











Modeling the Mutational Dynamics of Very Short Tandem Repeats

Amos Onn^{1,2,*},ⁱ, Tzipy Marx³, Liming Tao⁴,ⁱⁱ, Tamir Biezuner³,ⁱⁱⁱ, Ehud Shapiro³, Christoph A. Klein^{1,5},^{iv}, and Peter F. Stadler^{2,6–10},^v

- ¹ Chair of Experimental Medicine and Therapy Research, University of Regensburg, Universitätsstraße 31, D-93053 Regensburg, Germany, amos.onn@klinik.uni-regensburg.de, christoph.klein@klinik.uni-regensburg.de
- ² Bioinformatics Group, Faculty of Mathematics and Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany, amos.onn@uni-leipzig.de, studla@bioinf.uni-leipzig.de
- ³ Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, 234 Herzl Street, POB 26, Rehovot 7610001, Israel, tzipy.marx@gmail.com, tamir.biezuner@weizmann.ac.il, ehud.shapiro@weizmann.ac.il
- ⁴ Cellular Tissue Genomics, Genentech, 94080 South San Francisco, CA, USA, taoliming.too@gmail.com
- ⁵ Fraunhofer Institute for Toxicology and Experimental Medicine Regensburg, Am BioPark 13, D-93053 Regensburg, Germany
- ⁶ Max Planck Institute for Mathematics in the Sciences, Inselstraße 22 D-04103, Leipzig, Germany
- ⁷ Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Vienna, Austria
- ⁸ Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia
- ⁹ Center for non-coding RNA in Technology and Health, University of Copenhagen, Ridebanevej 9, DK-1870 Copenhagen, Denmark
- ¹⁰ Santa Fe Institute, 1399 Hyde Park Rd., 87501 Santa Fe, New Mexico, USA

Abstract. Short tandem repeats (STRs) are low-entropy regions in the genome, consisting of a short (1-6 bp) unit that is consecutively repeated multiple times. They are known for high mutational instability, due to so-called stutter-mutations, in which the number of units in the run increases or decreases. In particular, STRs with repeat unit length of 1-2 bp are prone to mutate even within several cell divisions. The extremely rapid accumulation of variation makes them interesting phylogenetic markers for retrospective single-cell lineage reconstruction. Here we model their mutational dynamics at the level of individual repeat unit motif and then aggregate length variations over many STR loci with the

ⁱ  0000-0002-8403-0754
ⁱⁱ  0000-0002-3034-925X
ⁱⁱⁱ  0000-0002-3642-0977
^{iv}  0000-0001-7128-1725
^v  0000-0002-5016-5191

*Corresponding author

aim of obtaining a very fast “molecular clock”. We calibrate our model based on several datasets with known lineage structure prepared from cultured cells. We find that the mutational dynamics of STRs are reasonably consistent for a given cell-line, but vary among different ones. This suggests that the dynamics are not entirely explained by mutations in caretaker genes, rather, various other factors play a role — possibly tissue origin and differentiation state. Further data and research is necessary to assess their relative effects.

Keywords: short tandem repeats, microsatellites, STR, MS, lineage, single cell

1 Introduction

Reconstructing the lineage of single cells is a problem of ongoing interest, with implications in various fields. Most methods typically employed for this purpose, however, require intensive intervention during the cell division process — either manipulation of the cells themselves [13], or some modification of their genetic material [9, 10, 11]. Hence they are not applicable to retrospective lineage reconstruction of human individual samples, where data are by definition limited to *a posteriori* observations of naturally-occurring somatic mutations, and intervention *in vivo* is not possible. While it is possible in principle to use single nucleotide polymorphisms (SNPs) for this purpose, the resolution of such data is limited by the comparably low mutations rates, around 10^{-8} per site per cell division [16], implying that even genome-wide sequencing with high coverage — and the associated high effort and cost — will capture only a very small number of somatic mutations [6, 18]. A possibly cost-saving alternative is to focus on “hot-spot” SNPs; however, this results in a panel of SNPs partially driven by selective pressure, thus invalidating the assumption of independence of mutations and skewing the results [7]. This is particularly relevant in the case of lineage-tracing of cancer, which is one of the major use-cases [5].

To overcome some of these problems, we have developed an assay for reconstructing lineage based on sequencing particular regions of the genome known as short tandem repeats (STRs, also known as microsatellites) [1, 12, 14]. These are low-entropy regions, where a short (1–6 bp) motif of nucleotides, repeats itself in direct consecution, with numerous (≥ 5) copies of the repeat unit following each other. Due to the low entropy, during DNA duplication, the strands may detach and reattach at an offset of one or several repeat units, leading to a so-called stutter mutation. As a consequence the number of repeat units, also referred to as the length of the STR, increases or decreases. The rate for these mutations is several orders of magnitude higher than that of random single point mutations: estimates range from 10^{-3} – 10^{-5} per site per cell division for di-repeats [17]. In addition, STRs are considered evolutionary neutral [3]. The extremely rapid accumulation of differences in repeat length makes STRs, in particular those with very short units, excellent candidates for tracing single cell lineage. The high mutability, however, also creates difficulties for the analysis of such data:

- (1) The STR length observable in sequencing data differ from the *in vivo* state because the unavoidable amplification steps during library preparation introduce additional stutter mutations.
- (2) Mutation rates differ substantially between loci and cell types.

The first point was addressed in a previous work [12]; our goal is to devise a continuous-time Markov model describing *in vivo* temporal evolution of STR length, focusing on second point. This model may be then used as a fast molecular clock, enabling estimations of distance in cell-divisions between single-cell samples. This distance can then be used for reconstructing lineage trees of these cells; the Markov model can also be used directly to assess and maximize the likelihood of the tree topologies. Such trees, which, in addition to a topological structure, also have reliable edge lengths, contribute toward a better understanding of the evolution of the analysed tissues.

Our aim is to allow for different rates of evolution at each locus. The size of our panel, consisting of thousands or tens of thousands of loci, however, makes it infeasible to optimise all parameters simultaneously. Therefore, we proceed in two steps. First, we estimate a separate base rate for the evolution of STR length at each locus, and then we aggregate across groups of loci of the same repeat unit motif. Then we iteratively repeat these two steps until the estimated parameters converge. We use eight artificial lineage trees constructed from four different cell-lines to estimate the evolutionary dynamics of STR evolution. Although there is coherence between data from the cell types, we also observe significant variations of the model parameters.

2 STR Data

STR data are produced by a specialized assay comprising the following steps:

- (1) Following DNA extraction from the sample, whole genome amplification is performed. Different protocols were used for different datasets: the Repli-G protocol in the DU145 dataset, the Ampli1 protocol [2] in the HESC dataset, and an in-house phi29-based protocol [15] in the two HCT116 datasets.
- (2) The amplified DNA is hybridized with a panel of duplex molecular inversion probes (MIP) that target genomic regions known to harbour a STR with unique flanking regions. For details about this panel see [14]. In the HESC dataset and the two HCT116 datasets, OM9 was used. In the DU145 dataset, OM6 was used.
- (3) The selected DNA is then gap-filled, ligated, prepared as standard Illumina libraries and sequenced [14].
- (4) The forward and reverse reads are merged using PEAR [19].
- (5) The paired-end reads are aligned to a custom index comprising the same loci as the hybridization panel. For each locus, target sequences with different number of repetitive units are enclosed between unaltered flanking sequences. Each mapped read therefore specifies a locus and STR length.
- (6) For each STR locus, a histogram of observed lengths is determined.

These STR length measurements cannot be used without corrections because the *in vitro* amplification steps during library preparation introduce further stutter mutations. The raw STR length data therefore follow a characteristic distribution. In previous work, a collection of reference distributions for the same panel of loci was generated for various lengths of the original alleles [12]. The distributions obtained from the sequencing data are then compared to the reference distributions taking 1 minus correlation as distance metric. In order to find the best match, we use k-d tree search [4]. A further difficulty arises for the fact that the cells of interest are derived of a diploid genome and hence for each locus there are (usually) two alleles. We therefore estimate a superposition of two reference distributions. The result of this procedure is the best estimate of the true *in vivo* STR length for each allele. In typical samples we assay 3000–8000 loci per sample, translating to 5000–10000 alleles.

Table 1: Overview of STR datasets used for model inference.

Tree	# of leaves	# of unique cells	Max. depth of leaf
DU145_A4	671	138	8
DU145_A4_deep	405	84	4
DU145_A4_rest	266	54	7
HCT116_MSI	80	80	3
HCT116_MSS	92	92	3
WIS_A8	90	35	1
WIS_D1	523	438	11
WIS_D11	33	15	1

In order to derive a model for converting STR length data into evolutionary distances we use a collection of datasets that were specifically prepared to investigate the use of STR length data. In brief, a culture is grown starting from a single cell. At predefined time points, single cells are picked from this culture. Some of these cells extracted and STR lengths data are measures, while other cells form the seed for new cultures. This process is then repeated for several generations to produce an artificial lineage tree. See [1] for more details about the preparation. For the present study we use the following eight datasets:

- a tree originated from a single DU145 cell (prostate cancer, displaying microsatellite instability) [1] (DU145_A4) as well as deep sub-clade of this tree (DU145_A4_deep) and the complement of this sub-clade (DU145_A4_rest)
- a tree originated from a single HCT116 cell (colon cancer, displaying microsatellite instability) [15] (HCT116_MSI)
- a tree originated from a single HCT116 cell with a functional version of the MMR gene hMLH1 reintroduced [8, 15] (HCT116_MSS)

- three trees originated from three related (same ampule) single HESC cells (first published here) (WIS_A8, WIS_D1, WIS_D11)

Sizes of the datasets are detailed in Tab. 1. For a sketch of the structure of the trees, see Fig. S1.

3 Model of STR Length Evolution

We model the temporal evolution of STR length as a continuous-time Markov chain. A practical difficulty is that STR lengths show large variations across loci and hence we need to consider a fairly large number of states. For the data described in the previous section we consider $5 \leq j, k \leq 38$, i.e., lengths between 5 and 38. To avoid complicated treatment of “partial” repeat units, we only count complete repeat units and therefore take the STR length to be an integer. In order to reduce the number of parameters that need to be estimated independently for each locus we assume that we can decompose the rate matrix for a given locus ℓ and repeat unit motif τ in the following simple product form:

$$\mathbf{R}(\ell, \tau) = \mu(\ell) \mathbf{R}(\tau) \quad (1)$$

Here $\mu(\ell)$ is a single mutation rate specific for each locus ℓ . In contrast, $\mathbf{R}(\tau)$ is a matrix common to all loci of the same motif that describe the *relative* rates of length changes. Note that we allow both $\mu(\ell)$ and $\mathbf{R}(\tau)$ to depend on the cell-line. These parameters need to be therefore inferred independently for each dataset. We further constrain the model by two additional plausible assumptions:

- (1) The Markov chain is reversible and hence has a stationary distribution;
- (2) This stationary distribution coincides with the empirical distribution of STR lengths in the human genome.

The first assumption is clearly an approximation since STR that become too short will not behave like STRs any more. Since the data we actually measure are far away from this limit, it is nevertheless a safe assumption to make in our setting. The short time-scale of stutter mutations strongly suggests that their length distributions are equilibrated in the human genome.

In order to reduce the number of parameters that need to be estimated we use a simple model for the relative rates \mathbf{R}_{jk} of a transition from length j to k , and assume that these values depend only on the length of the STR and the change in length. We first construct a symmetric, doubly-stochastic rate matrix as following:

$$\tilde{\mathbf{R}}_{jk} := \begin{cases} \exp(\gamma + \alpha(j+k) - \lambda|j-k|), & j \neq k \\ -\sum_{i \neq j} \tilde{\mathbf{R}}_{ji}, & j = k \end{cases} \quad (2)$$

The parameter γ is a scaling parameter. Such a symmetric rate matrix will lead to a uniform stationary distribution. It can be adjusted to enforce a given empirical stationary distribution by setting

$$\mathbf{R}_{jk}(\tau) := s_j^{-1} \tilde{\mathbf{R}}_{jk}(\tau) \quad (3)$$

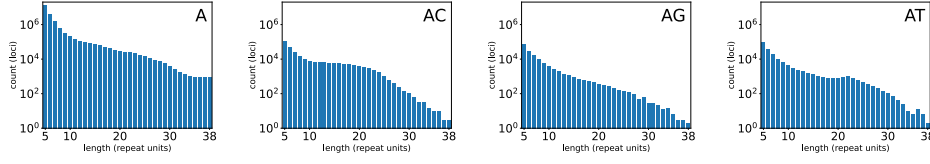


Fig. 1: Stationary distribution of STR length for the repeat unit motifs $\tau \in \{A, AC, AG, AT\}$ in the reference genome hg38.

where $s_j(\tau)$ is the proportion of STRs of repeat unit motif τ of length j in the genome. A short derivation of this correction can be found in Appendix A. The empirical distributions for the human genome are shown in Fig. 1.

In order to efficiently fit the model to the observed STR count data for a given sample, we proceed iteratively. We start with a uniform matrix of transition rates $\mathbf{R} = \mathbf{J} - N\mathbf{I}$, where \mathbf{I} is the identity matrix, \mathbf{J} is the matrix with entries 1, and N is the size of the matrices (in our case 34). For this transition rate matrix, we have the following closed-form solution for the probabilities S_{jk} that j is substituted by k after time t :

$$\begin{aligned} S_{kk}(t) &= S_{=} = \frac{1}{N} (1 - \exp(-Nt)) \\ S_{jk}(t) &= S_{\neq} = \frac{1}{N} (1 + (N-1)\exp(-Nt)) \quad \text{if } j \neq k \end{aligned} \quad (4)$$

We aim to use the maximum-likelihood estimator to obtain $\mu(\ell)$ for each locus. To this end, we compare the STRs in each pair of leaves u and v in the known trees. We write $d(u, v)$ for their divergence in the tree measured as the number of generations, more precisely, passages of the cultures separating the two cell colonies, plus 2 for the divergence of each sample from the founder of its colony. For two cells within the same colony, this number is still taken as 2; a short derivation of this can be found in Appendix B. Now we substitute the time as $t = \mu(\ell)d(u, v)$ in the above solution. Moreover, we write $n_{=}(\ell, u, v)$ and $n_{\neq}(\ell, u, v)$ for the number of equal and distinct alleles between the two samples, respectively. Note that $n_{=}(\ell, u, v) + n_{\neq}(\ell, u, v) = 2$ if u and v have two alleles each for locus ℓ , and $n_{=}(\ell, u, v) + n_{\neq}(\ell, u, v) = 1$ otherwise. The likelihood $\mathcal{L}(\ell)$ of observing a set pairs of STRs at given locus ℓ thus takes the form

$$\log L(\ell) = \sum_{u,v} n_{=}(\ell, u, v) \log S_{=}(\mu(\ell)d(u, v)) + n_{\neq}(\ell, u, v) \log S_{\neq}(\mu(\ell)d(u, v)) \quad (5)$$

A Newton-Raphson search is sufficient to determine $\max \log \mathcal{L}$ as a function of μ .

Given the estimated values of $\mu(\ell)$, we proceed by refining the model for the rate matrices $\mathbf{R}(\tau)$. We estimate a single matrix for all the loci of each repeat unit motif $\ell \in \mathcal{L}(\tau)$. Again this is done for each tree separately. For a given locus ℓ and two samples u, v , we count the number $a_{j,k}(\ell, u, v)$ of transitions $j \rightarrow k$ from allele length j in u to allele length k in v , as follows:

- If both u and v have a single allele, we count the single transition $u_1 \rightarrow v_1$.
- If both u and v have two distinct alleles, we order them as $u_1 < u_2, v_1 < v_2$, and count two transitions $u_1 \rightarrow v_1$ and $u_2 \rightarrow v_2$.
- If u has two alleles and v has one, we count a single transition of the allele of u closer to the one in v , i.e. denoting $i = \minarg\{|u_2 - v_1|, |u_1 - v_1|\}$ we count $u_i \rightarrow v_1$.
- If u has one allele and v has two, we count a single transition of the allele of u to the one in v closer to it, i.e. denoting $i = \minarg\{|u_1 - v_1|, |u_1 - v_2|\}$ we count $u_1 \rightarrow v_i$.

The likelihood function for a tree then takes the form

$$\log \mathcal{L}(\tau; \gamma, \alpha, \lambda) = \sum_{u,v} \sum_{\ell \in \mathcal{L}(\tau)} \sum_{j,k} a_{j,k}(\ell, u, v) \log \mathbf{S}_{jk}(\mu(\ell) d(u, v)) \quad (6)$$

with $\mathbf{S}(t) = \exp(t\mathbf{R}(\gamma, \alpha, \lambda))$

For each value of the three parameters γ , α , and λ we can calculate the rate matrix, and from it the transition probability matrices for the required time points, in order to evaluate $\log \mathcal{L}$. Optimization of $\log \mathcal{L}$ over the parameter space is done using L-BFGS-B [20]; finding the maximal log-likelihood value, we obtain the optimal parameters and our estimation of the rate matrices $\mathbf{R}(\tau)$.

We now proceed to re-estimate the rate coefficients of $\mu(\ell)$ based on the newly estimated rate matrices. The expression for the likelihood

$$\log \mathcal{L}(\ell) = \sum_{u,v} \sum_{j,k} a_{j,k}(\ell, u, v) \log \mathbf{S}_{jk}(\mu(\ell) d(u, v)) \quad (7)$$

with $\mathbf{S}(t) = \exp(t\mathbf{R}(\tau(\ell)))$

is somewhat more involved than in (5), but still analytical and therefore a Newton-Raphson search can again be used. After re-estimating μ , the rate matrices $\mathbf{R}(\tau)$ are re-estimated as well using (6) again. We observe that a third iteration of estimation of μ leads to no further noticeable changes.

4 Results

In the top row of Fig. 2 we can see a density plot of the distributions of the locus-specific rates μ , for the eight datasets listed in Sect. 2. They are grouped by cell-line: the three DU145 trees (one tree and two subtrees), the two HCT116 trees, and the three HESC trees. We note that all curves have a similar form, suggesting some underlying common distribution, with varying parameters. We also note some skews of the center of this distribution, corresponding to a difference in the base rate of mutation.

For a more detailed comparison of the rates we computed linear regressions of shared loci for each pair of trees. We begin by taking the correlations (R-values) as a measure of similarity of the distributions (Fig. 3, left panel). Comparing rate estimations for trees of different cell-lines, we observe a substantial difference,

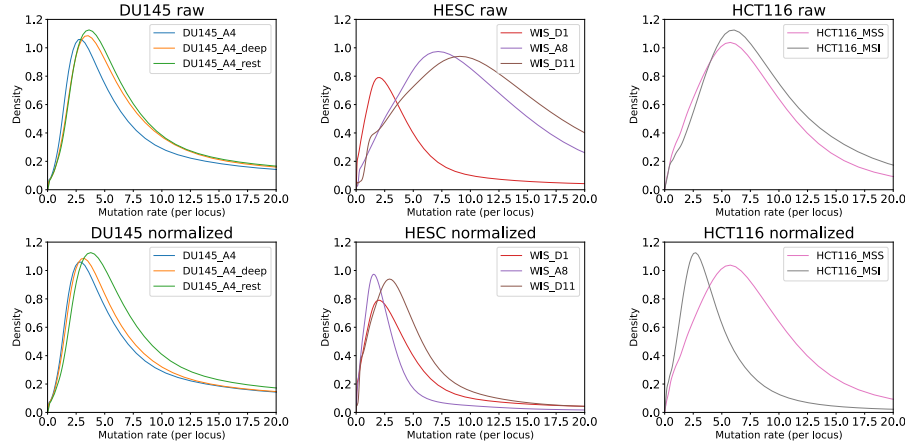


Fig. 2: Distribution of locus-specific rates μ estimated independently for the eight datasets, and grouped by cell-line: the three DU145 trees are subsets of the same tree; the three HESC trees are separately generated; the two HCT116 trees are grouped together, despite a genetic modification in the cell-line seeding HCT116-MSS. See also 1. Top row are the raw coefficients; bottom row they are scaled by the slope coefficients of linear regression within each group.

reflected through low correlations. Comparing the rate estimates for the same cell-lines, however, we observe very high correlation for each pair DU145 trees and each pair of HESC trees, respectively. The locus-specific rates for two HCT116 trees exhibit a lower correlation than the DU145 and HESC data. This is consistent with the fact that these are two related cell-lines, but with a major genetic modification concerning mismatch repair (MMR) [8]. However the R-values are still higher than the R-values computed by comparing completely different cell-lines. This suggests that the common origin of the two variants of HCT116 still plays a significant role.

Using the slope of the linear regression as an estimation for overall scaling of rate (Fig. 3, right panel), we continue to the scaling of the rates estimated for trees of the same cell-line. We note that the rates for the DU145 trees have very little scaling, as would be expected since the three trees are essentially subsets of the same tree. A 2.2-fold scaling of the rates of HCT116_MSI as compared to the HCT116_MSS tree is also as expected, since the variants displaying microsatellite instability should indeed be more quickly mutating. The large factor between one of the HESC trees (WIS_D1) and the two others (WIS_A8, WIS_D11), of respectively 4.7 and 3.1, is somewhat surprising, but might be an artifact arising from the shallowness of the two latter trees, which allows proper calibration of only the faster-evolving loci. This might also explain the lower correlation compared to DU145.

Scaling the locus rate distributions by these slopes, we obtain the bottom row of Fig. 2. For DU145 there is no large difference, but they are more closely

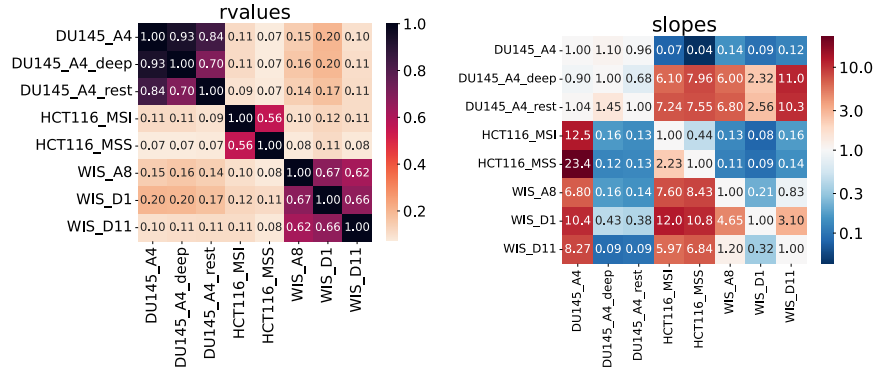


Fig. 3: Linear regressions of the locus-specific rate parameters $\mu(\ell)$ between pairs of samples: R-values (left) and estimated slopes (right). The slopes are column against row, so that a higher-than-1 slope means the tree of the column mutates more quickly than that of the row.

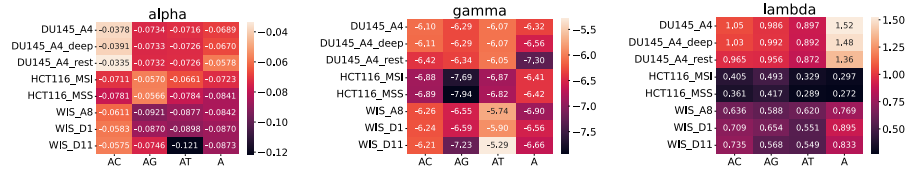


Fig. 4: Optimised rate-matrix model parameters, estimated separately for each tree and repeat unit motif.

aligned. For HESC this successfully corrects the shift seen in the top row, aligning the peaks of the density plot, and we can see the close relation of these distributions, despite the scaling on both axes. For HCT116 the peaks are not aligned, but the center of the distribution is. This highlights their lower correlation.

Let us now turn to the motif-specific parameters γ , α , and λ , as seen in Fig. 4. These were independently estimated for each repeat unit motif, and we considered only data for the four motifs $\tau \in \{A, AC, AG, AT\}$. Although the parameters have a straightforward interpretation in the symmetric rate model \mathbf{R} , they are confounded by the stationary length distributions in Fig. 1. Even though the negative values of α appear to suggest that the mutation rate decreases with repeat length, we observe that \mathbf{R} in fact contains higher values for large lengths.

Again, we see a similar pattern of coherence between cell types. The parameters for each given τ are very close among the DU145 trees, among the HESC trees, and also between the two HCT116 trees. Remarkably, the parameters for the HCT116 trees are significantly more similar to each other than the locus rates are. This suggests that the common origin indeed plays a role in the length transition rates, and that the modification in MMR gene hMLH1 might affect certain loci more than others, but not so much the per-motif transition rates.

5 Concluding Remarks

The main application for estimating the rate models described in the previous section is retrospective single-cell lineage reconstruction, in particular of sample sets consisting of various healthy cells and cancer cells of various stages. The variations in mutational behaviour among cell-lines, in particular the low correlation of locus specific mutation rates, however, suggest that cells of different differentiation states might also vary in their mutational behaviour. This presents a problem when attempting to reconstruct a tree composed of heterogeneous samples, which would require a unified model of mutation.

The datasets we analysed in this contribution provide us with limited insight as to the factors governing this difference in locus specific rates and of length transition rates. This is due to the fact, that they all belong to different individuals as well as different cell types. In order to further investigate those factors, we will require datasets controlling for these various factors — trees originating from the same cell type in different individuals, and trees originating from various cell types within a single individual. Such datasets may allow us to determine some commonalities and possibly develop a collection of models suitable for reconstructing lineage of various datasets. This will require in addition the development of a framework for estimating the likelihood of topologies which include shifts between mutational models.

In addition to the lineage reconstruction concerns, a better understanding of the factors determining the mutational behaviour of STRs might be of interest as of itself. Microsatellite instability (a marked increase in the mutation rate of STRs) is a well-studied phenomenon in cancer research, and further results of our method might contribute to this research.

Acknowledgments

A.O. and research in the Klein lab was supported by the Deutsche Forschungsgemeinschaft (DFG) (TRR-305/A01). Research in the Shapiro lab was supported by the European Union grants ERC-2014-AdG (project no. 670535) and EU-H2020-Health (project no. 874606). Research in the Stadler lab is supported by the German Federal Ministry of Research, Technology and Space (BMFTR) through DAAD project 57616814 (SECAI, School of Embedded Composite AI), the German Network for Bioinformatics Infrastructure, (de.NBI/RBC, grant W-de.NBI-018), and jointly by the BMFTR and the Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research *Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig*, project identification number: SCADS24B.

Author Contribution

Conceptualization, A.O. and P.F.S.; formal analysis and software, A.O.; wet-lab methodology, single-cell isolation, and sample preparation, T.M., L.T. and

T.B.; funding acquisition, E.S. and C.A.K.; wet-lab supervision, E.S.; supervision, C.A.K and P.F.S.

Data Availability

Data for the DU145 dataset is available as supplementary material for [14]. Additional data for that dataset, as well as data for the HESC dataset, will be gladly made available upon request by the corresponding author. Regarding the HCT116 datasets please contact the authors of [15] until that manuscript is published.

References

1. Biezuner, T., Spiro, A., Raz, O., Amir, S., Milo, L., Adar, R., Chapal-Ilani, N., Berman, V., Fried, Y., Ainbinder, E., Cohen, G., Barr, H.M., Halaban, R., Shapiro, E.: A generic, cost-effective, and scalable cell lineage analysis platform. *Genome research* **26**(11), 1588–1599 (2016). <https://doi.org/10.1101/gr.202903.115>
2. Czyż, Z.T., Klein, C.A.: Deterministic whole-genome amplification of single cells. *Methods in molecular biology (Clifton, N.J.)* **1347**, 69–86 (2015). https://doi.org/10.1007/978-1-4939-2990-0_5
3. Ellegren, H.: Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5**(6), 435–445 (Jun 2004). <https://doi.org/10.1038/nrg1348>
4. Freidman, J.H., Bentley, J.L., Finkel, R.A.: An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Mathematical Software* **3**(3), 209–226 (1977). <https://doi.org/10.1145/355744.355745>
5. Hess, J.M., Bernards, A., Kim, J., Miller, M., Taylor-Weiner, A., Haradhvala, N.J., Lawrence, M.S., Getz, G.: Passenger hotspot mutations in cancer. *Cancer Cell* **36**(3), 288–301.e14 (Sep 2019). <https://doi.org/10.1016/j.ccell.2019.08.002>
6. Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., Liu, X., Yao, X., Zheng, H., Yu, C., Li, Q., Gong, Z., Mao, M., Yang, X., Yang, L., Li, J., Wang, W., Lu, Z., Gu, N., Laurie, G., Bolund, L., Kristiansen, K., Wang, J., Yang, H., Li, Y., Zhang, X., Wang, J.: Single-cell exome sequencing and monoclonal evolution of a *jemζjak2i/emζ*-negative myeloproliferative neoplasm. *Cell* **148**(5), 873–885 (Mar 2012). <https://doi.org/10.1016/j.cell.2012.02.028>
7. Johnson, P.L.F., Hellmann, I.: Mutation rate distribution inferred from coincident snps and coincident substitutions. *Genome Biology and Evolution* **3**, 842–850 (05 2011). <https://doi.org/10.1093/gbe/evr044>
8. Koi, M., Umar, A., Chauhan, D.P., Cherian, S.P., Carethers, J.M., Kunkel, T.A., Boland, C.R.: Human chromosome 3 corrects mismatch repair deficiency and microsatellite instability and reduces n-methyl-n'-nitro-n-nitrosoguanidine tolerance in colon tumor cells with homozygous hmlh1 mutation. *Cancer research* **54**(16), 4308–12 (1994)
9. Kretzschmar, K., Watt, F.: Lineage tracing. *Cell* **148**(1), 33–45 (Jan 2012). <https://doi.org/10.1016/j.cell.2012.01.002>
10. Lu, R., Neff, N.F., Quake, S.R., Weissman, I.L.: Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nature Biotechnology* **29**(10), 928–933 (Oct 2011). <https://doi.org/10.1038/nbt.1977>

11. McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., Shendure, J.: Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**(6298), aaf7907 (2016). <https://doi.org/10.1126/science.aaf7907>, <https://www.science.org/doi/abs/10.1126/science.aaf7907>
12. Raz, O., Biezuner, T., Spiro, A., Amir, S., Milo, L., Titelman, A., Onn, A., Chapal-Ilani, N., Tao, L., Marx, T., Feige, U., Shapiro, E.: Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. *Nucleic Acids Research* **47**(5), 2436–2445 (01 2019). <https://doi.org/10.1093/nar/gky1318>
13. Sulston, J., Schierenberg, E., White, J., Thomson, J.: The embryonic cell lineage of the nematode *caenorhabditis elegans*. *Developmental Biology* **100**(1), 64–119 (1983). [https://doi.org/https://doi.org/10.1016/0012-1606\(83\)90201-4](https://doi.org/https://doi.org/10.1016/0012-1606(83)90201-4), <https://www.sciencedirect.com/science/article/pii/0012160683902014>
14. Tao, L., Raz, O., Marx, Z., Ghosh, M.S., Huber, S., Greindl-Junghans, J., Biezuner, T., Amir, S., Milo, L., Adar, R., Levy, R., Onn, A., Chapal-Ilani, N., Berman, V., Ben Arie, A., Rom, G., Oron, B., Halaban, R., Czyz, Z.T., Werner-Klein, M., Klein, C.A., Shapiro, E.: Retrospective cell lineage reconstruction in humans by using short tandem repeats. *Cell Reports Methods* **1**(3), 100054 (2021). <https://doi.org/https://doi.org/10.1016/j.crmeth.2021.100054>, <https://www.sciencedirect.com/science/article/pii/S2667237521001028>
15. Tao, L., Silverbush, D., Suva, M.: Decoding glioma plasticity via single-cell multi-omics (unpublished)
16. Wang, J., Fan, H., Behr, B., Quake, S.: Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* **150**(2), 402–412 (Jul 2012). <https://doi.org/10.1016/j.cell.2012.06.030>
17. Willems, T., Gymrek, M., Poznik, G., Tyler-Smith, C., Erlich, Y.: Population-scale sequencing data enable precise estimates of y-str mutation rates. *The American Journal of Human Genetics* **98**(5), 919–933 (May 2016). <https://doi.org/10.1016/j.ajhg.2016.04.001>
18. Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H., Liu, X., Cao, D., Guo, G., Hu, X., Gui, Y., Li, Z., Xie, W., Sun, X., Shi, M., Cai, Z., Wang, B., Zhong, M., Li, J., Lu, Z., Gu, N., Zhang, X., Goodman, L., Bolund, L., Wang, J., Yang, H., Kristiansen, K., Dean, M., Li, Y., Wang, J.: Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**(5), 886–895 (Mar 2012). <https://doi.org/10.1016/j.cell.2012.02.025>
19. Zhang, J., Kobert, K., Flouri, T., Stamatakis, A.: Pear: a fast and accurate illumina paired-end read merger. *Bioinformatics* **30**(5), 614–620 (10 2013). <https://doi.org/10.1093/bioinformatics/btt593>
20. Zhu, C., Byrd, R.H., Lu, P., Nocedal, J.: Algorithm 778: L-bfgs-b: Fortran sub-routines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* **23**(4), 550–560 (Dec 1997). <https://doi.org/10.1145/279232.279236>

Appendix A

The following result seems to be well known although we are not aware of a convenient reference. We therefore include a short proof for completeness. Denote by $\mathbf{1}$ the vector with all entries 1 and write $\mathbf{0}$ for the zero vector. Note that any rate matrix $\hat{\mathbf{R}}$ satisfies $\hat{\mathbf{R}}\mathbf{1} = \mathbf{0}$ since $\mathbf{R}_{jj} = -\sum_{k \neq j} \mathbf{R}_{jk}$ by definition. Moreover,

for a row vector \mathbf{p} , denote by $\text{diag}(\mathbf{p})^{-1}$ the diagonal matrix with entries \mathbf{p}_j^{-1} . One easily checks that $\mathbf{p} \text{diag}(\mathbf{p})^{-1} = \mathbf{1}$.

Lemma 1. *Let $\tilde{\mathbf{R}}$ be a symmetric rate matrix satisfying $\tilde{\mathbf{R}}\mathbf{1} = \mathbf{0}$, let \mathbf{p} be a strictly positive (row) vector and set $\mathbf{R} := \text{diag}(\mathbf{p})^{-1}\tilde{\mathbf{R}}$. Then \mathbf{p} is a (left) eigenvector of $\exp(t\mathbf{R})$ with eigenvalue 1 for all t .*

Proof. Symmetry of $\tilde{\mathbf{R}}$ implies $\mathbf{1}\tilde{\mathbf{R}} = \mathbf{0}$. The definition of \mathbf{R} yields $\mathbf{p}\mathbf{R} = \mathbf{p} \text{diag}(\mathbf{p})^{-1}\tilde{\mathbf{R}} = \mathbf{1}\tilde{\mathbf{R}} = \mathbf{0}$. We compute

$$\mathbf{p} \exp(t\mathbf{R}) = \mathbf{p} \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbf{R}^k = \mathbf{p}\mathbf{I} + \underbrace{\mathbf{p}\mathbf{R}}_{\mathbf{0}} \sum_{k=1}^{\infty} \frac{t^k}{k!} \mathbf{R}^{k-1} = \mathbf{p}$$

i.e., \mathbf{p} is indeed an eigenvector of $\exp(t\mathbf{R})$ with eigenvalue 1. \square

In particular, therefore, if \mathbf{p} is a strictly positive probability distribution, it is a stationary distribution of the continuous time Markov chain with rate matrix $\mathbf{R} = \text{diag}(\mathbf{p})^{-1}\tilde{\mathbf{R}}$, where $\tilde{\mathbf{R}}$ is any symmetric rate matrix.

Appendix B

The expected STR distance of two randomly selected cells in a culture can be estimated by assuming a constant duplication rate. Recall that each culture is started from a single cell in the experiments described in Sect. 2 (STR Data). Assuming all divisions are symmetric, the culture forms in the n -th generation (considering the seed cell a zeroth generation) a complete binary tree with $2^{n+1} - 1$ cells of which 2^n are leaves. In this case a given extant cell (leaf of the tree) has 2^h other extant cells within its clade of height h , and thus 2^{h-1} cells of distance exactly $2h$. Thus, denoting the depth of the entire colony as n , the sum of all distances is: $2 \sum_{h=1}^n h 2^{h-1} = (n-1)2^{n+1} + 2$, and hence the expected distance (divided by the number of other leaves, $2^n - 1$) is $2(n-1) + O(n2^{-n})$. Since we measure time in terms of passages of the culture, not replications, we obtain an average distance of $2(1 - \frac{1}{n})$ up a finite size correction of order 2^{-n} . For estimated n values of 10–30 this is close enough to be taken as 2.

Appendix C

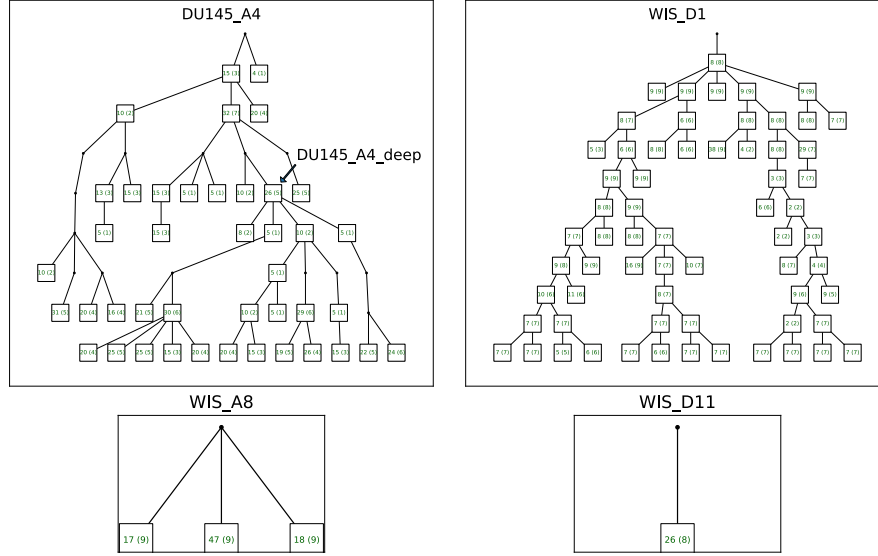


Fig. S1: Sketches of DU145 and HESC trees. Nodes represent colonies. The number of extracted samples is annotated in each node. The value in brackets gives the number of unique extracted cells. For DU145, the entire tree is DU145_A4, the marked clade is DU145_A4_deep, and the tree excluding that clade is DU145_A4_rest. See also Tab. 1 for an overview of the trees.