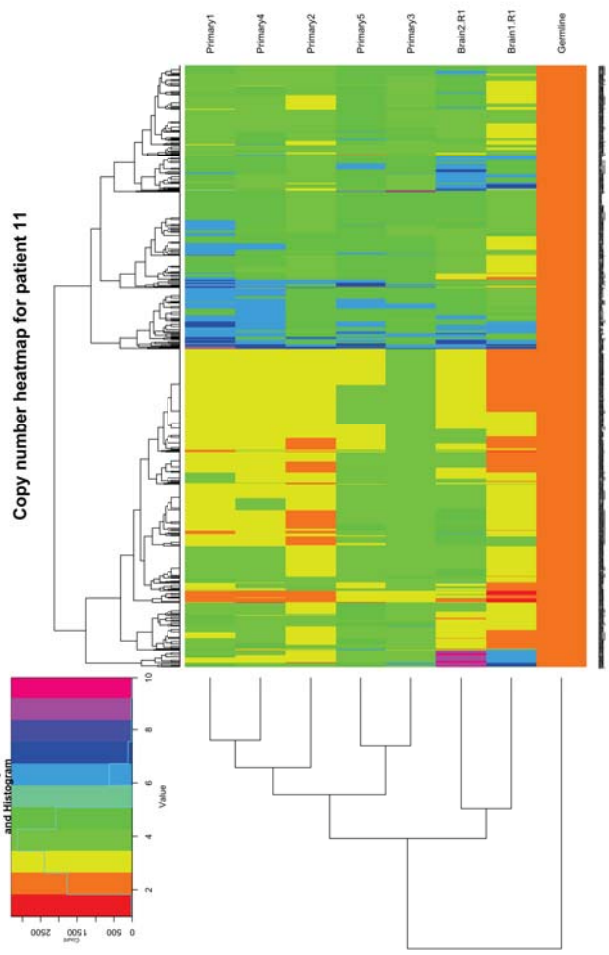


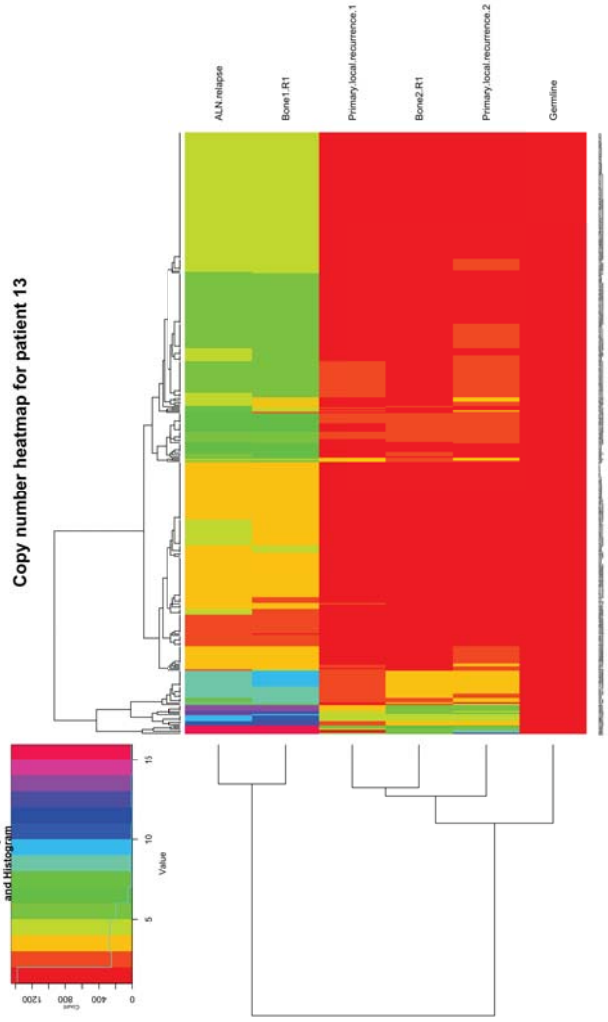


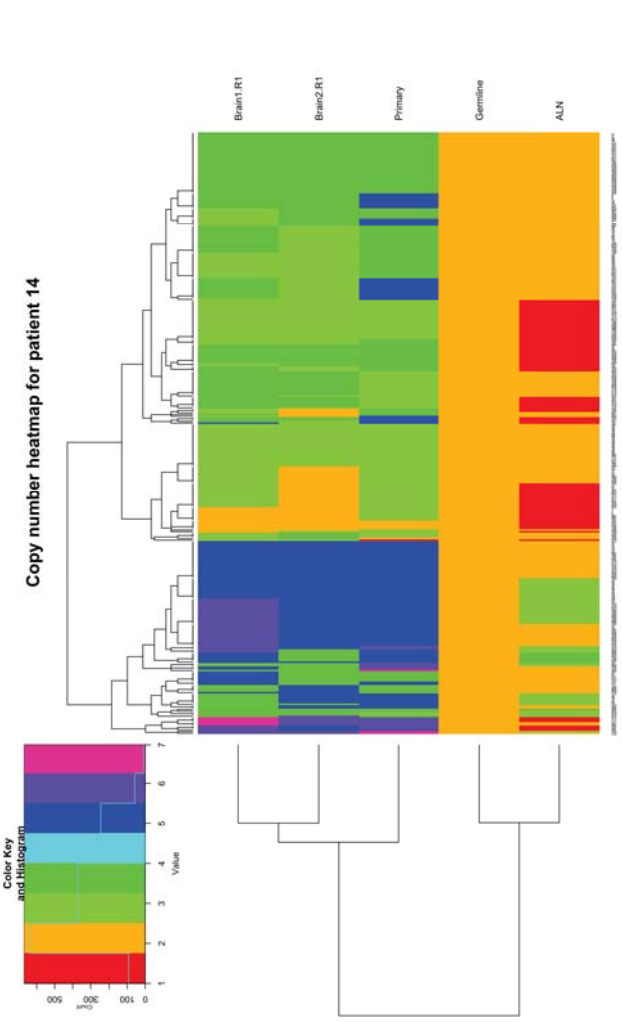


Supplementary Figure 8

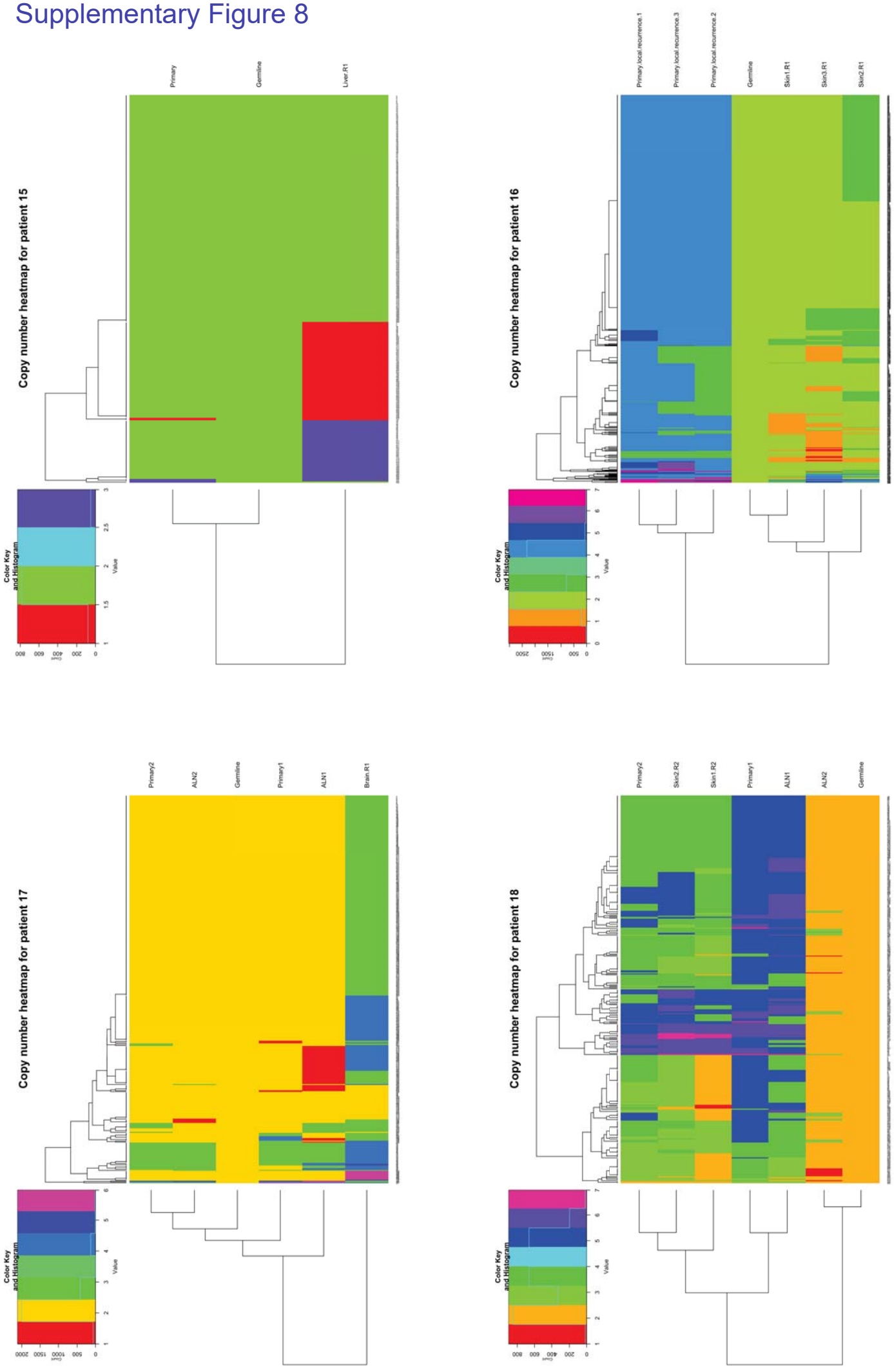




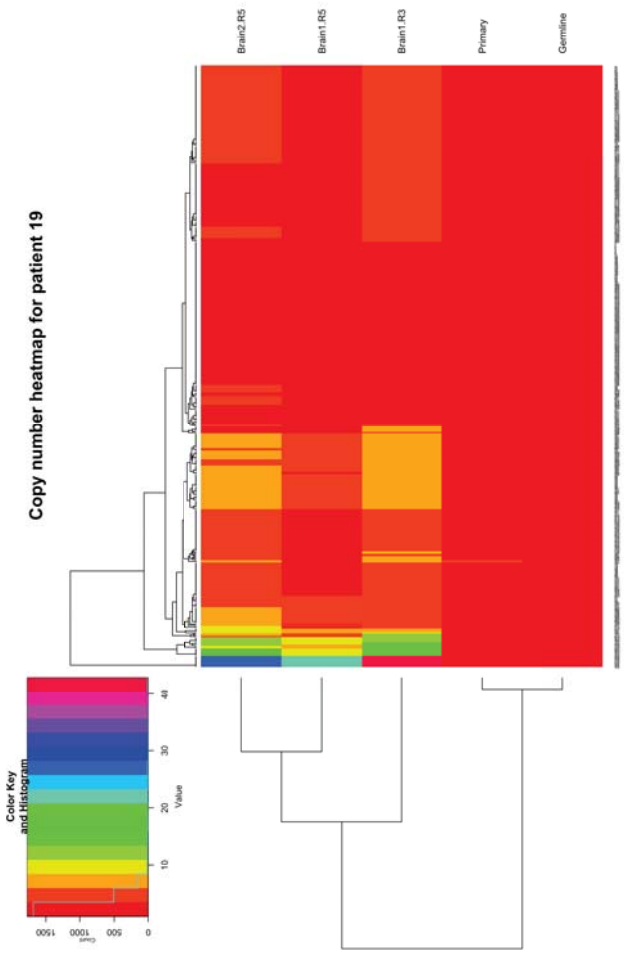


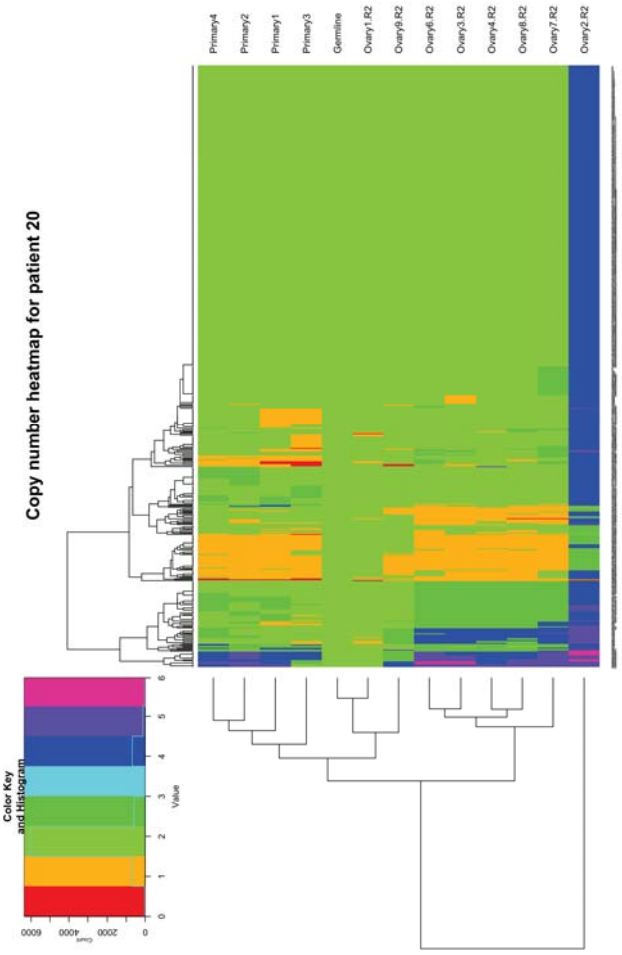


Supplementary Figure 8

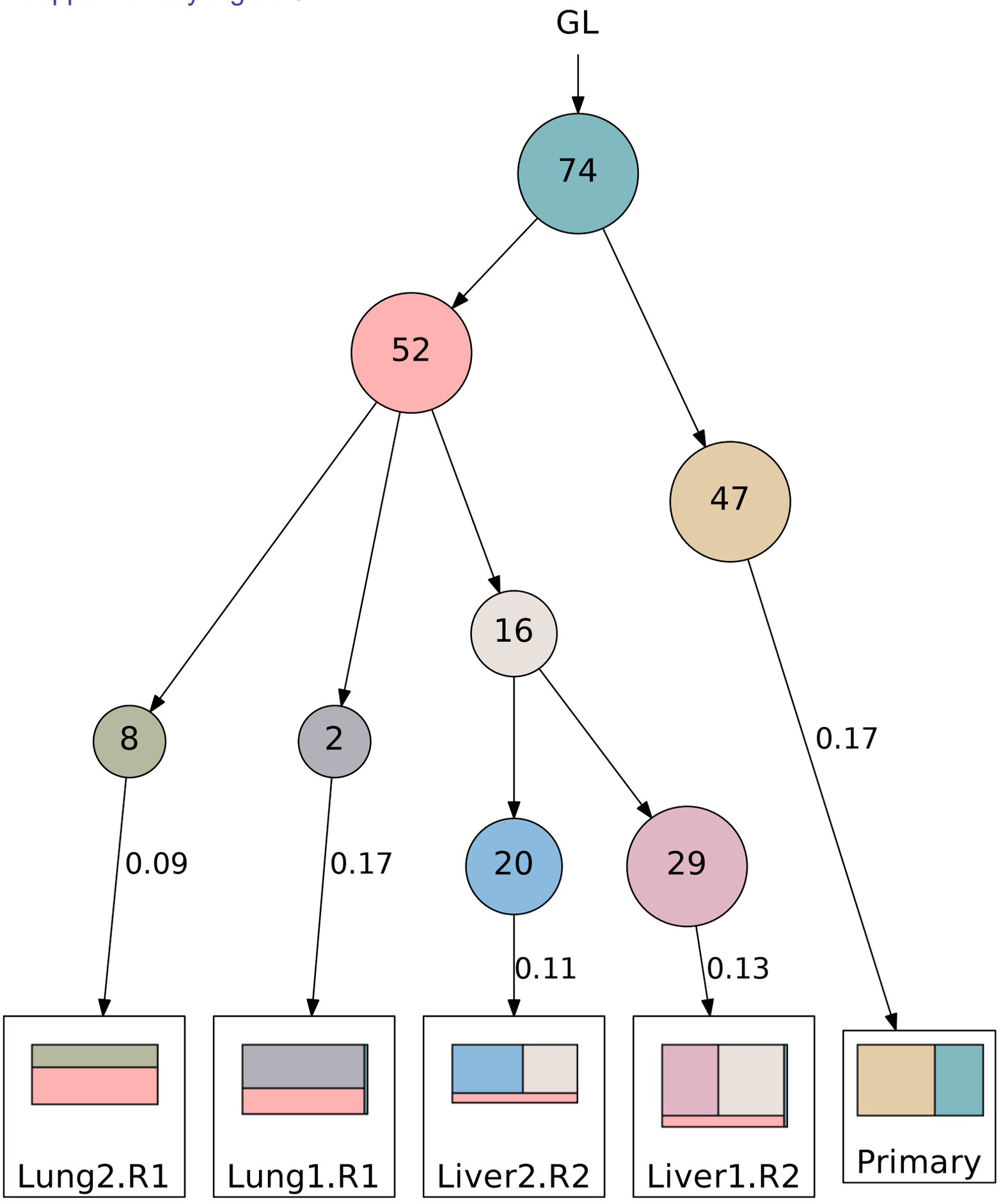


Supplementary Figure 8

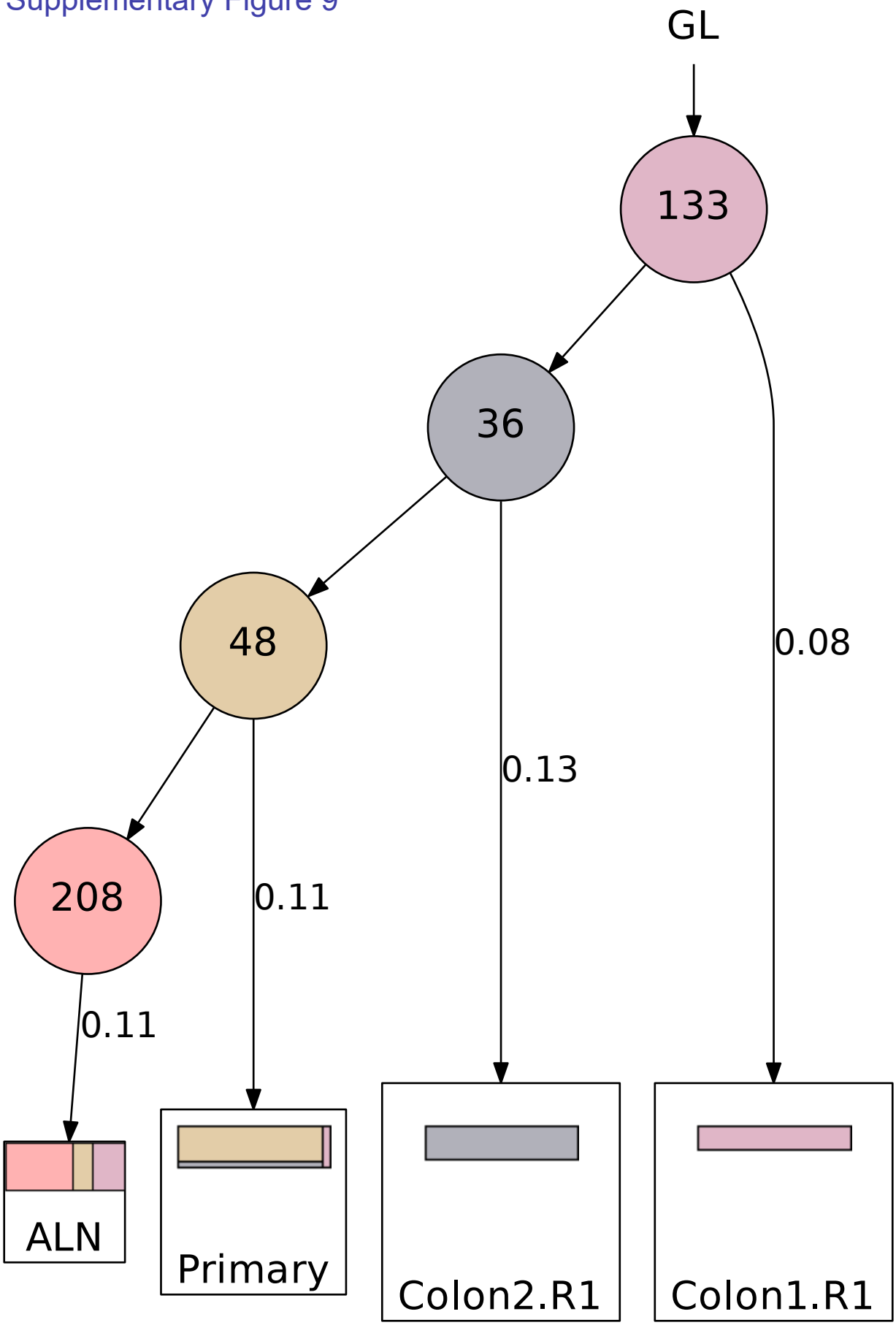




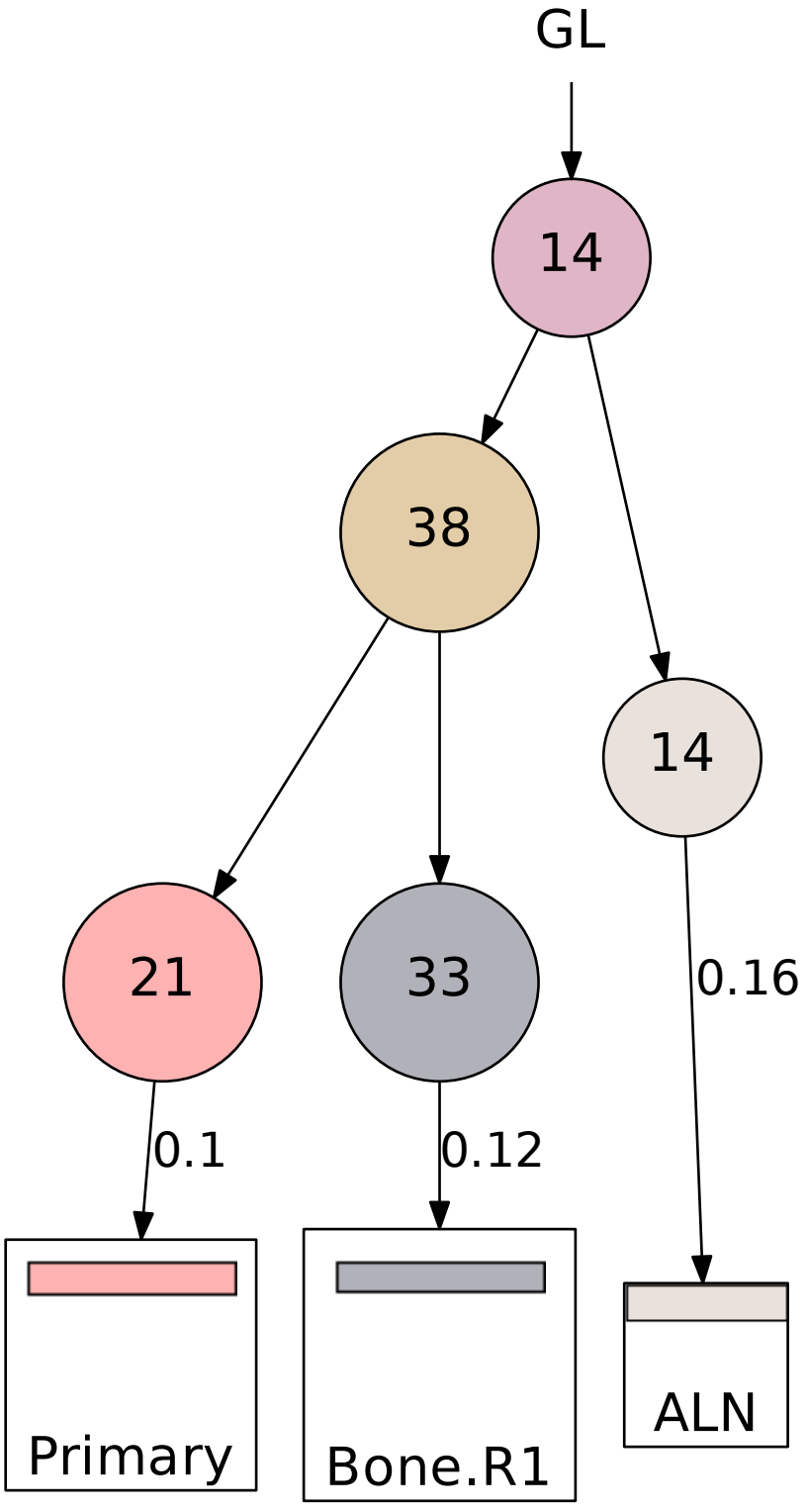
Supplementary Figure 9



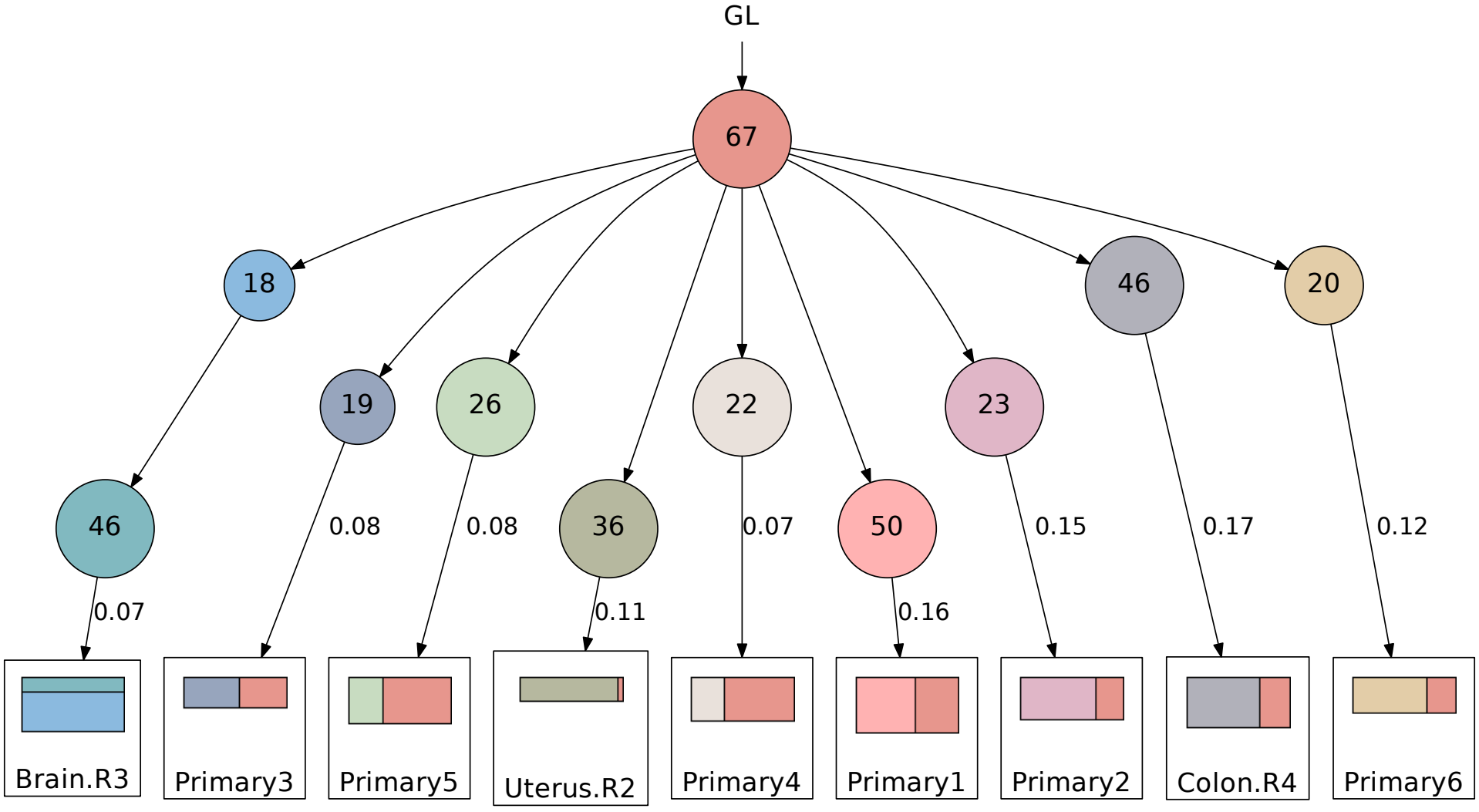
Supplementary Figure 9



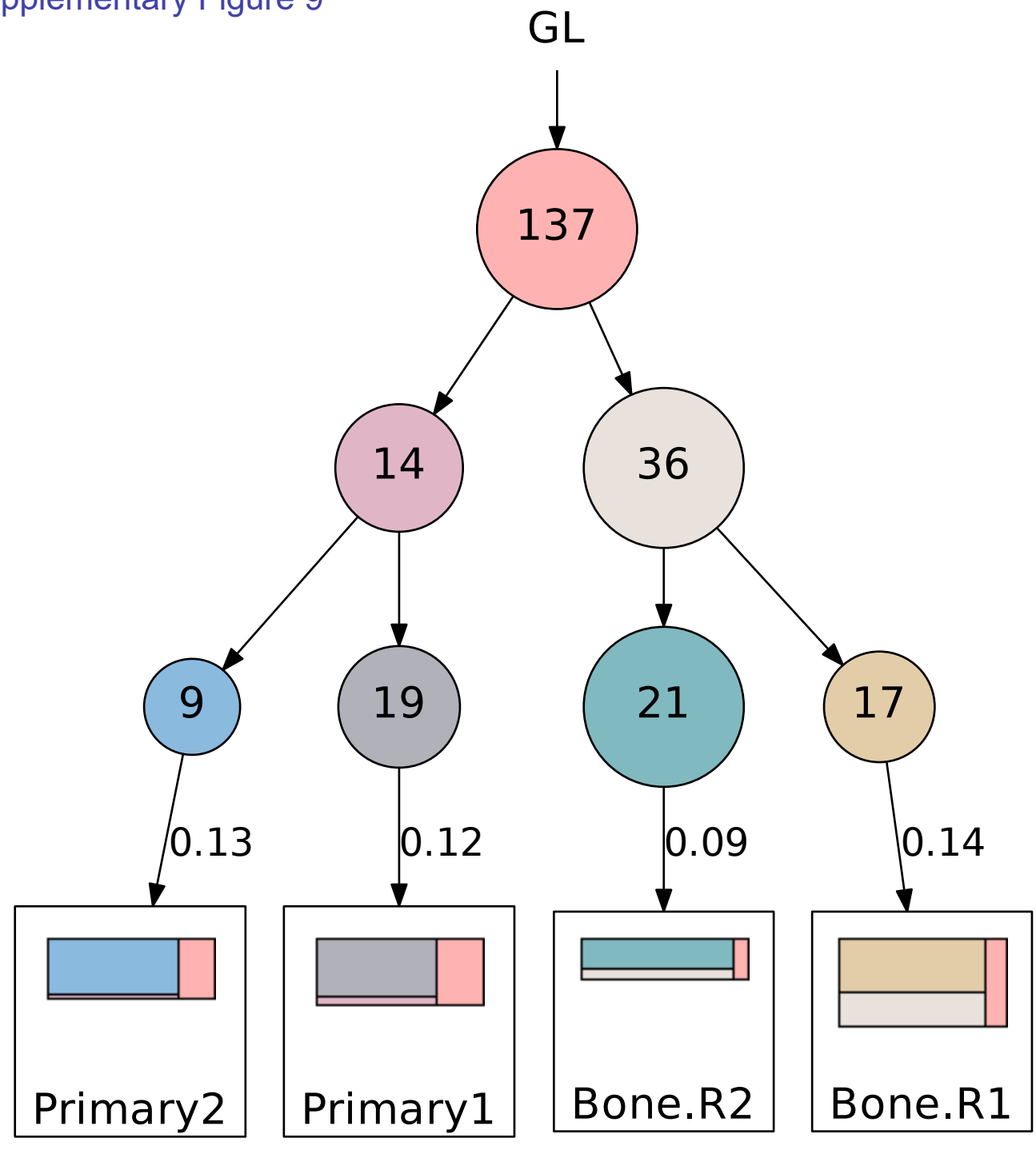
Supplementary Figure 9



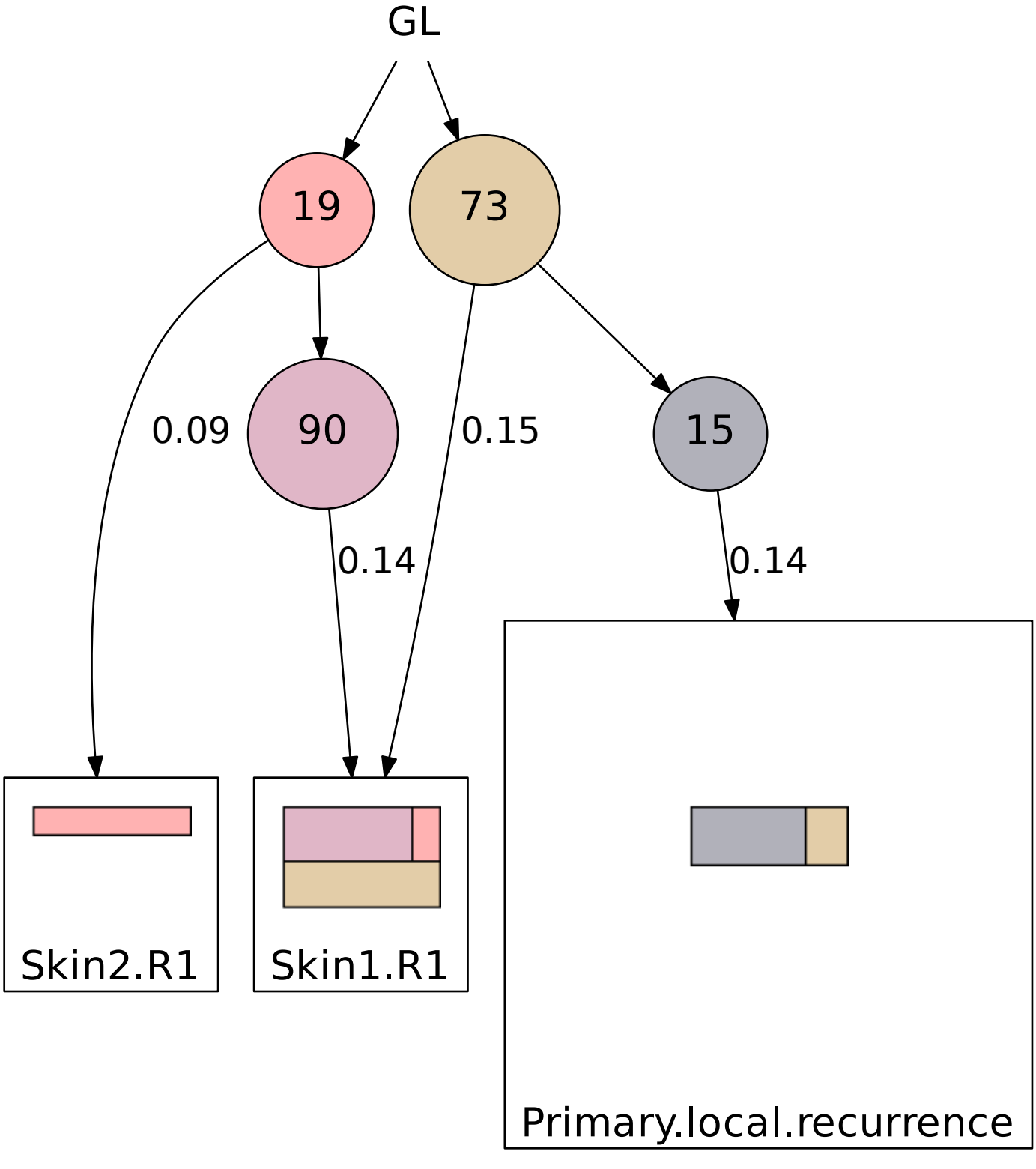
Supplementary Figure 9



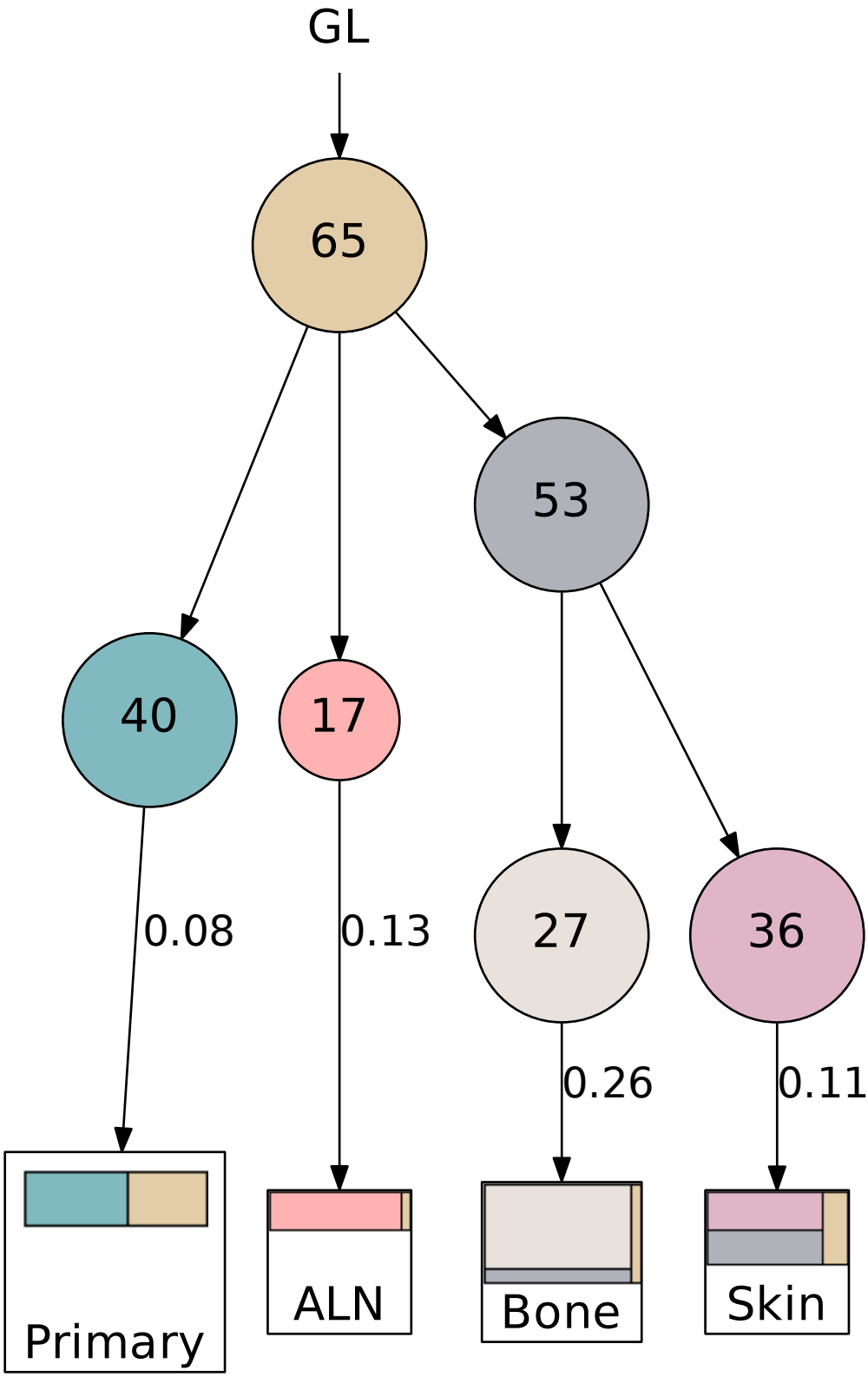
Supplementary Figure 9



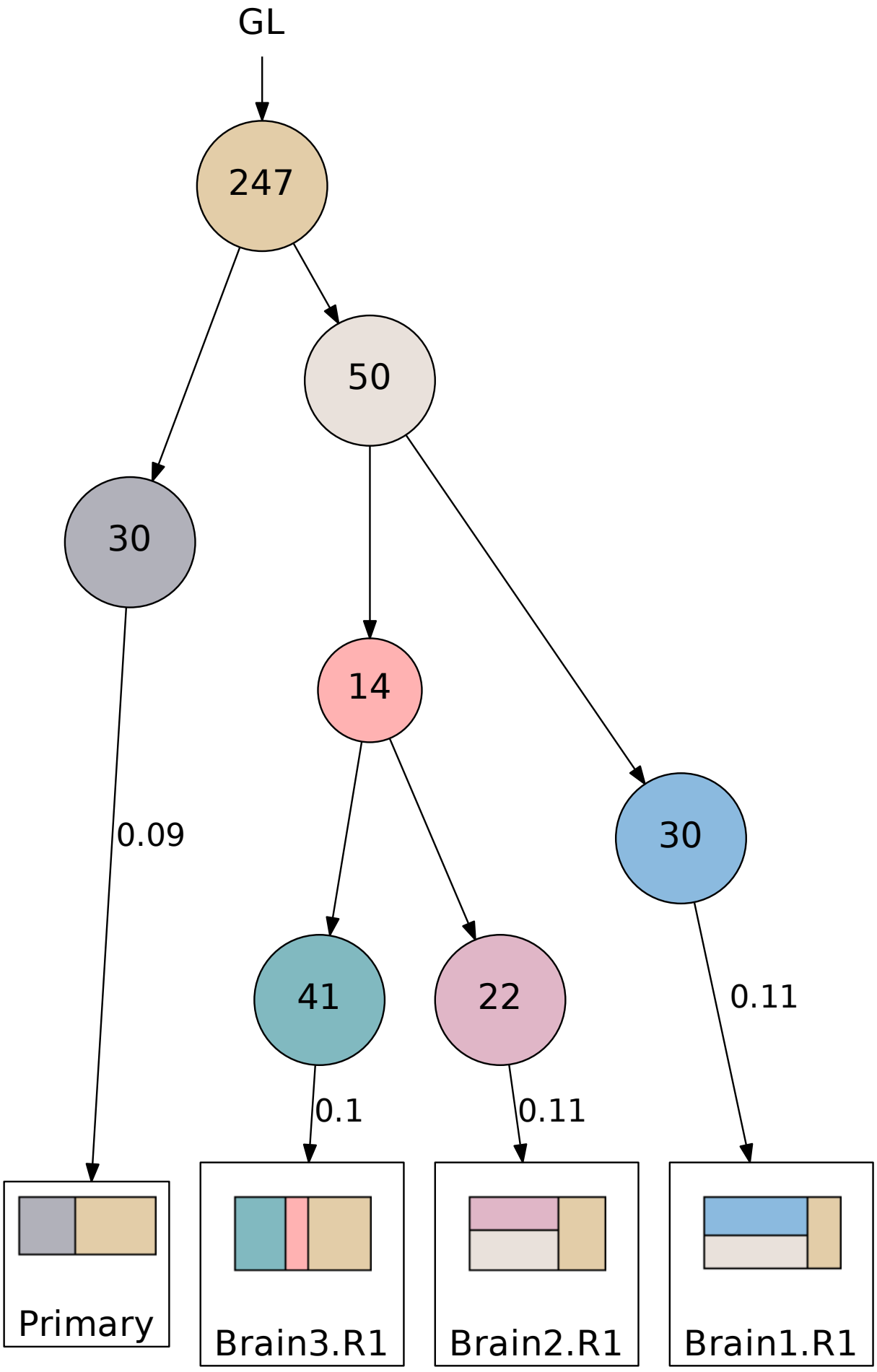
Supplementary Figure 9



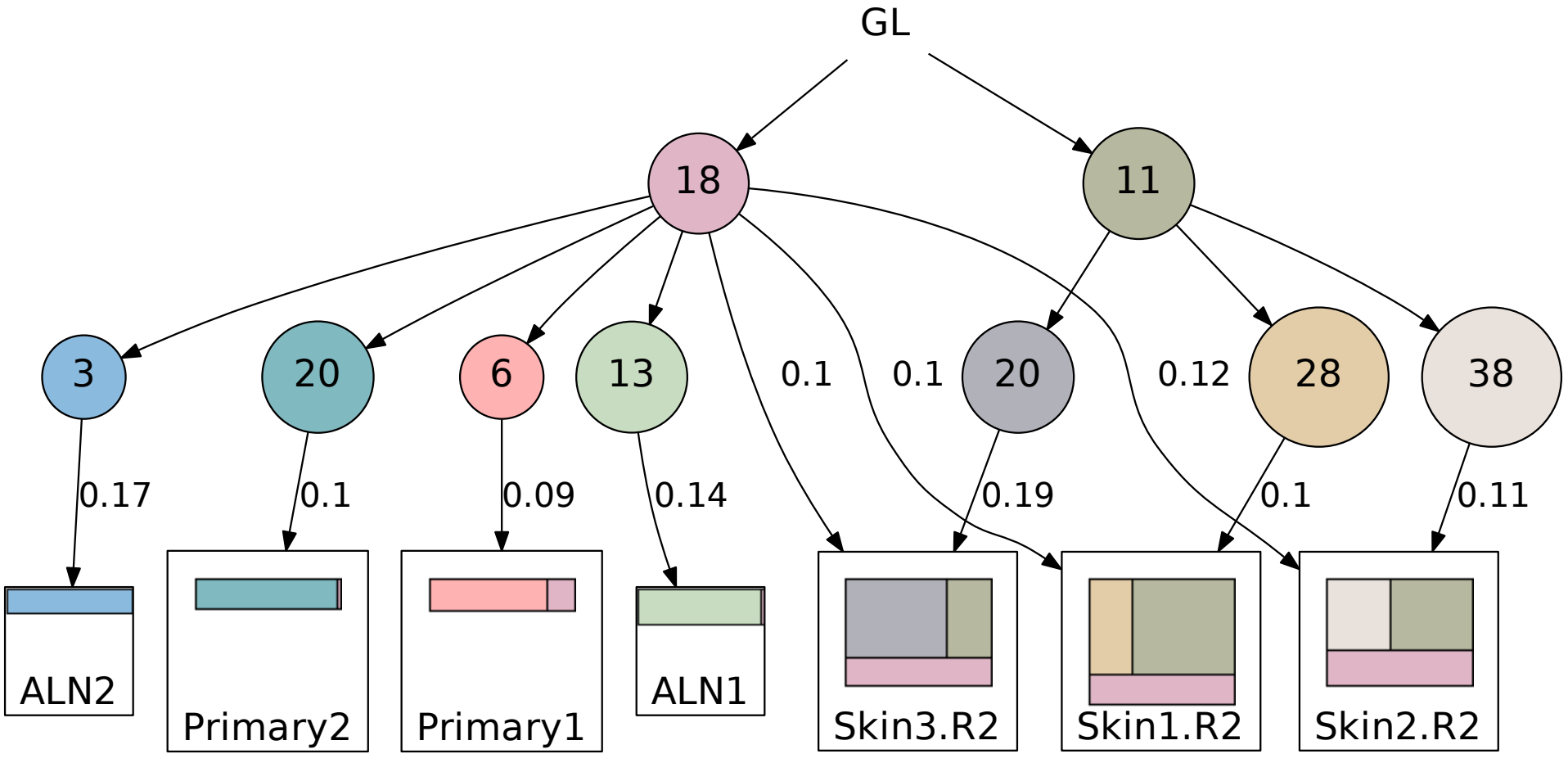
Supplementary Figure 9



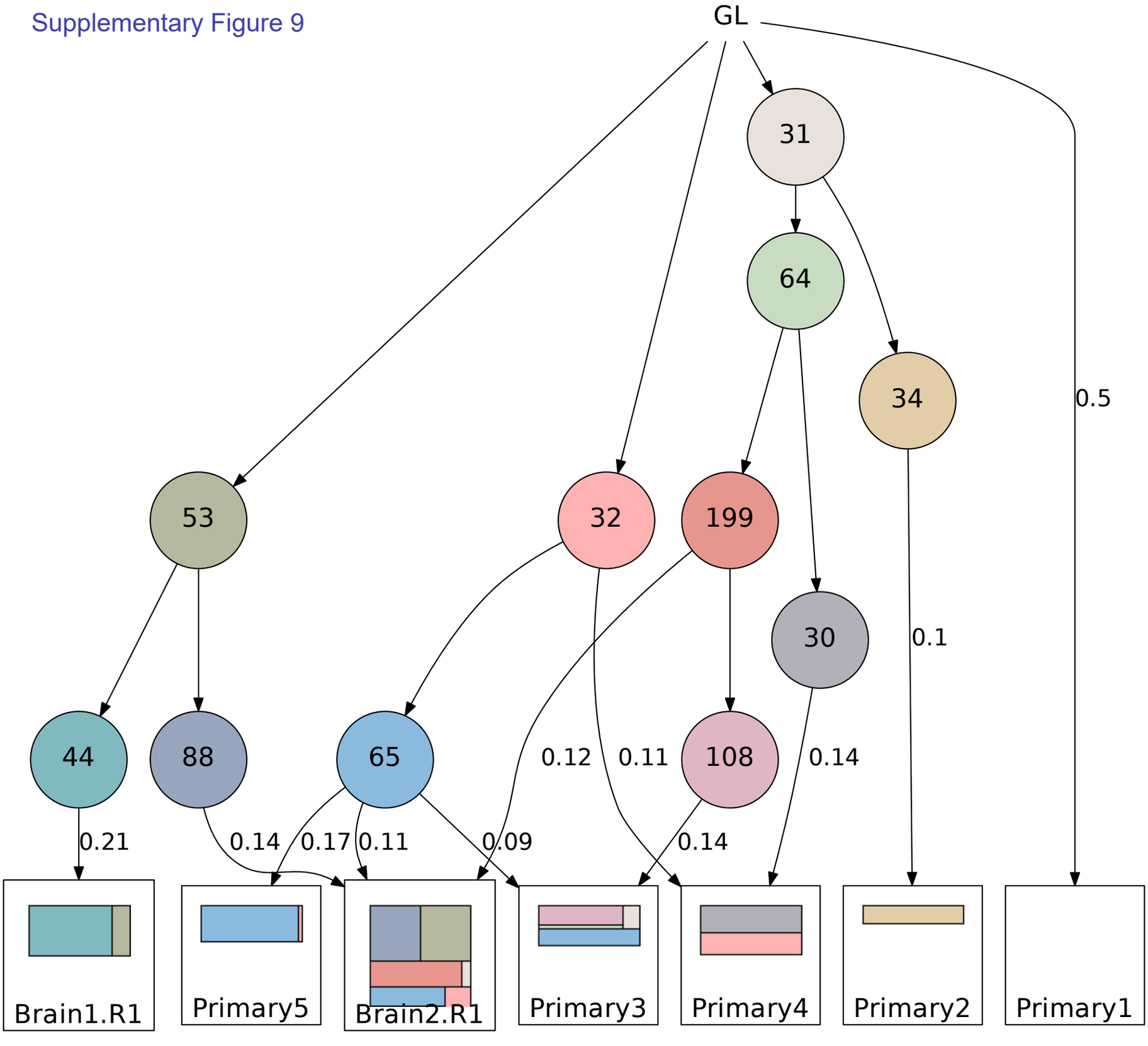
Supplementary Figure 9



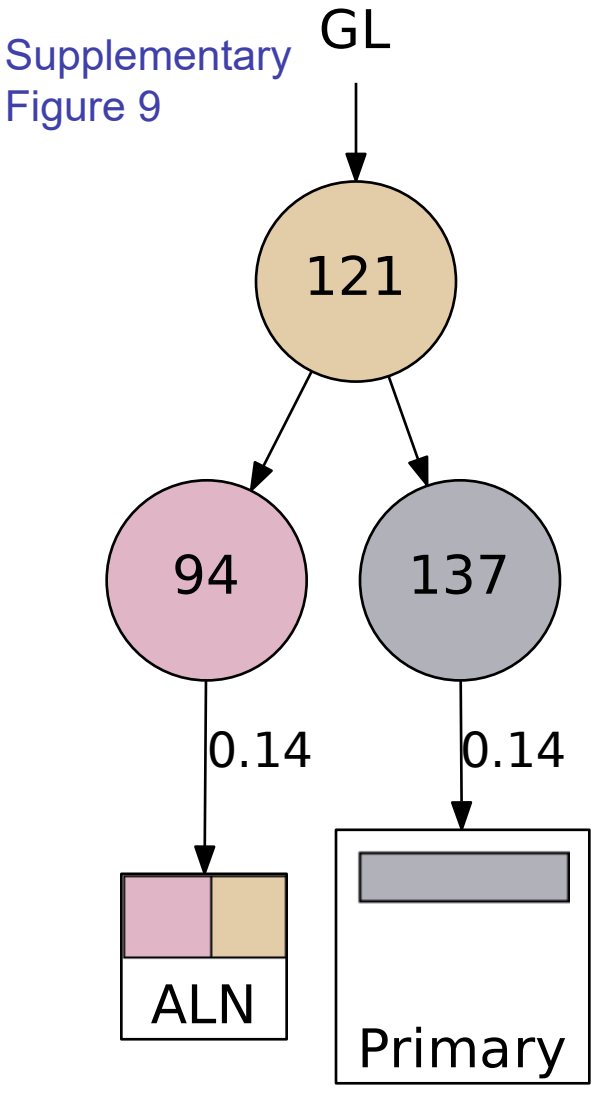
Supplementary Figure 9



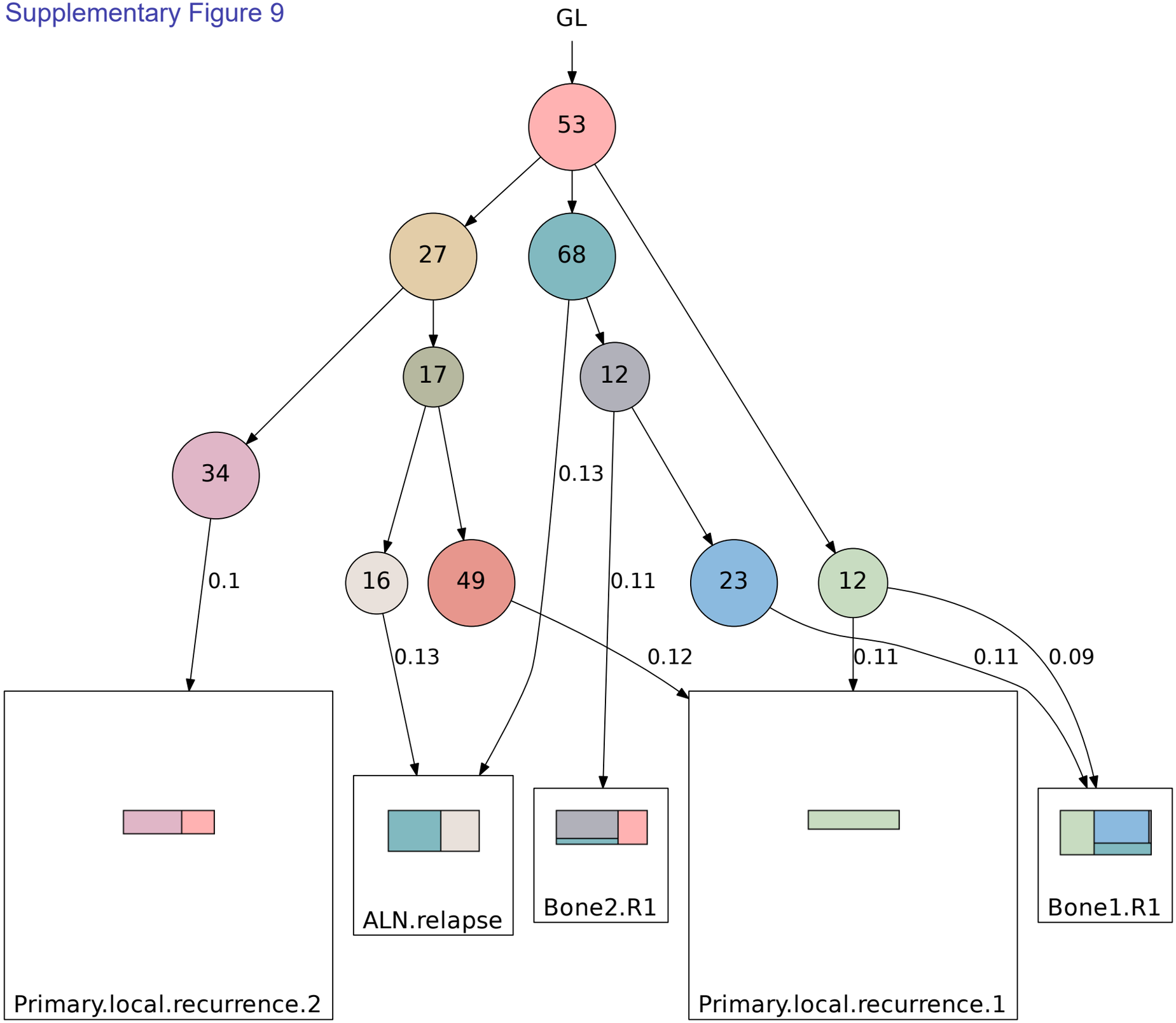
Supplementary Figure 9



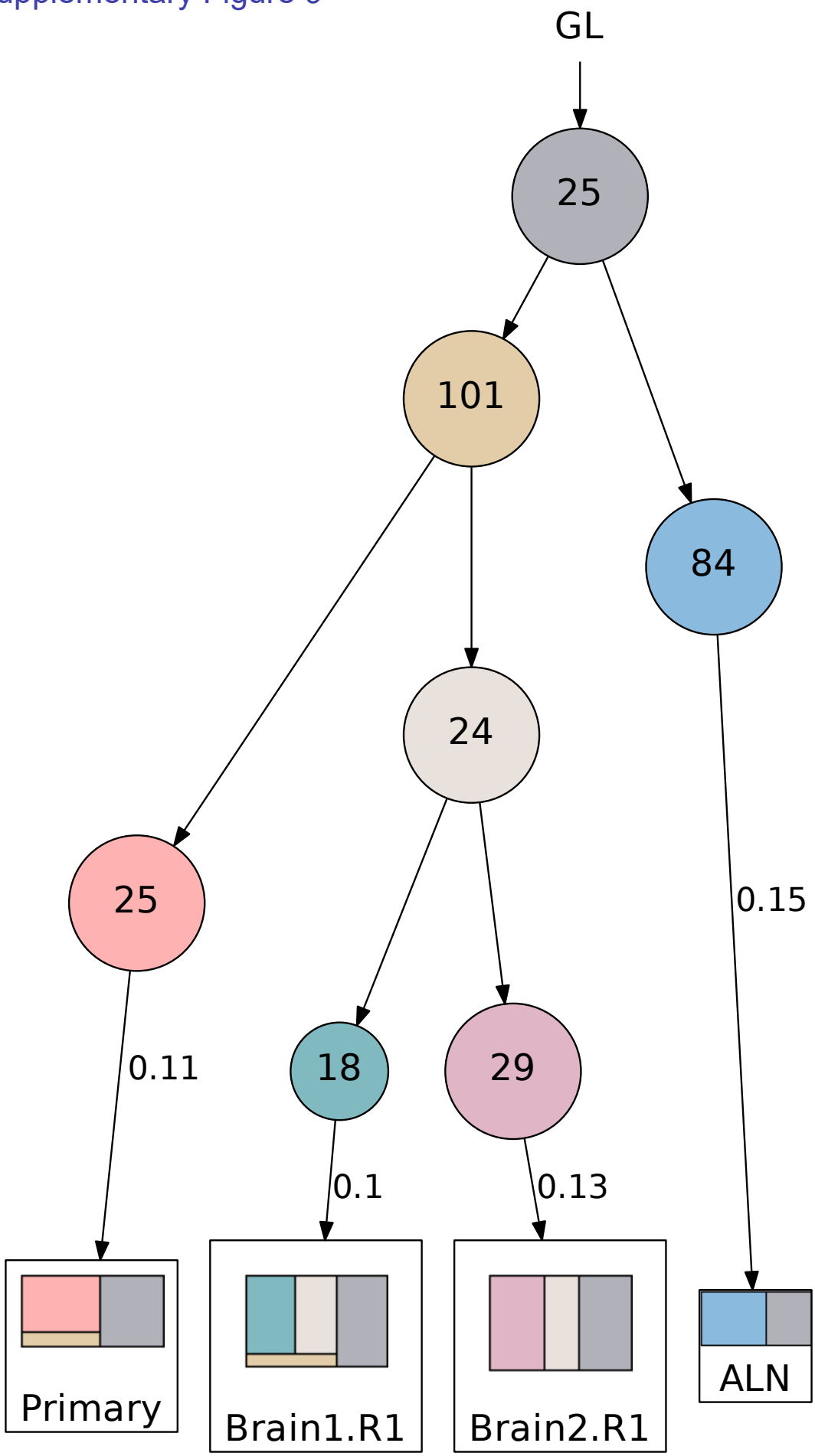
Supplementary Figure 9

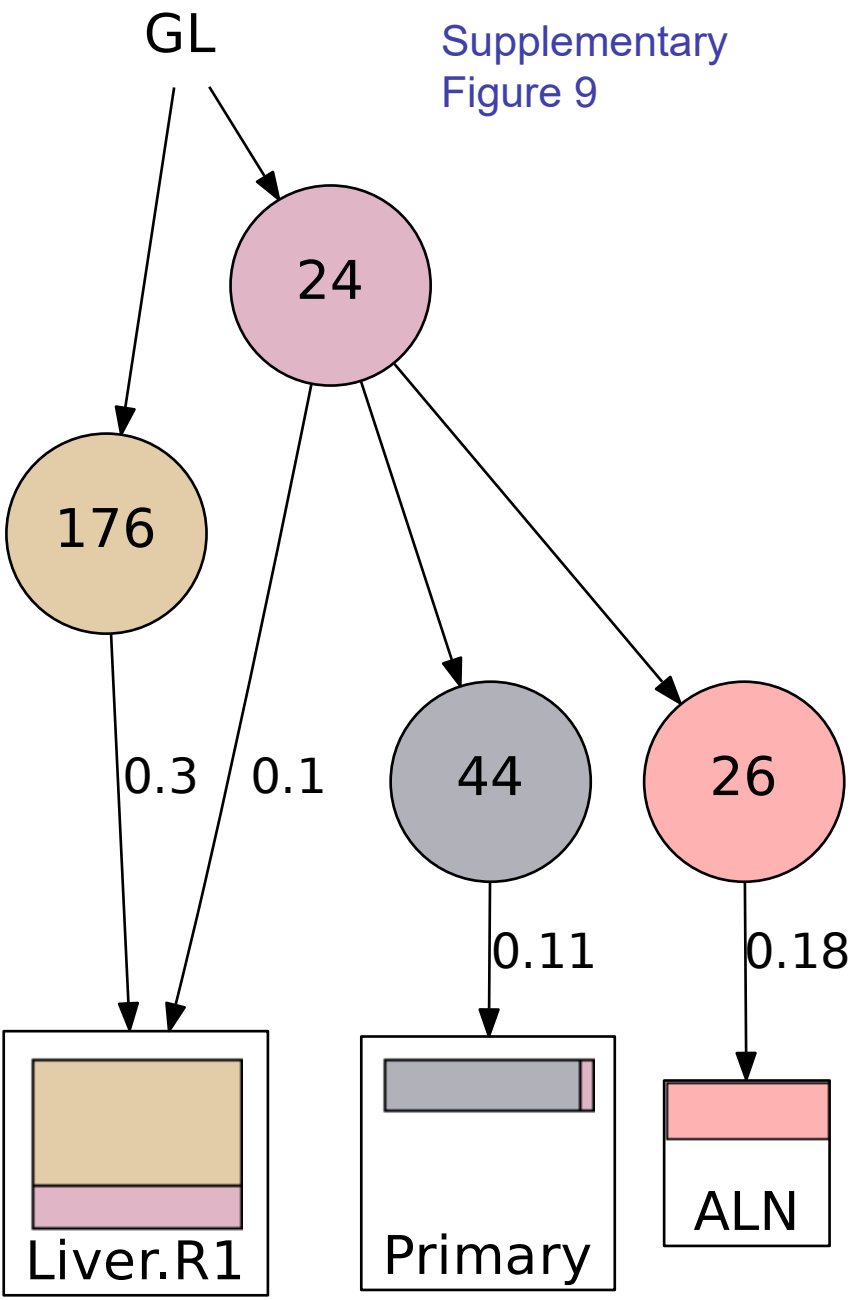


Supplementary Figure 9

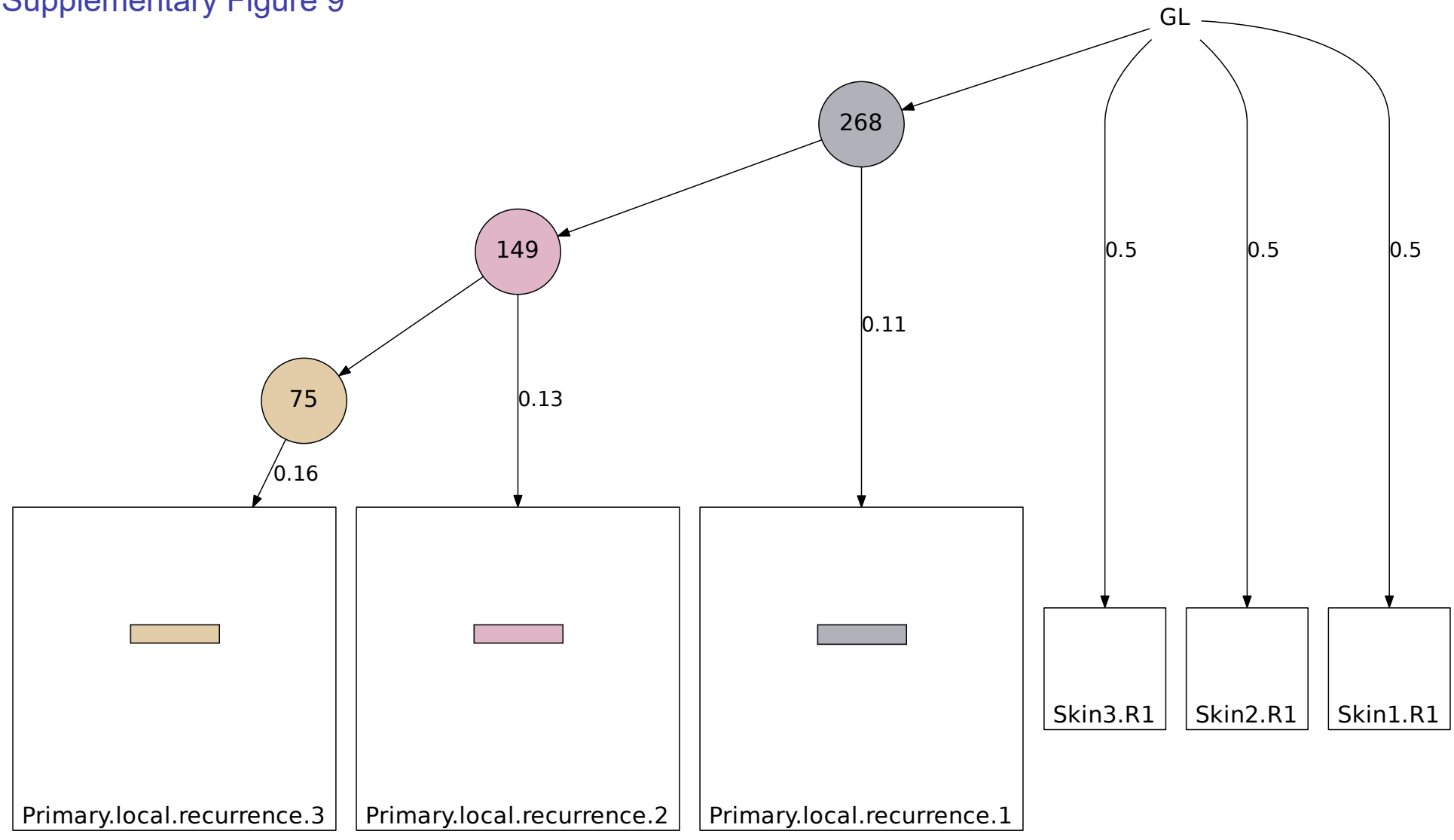


Supplementary Figure 9

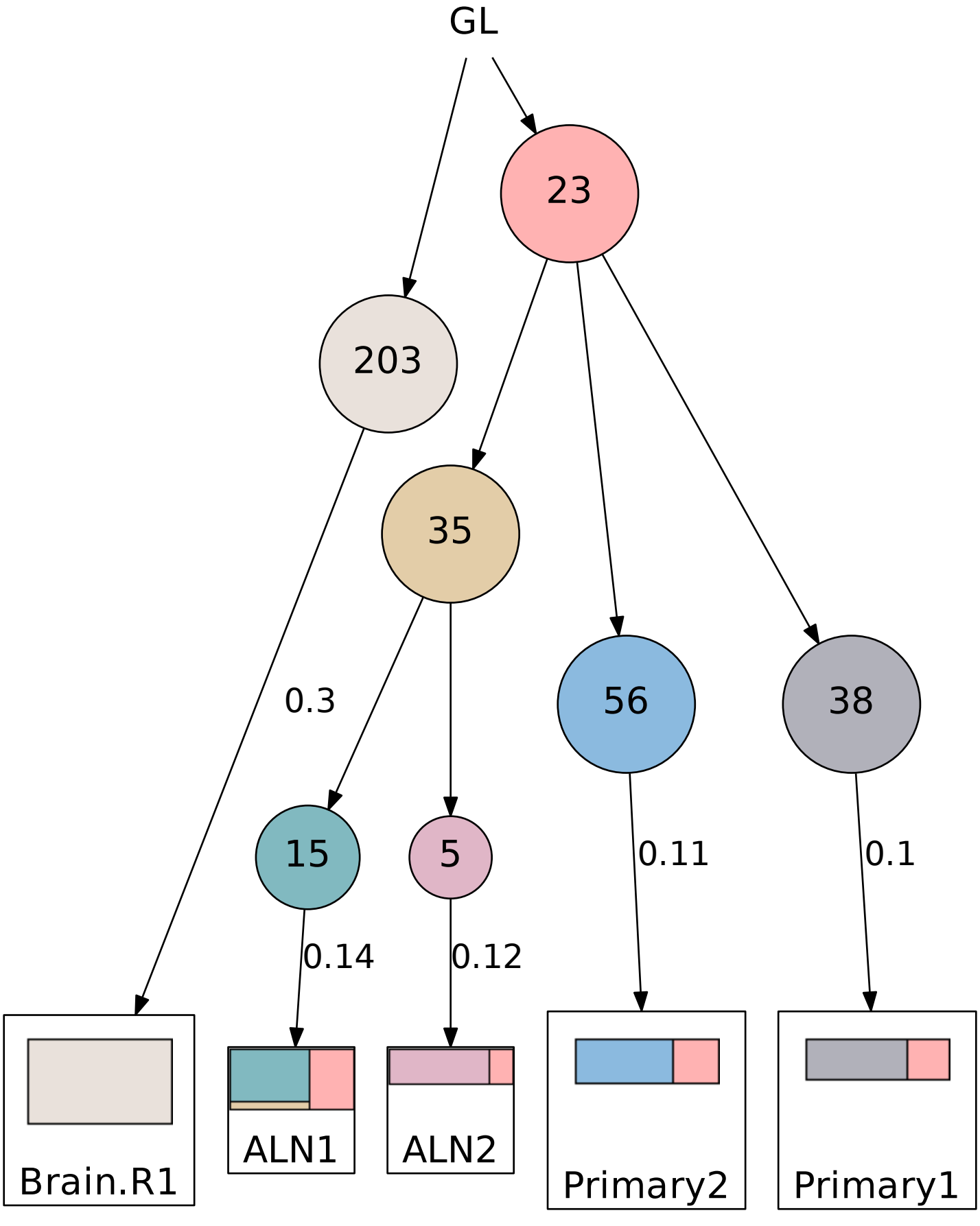



GL
Supplementary
Figure 9
24
176
0.3
0.1
44
26
0.11
0.18
Liver.R1
Primary
ALN

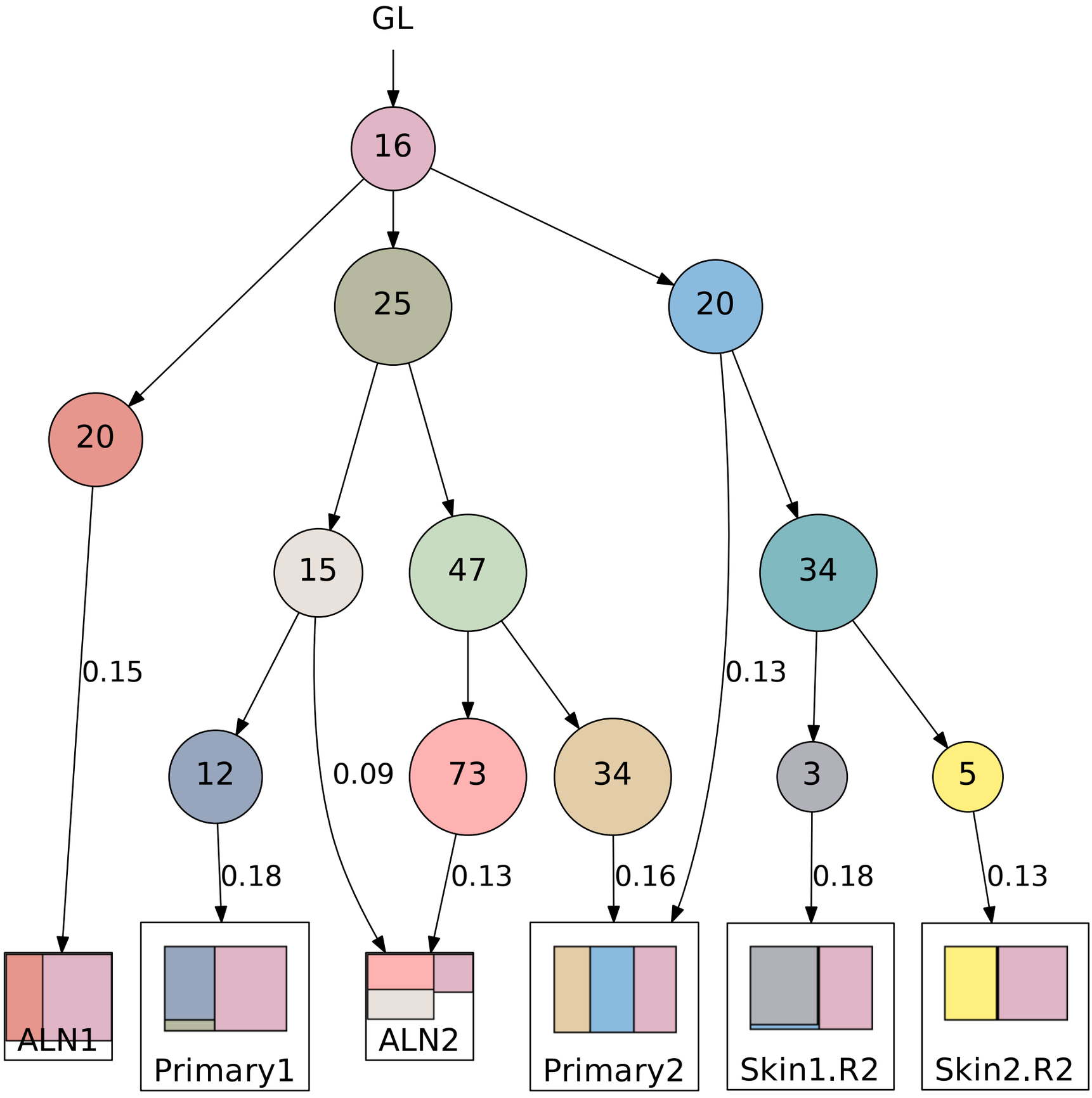
Supplementary Figure 9



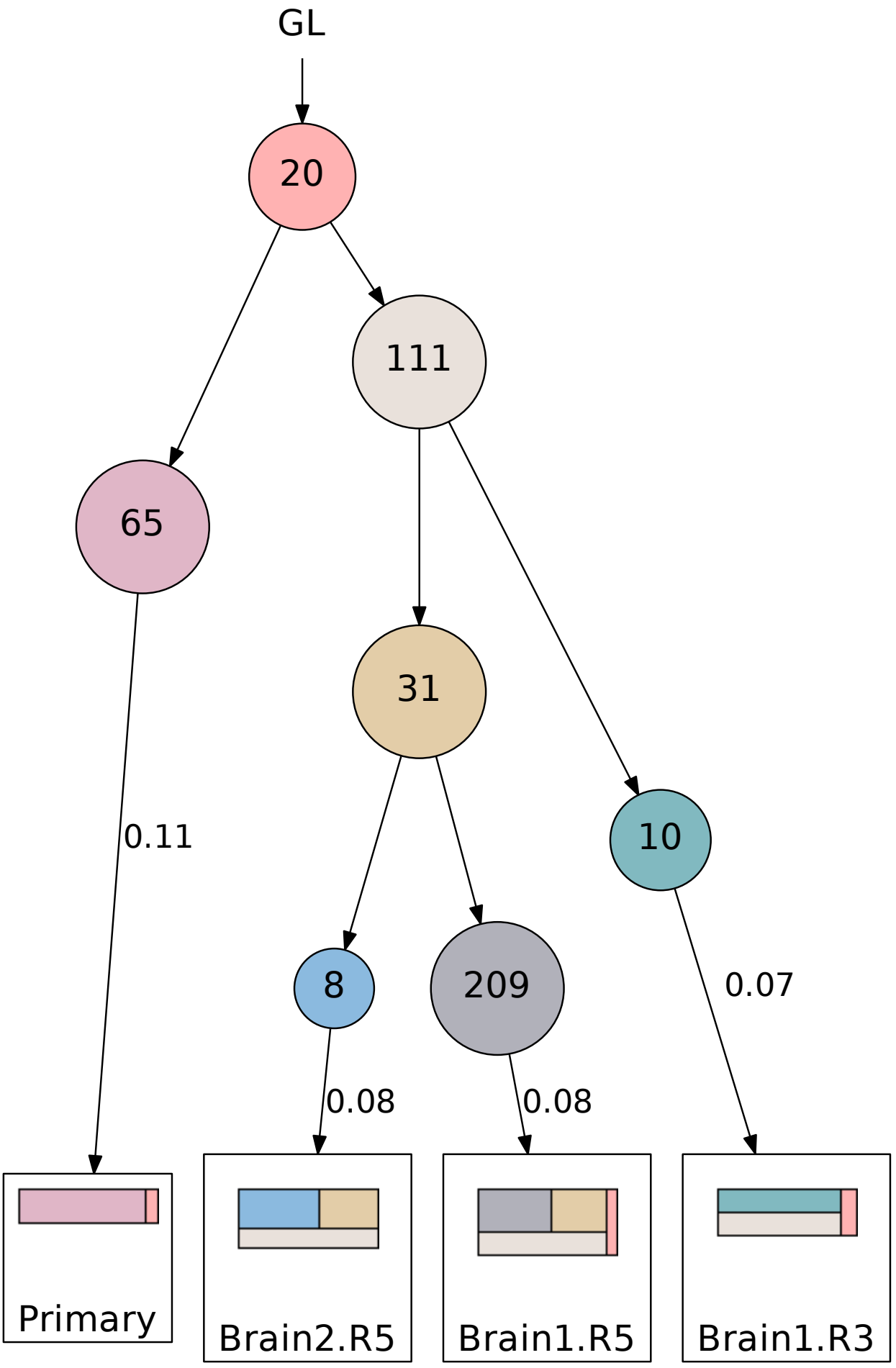
Supplementary Figure 9



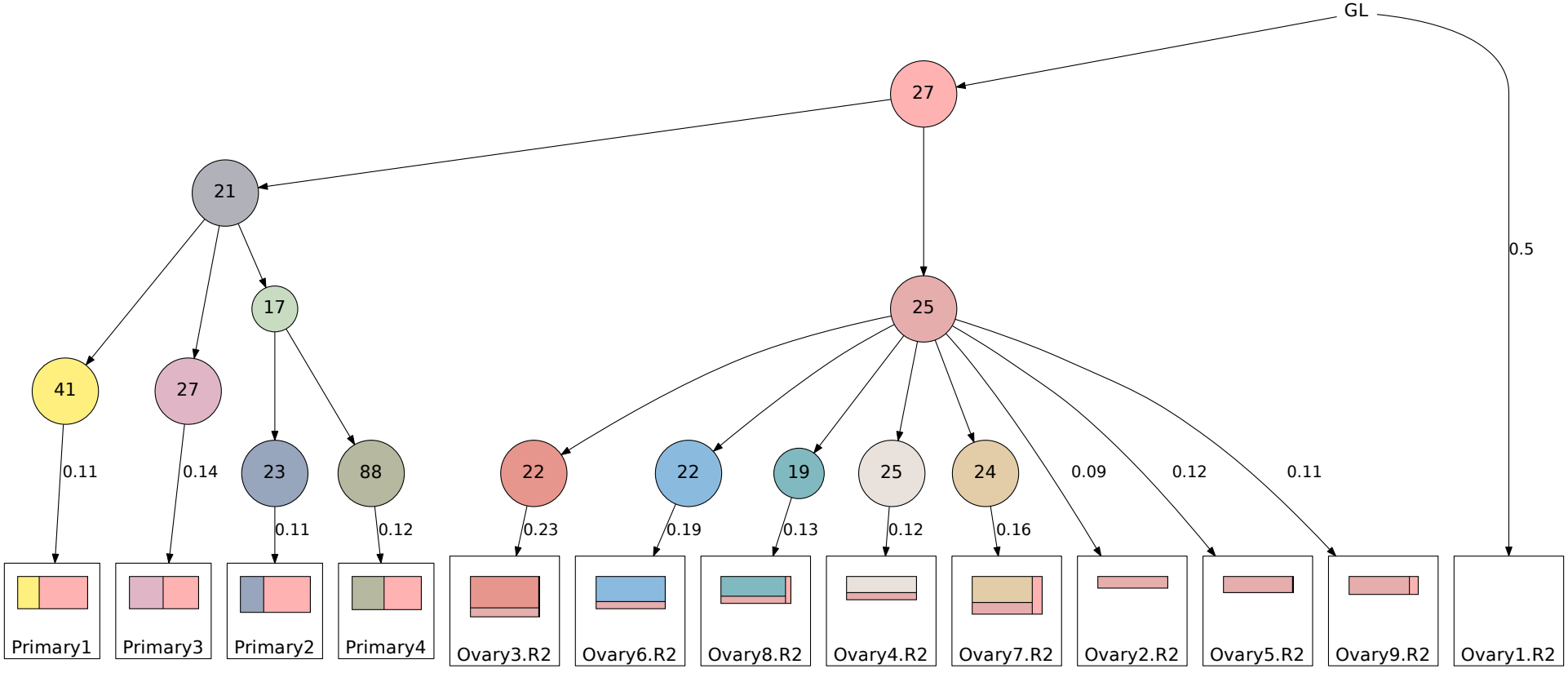
Supplementary Figure 9



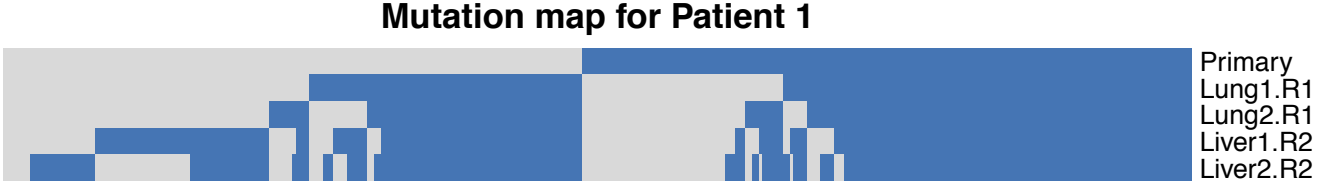
Supplementary Figure 9



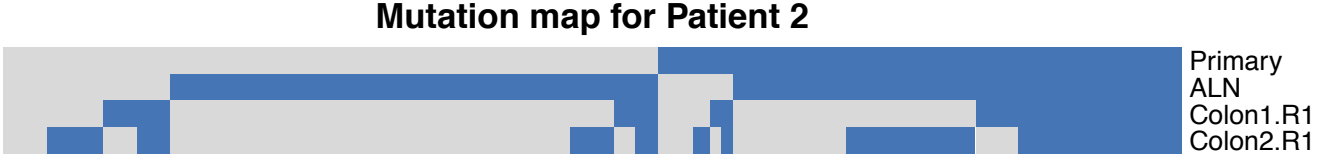
Supplementary Figure 9



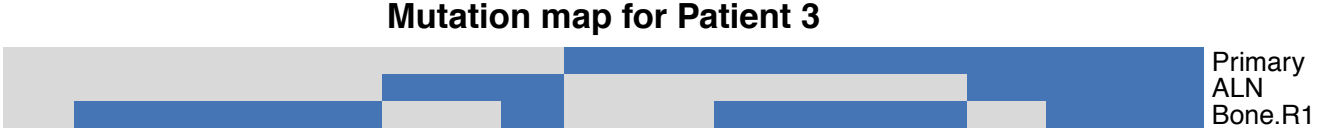
Supplementary Figure 10



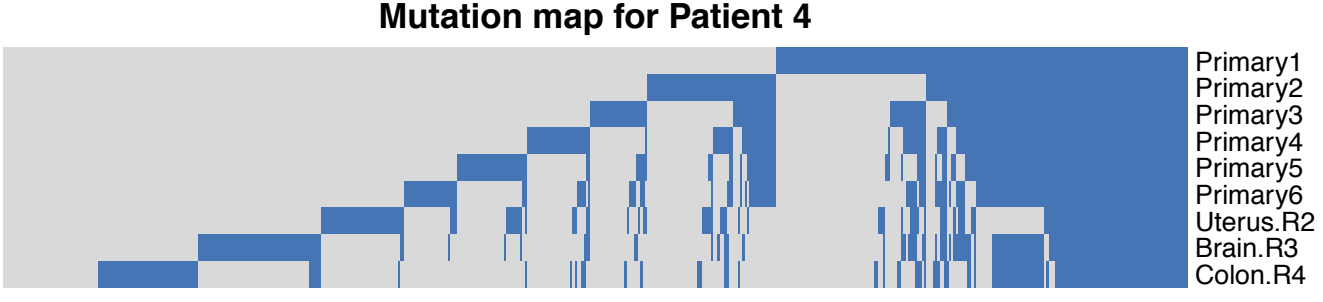
Supplementary Figure 10



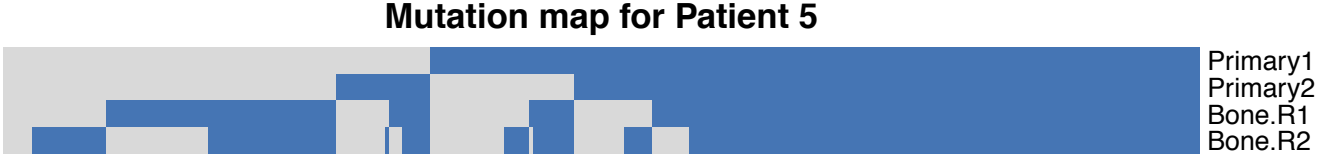
Supplementary Figure 10



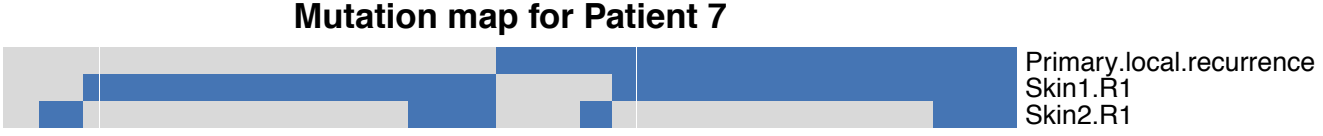
Supplementary Figure 10



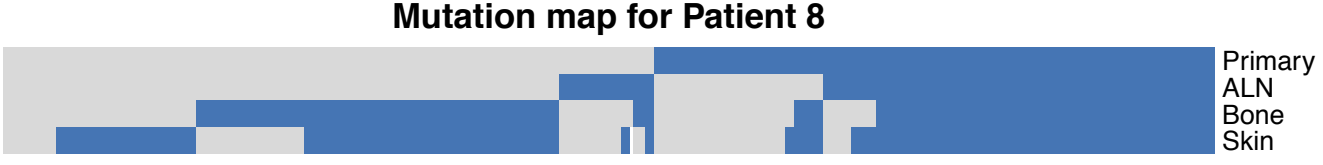
Supplementary Figure 10



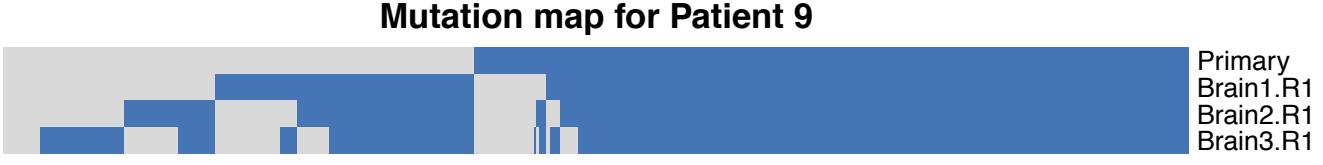
Supplementary Figure 10



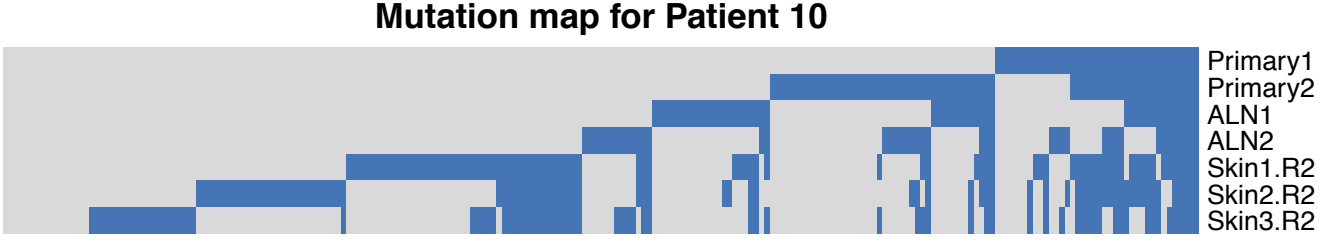
Supplementary Figure 10



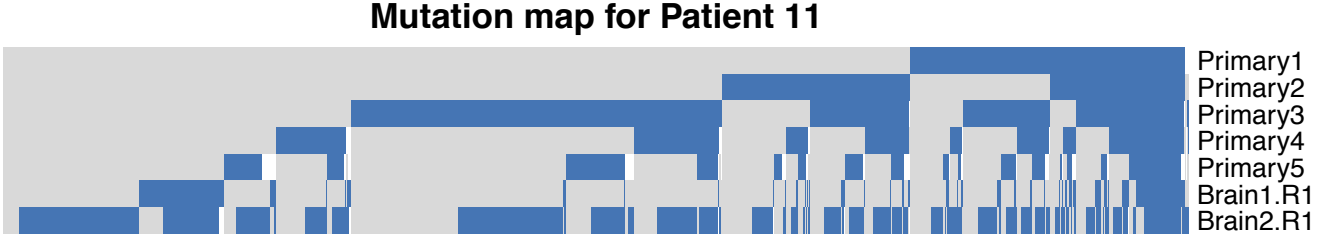
Supplementary Figure 10



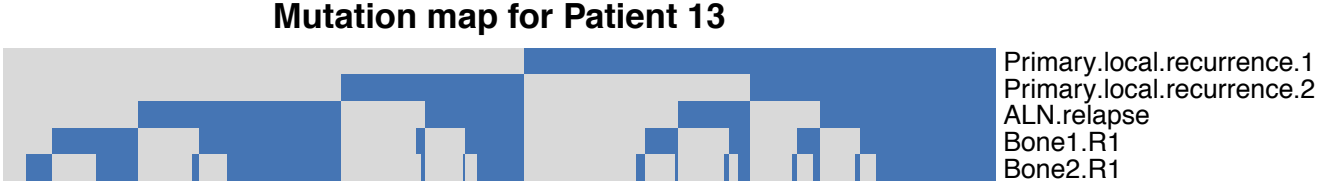
Supplementary Figure 10



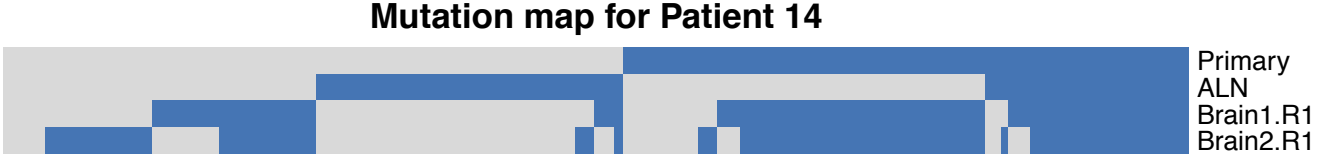
Supplementary Figure 10



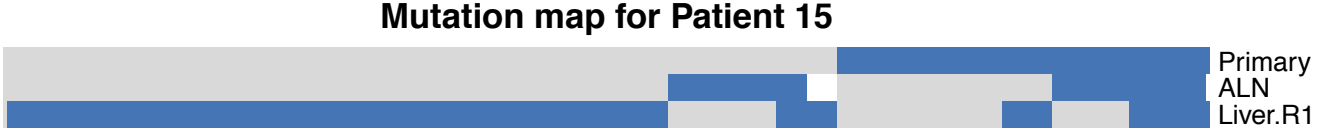
Supplementary Figure 10



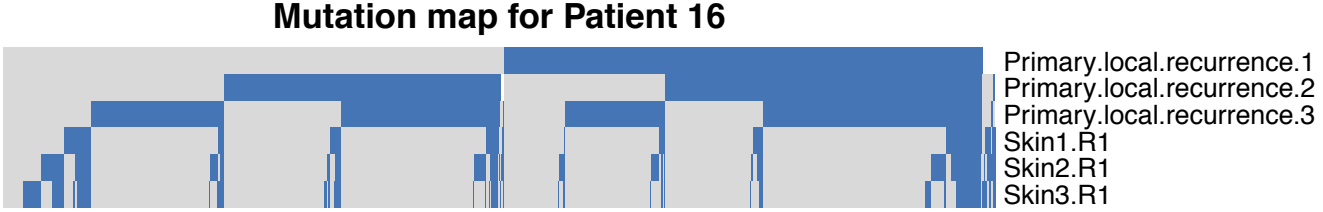
Supplementary Figure 10



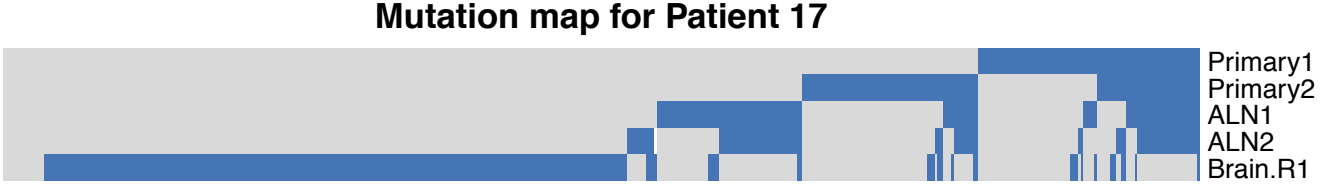
Supplementary Figure 10



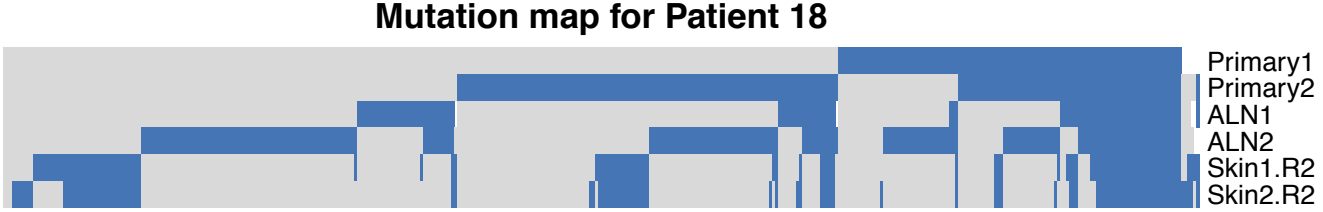
Supplementary Figure 10



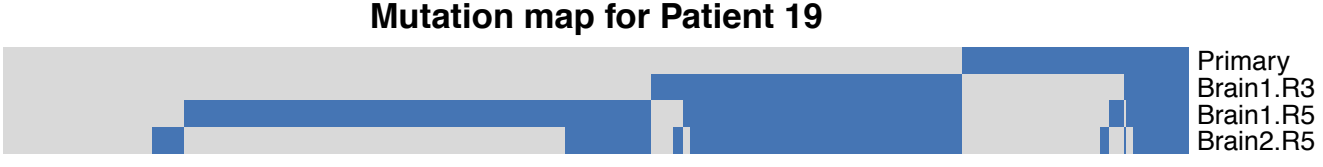
Supplementary Figure 10



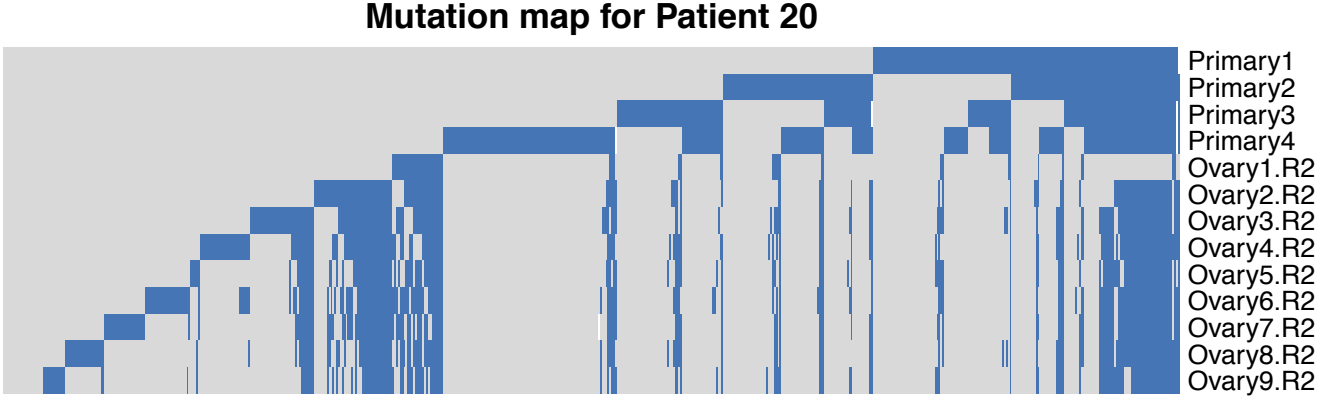
Supplementary Figure 10



Supplementary Figure 10



Supplementary Figure 10



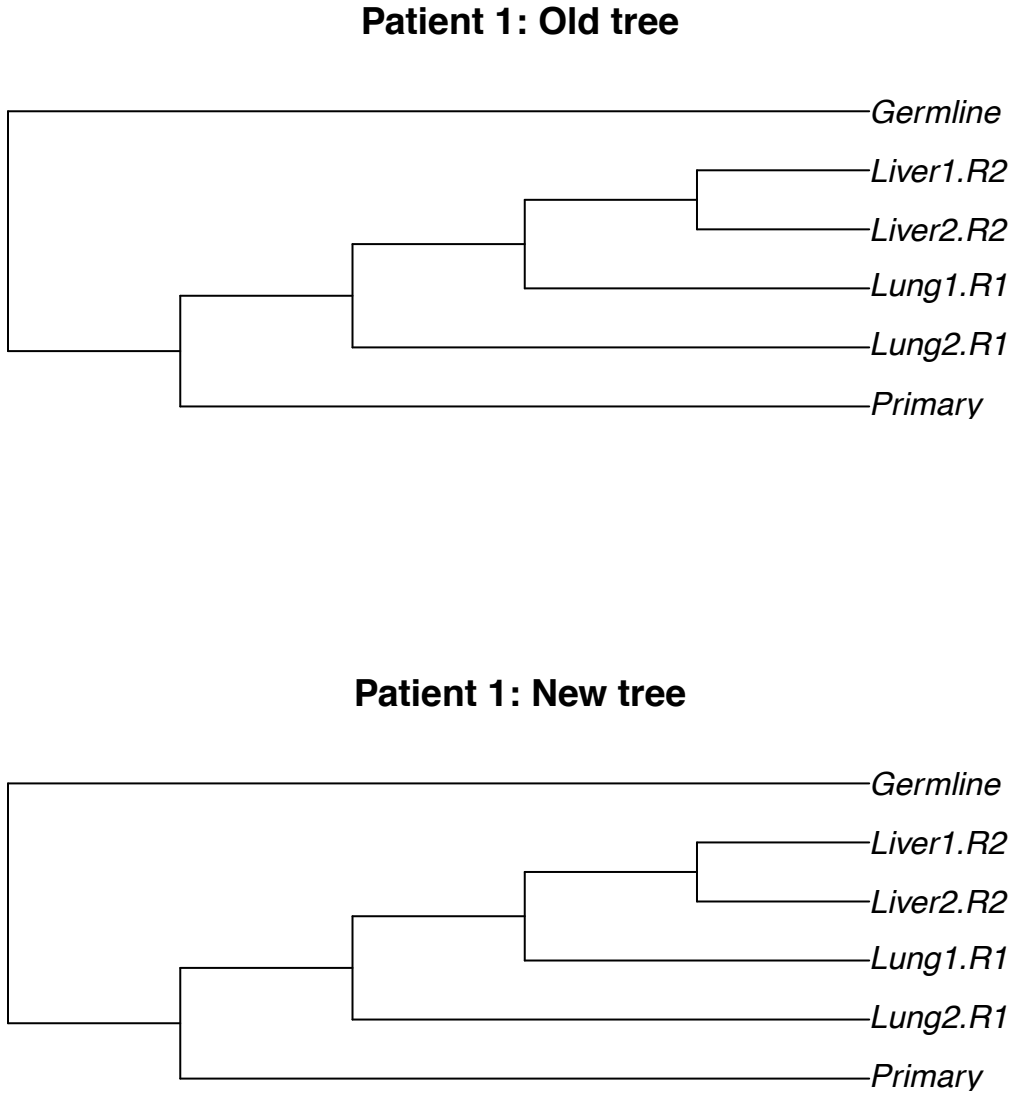
Supplementary Figure 11

Validation of phylogenetic trees using more conservative mutation-filtering criteria

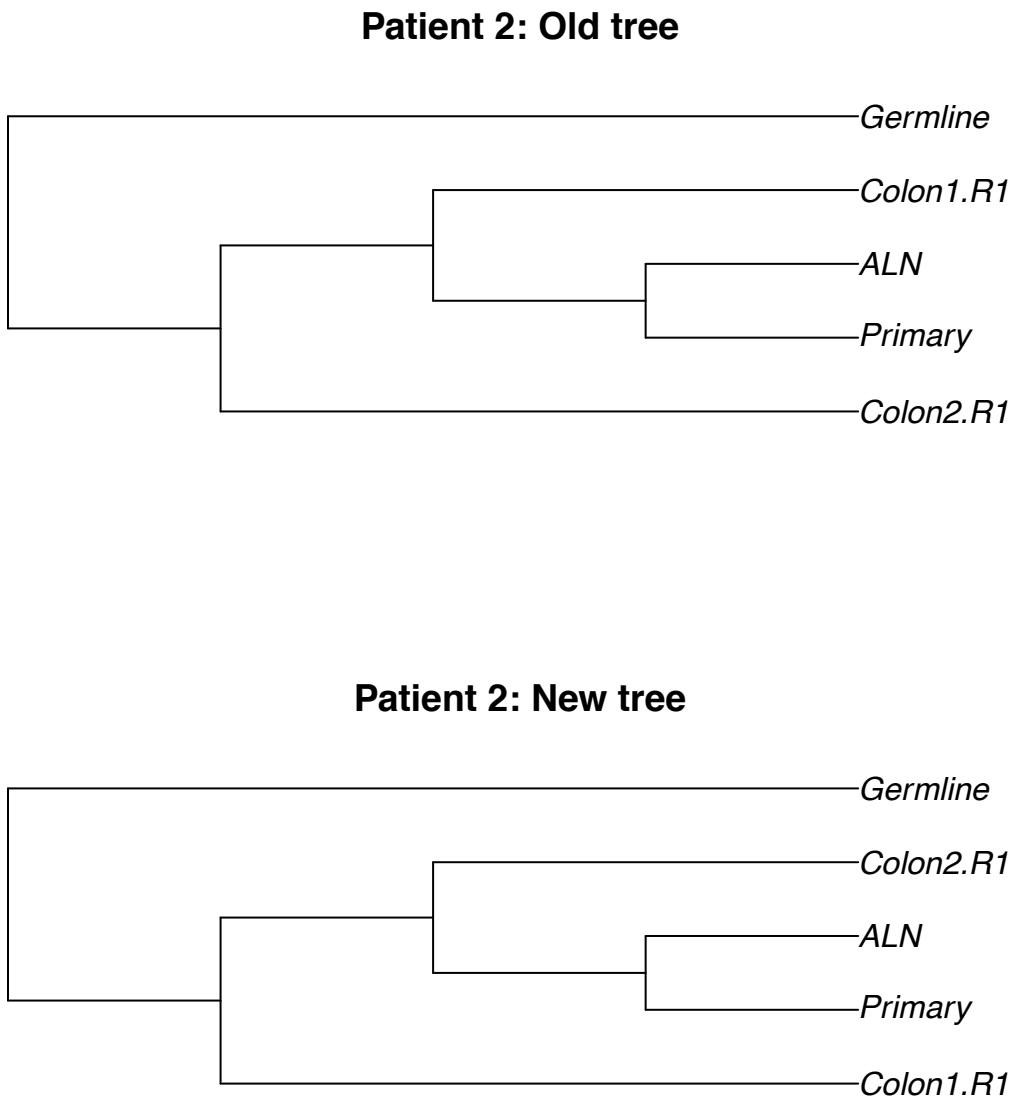
For the phylogenetic trees reported in the manuscript (Figure 2, 3, 4 and Supplementary Figure 6), we used mutations obtained using mutation-calling criteria as described in the Methods section. Next, we tested the robustness of these results by removing mutations affected by variable coverage and/or different tumor purity among samples. For this, we used mutation-filtering criteria as described in "Validation of phylogenetic trees" subsection of the Method section.

In the following pages, we present, for each patient, a side-by-side comparison of the tree reported in the manuscript (termed here **Old tree**) vs. the one reconstructed using the more conservative mutation selection criteria (termed here **New tree**).

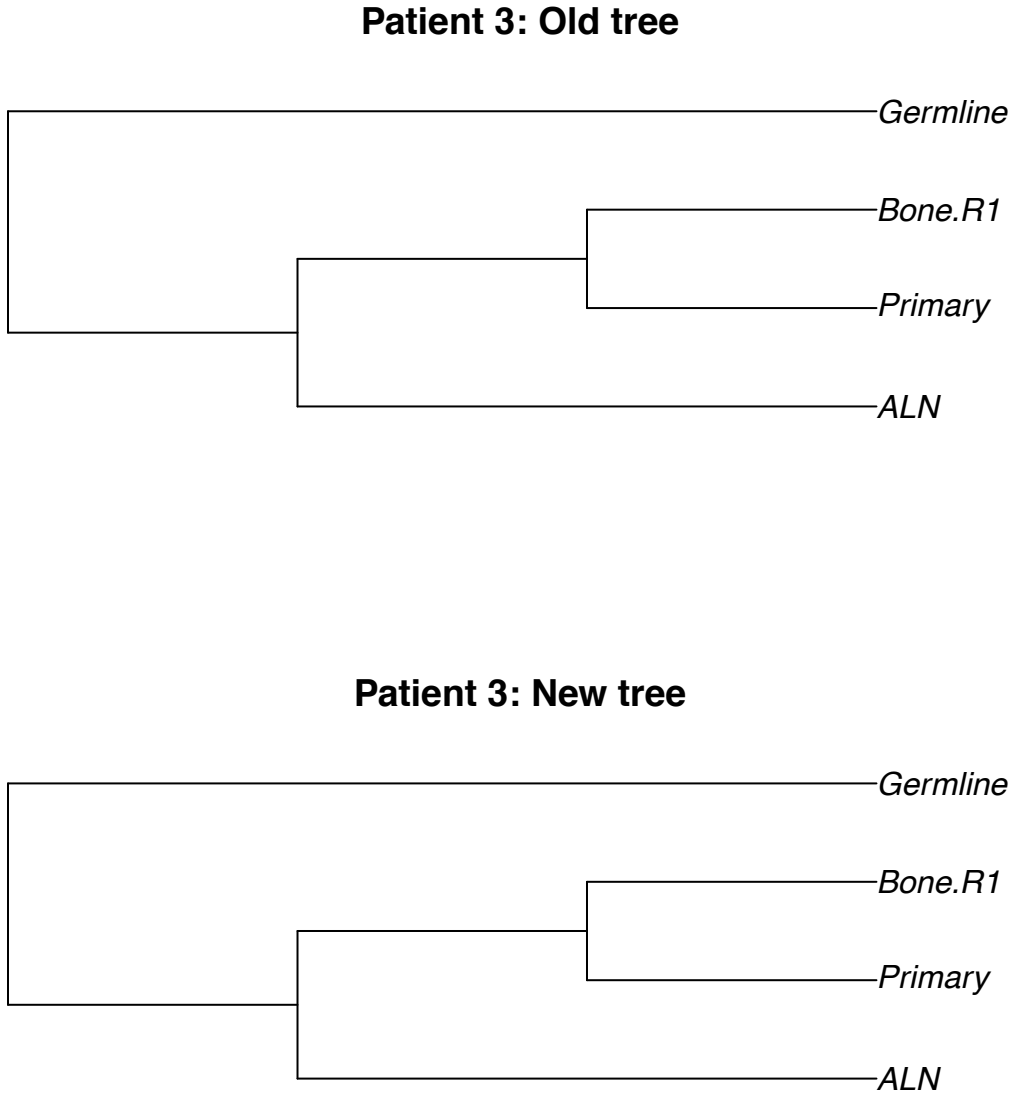
Supplementary Figure 11



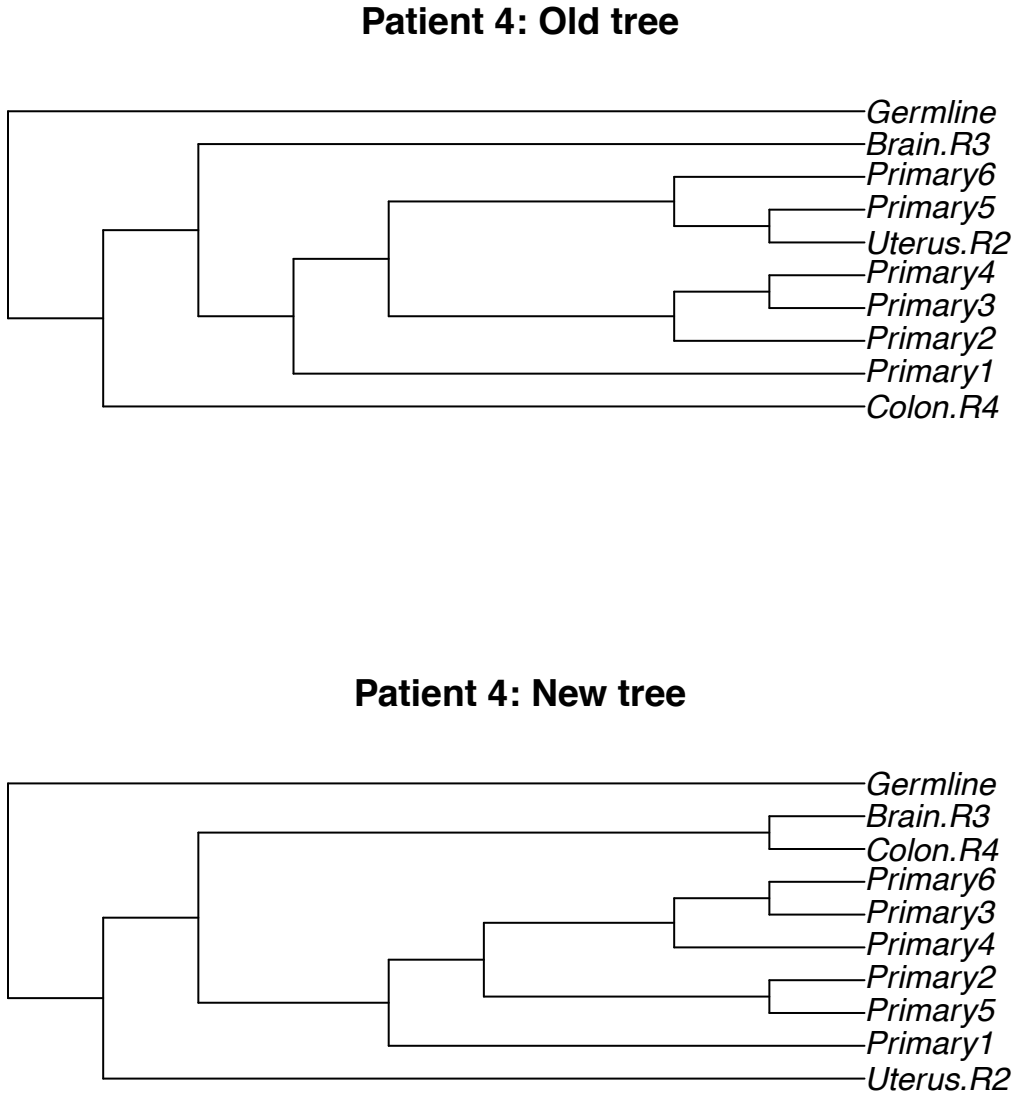
Supplementary Figure 11



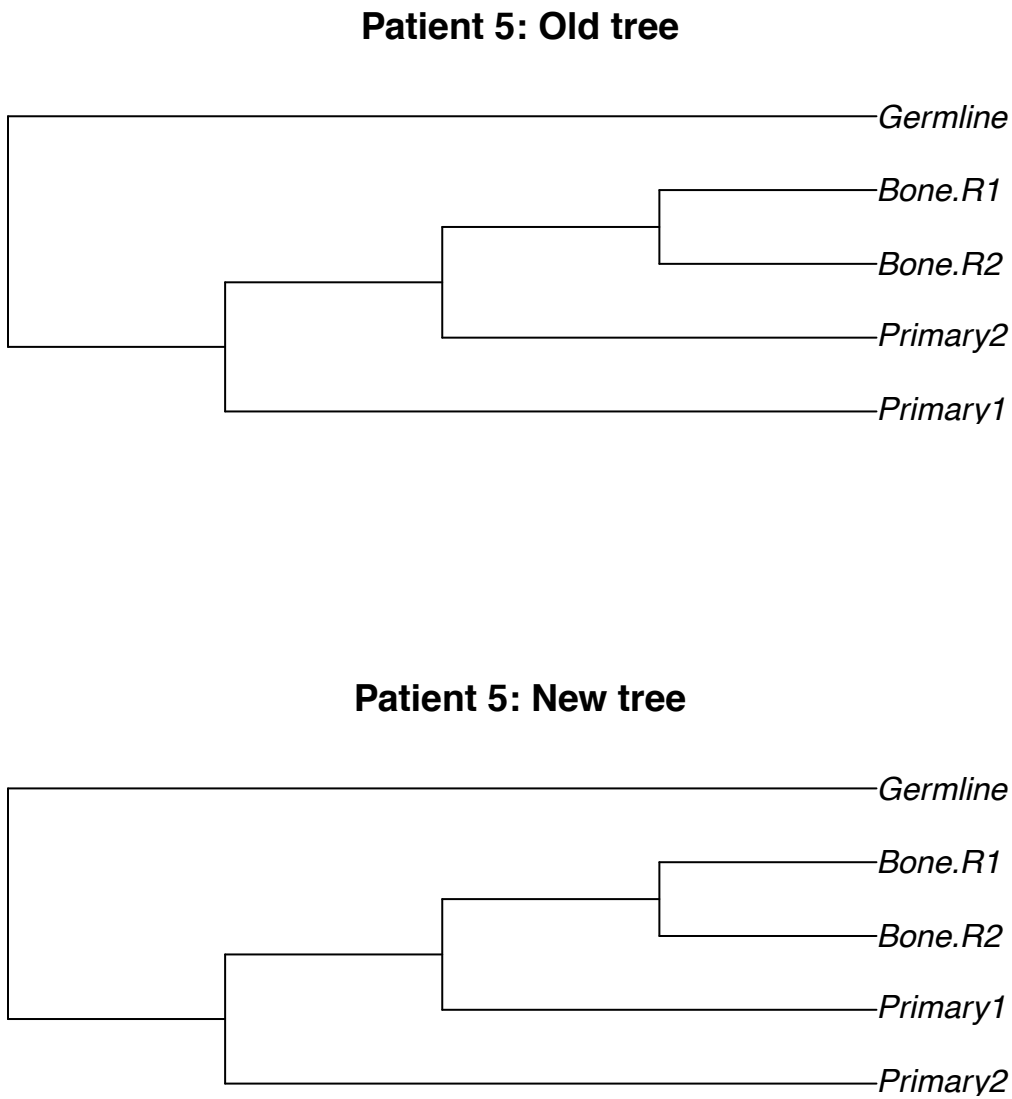
Supplementary Figure 11



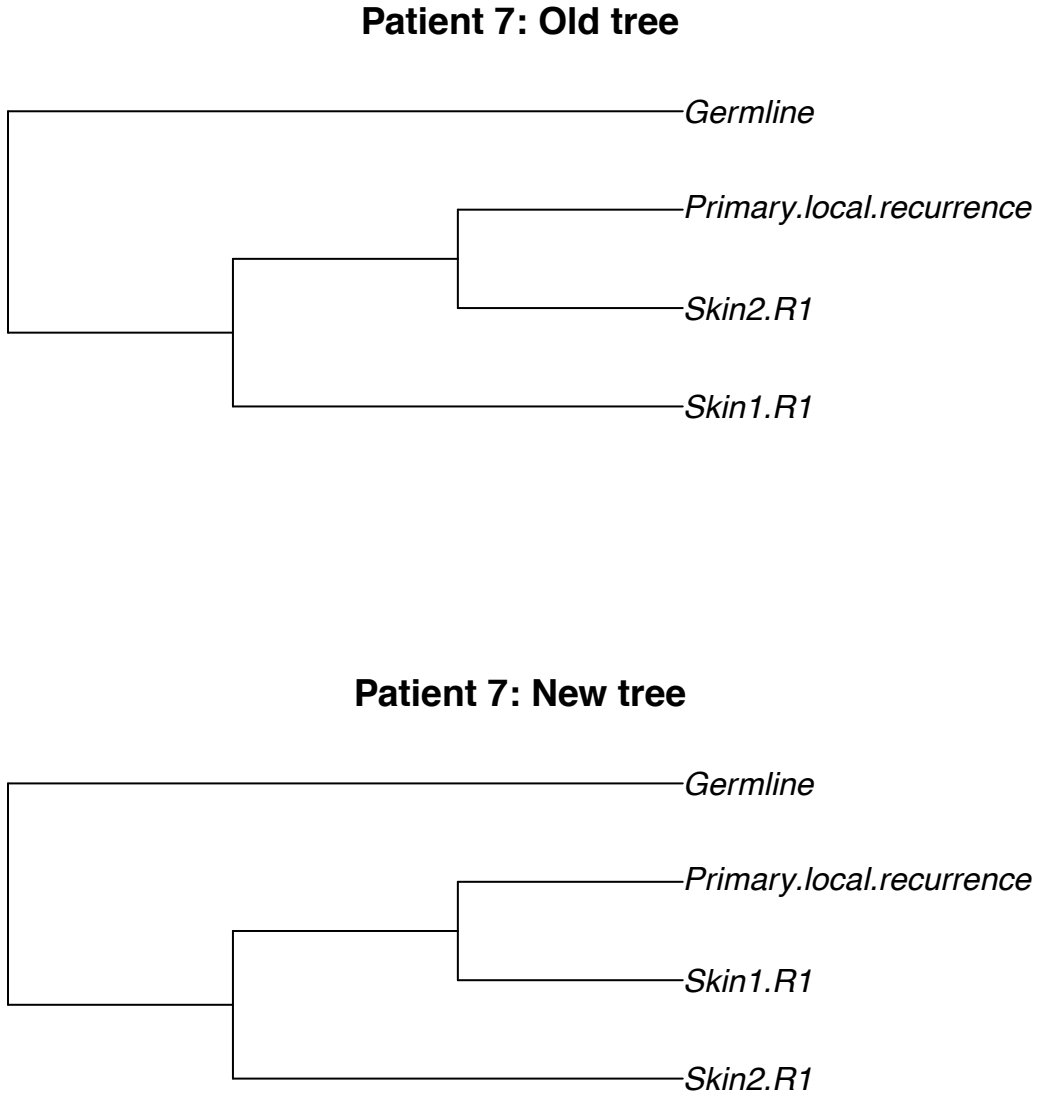
Supplementary Figure 11



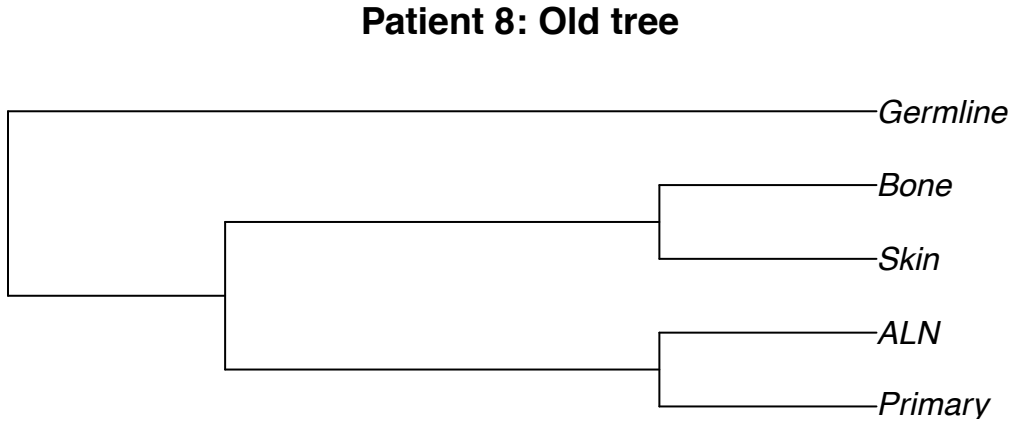
Supplementary Figure 11

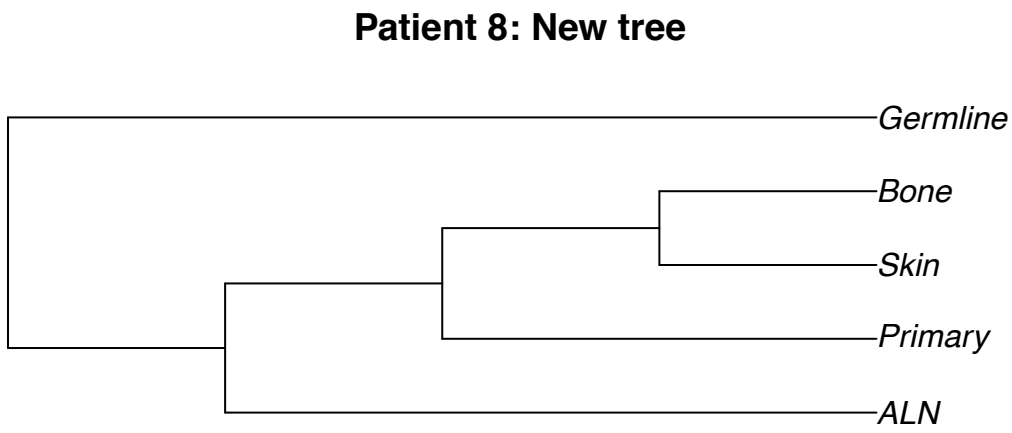


Supplementary Figure 11



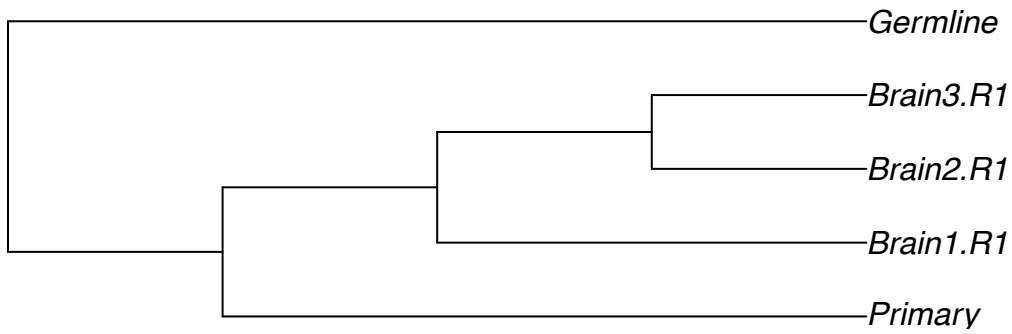
Supplementary Figure 11



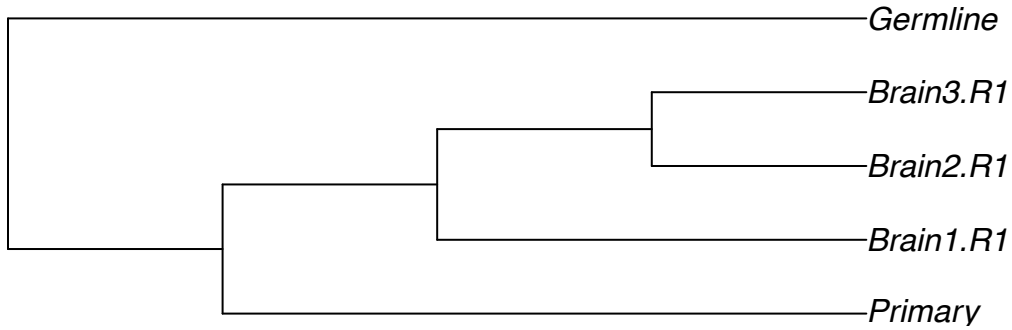


Supplementary Figure 11

Patient 9: Old tree

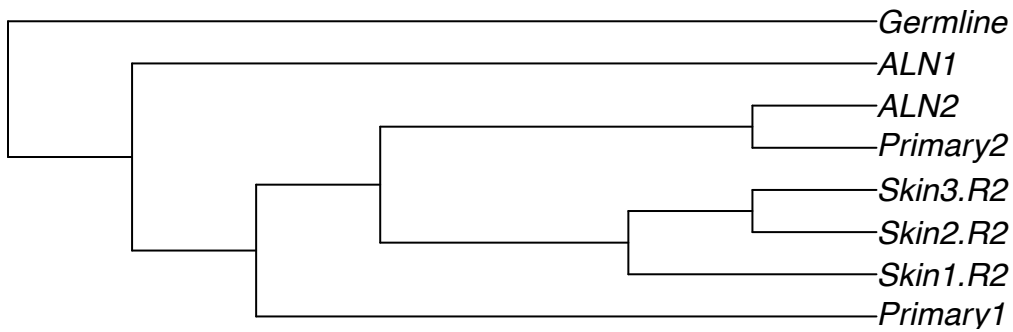


Patient 9: New tree

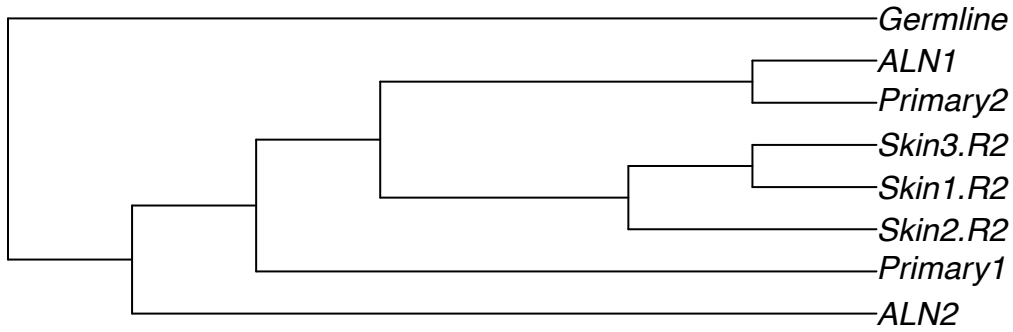


Supplementary Figure 11

Patient 10: Old tree

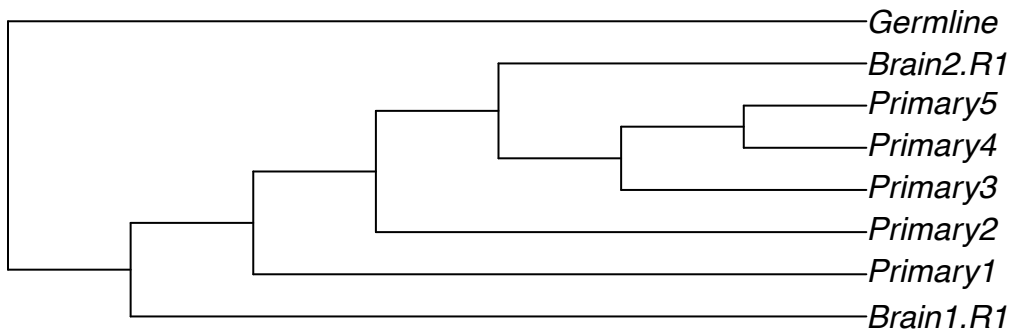


Patient 10: New tree

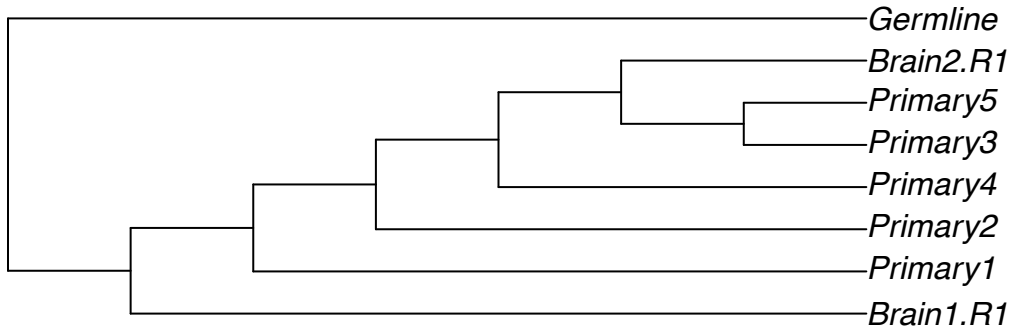


Supplementary Figure 11

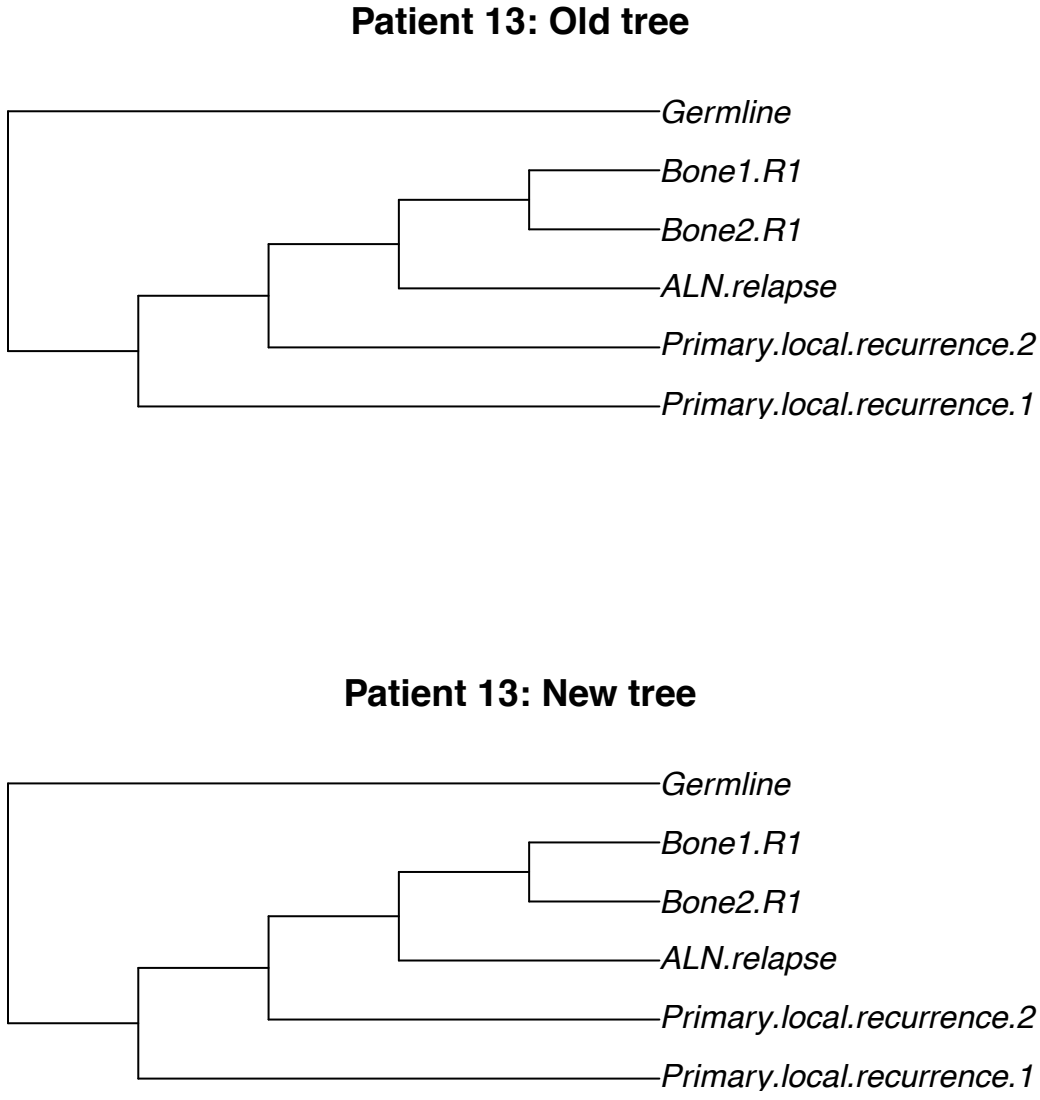
Patient 11: Old tree



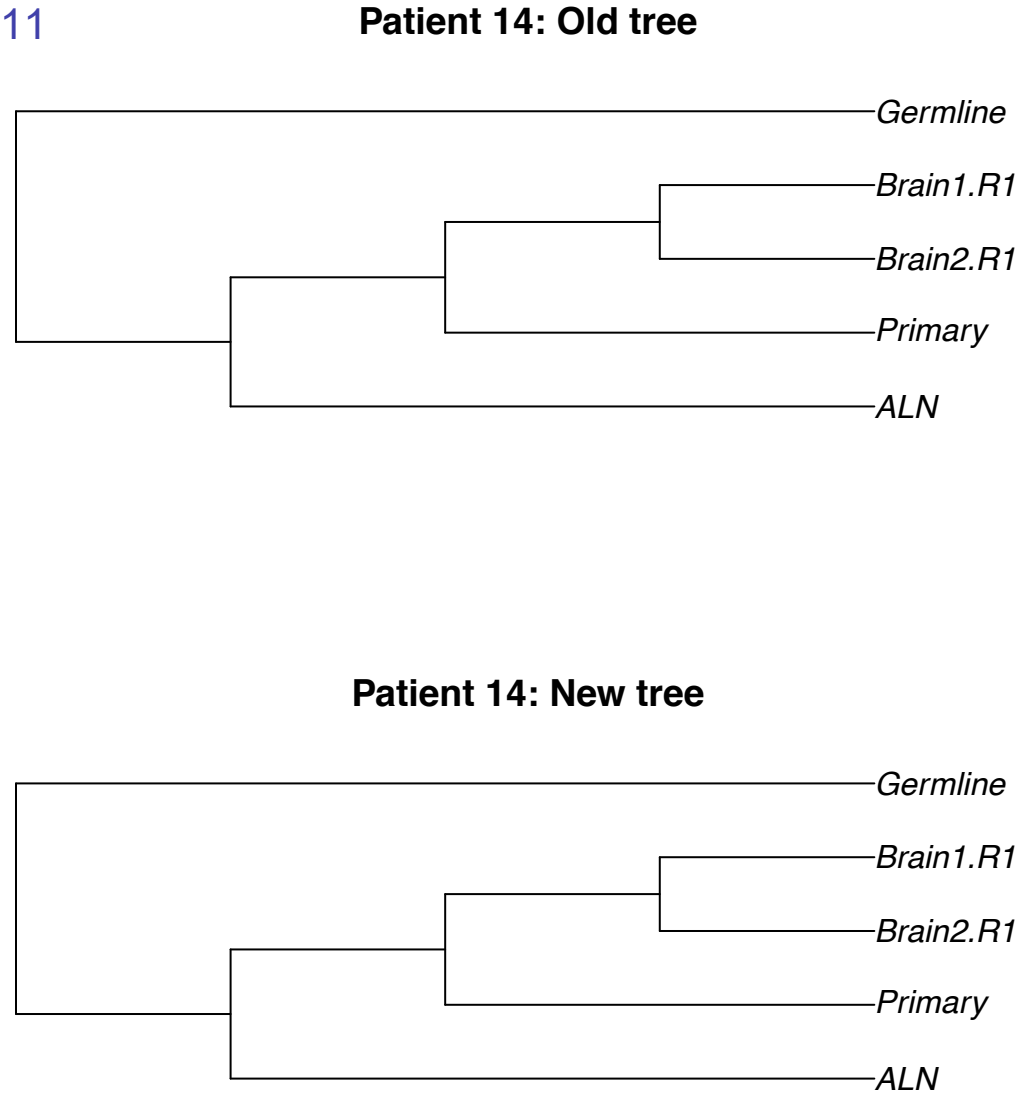
Patient 11: New tree



Supplementary Figure 11

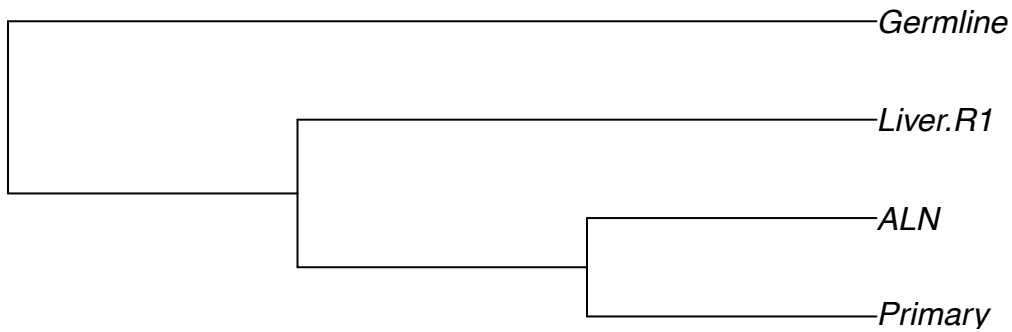


Supplementary Figure 11

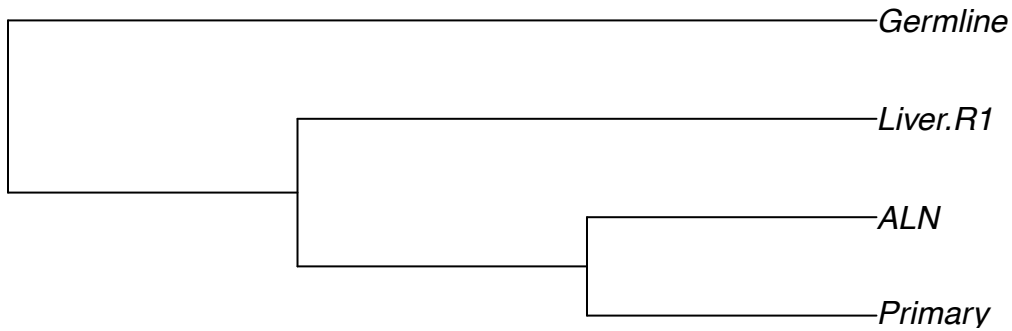


Supplementary Figure 11

Patient 15: Old tree

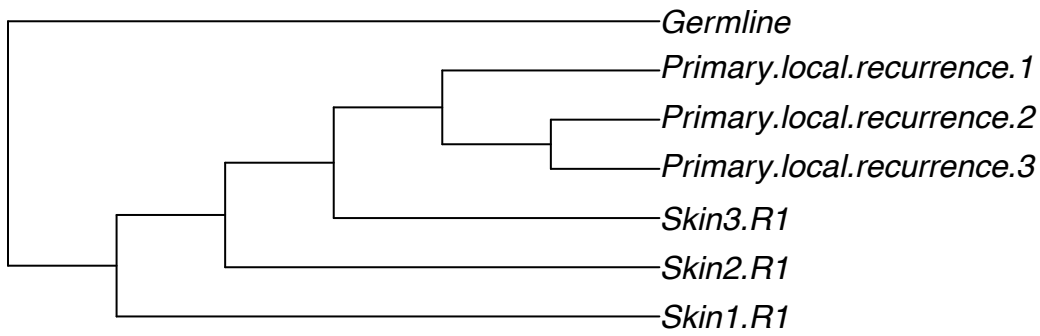


Patient 15: New tree

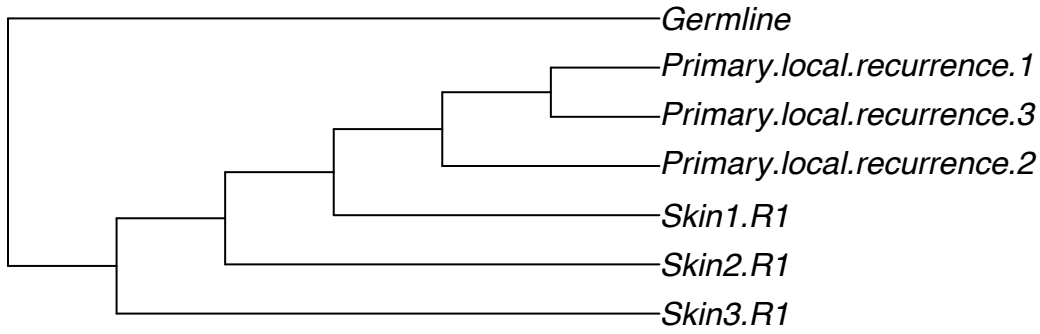


Supplementary Figure 11

Patient 16: Old tree

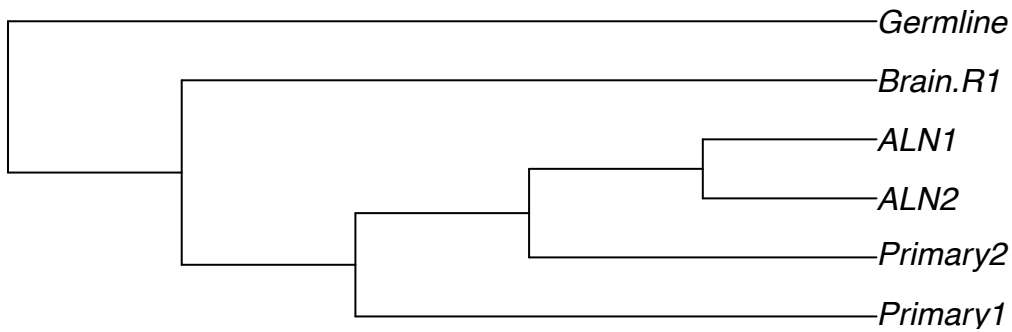


Patient 16: New tree

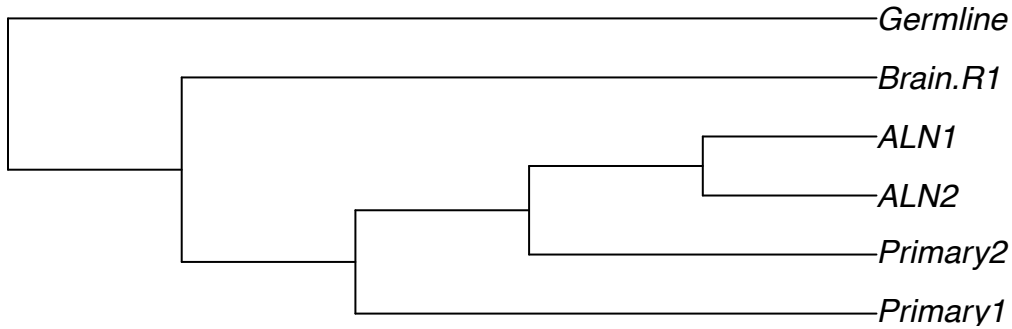


Supplementary Figure 11

Patient 17: Old tree

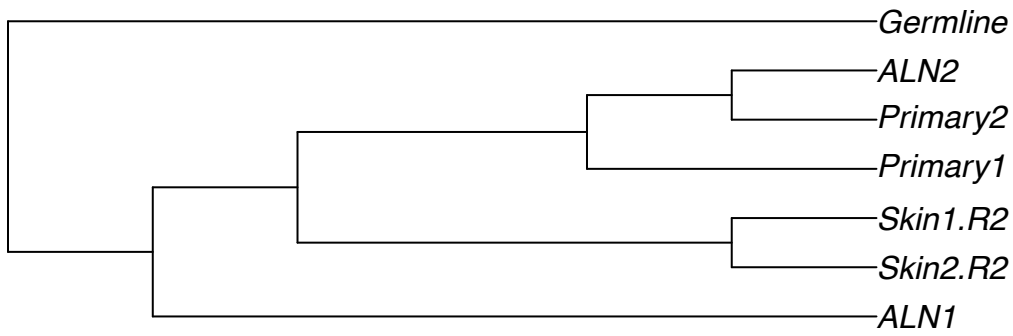


Patient 17: New tree

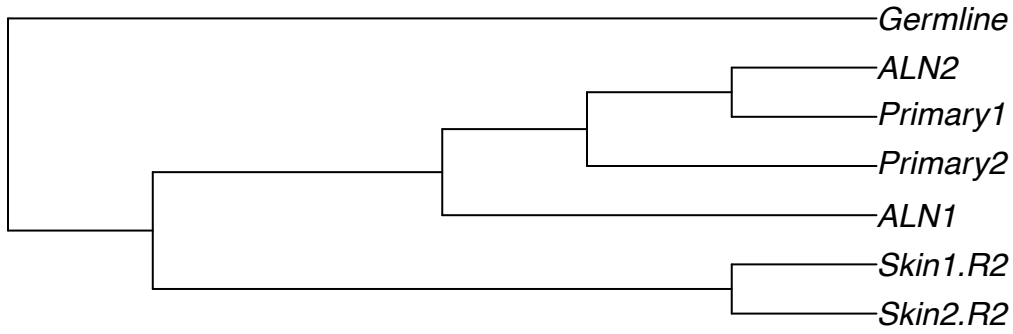


Supplementary Figure 11

Patient 18: Old tree

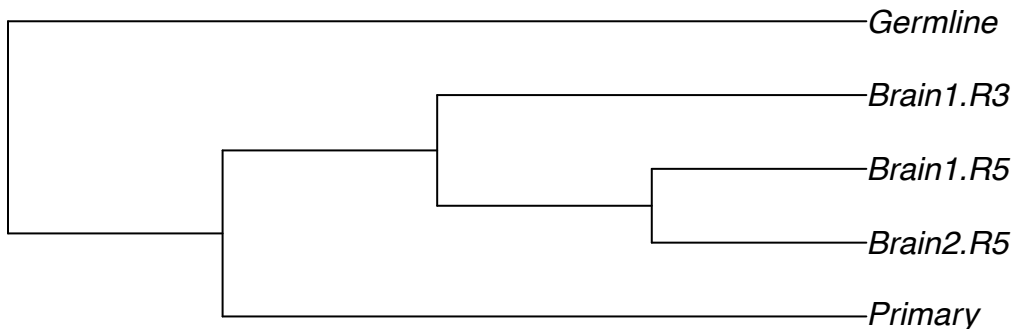


Patient 18: New tree

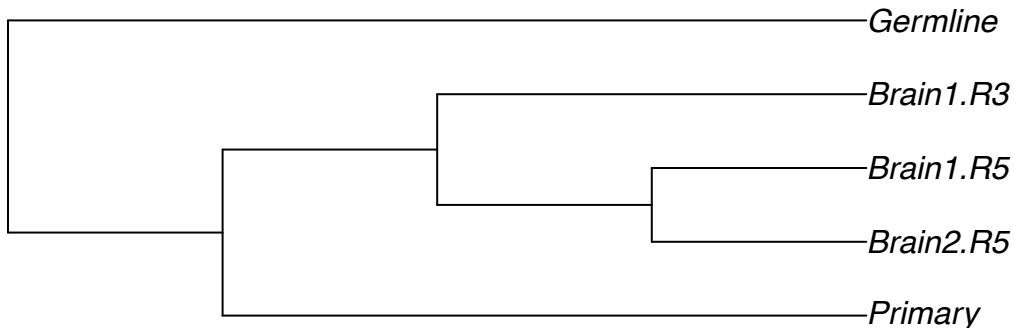


Supplementary Figure 11

Patient 19: Old tree

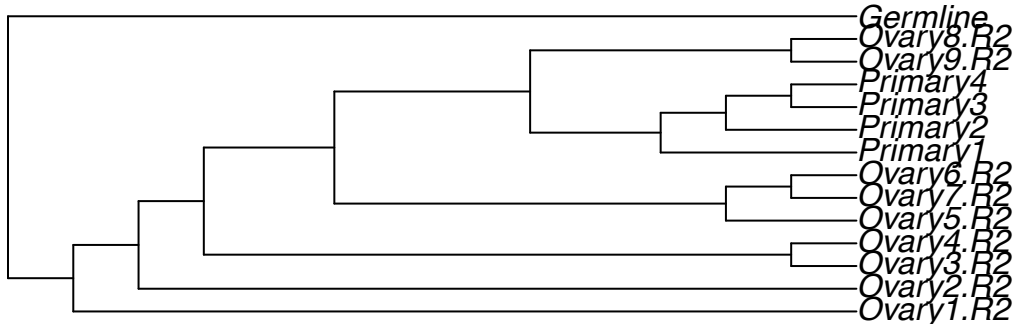


Patient 19: New tree

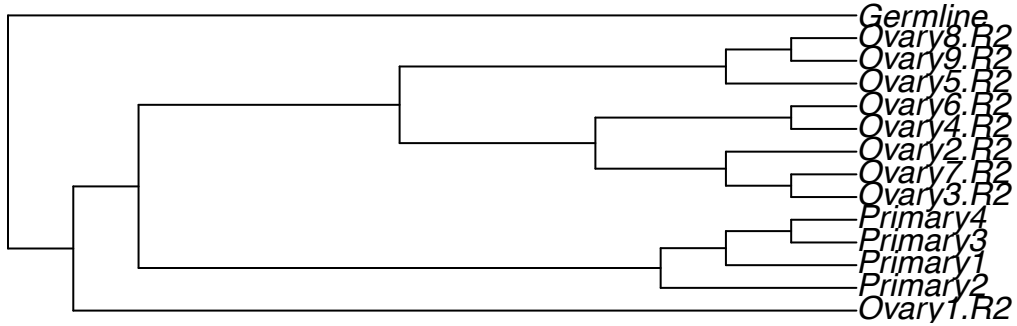


Supplementary Figure 11

Patient 20: Old tree



Patient 20: New tree



Supplementary Figure 12

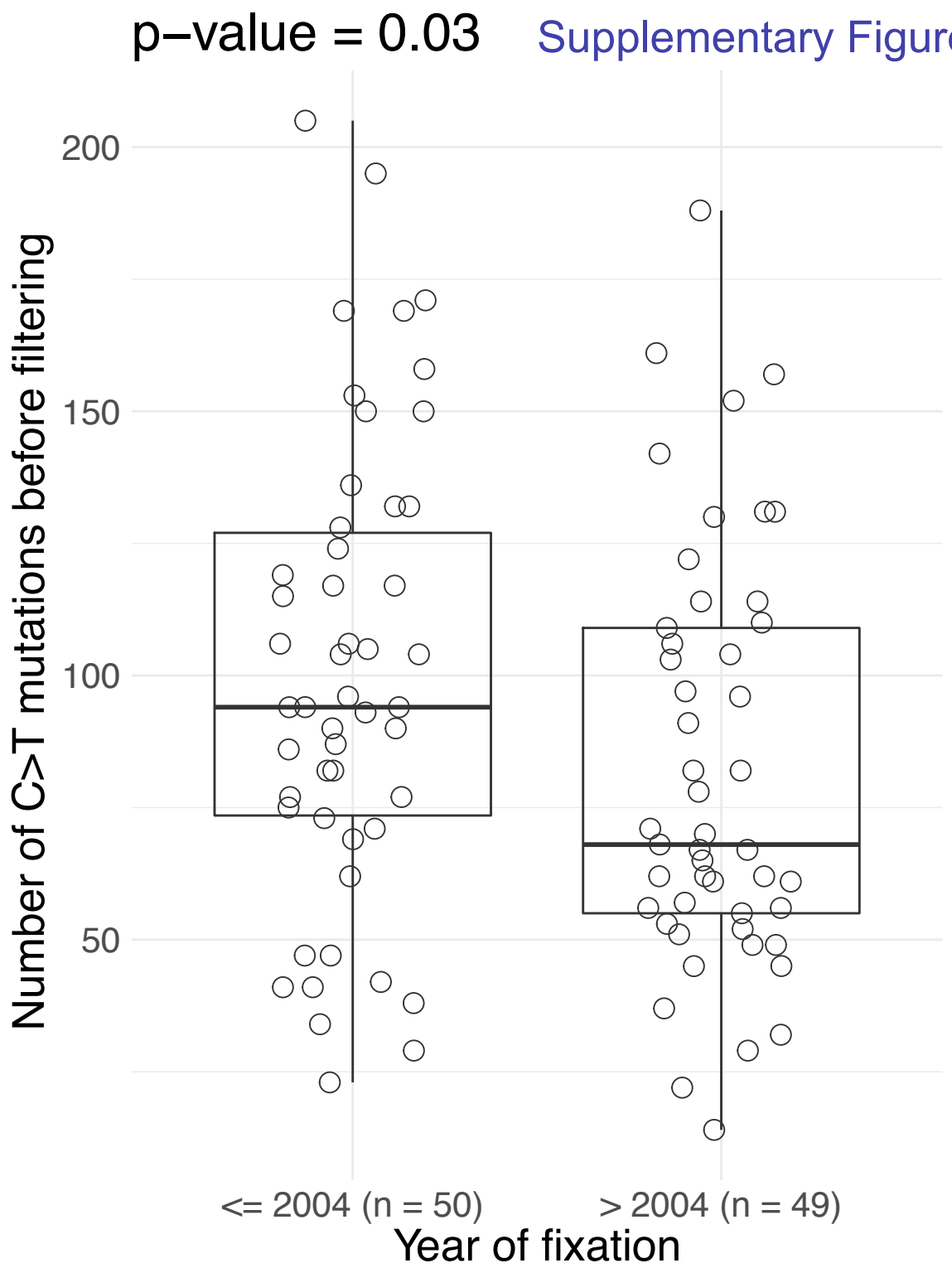
Efficient filtering of FFPE-related C>T/G>A artifacts

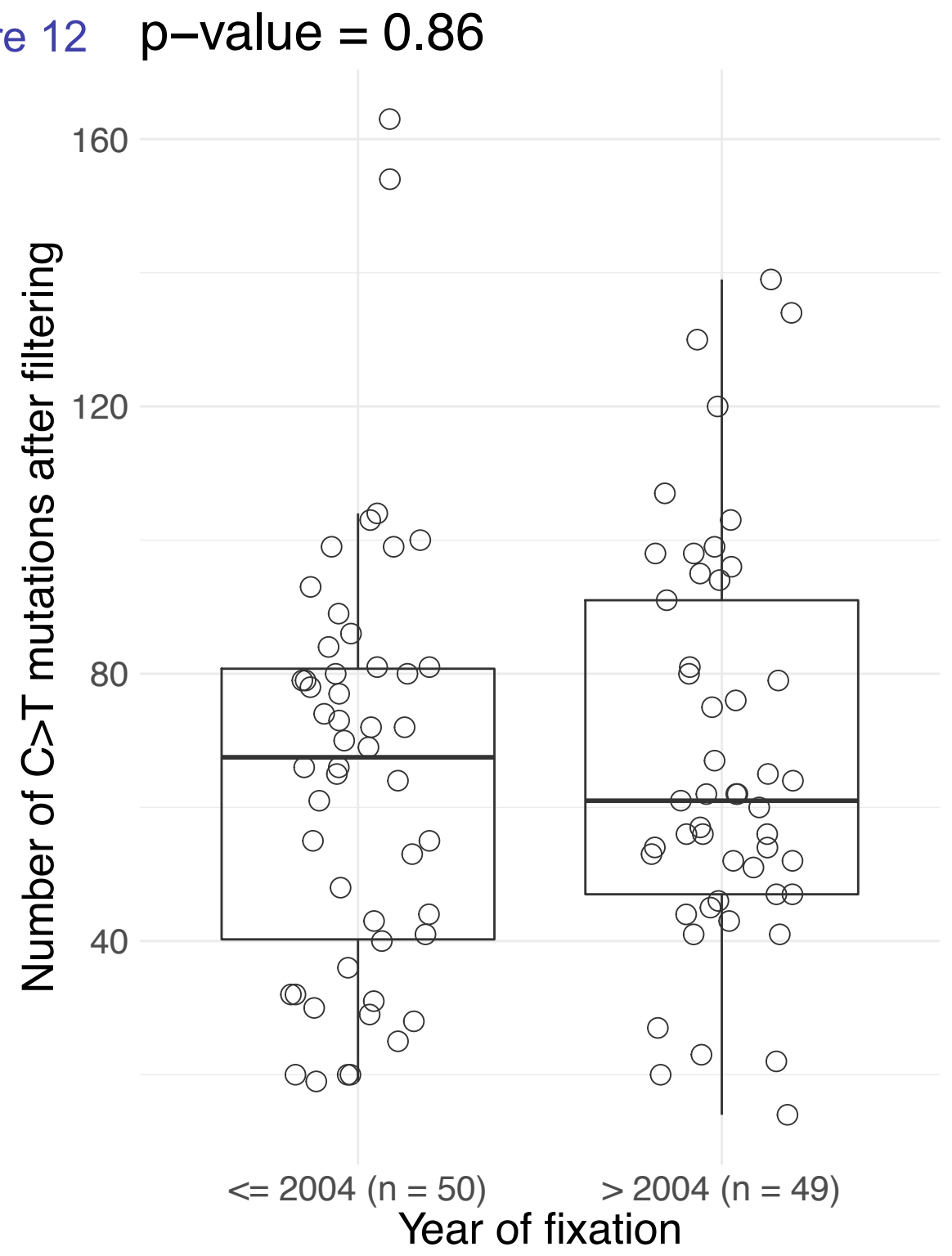
To account for potential artifacts induced by formalin-fixed paraffin embedded (FFPE) samples, we employed mutation-filtering criteria described in the "Variant calling, filtering, and copy number alteration detection" subsection in Methods. Apart from other analysis, this also insured that the age of FFPE samples could not negatively influence the signature analysis.

To show how effective our filtering was, we divided our samples into two groups, i.e. <=2004 and >2004. Then we compared the number of C>T/G>A substitutions between the two groups before and after the filtering. According to the results, the significant difference in number of C>T/G>A substitutions seen before filtering was efficiently canceled after filtering.

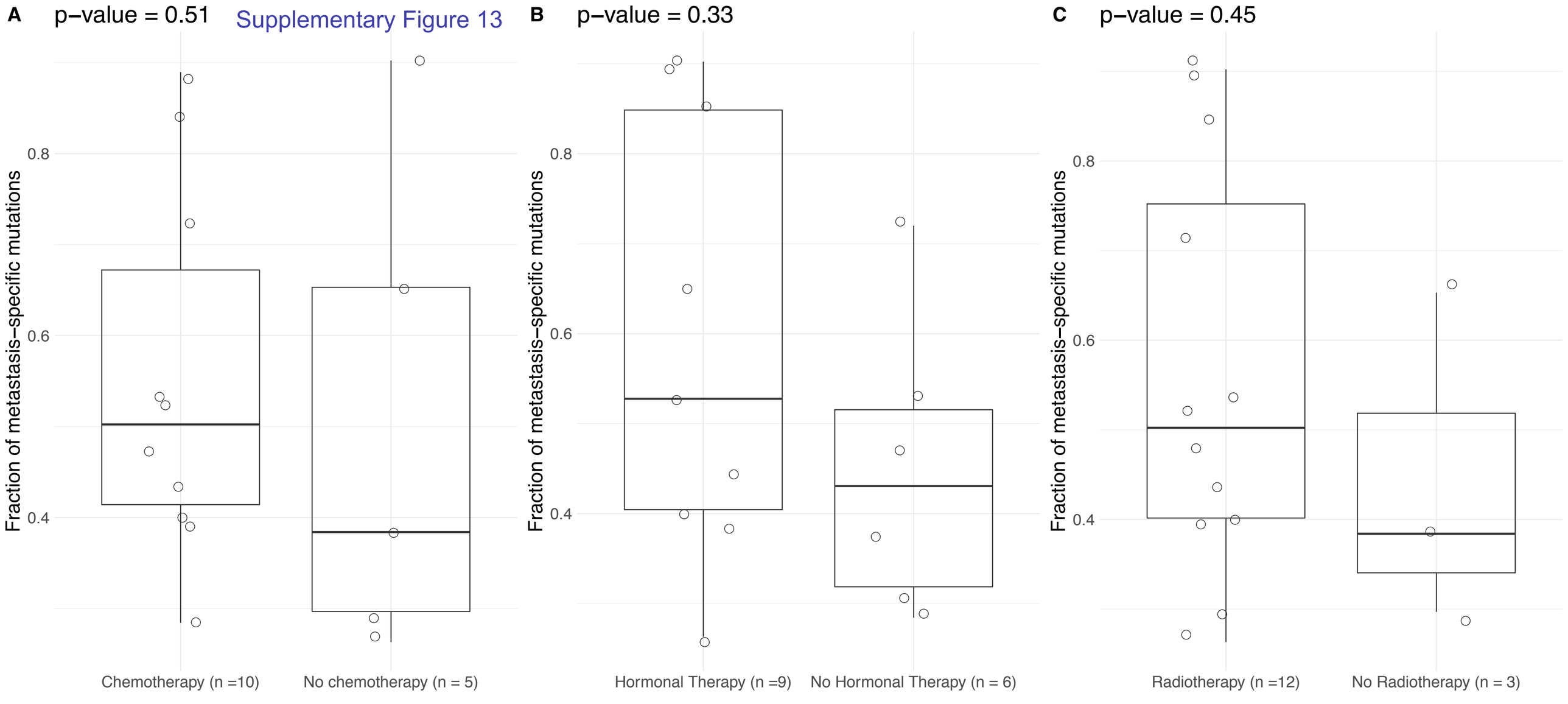
The figure shows Number of C>T mutations before and after filtering. The p-values reported are two-sided.

Supplementary Figure 12



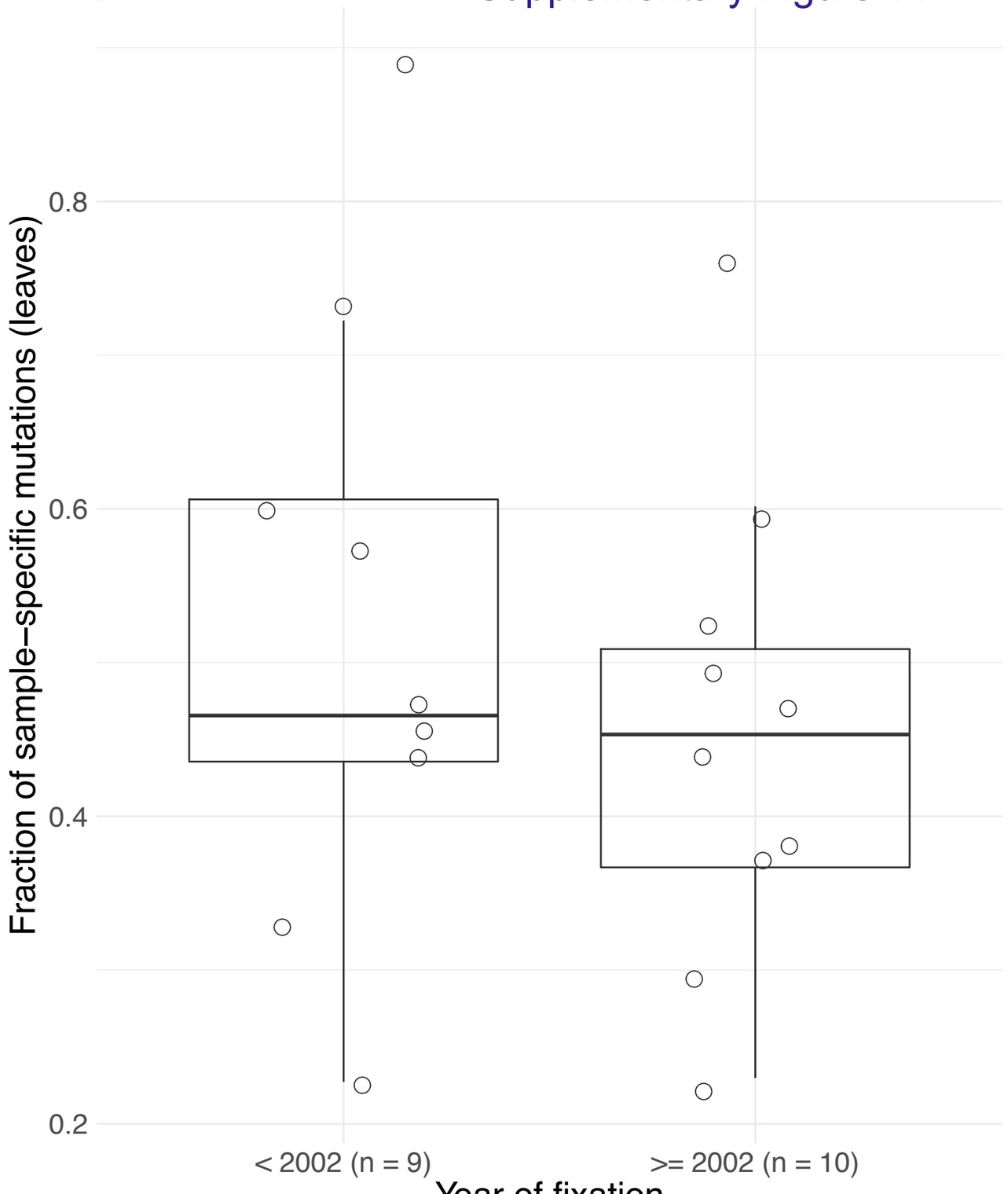


Supplementary Figure 13

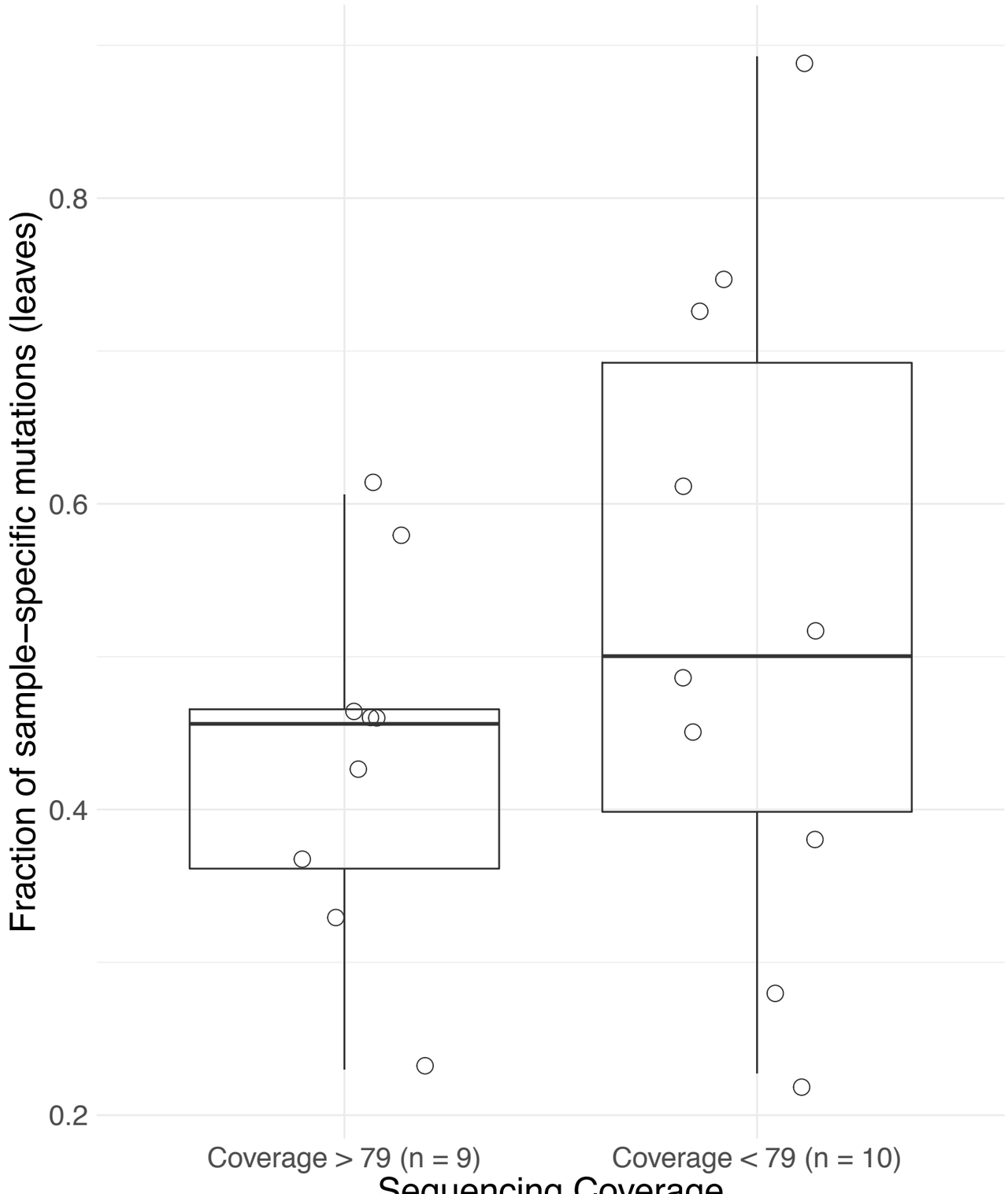


Supplementary Figure 14

A p–value = 0.55



B p–value = 0.36



SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1: Exome sequencing coverage statistics. Stacked bar showing the percentage of target regions covered at certain coverage. Each bar represents one sample and the bars are grouped by patient.

Supplementary Figure 2: Schematic representation of treatment history, number of relapses, relapse locations, sequenced relapsed sites, PAM50 intrinsic molecular subtypes for primary tumors and survival timeline of patients in our cohort. Color bands, whose length is proportional to the timescale, represent different treatment types. Each molecular subtype is represented by its own specific color. Failed and un-sequenced samples are colored grey. L+, positive axillary lymph node; Loc, Local relapse; Contr: contralateral event; BL, Basal like; LA, Luminal A; LB, Luminal B; H2, Her 2 enriched; NBL, Normal breast like CT, chemotherapy; RT, radiotherapy; HT, hormonal therapy; M1, metastasis 1; M2, metastasis 2; M3, metastasis 3; M4, metastasis 4;

Supplementary Figure 3: Analysis pipeline for investigating tumor progression models in breast cancer. Given the exome-sequencing data, Mutect was used for calling somatic mutations while AscatNGS was used for estimating tumor purity and copy number aberrations. The input to phylogenetic reconstruction, using Dollo Parsimony, consisted of a binary matrix obtained by first weighing the mutant allele frequency by tumor purity and then thresholding the resulting values by 0.05. To infer the statistical support of internal vertices, non-parametric bootstrapping was used. The phylogenetic analysis resulted in a tree with bootstrap support. The input to subclonal reconstruction (using PyClone) consisted of mutant allele frequency, copy number aberrations and tumor purity data. The subclonal analysis resulted in inferred clusters, represented here as density plot, which shows the cellular prevalence of each cluster (or subclone) in each sample. Finally the output from phylogenetic and subclonal analysis is integrated as a tree containing the subclonal information as colored (single clone) or dotted (multiple clones) lines along its edges. Edge lengths in the tree are scaled by number of substitutions while internal vertices are marked with bootstrap support values.

Supplementary Figure 4: Separating property in the tumor tree. (**A**) Inferring the role of axillary lymph node in seeding distant metastasis, based on the separating property in tumor tree. We observe that Germline-to-Primary path (color red) is separating the path from “axillary lymph node” to “Metastasis 1”, “Metastasis 2a”, and “Metastasis 2b”. Thus, we infer that Primary, rather than Lymph node, has seeded distant metastases. (**b**) Inferring linear progression based on the separating property in tumor tree. We observe that Germline-to-Primary path (colored red) is not separating the path from “Metastasis 1” to the two blocks of “Metastasis 2”, namely “Metastasis 2a” and “Metastasis 2b”. Thus we infer that “Metastasis 1”, rather than Primary, has seeded “Metastasis 2”.

Supplementary Figure 5: Pairwise mutation heatmaps for each patient in the cohort. For each patient, the fraction of shared and specific mutations is presented in the left and right column respectively. The heatmap in the left column illustrates, for row i and column j, the fraction of shared mutations between i and j divided by the total

mutations in both samples. The heatmap on the right illustrates, for row i and column j, the fraction of specific mutations present in sample i but absent in present j. Pairwise mutation heatmaps are not given for patient 6 and 12 due to low number of samples. **(A)** Patient 1 **(B)** Patient 2 **(C)** Patient 3 **(D)** Patient 4 **(E)** Patient 5 **(F)** Patient 7 **(G)** Patient 8 **(H)** Patient 9 **(I)** Patient 10 **(J)** Patient 11 **(K)** Patient 13 **(L)** Patient 14 **(M)** Patient 15 **(N)** Patient 16 **(O)** Patient 17 **(P)** Patient 18 **(Q)** Patient 19 **(R)** Patient 20.

Supplementary Figure 6: Phylogenetic trees and subclonal information for each patient in the cohort. For each patient, subclonal information is embedded in the phylogenetic, which is presented as subfigure I. In the tree, edge lengths are proportional to the number of mutations, with the actual number given in parenthesis for each edge. The list of alterations in putative driver genes are given for each edge, with mutations, amplifications and deletions shown in black, red and blue color respectively. The information about individual subclones is given in tabular format which includes cluster ID of the subclone, the color used for subclone in phylogenetic tree (and the density plot), number of mutations in the subclone and list of putative driver genes included in the subclone. The density plot is given as subfigure II, which shows the cellular prevalence of each subclone in each sample. In the density plot, the cluster IDs along with the number of mutations are given on x-axis while their cellular prevalence in samples are given on y-axis. Figures for the patients are given in numeric order and exclude patients 6 and 12 since the number of samples, in both cases, is less than 3 (minimum number for reconstructing a phylogenetic tree). **(A)** Patient 1 **(B)** Patient 2 **(C)** Patient 3 **(D)** Patient 4 **(E)** Patient 5 **(F)** Patient 7 **(G)** Patient 8 **(H)** Patient 9 **(I)** Patient 10 **(J)** Patient 11 **(K)** Patient 13 **(L)** Patient 14 **(M)** Patient 15 **(N)** Patient 16 **(O)** Patient 17 **(P)** Patient 18 **(Q)** Patient 19 **(R)** Patient 20.

Supplementary Figure 7: (A) i. Residuals sum of squares (RSS) as a function of number of signatures attempted. Dots represent mean values and bars represent standard errors. **ii.** Explained variance as a function of number of signatures attempted. Dots represent mean values and bars represent standard errors. **(B)** A heatmap of Euclidean distances between the extracted four signatures (x-axis) and the published signatures (y-axis). **(C)** Barplots showing the frequencies of six classes of substitutions in both the transcribed strand (red) and the untranscribed strand (blue) across the four extracted signatures.

Supplementary Figure 8: Heatmap showing the copy number landscape across samples for each patient. To visualize the varying landscape of copy numbers between different samples, copy number heatmap for each patient is given.

Supplementary Figure 9: Lineage analysis for all patients performed using LICHeE. In order to validate our tumor progression results, we used LICHeE to reconstruct lineage trees for all patients (except patient 6, which has only a single sample) as described in “Validation of phylogenetic trees” subsection in Methods. LICHeE uses variant allele frequencies of somatic mutations to reconstruct multi-sample cell lineage trees and infer the subclonal composition of the samples. **(A)** Patient 1 **(B)** Patient 2 **(C)** Patient 3 **(D)** Patient 4 **(E)** Patient 5 **(F)** Patient 7 **(G)**

Patient 8 **(H)** Patient 9 **(I)** Patient 10 **(J)** Patient 11 **(K)** Patient 12 **(L)** Patient 13 **(M)** Patient 14 **(N)** Patient 15 **(O)** Patient 16 **(P)** Patient 17 **(Q)** Patient 18 **(R)** Patient 19 **(S)** Patient 20.

Supplementary Figure 10: Mutations heatmap for all patients to visualize the shared and specific mutations among samples

Supplementary Figure 11: Comparison of phylogenetic trees reconstructed using two different mutation-filtering criteria. Validation of phylogenetic trees by removing mutations affected by variable coverage and/or different tumor purity among samples as described in "Validation of phylogenetic trees" subsection of the Method section. For each patient, a side-by-side comparison of the tree reported in the paper (termed here **Old tree**) vs. the one reconstructed using the more conservative mutation selection criteria (termed here **New tree**) is presented.

Supplementary Figure 12: Efficient filtering of FFPE-related C>T/G>A artifacts. Boxplots showing the number of C>T/G>A mutations (Y-axis) in two groups of samples (x-axis) defined based on sample age. The comparison was performed both before applying filtering (A) and after applying filtering (B). P-values are computed based on two-sided Mann-Whitney test.

Supplementary Figure 13: Effect of treatment on fraction of metastasis-specific mutations. Boxplots showing the proportions of metastasis-specific mutations (Y-axis) between two groups of patients divided based on treatment history. Each type of treatment was tested separately: (A) chemotherapy, (B) hormonal therapy and (C) radiotherapy. Only patients where at least one primary and one distant metastasis sample have been sequenced (15 patients) were considered for this comparison. For each patient, the fraction of metastasis mutations that are not detected in primary tumor was computed for each distant metastasis site. In patients where more than one distant metastasis site were sequenced, we chose the highest fraction. P-values were computed using two sided Mann-Whitney test.

Supplementary Figure 14: Effect of sample age and coverage on fraction of sample-specific mutations (length of leaves in phylogenetic trees). Boxplots showing the proportion of sample-specific mutations (Y-axis) for two groups of patients divided based on time of first sample acquisition (A) and based on average sample coverage (B). P-values are computed based on two-sided Mann-Whitney test.