nature portfolio

https://doi.org/10.1038/s41586-025-09374-4

Supplementary information

Fluctuating DNA methylation tracks cancer evolution at clinical scale

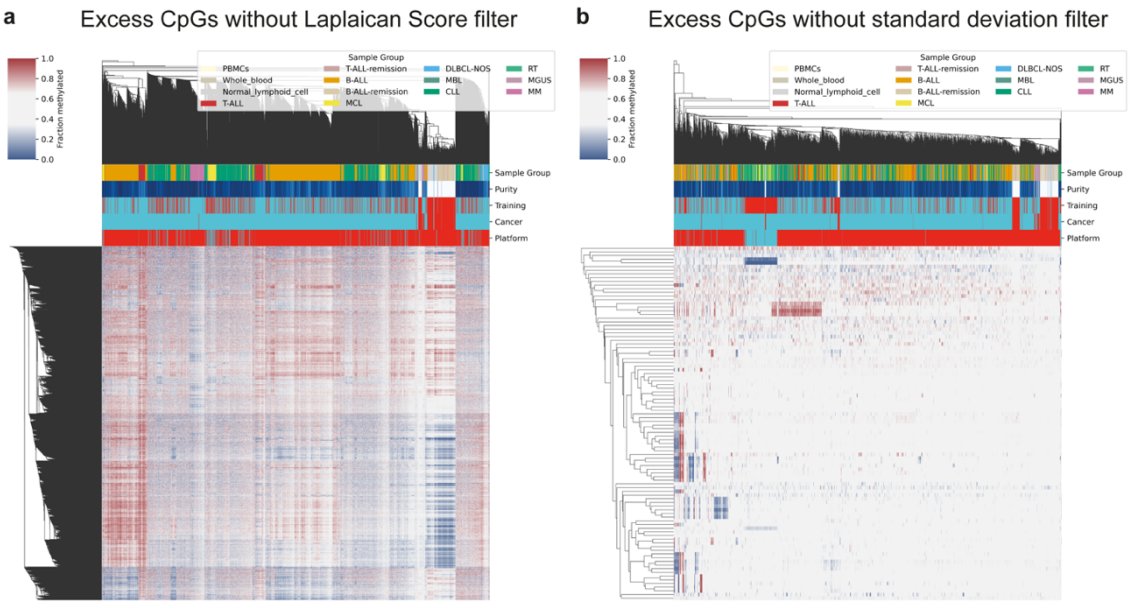
In the format provided by the authors and unedited

Supplementary Information

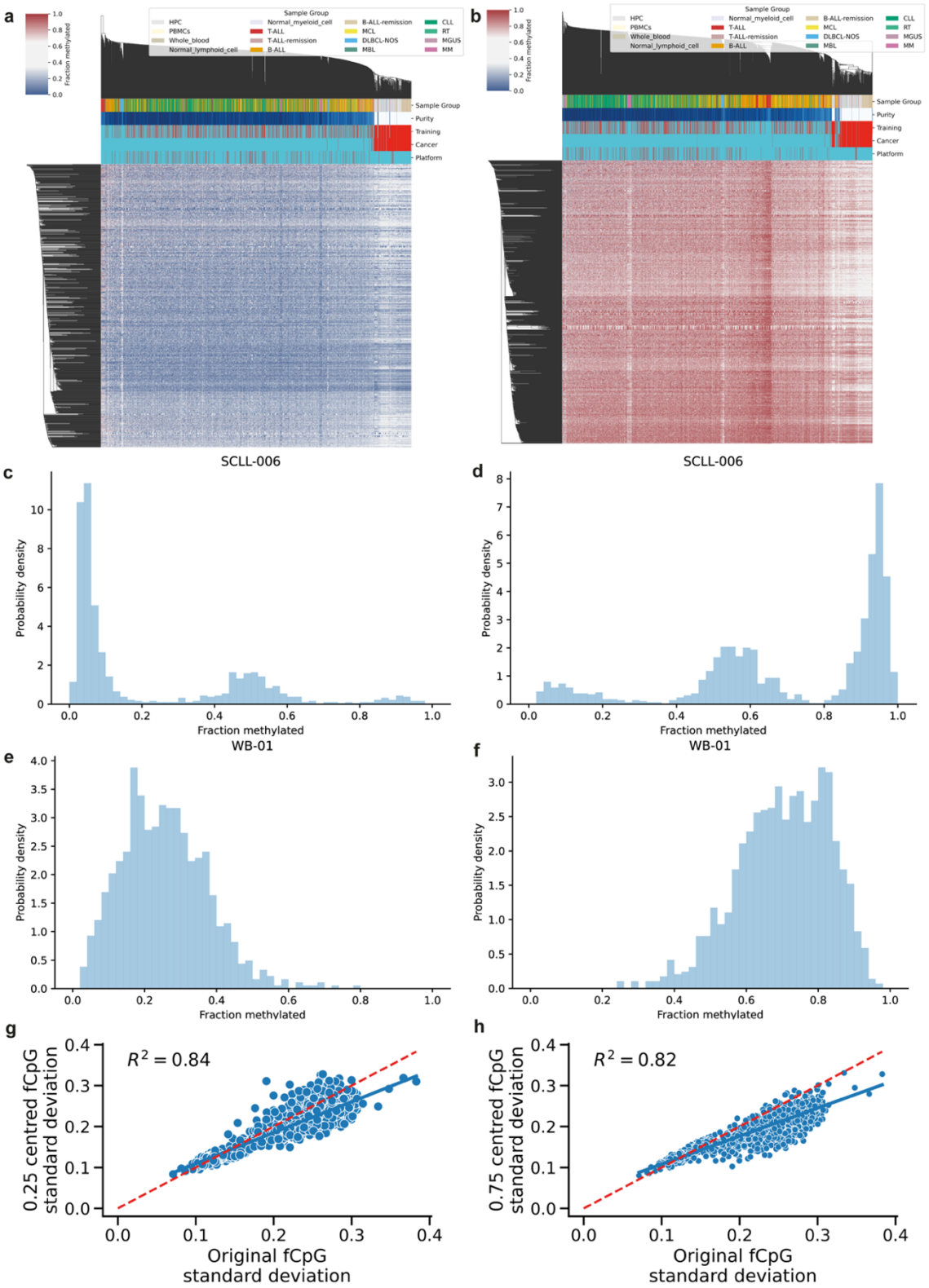
Fluctuating DNA methylation tracks cancer evolution at clinical scale

Gabbutt & Duran-Ferrer, et.al.

Supplementary Figures

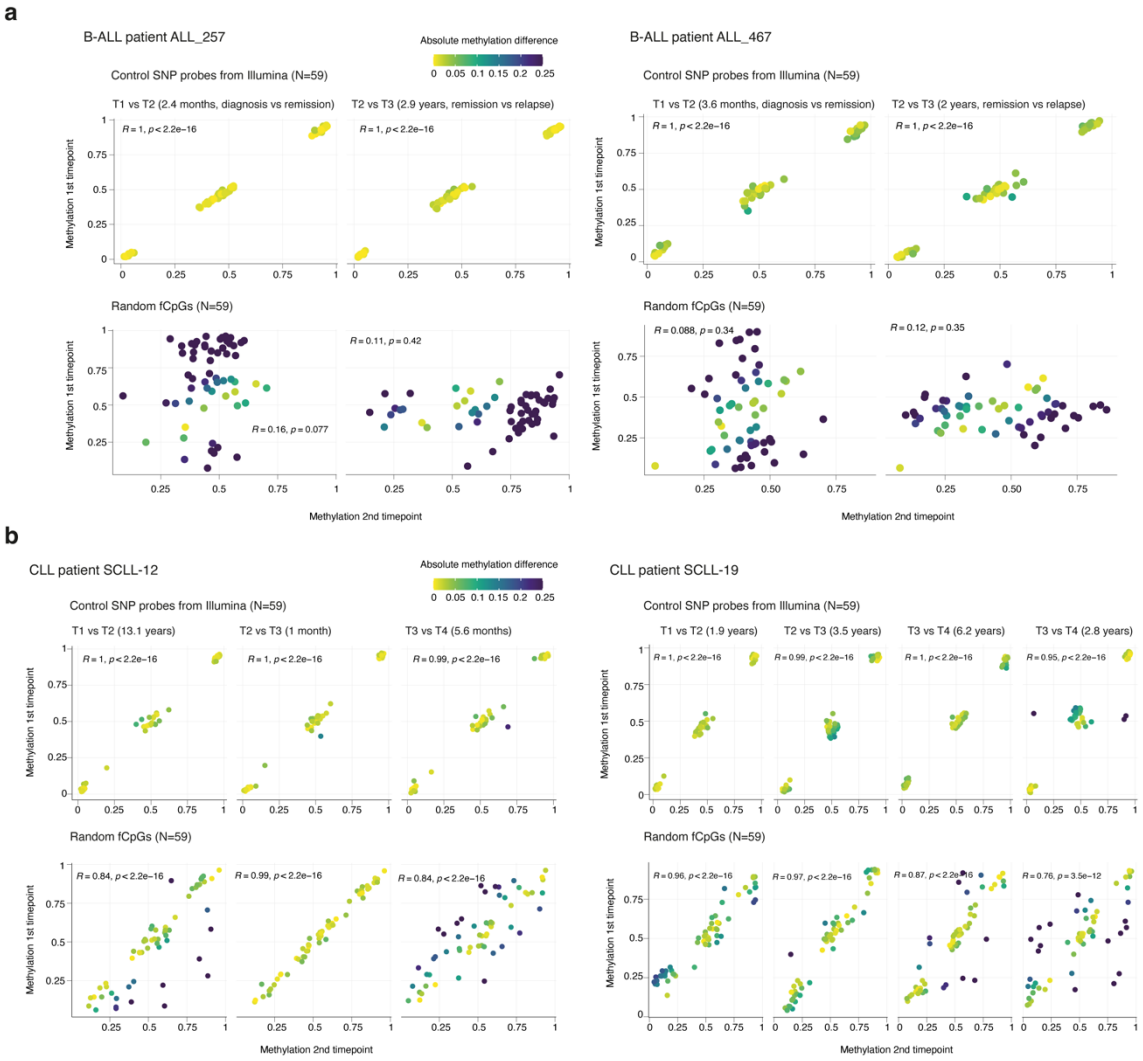


Supplementary Figure 1 – The effect of different aspects of the fCpG selection process.
a-b: Clustered heatmaps (hierarchical with average linkage and a Euclidean metric) showing CpGs that are excluded from the fCpG selection by the Laplacian Score filter (a) or the standard deviation filter (b).

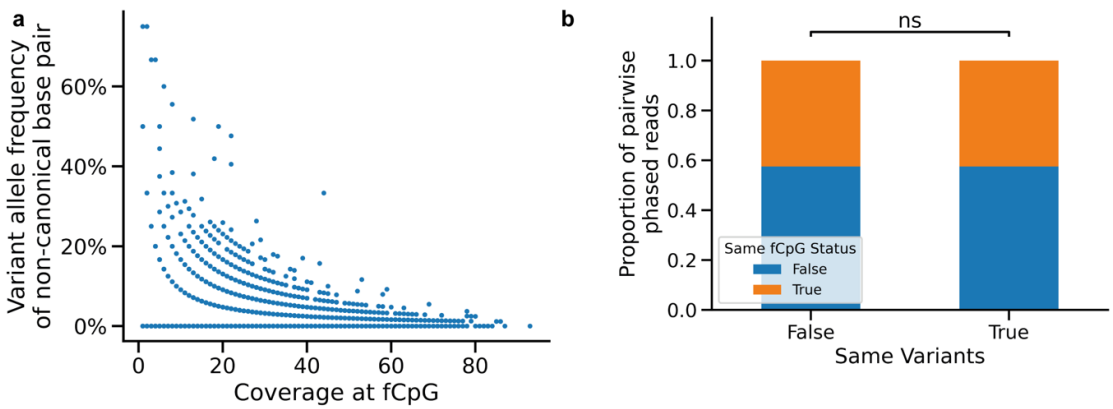


Supplementary figure 2 – Selection of biased fCpGs

a-b: Heatmaps showing the fCpG methylation distributions for fCpGs selected where the mean methylation between 0.15 and 0.35, or 0.65 and 0.85 respectively. **c-d:** Example histograms showing the distribution of biased hypo-/hyper-methylated fCpGs in a CLL sample (SCLL-006). **e-f:** Example histograms showing the distribution of biased hypo-/hyper-methylated fCpGs in a normal whole blood sample (WB-01). **g-h:** Scatterplots showing the relationship between the fCpG standard deviation in between the original unbiased fCpGs and the biased hypo-/hyper-methylated fCpGs respectively.

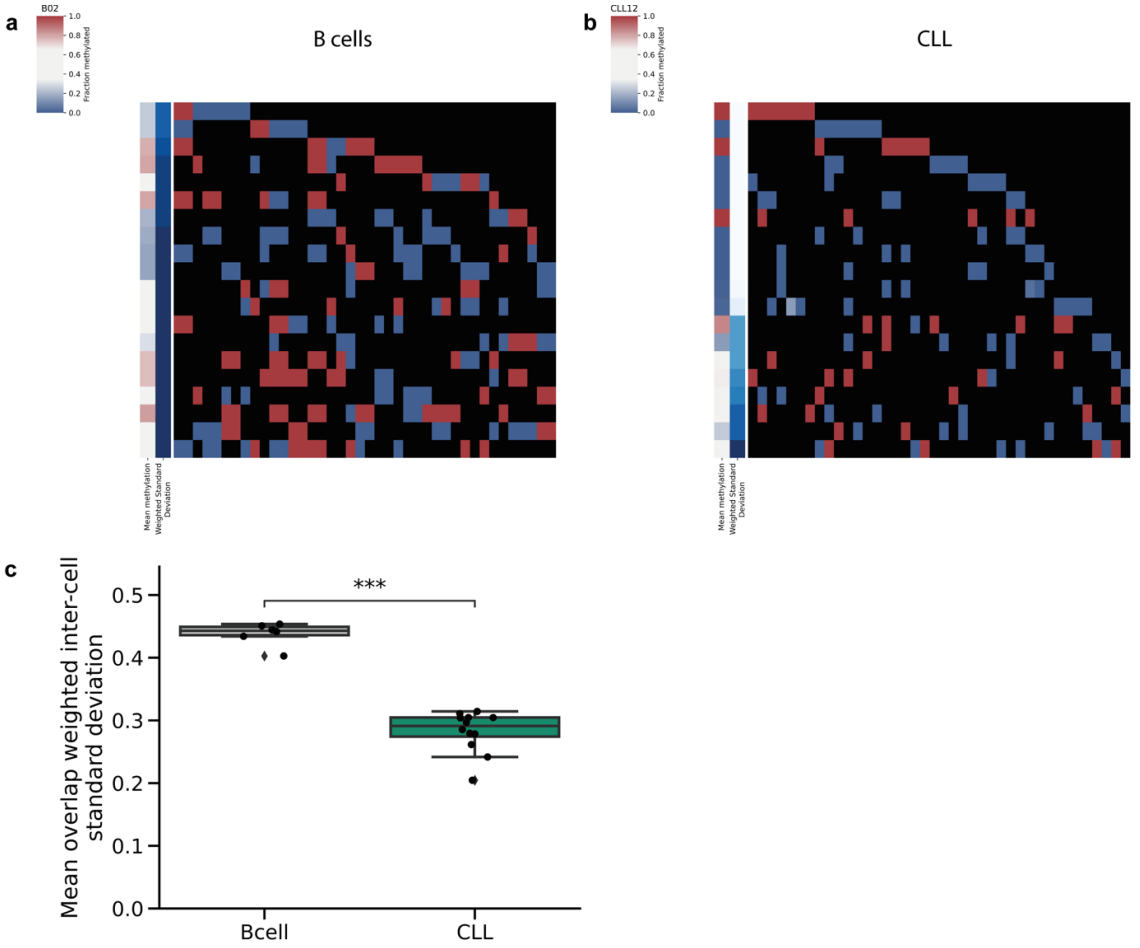


Supplementary figure 3 – fCpGs and SNPs show distinct methylation dynamics in longitudinal samples
a: Two B-ALL patients with longitudinal samples including diagnosis, remission and relapse. Pairwise comparisons between diagnosis and remission and remission and relapse are shown for 59 control SNP probes (top panels) and 59 randomly selected fCpGs (bottom panels). **b:** Same as in **a** but for two CLL patients in diagnosis, progression, relapse and Richter transformation (RT) in the last sample timepoints. Using matched WGS, we show that there are changes in the tumor subclonal composition in T1 vs T2 and T3 vs T4 in SCLL-12, whereas in SCLL-019 the biggest changes occur between T3 vs T4 and T4 vs T5 (fig 4). Fig. 4 also show the methylation values of the full fCpGs set, and supplementary table 15 the full clinical annotation of these 2 patients.



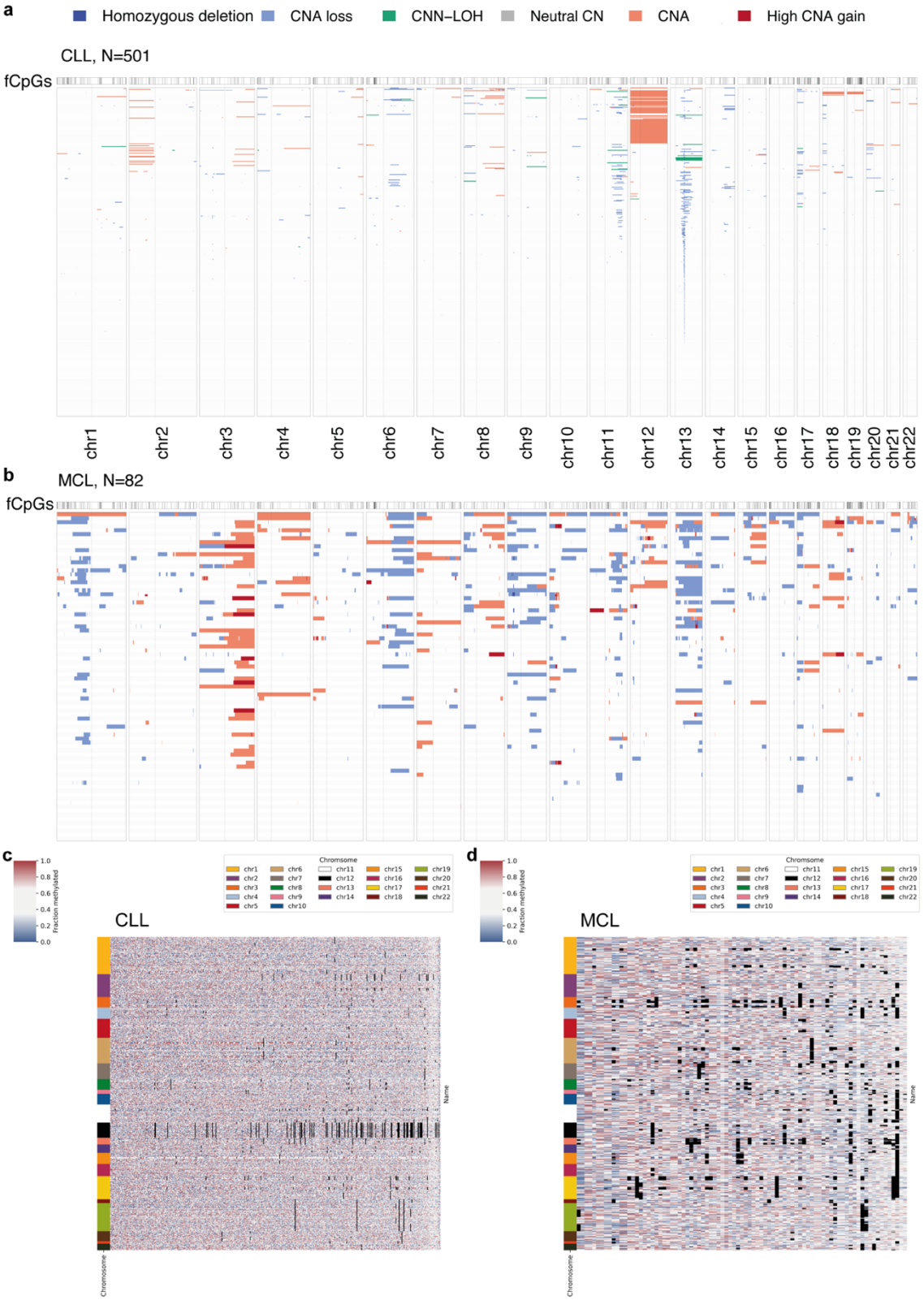
Supplementary figure 4 – Nanopore methylation calls show fCpG methylation is not due to genetic confounding.

A: A scatterplot showing that the variant allele frequency (VAF) of non-canonical fCpG calls is inversely proportional to the coverage at that locus. **B:** The proportion of pairwise phased read comparisons in which the same fCpG site is in the same methylation state in normal B cells, split by whether the pattern of variants in a 50bp window surrounding that fCpG site is the same.

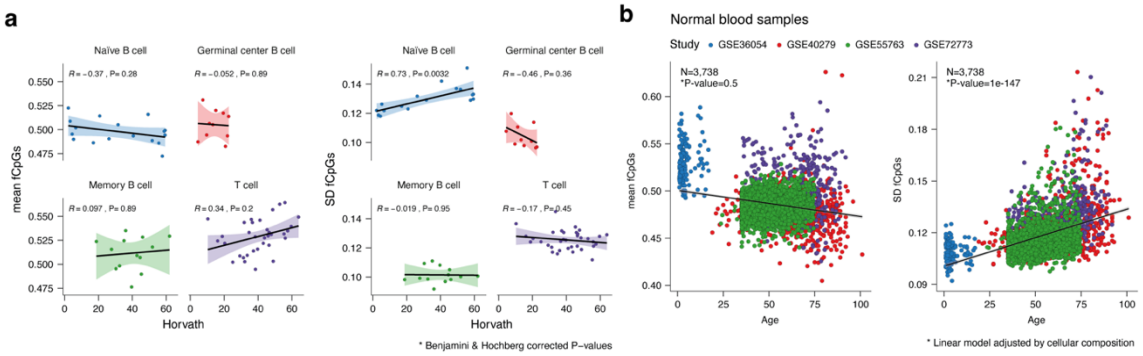


Supplementary figure 5 – Single cell reduced representation bisulfite sequencing (scRRBS) data validates fCpGs as an evolving barcode

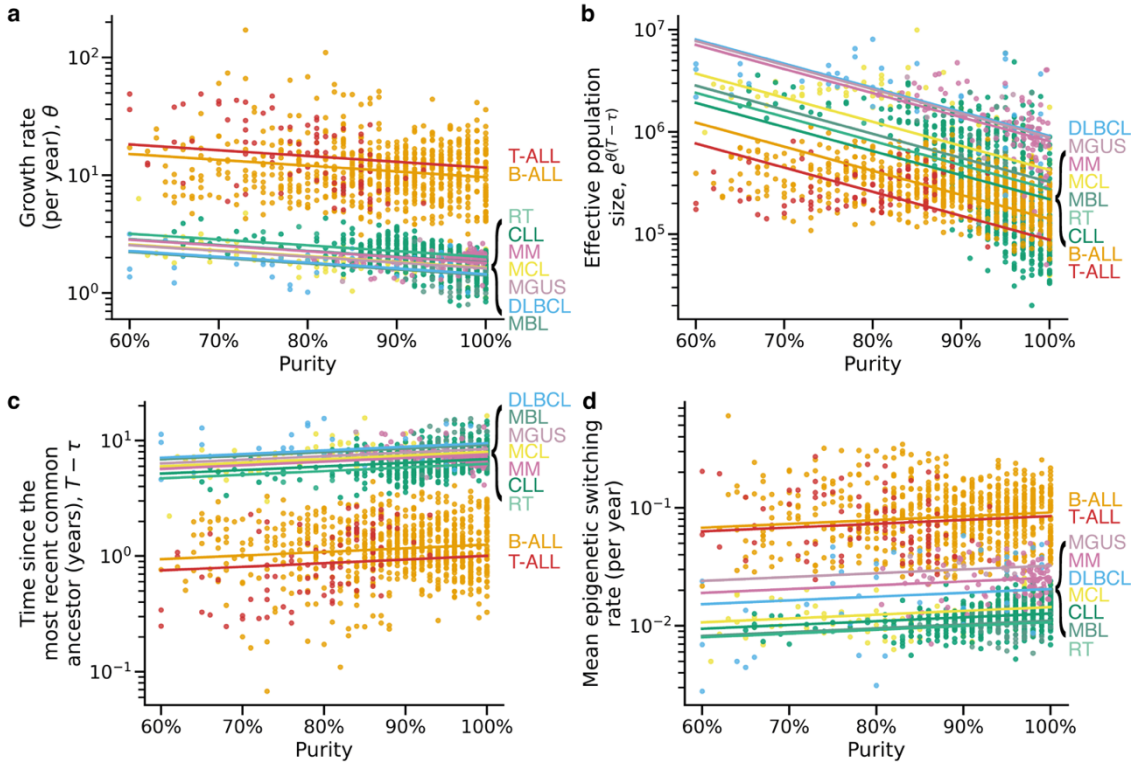
a-b: Heatmaps displaying the methylation status of 20 fCpGs (rows) from 40 single cells (columns) from a normal B cell sample (**a**) and a CLL sample (**b**). Red and blue represent hyper/-hypomethylation respectively, whereas black represents missing data. Rows are annotated with the mean methylation across the 40 cells (left) and the standard deviation at that fCpG weighted by the number of fCpGs with data available (colour scale - white to blue). **c:** Boxplot comparison of the weighted standard deviation across the population of cells in normal B cell and CLL samples.



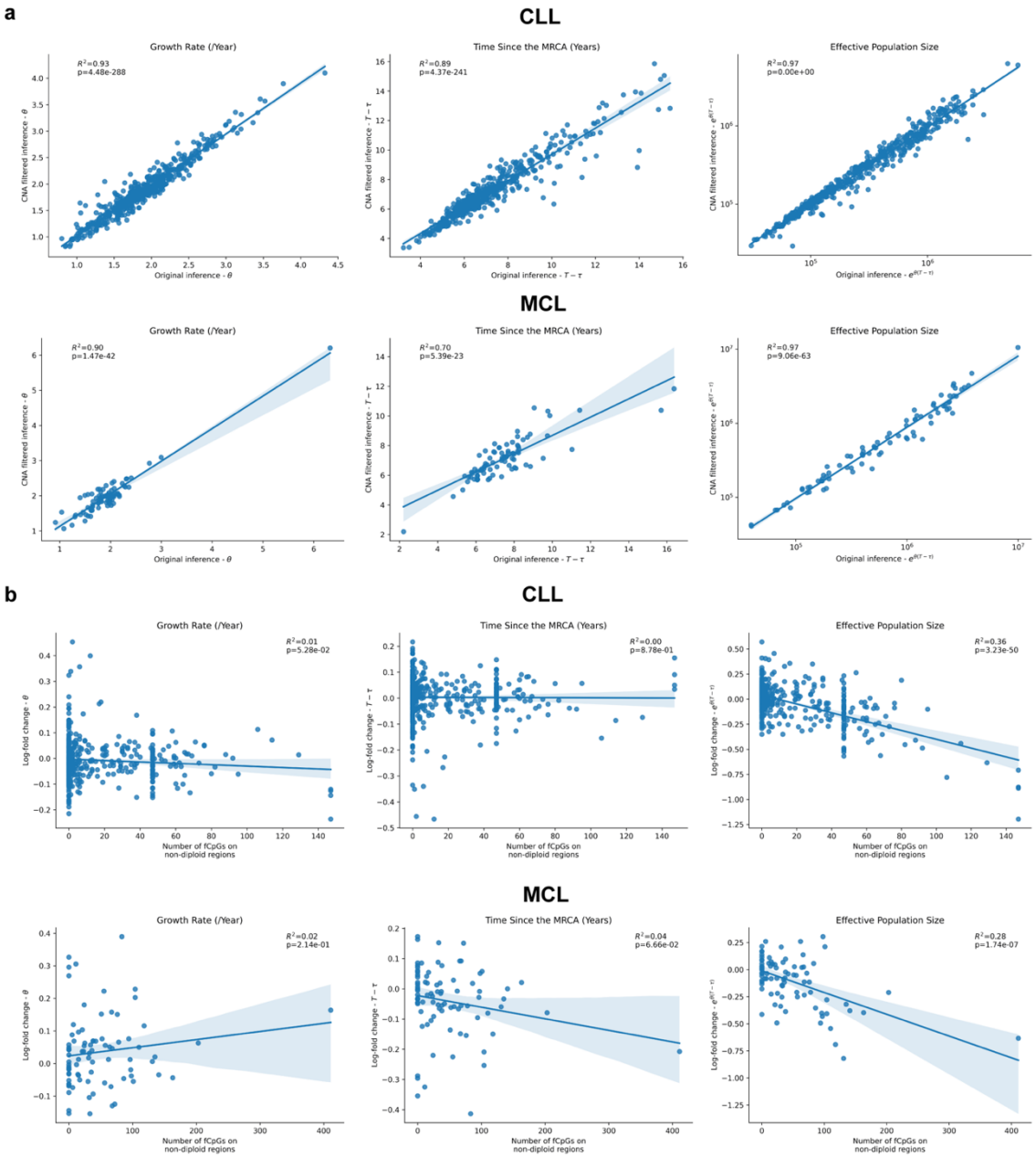
Supplementary Figure 6 – Copy number alterations in CLL and MCL and fCpG location.
a-b: Copy number profiles in the CLL (**a**) and MCL (**b**) cohorts. Rows represent individual patient samples; columns represent genomic locations. Diploid regions are shown in white, gains in red and losses in blue
c-d: Heatmaps of fCpGs organised by genomic location in the CLL (**c**) and MCL (**d**) cohorts, with fCpGs present on non-diploid regions coloured in black.



Supplementary Figure 7 – Dynamics of fCpGs in normal lymphoid cells during aging.
a: Mean and standard deviation of fCpGs as a function of age estimated by the Horvath clock in normal lymphoid cells included in the study. **b:** Mean and standard deviation of whole-blood samples in 4 independent studies covering the entire human lifespan. P-values are adjusted by cellular fractions calculated with *EpiDISH* package using 12 cell types.

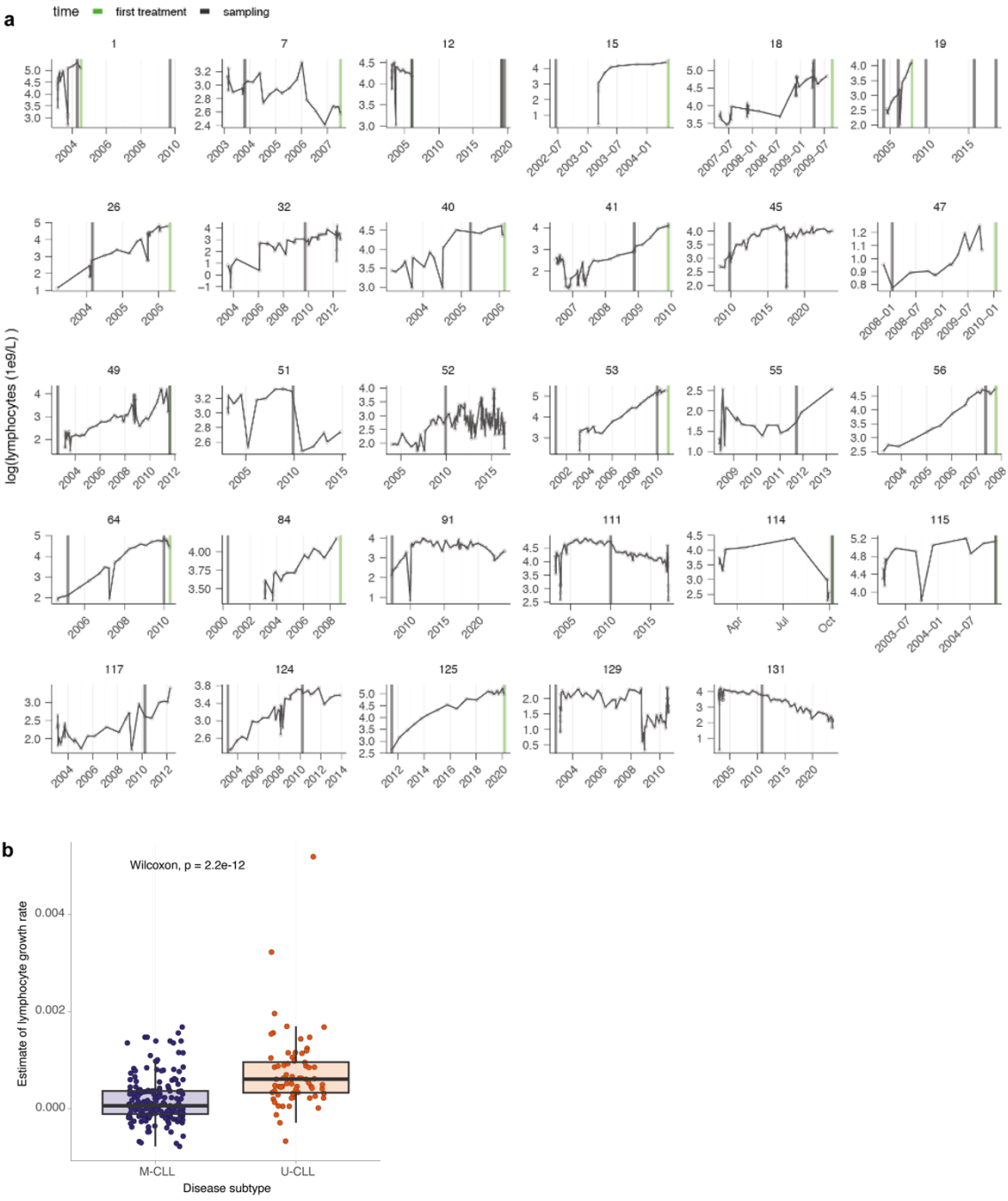


Supplementary Figure 8 – Comparisons between disease entities are robust to varying tumour purities
a-d: Multivariate linear regression between purity and the different disease entities vs the distribution of inferred growth rate (**a**), effective population size (**b**), time since the most recent common ancestor (**c**), and mean epigenetic switching rate (**d**), separated by cancer types. Points are coloured and labelled according to disease entity, and the predictions from the multivariate linear regression are overlaid.

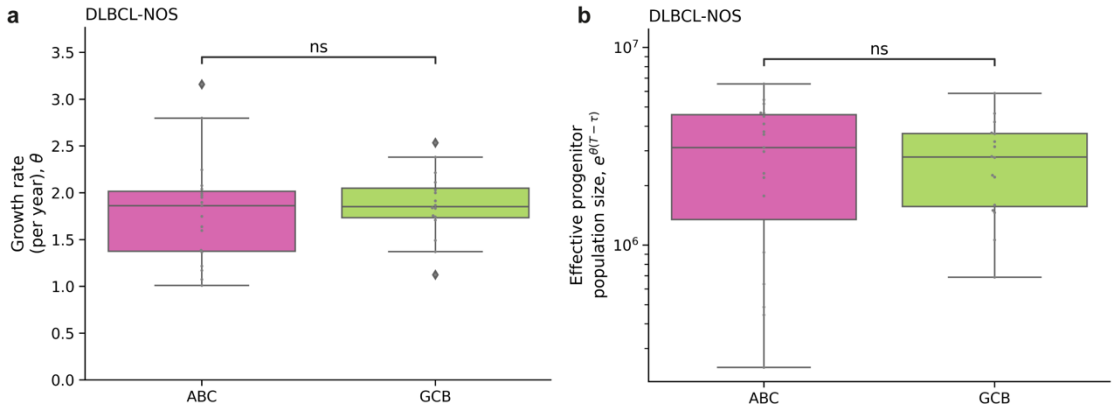


Supplementary Figure 9 – EVOFLUx in robust to removal of fCpGs present on non-diploid genomic regions.

a: Regression plots between the parameters inferred by running EVOFLUx on all 978 fCpGs (x-axis) vs just those fCpGs present on diploid regions (y-axis) in the CLL (top) and MCL (bottom) cohorts. **b:** Regression plots between the number of fCpGs located on non-diploid regions (x-axis) vs the log-fold change in the parameter inferred from the inference run on fCpGs on diploid regions compared to the original inference (y-axis) in the CLL (top) and MCL (bottom) cohorts.

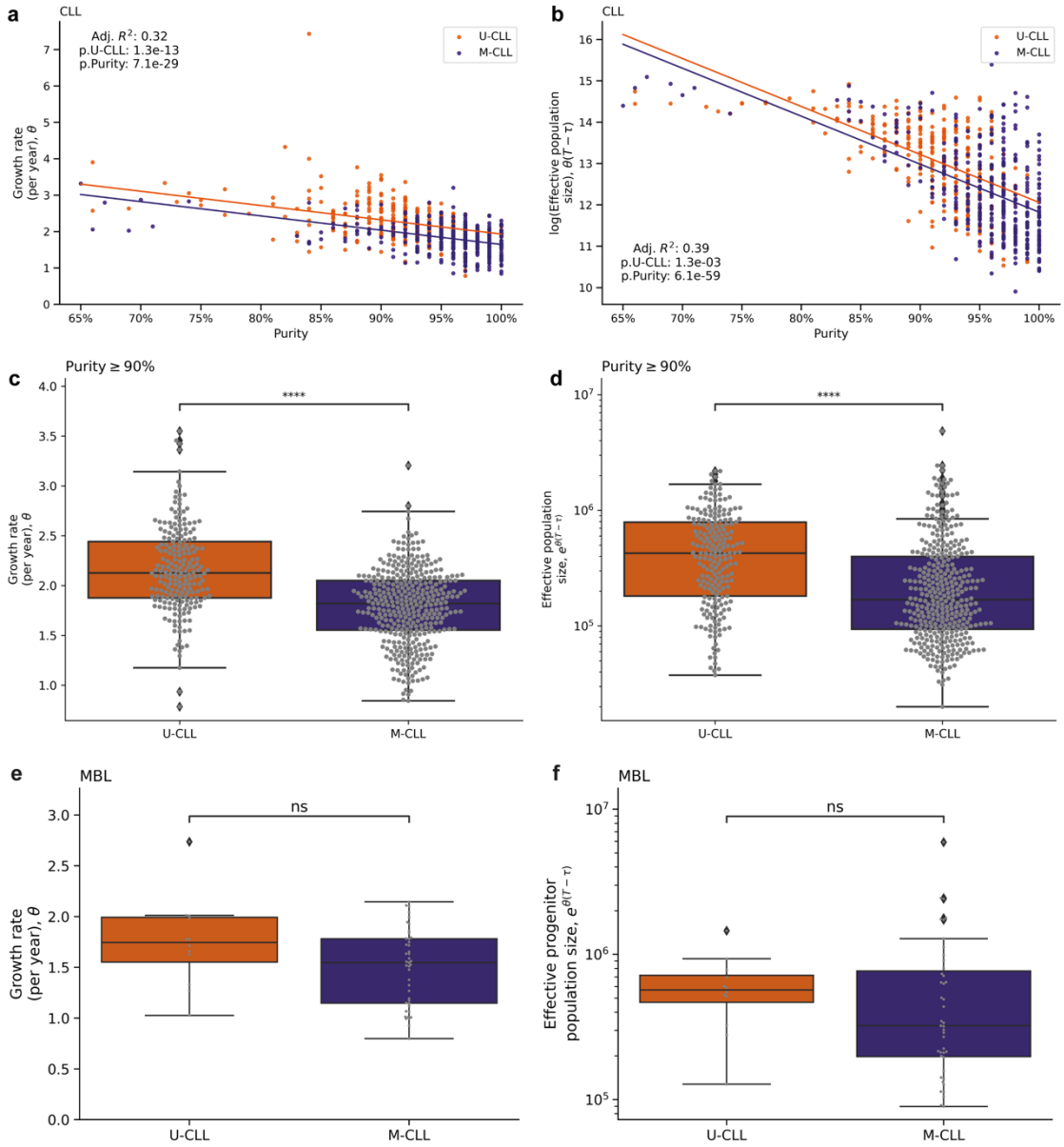


Supplementary figure 10 – Contemporary growth rates estimated via longitudinal lymphocyte counts
a: Example time courses showing the longitudinal lymphocyte counts (a proxy for the number of CLL cells in the blood) prior to treatment (green vertical line). **b:** Comparison of the rate of change in lymphocyte count between U-CLL and M-CLL (p=2.2e-12).



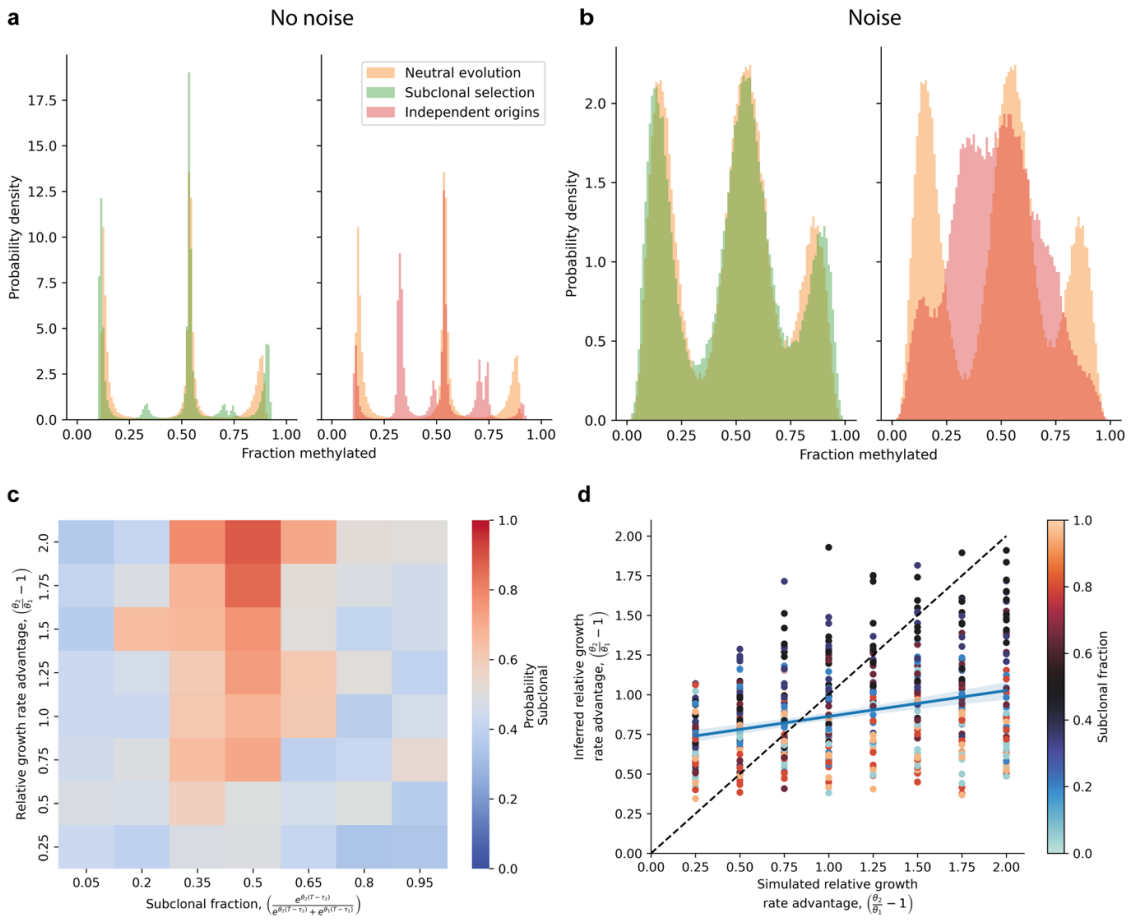
Supplementary Figure 11 – Additional subtype comparisons of the inferred evolutionary parameters in DLBCL.

The inferred growth rate (**a** and **b**) and effective population individual cancer samples separated by molecular subtype in diffuse large B-cell lymphoma (DLBCL-NOS). Differences between subtypes were tested using Mann-Whitney U tests.

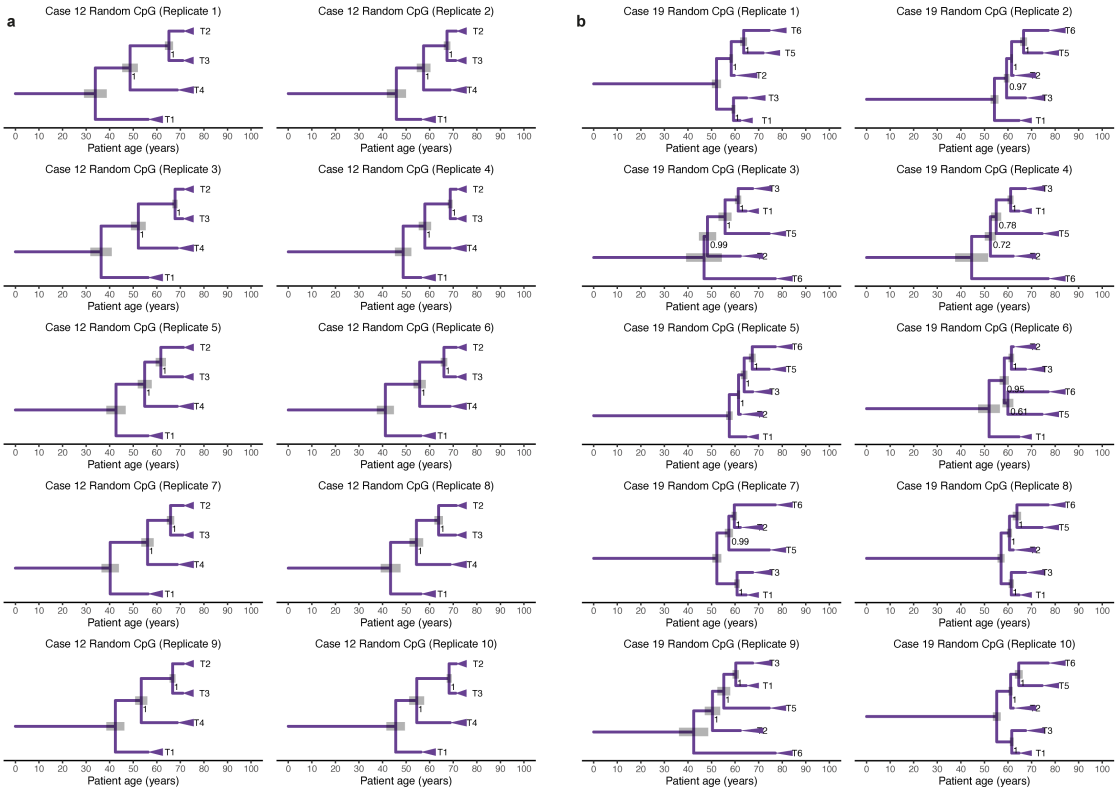


Supplementary Figure 12 – The effect of purity on the inferred EVOFLUx parameters in CLL and additional comparison of evolutionary variables in MBL.

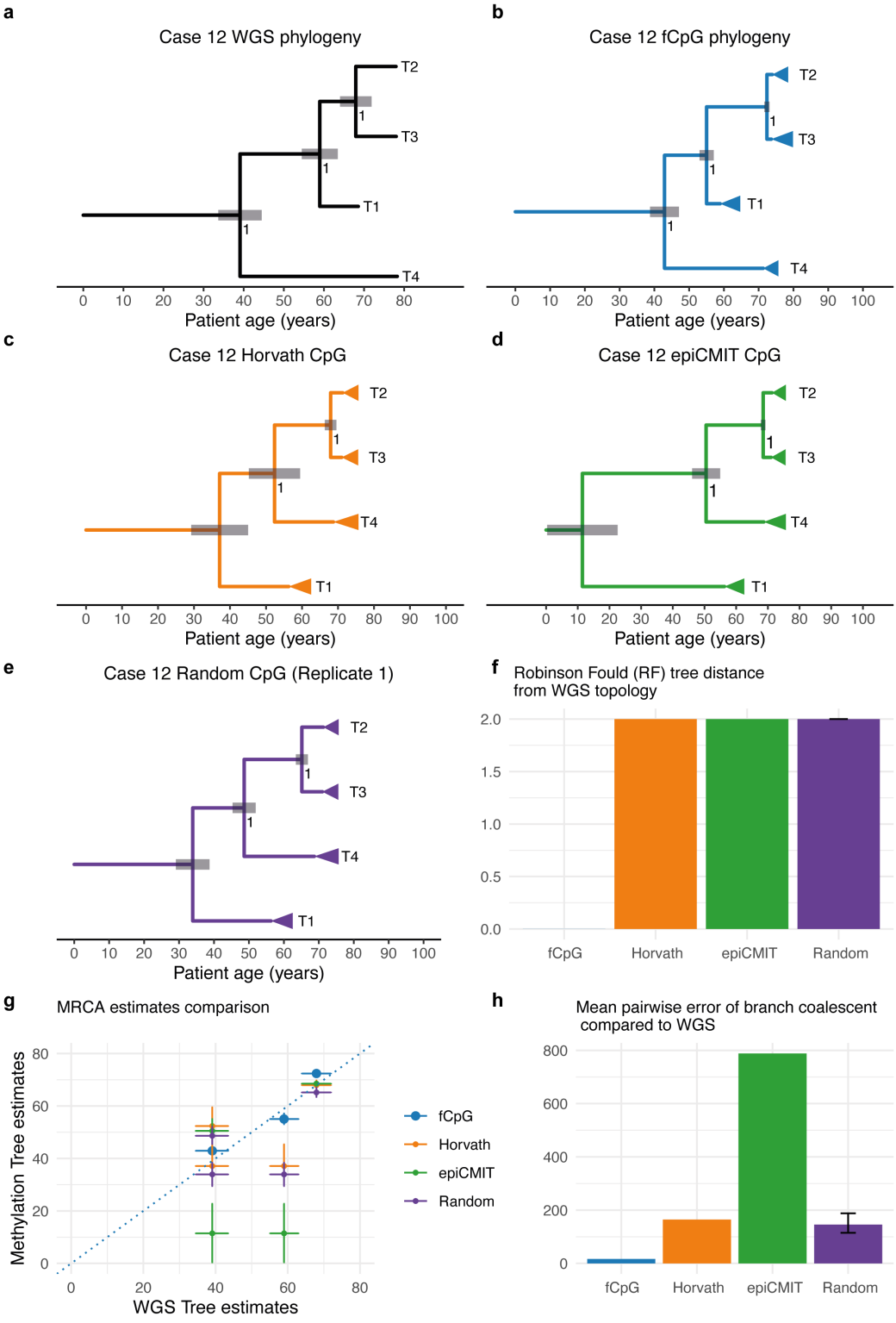
a-b: Linear regression of purity against growth rate (**a**) and effective population size (**b**), separated by IGHV mutational status. **c-d:** A comparison of the inferred growth rate (**a**) and effective population size (**b**) between IGHV subtypes in the subset of CLL samples with purity greater than 90%. **e-f:** Growth rate and effective population size in MBL. Differences between subtypes were tested using Mann-Whitney U tests. The low statistical power in MBL is driven by the low number of U-CLL samples, due to the enrichment of the M-CLL subtype in MBL.



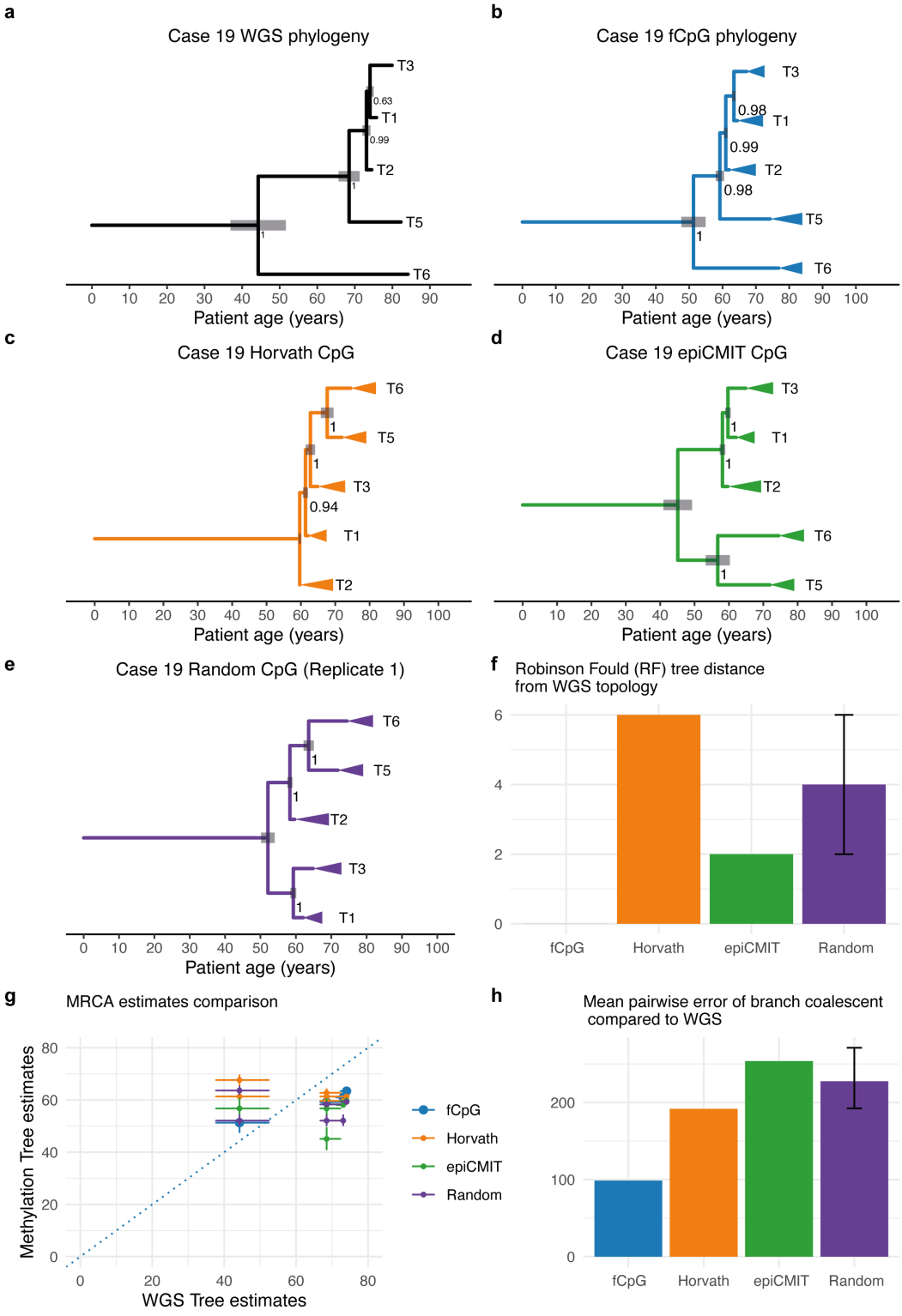
Supplementary Figure 13 – Simulations of non-neutral evolutionary dynamics
a-b: Simulations comparing the fCpG methylation distribution in the neutral (orange) case vs subclonal selection (left, green) and independent clonal origins (right, red) in the case with no noise (**a**) and with noise added (**b**). **c:** A heatmap of showing how the subclonal model weight of a simulation varies with the fraction of the population made up of the fitter subclone and the growth advantage of the fitter subclone. **d:** A comparison of the inferred relative growth advantage of a simulation vs the actual ground truth value, where the colour of the point reflects the fraction of the subclonal population.



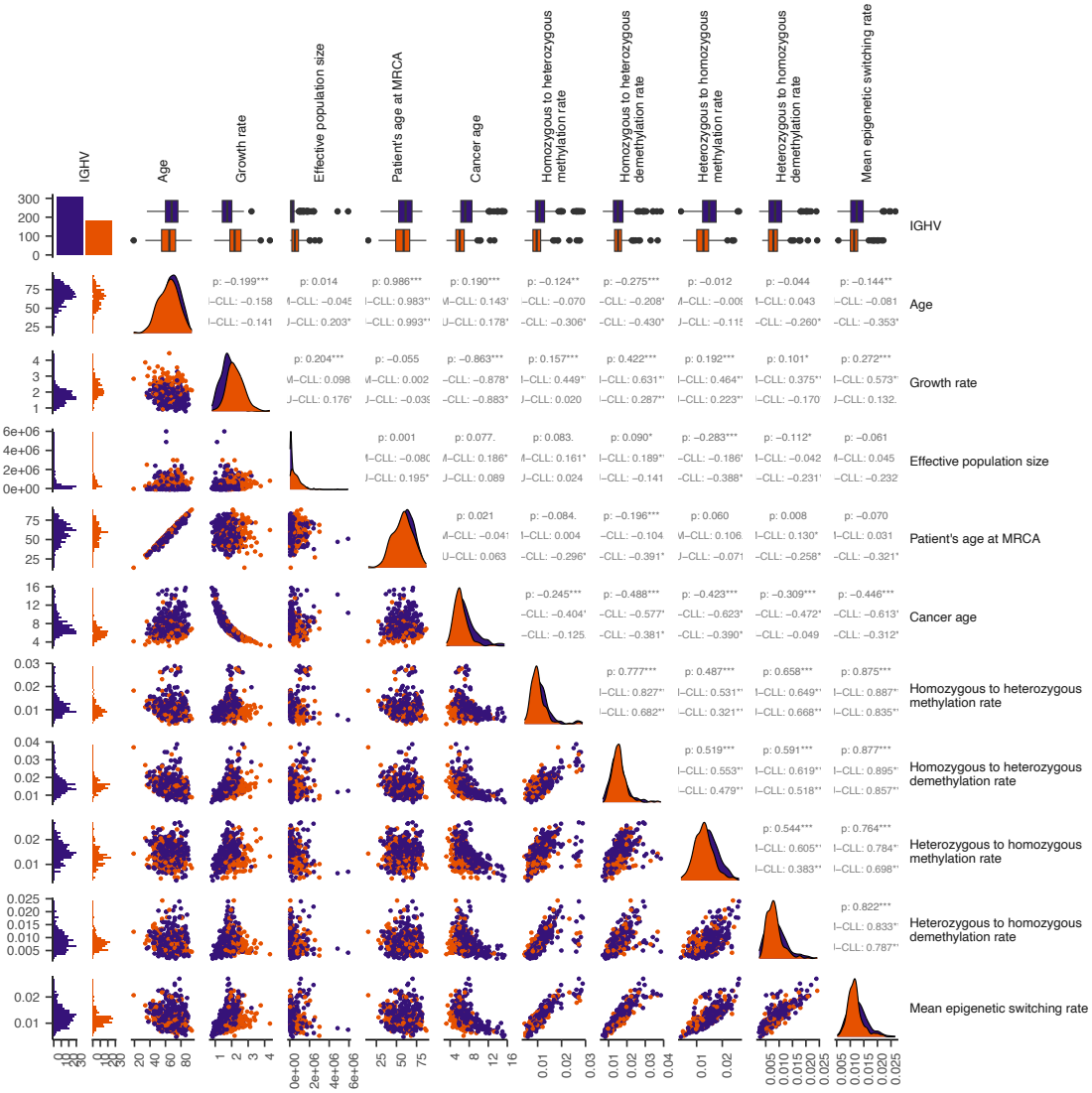
Supplementary Figure 14 – Phylogenetic trees based on random CpGs
Phylogenetic trees reconstructed from random subsets of 1,000 CpGs (10 replicates) for case 12 (**a**, fig 4a) and case 19 (**b**, fig 4b).



Supplementary Figure 15 – WGS SNV phylogenies validate fCpG phylogenies (Case 12)
a-e: Phylogenies reconstructed on matched WGS SNVs (**a**), fCpGs (**b**), Horvath's CpGs (**c**) , epiCMIT (**d**) and a random subset of 1,000 CpGs (**e**). **f:** Comparison of the Robinson Fould (RF) distance between the WGS SNV tree and the methylation-based trees. **g:** A scatterplot showing the inferred tip coalescent times (posterior median and 95% CI) for the WGS SNV tree vs each of the methylation-based trees. **h:** Comparison of the mean error between the tip coalescent times between the WGS SNV tree and the methylation-based trees.

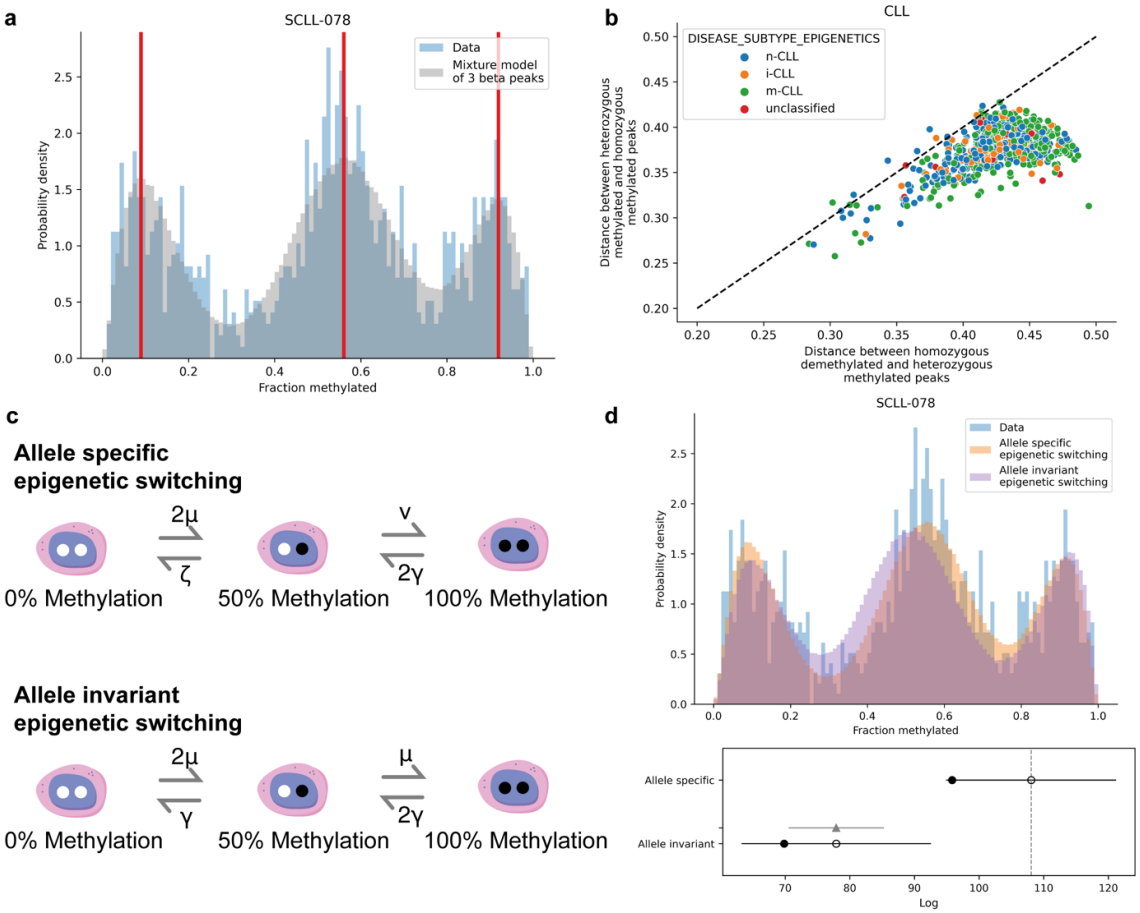


Supplementary Figure 16 – WGS SNV phylogenies validate fCpG phylogenies (Case 19)
a-e: Phylogenies reconstructed on matched WGS SNVs (**a**), fCpGs (**b**), Horvath's CpGs (**c**) , epiCMIT (**d**) and a random subset of 1,000 CpGs (**e**). **f:** Comparison of the Robinson Fould (RF) distance between the WGS SNV tree and the methylation-based trees. **g:** A scatterplot showing the inferred tip coalescent times (posterior median and 95% CI) for the WGS SNV tree vs each of the methylation-based trees. **h:** Comparison of the mean error between the tip coalescent times between the WGS SNV tree and the methylation-based trees.



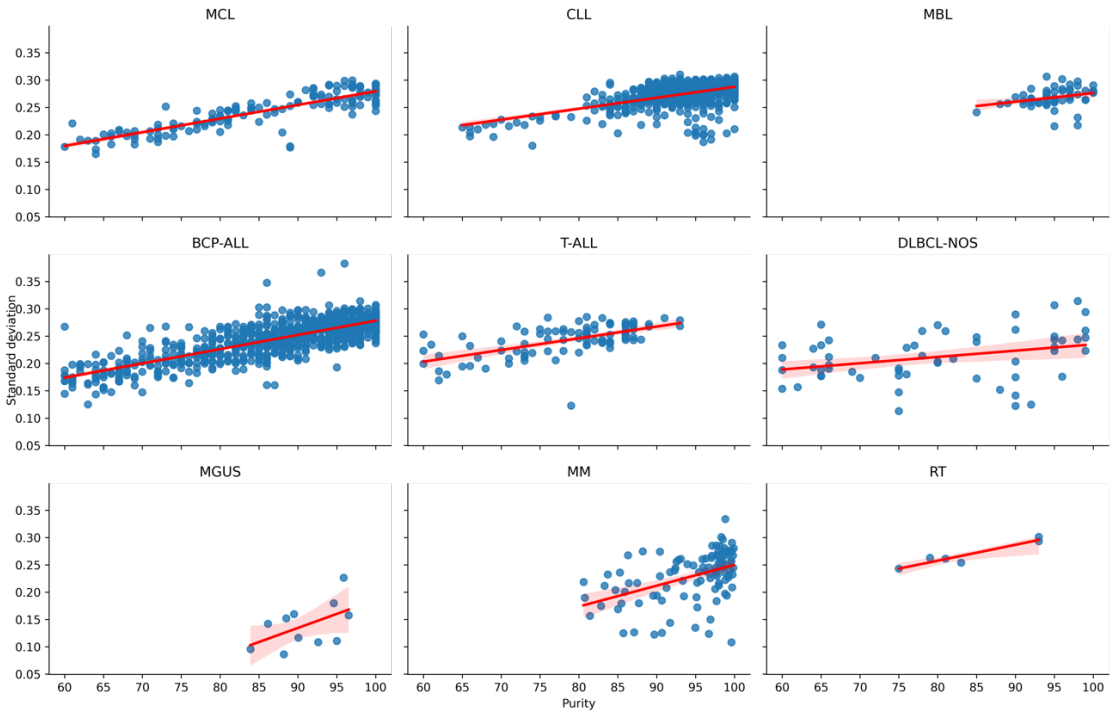
Supplementary Figure 17 – Correlations in CLL evolutionary parameters

A pairs plot showing the marginal relationships between the observed evolutionary parameters in the discovery cohort as scatter plots (lower left triangle), histograms (leading diagonal) and summary correlation coefficients (stars indicate significance). Patients separated by IGHV mutational status (unmutated orange, mutated purple).

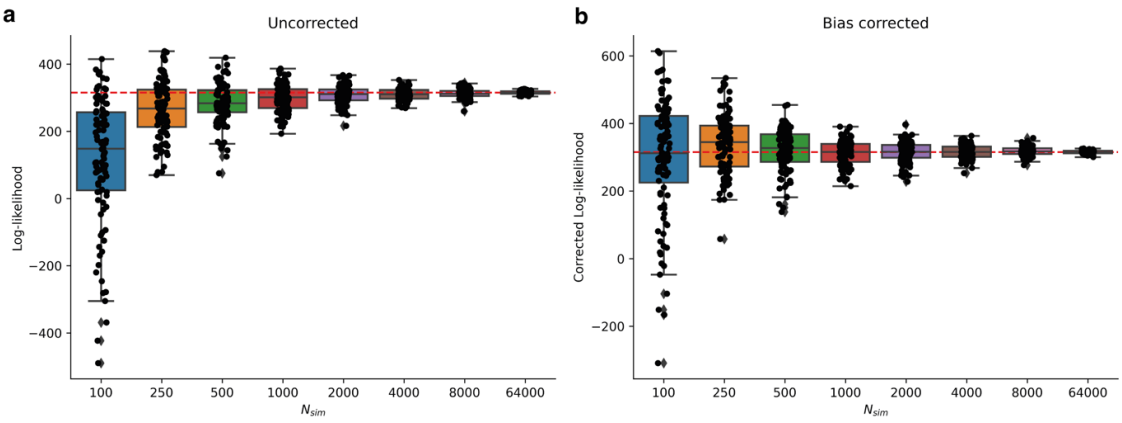


Supplementary Figure 18 – Allele-specific epigenetic switching

a: A histogram displaying the fCpG methylation value distribution of a CLL sample (blue) with a posterior predictive of a beta mixture model (with 3 components) overlaid (gray). The vertical red lines indicate the mode of the 3 inferred beta mixture components. **b:** A scatterplot comparing the distance between the inferred modes of the two homozygous peaks and the heterozygous peak in the CLL cohort. The distance between the homozygous unmethylated and the heterozygous peak is larger in the vast majority of samples. **c:** A schematic illustrating two competing models, one in which the methylation of a given allele depends on the methylation status of the other (top, 4 parameters) and a model in which the two alleles are independent (bottom, 2 parameters). **d:** (top) A histogram displaying the fCpG methylation value distribution of a CLL sample (blue) with the model fits of the allele specific (orange) and allele invariant epigenetic (lilac) switching models overlaid. (bottom) A leave one out cross validation model comparison, showing the allele specific model is strongly preferred. **E:** Scatterplots showing the association between tumour purity and the standard deviation of fCpG methylation in lymphoid cancers.



Supplementary Figure 19 – Purity affects the distribution of fCpG methylation
Scatterplots showing the association between tumour purity and the standard deviation of fCpG methylation in lymphoid cancers.



Supplementary Figure 20 – Loglikelihood correction
a: Repeated calculations (n=100) of the uncorrected synthetic loglikelihood (LL) value of simulated data calculated at the parameter values used to simulate the data, as a function of the number of simulations (n_{sim}) used in the LL calculation. Using a finite n_{sim} systematically underestimates the true LL. **b:** The bias corrected synthetic LL, the mean of which does not vary as a function of n_{sim}.

Supplementary Methods

Impact of fCpG selection criteria

To evaluate the impact of the different fCpG selection criteria we used two approaches. First, we repeated the fCpG selection process without including the LS or standard deviation filters and studied those CpGs which were thus excluded by that specific filter. We found 5,336 CpGs that passed the other filters but were excluded by the LS filter (supplementary fig. 1a). These CpGs show clear disease specific clustering and stable hypo-/hypermethylation in normal cells. The 96 CpGs excluded by the standard deviation filter had intermediate methylation in every sample, cancer or normal (supplementary fig. 1b). These could either represent CpGs that are stably maintained at 0.5 methylation (such as imprinted loci), or those that undergo epimutations exceptionally quickly.

Our fCpG selection process selects for CpGs with similar methylation and demethylation rates. We searched for CpGs with a bias in (de)methylation rates by repeating the fCpG derivation using mean methylation filters of $0.15 \leq \bar{x} \leq 0.35$ and $0.65 \leq \bar{x} \leq 0.85$ to produce hypo- and hyper-fCpGs (supplementary fig. 2a-b). Within individual cancer samples, the hypo- and hyper-fCpGs still gave rise to trimodal distributions, but with the occupancy of the peaks biased towards 0% and 100% methylated_respectively (supplementary fig. 2c-d). Normal blood samples had unimodal distributions centred around 0.25/0.75 respectively (supplementary fig. 2e-f). The standard deviation of the hypo- and hyper-fCpG methylation distribution were highly correlated with the standard deviation of the original fCpGs (R2=0.84 and R2=0.82 respectively, supplementary fig. 2g-h). Hence, fCpGs are a subset of a larger set of CpGs prone to stochastic changes in methylation state.

Association between fCpG distribution, age and disease state

We did not find any association between age and mean fCpG methylation in any of our normal blood samples or in independent cohorts of normal blood cells (N=4,099, p>0.05, extended data fig 3b, supplementary fig 7). When we used the standard deviation of each sample's fCpG methylation distribution as a proxy for clonal expansion (higher standard deviation, more "W-like" the distribution), we observed a significant correlation with age in several normal cell subpopulations, consistent with aging-related clonal expansions in the hematopoietic system (p<0.05, extended data fig 3b, supplementary fig 7). As expected, highly enriched clonal populations of lymphoid cancers had a much higher standard deviation than both peripheral blood cells and sorted B and T lymphocytes regardless the age of the donor (extended data fig 3c).

EVOFLUx – a novel method to simulate and infer the evolutionary history of well-mixed cancers from methylation-based lineage markers

In our previous work, an analytic hidden Markov model was derived to describe the distribution of fluctuating methylation clocks in a fixed effective population of N_e cells, replacing each other at a rate λ per stem cell and with each fCpG loci allowed to alter its methylation state at a methylation/demethylation rate μ/γ respectively. In principle, the mathematics developed there applies for all values possible cell numbers; however, the number of possible states (i.e., the different possible combinations of heterozygous and

homozygous (de)methylated cells at a specific CpG locus) grows as $\sim {N_e}^2$, hence this approach is inappropriate for large populations of cells. Furthermore, the previous approach assumed a fixed number of cells, rather than a growing population.

Hence, to extend our analysis to large, growing cell populations, such as those found within a blood cancer, an alternative approach to modelling the probability distribution that does not scale with the number of stem cells must be taken. Instead, we took a stochastic approach to approximate the fCpG methylation distribution in an arbitrarily large, exponentially growing population. Rather than considering the individual states and the possible flows between them, we constructed equivalent stochastic jump differential equations. Finally, whereas previously it was the stochastic replacement of one stem cell by another that counteracted the ongoing (de)methylation, effectively resynchronising the methylation clocks upon a random clonal expansion; in the case of a growing cancer this neutral drift is negligible when compared to the exponential growth of each lineage within the tumour.

In summary, then, our model consists of a single cell that undergoes random (de)methylation until time τ, at which point it begins exponentially growing at rate θ per year until the patient's age at the time of sampling, T, is reached. We allow for the possibility of allele specific methylation, and therefore require four epigenetic switching parameters to describe the system (fig 2A):

- μ – Homozygous to heterozygous methylation rate (per year)
- υ – Heterozygous to homozygous methylation rate (per year)
- γ - Homozygous to heterozygous demethylation rate (per year)
- ζ – Heterozygous to homozygous demethylation rate (per year)

Importantly, the time of cancer emergence should be thought of as the time of occurrence of the most recent common ancestor (MRCA) of the sampled extant cancer population, rather than the cancer age *per se*, as there may have been previous clonal expansions prior to the "final" clonal sweep that produced a clinically-detectable cancer, and the MRCA of the sample is inevitably the same or more recent than the MRCA of the whole cancer population.

We employed four separate methylation switching parameters as these were necessary to reproduce the observed data. In CLL patient data, the distance between the homozygous unmethylated peak and the heterozygous methylated peak was almost always lower than the distance between the homozygous methylated peak and the heterozygous methylated peak (supplementary fig 18a-b). A model where the homozygous to heterozygous rate is the same as the heterozygous to homozygous rate (i.e. $\upsilon = \mu$ and $\zeta = \gamma$, supplementary fig 18c) provided a significantly worse fit to the data (accounting for the higher degrees of freedom in the 4-parameter model) than one in which the parameters are free to vary (supplementary fig 18d).

We assume that each methylation site is equally likely to be homozygous methylated or unmethylated at $t = 0$, as in our previous work. We shall use the notation that there are m cells with both alleles methylated, k cells with one allele methylated and w cells with

neither allele methylated. To calculate the methylation probability distribution of the single transformed cell at $t = \tau$, the system is evolved according to the following set of differential equations:

$$\frac{d}{dt}\begin{pmatrix} p(m;t) \\ p(k;t) \\ p(w;t) \end{pmatrix} = \begin{pmatrix} -2\gamma & \upsilon & 0 \\ 2\gamma & -(\zeta + \upsilon) & 2\mu \\ 0 & \zeta & -2\mu \end{pmatrix}\begin{pmatrix} p(m;t) \\ p(k;t) \\ p(w;t) \end{pmatrix}$$

To calculate the initial state of the transformed cell at a particular locus at $t = \tau$, the state (m, k, w) is drawn from a categorical distribution with probabilities $\big(p(m; t = \tau), p(k; t = \tau), p(w; t = \tau)\big)$, which may be calculated using matrix exponentiation. The subsequent evolution of the system for $t > \tau$ is then modelled using discrete time stochastic jump equations with a time step, δt, that is sufficiently small that the probability of double jumps is negligible:

$$\delta t = \frac{0.1}{\max(\theta, 2\mu, 2\gamma, 2\upsilon, 2\zeta)}$$

The exponential growth and the ongoing (de)methylation of the cancer cell population are treated in two steps, first the population of the cancer grows according to the deterministic exponential growth equation (rounded to the closest integer):

$$N_e(t) = e^{\theta(t-\tau)}$$

With the new cells assigned to each population according to the multinomial distribution:

$$\big(m_g, k_g, w_g\big) \sim \text{Multinomial}\left(N_e(t + \delta t) - N_e(t), \left(\frac{m(t)}{N_e(t)}, \frac{k(t)}{N_e(t)}, \frac{w(t)}{N_e(t)}\right)\right)$$

$$\big(m'(t), k'(t), w'(t)\big) = \big(m(t) + m_g, k(t) + k_g, w(t) + w_g\big)$$

The flow from the homozygous methylated to the heterozygous methylated population due to ongoing demethylation can then be calculated as:

$$\delta_{m\to k;t} \sim \text{Binomial}(m'(t), 2\gamma\delta t)$$

Similarly, the flow from the homozygous unmethylated to the heterozygous methylated population is:

$$\delta_{w\to k;t} \sim \text{Binomial}(w'(t), 2\mu\delta t)$$

To calculate the flow from the heterozygous population to the homozygous unmethylated and methylated populations, we first calculate the total number of cells that change state and then work out whether they flow to the homozygous methylated or unmethylated population:

$$\delta k_t \sim \text{Binomial}(k'(t), (\upsilon + \zeta)\delta t)$$

$$\delta_{k\to m;t} \sim \text{Binomial}\left(\delta k_t, \frac{\upsilon}{\upsilon + \zeta}\right)$$

$$\delta_{k\to w;t} = \delta k_t - \delta_{k\to m;t}$$

Updating the system at time $t + \delta t$ is then:

$$m(t+\delta t) = m'(t) + \delta_{k\to m;t} - \delta_{m\to k;t}$$
$$k(t+\delta t) = k'(t) + \delta_{k\to m;t} + \delta_{w\to k;t} - \delta k_t$$
$$w(t+\delta t) = w'(t) + \delta_{k\to w;t} - \delta_{w\to k;t}$$

Note that the reason for this formulation, rather than simply calculating the number of cells that have switched from one state to another via each process (e.g. following a Poisson process) separately, is to strictly ensure that the total cell number is conserved and each population is strictly positive. The methylation fraction at time t can be simply calculated as:

$$\beta_c(t) = \frac{k(t) + 2m(t)}{2N_e(t)}$$

The above assumes a perfectly pure tumour sample, however, there will inevitably be a degree of contamination from normal cells. To account for this contamination, we assumed that the contaminating normal cells did not share a recent common ancestor, so the measured methylation of a tumour sample with purity ρ would be the weighted average of the cancer methylation at that locus, $\beta_c(t)$, and the average methylation of the contamination normal cells, Λ:

$$\beta(t) = \rho\beta_c(t) + (1-\rho)\Lambda$$

This assumption was informed by the observation that the standard deviation of the different cancer sample's fCpG methylation distributions correlated linearly with the tumour purity (supplementary fig 19).

Using the recursive relationship above, a single stochastic path can be generated. The methylation fraction probability distribution can then be numerically estimated by generating a series of n_{sim} independent samples from the above process, the distribution of which will approximate the underlying hidden Markov model distribution.

The Likelihood Function for a Growing Population

The stochastic model above allowed samples to be drawn from the distribution of fCpG methylation values given a set of parameters $\{\theta, \tau, \mu, \gamma, \upsilon, \zeta, \rho, \beta_n\}$. We can use these samples to form an estimate of the likelihood function for a given set of data, $\vec{y}$. First, we can bin the n_{sim} stochastically generated $\beta(t)$ into J bins with width $\frac{1}{J}$, to estimate a probability mass function (p.m.f) $\mathrm{P}\left(\beta_j|\theta, \tau, \mu, \gamma, \upsilon, \zeta, \Lambda\right)$ for bins with centres $\beta_j = \frac{2j-1}{2B}$ $(j = 1, 2, \dots, J)$.

To account for the noise introduced by the methylation array, we follow a similar approach as in , first employing a linear transform to account for the background noise of the array: $x_j = (\varepsilon - \Delta)\beta_j + \Delta$. Then, we can model the measured methylation fraction values y_i as a mixture of beta distributed random variables, drawn from a beta distribution with mean x_j and precision κ, weighted by the p.m.f. $\mathrm{P}\left(\beta_j|\theta, \tau, \mu, \gamma, \upsilon, \zeta, \Lambda\right)$. Assuming that each measured fCpG site is independent, such that the likelihood of a set of fCpG loci is the product the per-fCpG likelihood, the likelihood is then:

$$\mathrm{P}\left(y_i|\beta_j,\Delta,\varepsilon,\kappa\right)=\frac{{y_i}^{\kappa x_j-1}(1-y_i)^{\kappa(1-x_j)}-1}{B\left(\kappa x_j,\kappa(1-x_j)\right)}$$

$$\mathcal{L}(\theta,\tau,\mu,\gamma,\upsilon,\zeta,\beta_n,\Delta,\varepsilon,\kappa|\vec{y})=\prod_{i=1}^{n_{sim}}\sum_{j=1}^{J}\mathrm{P}\left(y_i|\beta_j,\Delta,\varepsilon,\kappa\right)\mathrm{P}\left(\beta_j|\theta,\tau,\mu,\gamma,\upsilon,\zeta,\Lambda\right)$$

It is often convenient to work with the log-likelihood, rather than the likelihood directly. Employing the logsumexp function, $LSE(x_1,\dots,x_n)=\log(e^{x_1}+,\dots,+e^{x_n})$, the log-likelihood is:

$$\log\left(\mathcal{L}(\theta,\tau,\mu,\gamma,\upsilon,\zeta,\beta_n,\Delta,\varepsilon,\kappa|\vec{y})\right)=\sum_{i=1}^{N}LSE\left(\log\left(\mathrm{P}\left(y_i|\beta_j,\Delta,\varepsilon,\kappa\right)\right)+\log\left(\mathrm{P}\left(\beta_j|\theta,\tau,\mu,\gamma,\upsilon,\zeta,\Lambda\right)\right)\right)$$

Here, we have modelled the growth rate and the time at which the MRCA emerged as independent; however, as we are assuming exponential growth, if we vary these two parameters independently, we implicitly assume that the total cancer burden can take on implausibly high values. To penalise combinations of θ and τ that yield such high values, we add a penalisation factor for cancer burdens outside the plausible range $\left\{{N_e}^{min},{N_e}^{max}\right\}$.

$$\mathcal{L}_{penalise}=\ e^{-{N_e}^{min}/N_e}-e^{-{N_e}^{max}/N_e}$$

Note that this is equivalent to placing a joint dependent prior over θ and τ.

Stochastic likelihood calculation correction

The log-likelihood calculation above differs from a classical likelihood calculation in that $\log(\mathcal{L})$ is not an exact value, but rather an approximation of the "true" log-likelihood which should become more exact as $n_{sim}\to\infty$. To test whether this estimator is systematically biased, we simulated a set of $\vec{y}$ data with a set of known parameters, then calculated the $\log(\mathcal{L})$ with the same set of parameters for a range of n_{sim} values (supplementary fig 20a).

We find that at low n_{sim} values the approximation tends to systematically underestimate the log-likelihood, whilst as n_{sim} increases, the calculated log-likelihood tends to asymptote to a single value. The increase in the mean estimated log-likelihood with n_{sim} can be rationalised if one considers that we are attempting to estimate a continuous pdf by sampling from it, and as n_{sim} increases, the stochastic noise from sampling is reduced and the approximation improves. In this sense, the fit at the peak of the log-likelihood distribution truly is more accurate, on average, as n_{sim} increases, hence the calculated log-likelihood is higher. Unsurprisingly, the standard deviation of the calculated log-likelihood values falls as the number of simulated paths increases.

We found that the bias associated with having a finite n_{sim} varied as $\log(\mathcal{L}_{bias})\sim\frac{1}{n_{sim}}$. By varying different parameters, we found the log-likelihood bias could be estimated as:

$$\log(\mathcal{L}_{bias}) \approx -\frac{0.5 + 0.005\kappa^{4/3}}{n_{sim}}$$

Correcting for the bias yields log-likelihood estimates with a mean that does not vary with n_{sim} (supplementary fig 20b). Hence, the bias-corrected cancer burden penalised log-likelihood is:

$$\log(\mathcal{L})' = \log(\mathcal{L}) + \log\left(\mathcal{L}_{penalise}\right) - \log(\mathcal{L}_{bias})$$

Validation of EVOFLUx inference using simulated data

To evaluate the accuracy of the inference, we simulated fCpG data and tested the ability of EVOFLUx to recover the known input parameters. These values were inferred with good confidence, with the 95% credible intervals encompassing the ground truth values (extended data fig 4a), significant narrowing of the posterior compared to the prior and with the posterior predictive distribution well-matching that of the original simulated dataset (extended data fig 4b).

To interrogate our modelling assumptions that growth rate was constant and that the epigenetic switching rate of each fCpG was the same, we simulated data from models where these assumptions have been weakened:

1) The growth rate of the cancer following a logistic growth model, rather than exponential. This was parameterised using the carrying capacity (K, the maximum population size), such that the population size followed the growth equation $S = \frac{K}{1+(K-1)e^{-\theta(T-\tau)}}$. In the limit $K \to \infty$, this growth equation reduces to the exponential growth equation used in our original model, $S = e^{\theta(T-\tau)}$.

2) Heterogeneity in the fCpG switching rate by fCpG site. We assumed that rather than each fCpG having the same epigenetic switching rates $\{\mu, \gamma, \upsilon, \zeta\}$, the jth fCpG had switching rates equal to the mean switching rates multiplied by a multiplicative factor drawn from a lognormal distribution with mean 1 and standard deviation sigma σ. In the limit $\sigma \to 0$, this model reduces to the original homogenous switching rate model.

We ran our EVOFLUx inference on these simulated datasets and tested how each parameter varied with the carrying capacity/ fCpG rate heterogeneity respectively.

In the case of logistic growth, there was no difference between the distributions of fCpG methylation as the carrying capacity is varied (extended data fig 4c). This is because the shape of the fCpG distribution, bar any ongoing subclonal selection, is driven by the early clonal dynamics of the cancer and the ongoing desynchronisation of fCpGs. Unsurprisingly, the inferred evolutionary parameters do not vary as a function of the carrying capacity (extended data fig 4d-e).

The distribution of fCpG methylation does change as a function of fCpG switching rate heterogeneity (extended data fig 4f). Specifically, as the heterogeneity increases, the flanking clonal peaks of fCpGs homozygous methylated and demethylated in the founder

cell shift to be closer to 1 and 0 respectively. This is due to the skew inherent in the lognormal distribution employed to model the heterogeneity – the majority of heterogenous fCpGs have a switching rate lower than the mean of all fCpGs and thus shift less towards the steady state methylation value. In this case, whilst there is a weak correlation between the inferred parameters and the degree of fCpG switching rate heterogeneity, the ground truth value of the biological parameters of primary interest (i.e. the growth rate and MRCA time) lie within the 95% credible interval as we vary the fCpG rate heterogeneity over a plausible regime (fig. extended data fig 4g-h).

Finally, we repeatedly down-sampled the number of fCpGs for an example CLL sample and re-ran EVOFLUx (extended data fig 5f). We found that whilst the variance was higher when fewer fCpGs were randomly sampled, the mean of the parameter estimates were largely insensitive to the number of fCpGs, particularly when the number of fCpGs was greater than ~600 (extended data fig 5g-i). Interestingly, the effective population size does show a clear bias with decreasing numbers of fCpGs.

Subclonal selection is detectable only within a limited region of parameter space

We used computational simulation to explore the detectability of subclonal architecture with fCpG data. Only strong selection occurring at an intermediate time point was detectable (supplementary fig 13c), analogous to limitations of subclone detection in genetic sequencing data. Early or weakly selected subclones were at undetectably low frequencies, while very strongly selected early subclones rapidly swept to fixation. Very early subclones had insufficient time to develop a distinguishable fCpG pattern; in this scenario, EVOFLUx could accurately identify a subclone was present but could not accurately infer its growth dynamics (supplementary fig 13d).