

RESEARCH ARTICLE

Bulk Genotyping of Biopsies Can Create Spurious Evidence for Heterogeneity in Mutation Content

Rumen Kostadinov¹, Carlo C. Maley^{2*}, Mary K. Kuhner^{3*}

1 Pediatric Oncology, Johns Hopkins School of Medicine, Baltimore, Maryland, United States of America, **2** Center for Evolution and Cancer, University of California, San Francisco, San Francisco, California, United States of America, **3** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America

* Current address: School of Life Sciences, Arizona State University, Phoenix, Arizona, United States of America

* mkkuhner@uw.edu



OPEN ACCESS

Citation: Kostadinov R, Maley CC, Kuhner MK (2016) Bulk Genotyping of Biopsies Can Create Spurious Evidence for Heterogeneity in Mutation Content. *PLoS Comput Biol* 12(4): e1004413. doi:10.1371/journal.pcbi.1004413

Editor: Edwin Wang, University of Calgary Cumming School of Medicine, CANADA

Received: January 19, 2015

Accepted: March 18, 2016

Published: April 22, 2016

Copyright: © 2016 Kostadinov et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data are archived in Dryad, doi:[10.5061/dryad.hf93c](https://doi.org/10.5061/dryad.hf93c).

Funding: This work was funded by the following: Research Scholar Grant 117209-RSG-09-163-01-CNE from the American Cancer Society (RK and CCM), <http://www.cancer.org/>. Congressionally Directed Medical Research Programs: Breast Cancer Research Program Breakthrough Award BC132057 (CCM), <http://cdmrp.army.mil/bcrp/>. National Institutes of Health grants NCI R01 CA140657 (RK), NIH R01 CA149566 (CCM), NIH R01 CA170595 (CCM), NIH R01 CA140657 (CCM and MKK), and NIH P01 CA91955 (CCM and MKK), <http://www.nih.gov>. The

Abstract

When multiple samples are taken from the neoplastic tissues of a single patient, it is natural to compare their mutation content. This is often done by bulk genotyping of whole biopsies, but the chance that a mutation will be detected in bulk genotyping depends on its local frequency in the sample. When the underlying mutation count per cell is equal, homogenous biopsies will have more high-frequency mutations, and thus more detectable mutations, than heterogeneous ones. Using simulations, we show that bulk genotyping of data simulated under a neutral model of somatic evolution generates strong spurious evidence for non-neutrality, because the pattern of tissue growth systematically generates differences in biopsy heterogeneity. Any experiment which compares mutation content across bulk-genotyped biopsies may therefore suggest mutation rate or selection intensity variation even when these forces are absent. We discuss computational and experimental approaches for resolving this problem.

Author Summary

Researchers who take multiple samples from a cancer or pre-cancer tissue and find that some samples show far more mutations than others are likely to conclude that the high-mutation samples reflect cells with an abnormal mutation or growth rate. We considered the common practice of testing a bulk sample for mutations, which finds only mutations that are common within the sample. Our computer simulations show that even when all cells have identical mutation and growth rates, testing bulk samples frequently leads to spurious detection of rate differences. This can lead to false conclusions about the causes and progress of cancer. We discuss possible solutions involving either genetic testing of single cells or the use of computer algorithms to detect rare mutations within a sample.

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The somatic genotypes of cancerous and pre-cancerous tissues are frequently assayed by taking biopsies containing a substantial number of cells and genotyping each biopsy as a whole (via SNP chip, exome or genome sequencing, or other techniques). For example, in a study of Barrett’s esophagus genotypes were derived from biopsies containing approximately one million epithelial cells (e.g. [1]). We will refer to this type of data collection as bulk-biopsy genotyping. It is generally much less expensive and technically difficult than single-cell genotyping. In this study we examine the use of multiple biopsies from a single tissue or tumor. It is tempting to think that differences in mutation content observed in bulk-biopsy genotyping reflect underlying differences in the number of mutations per cell, which could be informative about the spatial or temporal evolution of the tissue. But is this really true?

The roles of mutations in cancer have often been assessed by analyzing single tumor samples or paired tumor/normal samples from many patients. However, recently it has been recognized that multiple samples from a single patient, separated in time or space, offer additional information. Spatial heterogeneity of clones implies that sampling a single region of a neoplasm may not be representative of the entire neoplasm. The force of natural selection may vary in different parts of a tissue (edge versus center, primary tumor versus metastasis) or over time (early versus late progression, before versus during or after chemotherapy). Multiple samples from a single individual also offer the possibility of phylogenetic analysis to infer relationships among different lineages and reconstruct past events in the history of the tissue. Table 1 shows a sampling of recent studies in which multiple cancer samples per patient were obtained, and phylogenetic methods were either used or could have been used. These studies considered both spatial separation—different parts of a tumor or neoplasm, a tumor and its metastases— and temporal separation— samples taken at different times, such as early and late in progression to cancer, or before and after chemotherapy. They show the potential power of the multiple-sample approach, which we expect will become increasingly important as genotyping costs decrease.

Table 1. Studies with multiple-sample data suitable for phylogenetic analysis. WGS, whole genome sequencing.

Phylogenetic analysis used			
Cases	Condition	Method	Reference
22	CLL	Targeted resequencing	[2]
1	Barrett’s esophagus	SNP arrays	[3]
2	Breast cancer	Single-cell WGS	[4]
30	ALL	FISH	[5]
4	Renal-cell carcinoma	WGS	[6]
21	Breast cancer	WGS and SNP arrays	[7]
1, 2	DS-ALL and ALL	WGS, single cell targeted mutational profiling	[8]
2	Breast cancer	Single cell WGS	[9]
13	Barrett’s esophagus	SNP arrays	[1]
23	Glioma	Exome sequencing	[10]
15	Colorectal adenoma and carcinoma	Multiple methods	[11]
Phylogenetic analysis feasible			
Cases	Condition	Method	Reference
7	Pancreatic cancer	Exome sequencing and SNP arrays	[12]
21	ALL	SNP arrays	[13]
7	MDS and sAML	WGS	[14]
8	Acute myeloid leukemia	WGS	[15]
258	Barrett’s esophagus	SNP arrays	[16]

doi:10.1371/journal.pcbi.1004413.t001

Existing methods do not have the resolution to detect all variants present in a million-cell sample. Variants present in just a few cells will go undetected. Peiffer et al. [17] found minimum frequencies of 33% to 50% for reliable detection of copy-number variants in heterogeneous tumor data using a SNP array. Deep sequencing can detect single nucleotide variants in cancer samples at lower frequencies, down to 1% [18] for 40x sequencing, but the threshold for reliable detection is much higher: even for an average of 2000x sequencing, single nucleotide variants were reproducibly detected only at >15% allele frequency and indels at >5% allele frequency [19]. Even when low-frequency variants are detected, they are often disregarded as they are difficult to quantify and assign to haplotypes. Thus, the result of bulk-biopsy genotyping is generally a survey of locally high-frequency variants only.

In a well-mixed tissue such as blood, variants which are at high frequency in one sample will generally be at high frequency in all samples. In such tissues, bulk genotyping will miss low-frequency variants, and will thus be biased toward detecting older rather than younger mutations. However, this bias will affect all samples equally and will not tend to produce spurious evidence of non-neutrality.

However, solid tissues are not well-mixed. We will consider the behavior of bulk-biopsy genotyping in a simulated tissue similar to Barrett's esophagus (BE): a sheet of tissue rolled into a cylinder, with very limited mobility of cell lineages except during initial development. While our simulations are inspired by BE, our conclusions should apply directly to neoplasms in two-dimensional epithelial sheets such as colon, skin, bladder and lung, and conceptually similar effects are also likely in three-dimensional tumors. The key factor is growth with limited mixing.

An increasingly common objective in taking multiple biopsies from a neoplastic tissues is to look for evidence of natural selection or heightened mutation acting on specific clones. This is distinct from standard methods of detecting selection or enhanced mutation via comparison of single samples from many different tumors. A straightforward statistical approach to detecting perturbing forces from multi-sample data would be to infer the evolutionary tree connecting samples from the same individual, and test if that tree conforms to a molecular clock. We simulate this experiment on data which do have a molecular clock, and show that bulk-biopsy genotyping very often leads to the spurious rejection of the clock, and thus to a conclusion of non-neutrality, even when the underlying data are completely neutral.

We emphasize that the bias we observe is not specific to the use of a phylogeny-based molecular clock test, but will influence any formal or informal comparison of apparent mutation content differences among biopsies. For example, if researchers use bulk genotyping to identify a biopsy with an unusually high number of mutations, and conclude that the highly mutant biopsy represents a genetically unstable lineage, they are implicitly assuming that bulk data have a molecular clock in the absence of perturbing forces. As we will show, this is not the case.

Methods

Model of the BE segment

We model the BE segment as a 300 x 300 grid of crypts rolled into a cylinder, approximating the size of a typical BE segment. We treat a crypt as the basic replicative unit, since genetic drift is expected to rapidly homogenize the genotype within each crypt. Crypts have a birth rate representing crypt fission. According to Totafurno's model of the crypt cycle, when stem cells double in number crypt fission is triggered, which results in halving the doubled stem cell population into two new daughter crypts [20]. In this study, we model the crypt fission cycle by allowing a crypt to either eliminate a neighboring crypt or fill in a space lacking crypts. Crypts also have a death rate, a dead crypt leaving an empty space.

The BE segment is thought to expand from the gastro-esophageal junction. An alternative possibility is that cells gain a BE phenotype and spread clonally from squamous duct glands that are situated throughout the esophagus [21]. In both scenarios, BE segments must be rapidly established, since BE segments have not been endoscopically observed in the process of expanding. However, when we simulated the esophagus beginning from a uniform field of non-mutant cells no tree structure arose even after many simulated decades. We do not show results for this case as it is trivially predictable from coalescent theory: our simulations cover approximately 50 generations, and a thoroughly mixed population of size 90,000 will have an average of 1765 distinct ancestors 50 generations ago and will thus appear as approximately 1765 unrelated patches. A less thoroughly mixed sheet of cells will be even patchier. This is not consistent with actual BE data [1] which show mutations shared among biopsies. It is possible that BE arises *in situ* and is then “overwritten” by an early selective sweep; but if so, this seems little different from the BE segment itself arising by growth from one or a few ancestral crypts. We therefore model the establishment of BE as an expansion from the gastro-esophageal junction.

Simulations

To simulate BE data we used the agent-based forward simulator of [22]. While this simulator provides for loci whose mutant alleles modify the growth or mutation rates, in the majority of experiments presented here we used a purely neutral model. We simulated 1000 neutral loci for phylogeny inference. Mutations were scored as number of changes from ancestral state; there was no back mutation. We considered neutral mutation rates per locus per crypt per year (μ) of 0.001 and 0.002. Data with the lower rate are fairly sparse, while data with the higher rate are highly polymorphic. We set the probability that a dividing crypt could displace a neighbor at 1. The crypt birth rate was 0.02 and death rate 0.001. We also did simulations with 100 neutral loci and mutation rates of 0.001, 0.002, and 0.004, presented in Supporting Information.

For illustrative purposes we also did a small number of simulations with five potentially selected loci, each having a mutation rate of 10^{-7} per locus per crypt per year and a twofold selective advantage for the mutant type over the wild type.

The mutation rates given here do not correspond directly to per-cell mutation rates since, when a mutation arises in a crypt, it may be lost rather than fixed. The per-cell mutation rate would be higher by a factor of the mean number of stem cells in the crypt. In any case our mutation rates are chosen in order to give ample mutations for phylogenetic analysis with a limited number of loci. Real data would have fewer mutations per locus, but far more loci. We do not expect this to substantially change the results.

Numerical estimates of BE crypt birth and death rates are not available. Our chosen numbers, which were roughly inspired by values measured for human colon crypts [23], produce an initial spread which is slower than in BE, and a subsequent steady state which probably has faster turnover. However, this is conservative for our conclusions: a faster initial spread and slower subsequent turnover with the same expected amount of mutation would show even greater distortion of the molecular clock. We believe that the details of our parameter choice will not affect our qualitative conclusions as long as the pattern of rapid spread followed by slow turnover is conserved.

We started with a single randomly placed crypt and simulated 20 years of growth. This was generally enough to allow crypts to fill the lower esophagus. A small proportion of simulations resulted in the death of the nascent Barrett’s epithelium; these were discarded. We then randomly chose 10 biopsies which were squares of 10x10 crypts, constrained not to overlap. Rarely, a biopsy was found to contain no live crypts; in such cases the entire simulation was discarded. Additional simulations were run to replace discarded simulations.

These simulation conditions imply a molecular clock, as the mutation rate is the same in all crypts. We tested for presence of a clock in single crypt samples and in biopsies of different sizes using PAUP* 4.0 [24]. For analytic purposes we treated all loci with one or more mutations as one state, and loci with zero mutations as another state. This corresponds to the presence/absence scoring typically used for BE data. To enable use of available phylogenetic software, these states were coded as purine and pyrimidine ambiguity codons. We tested both estimation of the state frequencies from the data using the EMPIRICAL algorithm in PAUP* (results shown in paper), and setting the frequencies equal (results shown in Supporting Information). We performed maximum-likelihood analyses of the recoded data with and without the clock constraint, and assessed the difference in log-likelihoods using a likelihood ratio test [25, 26] with a 5% significance cutoff. When multiple tied trees were produced, we used the first listed tree for analysis.

This use of the likelihood ratio test can be criticized as it assumes that the clocklike and non-clocklike best trees had the same topology [26], which was not always the case. We applied the test to all pairs of trees, even those differing in topology. Our argument is that when the topological difference is trivial (rearrangement across branches of near-zero length) the result of the test will be almost exactly the same as it would for identical topologies; and when the topological difference is non-trivial rejection of the clock is justified even though an exact statistical test is not available.

To measure the influence of biopsy size on detection of rate heterogeneity, we subsampled our biopsies. That is, to produce a 4x4 biopsy we took a 4x4 subsample from the original 10x10 biopsy. To avoid dead crypts, we examined subsamples in turn starting in the upper left and chose the first one in which at least 1 live crypt was found.

To measure the influence of detection thresholds, we used the same sets of simulated biopsies, but varied the cutoff used to establish the biopsy “genotype.” For example, when the cutoff was 30%, we scored a mutation as present if it appeared in 30% or more of the sampled living crypts from the biopsy, and absent otherwise.

In the simulations with 100 loci and $\mu = 0.001$, which had the smallest amount of information per phylogeny, a few cases with large biopsies and stringent cutoffs could not be run. Stringent cutoffs can generate biopsies with no detectable mutations, and having too many such biopsies in a single tree causes failure of the phylogeny analysis. Such runs were discarded. No more than 15/500 runs failed for any combination of conditions; the number of failed runs for each condition are given in the legends to S5 and S6 Tables.

Our simulated data is archived on Dryad at <http://dx.doi.org/10.5061/dryad.hf93c>.

Results

Our simulations were inspired by Barrett’s esophagus (BE), a neoplastic condition in which the lower esophagus is colonized by a tissue organized into crypts. We treat crypts as the fundamental unit of our simulation, and assume that all spread of genotypes results from reproduction (fission) of crypts which either replace their neighbors or spread into unoccupied areas. The details of the simulator are described in [22].

At the beginning of the simulation each crypt began with an identical genome of 100 or 1000 loci. Mutations in these loci were selectively neutral: they were used solely to infer the relationships among biopsies.

The first striking effect of bulk sampling was seen when the simulation was seeded with a completely filled grid of crypts. At the end of the simulation the tissue consisted of tiny patches of related crypts, each patch unrelated to its neighbors. This reflects the very low gene flow in a static crypt-organized tissue without natural selection. In a tissue of this kind, bulk genotyping would lead to the incorrect conclusion that there are few or no mutations present.

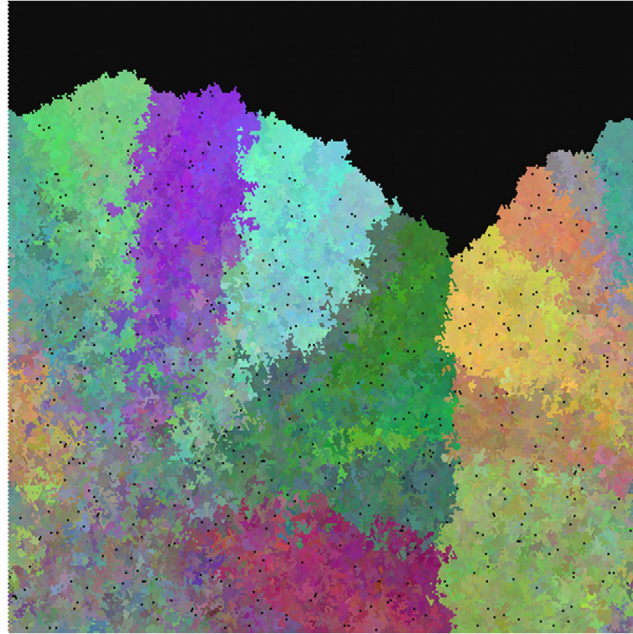


Fig 1. Simulated BE segment. Black regions represent not-yet-colonized areas of the esophagus, and small black dots represent crypts which have died and not yet been replaced. Colors indicate the first three principal components of a PCA analysis of crypt genotypes at 1000 neutral loci, essentially labeling distinct clones or cell lineages. Initial seeding of the esophagus was along the lower border of the figure. The left and right sides of the image wrap around.

doi:10.1371/journal.pcbi.1004413.g001

Bulk biopsy sampling of actual BE segments shows abundant mutations [1]. We therefore considered a theory of BE origin in which it spreads from a few crypts. We represented this by seeding the simulation with a single randomly placed crypt. Biopsies sampled from such a tissue did contain genetic variants detectable with bulk genotyping, consistent with actual BE data.

The spatial distribution of mutations in real BE segments is poorly known, as normally only a few biopsies are analyzed per individual. In our simulations we could readily examine the entire pattern, as well as taking simulated biopsies. The simulated BE segments developed a strongly sectorial pattern, with small diverse patches of cells near the original seeding area, and larger, more homogeneous patches far from it. Sharp borders between genetically distinct lineages were seen; these borders ran vertically along the simulated esophagus, roughly parallel to the direction of tissue growth. A typical example, captured partway through colonization of the simulated esophagus, is shown in Fig 1.

These patterns reflect the effect of “gene surfing” [27]. Gene surfing is a phenomenon in population genetics, observed when a population is rapidly expanding into a new geographical region but the mobility of individuals is limited. Colonization is therefore driven by a few individuals on the leading edge of the population, and their genotypes will be disproportionately represented in the newly colonized area. Patterns visually similar to our simulations can be seen when two different strains of bacteria are mixed and seeded onto a plate: sectors of pure strains are generated by replication of the few individuals on the colony edge [28] even in the absence of any selective advantage.

Our simulations, seeded with a single crypt, thus produced data that were broadly consistent with observations of actual BE segments. We next asked whether biopsies sampled from these purely neutral simulations would pass tests for neutrality. Based on ten biopsy samples from

each of 100 simulated BE segments, we inferred phylogenetic trees and tested whether those trees rejected the molecular clock at the 5% level. We considered biopsies of sizes from 1 (a single crypt) to 10x10 (100 crypts). For biopsies of size greater than 1, we also considered detection cutoffs from 10% (mutations present in 10% or more of crypts were scored) to 100% (only mutations present in all crypts were scored).

If biopsy sampling provided a phylogenetically unbiased sample of mutations occurring in our data, we would expect to see a molecular clock in our inferred trees with any size of biopsy. The proportion of inferences (out of 500) rejecting the molecular clock are shown in Figs 2 and 3 and are presented in table form in S1 and S2 Tables. In these figures, the white color seen at the left-hand edge (single-crypt samples) represents an acceptable clock rejection rate of 5%. (Note that detection cutoff does not affect the results from single crypts, and thus all of the left-hand results represent the same analyses.) All larger biopsy sizes, even 2x2 biopsies with only 4 crypts, rejected the clock at high rates for all conditions studied.

The choice of cutoff had a noticeable impact on clock rejection. Cutoffs in the 30%-50% range were better than higher or lower cutoffs; the larger the biopsy, the lower the optimal cutoff. However, no cutoff tested restored the clocklike nature of the underlying data.

Superficially satisfactory results can be obtained by using only 100 neutral loci and inferring frequencies of the mutant and non-mutant states (S3 Fig). However, this apparent improvement merely represents lack of statistical power to detect clock violations, as seen by the dramatic worsening of results with 1000 neutral loci and the same model (Fig 2).

Use of equal frequencies of mutant versus non-mutant states produces higher clock rejection: results are shown for completeness in S1 and S2 Figs for 1000 neutral loci and S4, S5 and S6 Figs for 100 loci.

We show a randomly selected pair of inferred trees from the simulation of S6 Table in Fig 4. The topologies of the two trees differed in ordering of the short bottommost branches. Larger discrepancies were seen in the branch lengths. The single-crypt tree (A) showed some heterogeneity of branch lengths, but it was well within the expected range for a data set of this size, and the clock was not rejected. The 10x10 crypt tree (B) was much more distorted, and rejected the molecular clock. It would be tempting to conclude that biopsy 10, in particular, had a higher mutation rate than biopsy 8; yet they arose from a simulation with perfectly equal rates.

Discussion

We have presented our results in terms of rejection of the molecular clock in a formal test. However, their significance is not limited to such tests. When we first examined bulk-genotyped BE data for multiple biopsies per patient we saw a striking difference in the mutation content of different biopsies. It was natural to read this as a difference in the underlying mutation rate. After further thought we realized that it could also reflect a difference in the growth rate, since rapidly growing cells will form more uniform samples and therefore appear to have more mutations. Only after performing simulations did we discover that heterogeneity in apparent mutation content is a general feature of this type of data and should be expected even when neither mutation rate nor growth rate varies. Comparison of Figs 2 and S5 shows that the more informative the data, the stronger this tendency to reject the clock. The clock test formalizes a scientist's intuition, but both the test and the intuition are liable to error in this case.

We stress that our findings do not challenge the important role non-neutral processes play in the development of cancer. Instead, they warn us against drawing conclusions about non-neutrality that cannot be supported.

Two factors combine to produce this spurious evidence for non-neutrality. Gene surfing causes biopsies taken near the origin of the growing population to be much more

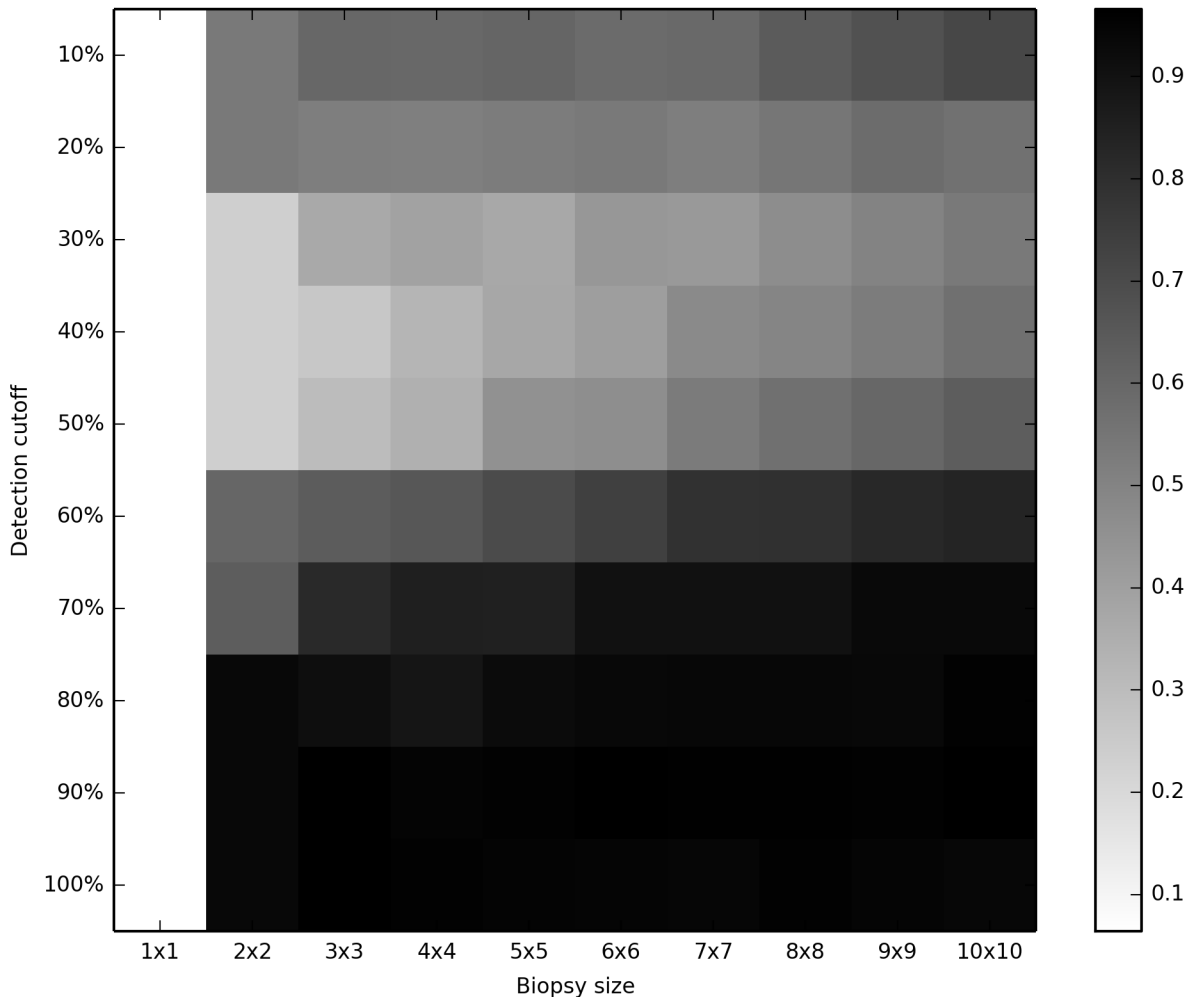


Fig 2. Rejection of the clock with 1000 neutral loci, $\mu = 0.001$. Heat-map showing the percentage of cases rejecting the molecular clock at the 5% level; lighter colors indicate lower rejection of the clock.

doi:10.1371/journal.pcbi.1004413.g002

heterogeneous than those taken far from the origin (see Fig 1). Bulk-biopsy sampling then translates this difference in diversity into a difference in detectable mutation content: a homogeneous cell sample will have more high-frequency mutations than a heterogeneous one, and bulk sampling detects only high-frequency mutations.

This is most easily understood by considering common ancestry. Consider, as an example, a detection cutoff of 50%. Mutations which reach this cutoff must exist in 50% or more of the cells in the biopsy, and thus, barring convergent evolution, must be inherited from a common ancestor of 50% of the cells. If this common ancestor existed early in the development of the tissue, it likely had relatively few mutations, so few mutations will be shared by its descendants. If it existed more recently, it likely had more mutations (since mutations accumulate over time) and its descendants will have more shared mutations. Cells from a biopsy whose common ancestor is ancient will, individually, have just as many mutations, but a much larger proportion will be at low frequency in the biopsy. Such mutations are difficult to detect with bulk genotyping.

We did not model complicating factors in analysis of bulk data such as differences in ploidy among lineages or typing errors. However, when a bias is present in analysis of clean, error-free genotypic data, there is no reason to believe that better results would come from dirtier data.

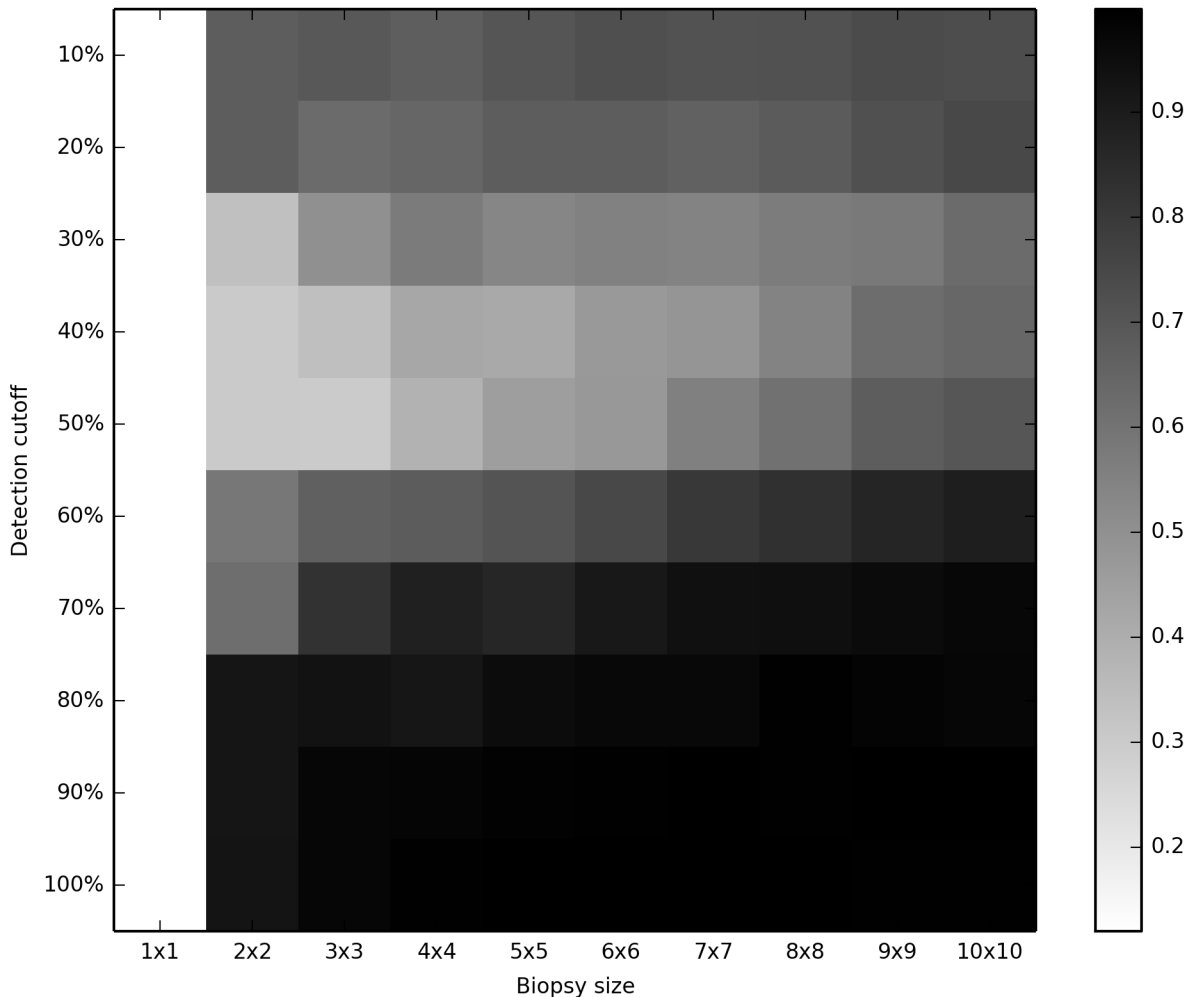


Fig 3. Rejection of the clock with 1000 neutral loci, $\mu = 0.002$. Heat-map showing the percentage of cases rejecting the molecular clock at the 5% level; lighter colors indicate lower rejection of the clock.

doi:10.1371/journal.pcbi.1004413.g003

Regrettably, no tested frequency cutoff rule was successful in resolving this problem. We suggest three possible approaches.

- (1) Examine single-cell or single-crypt samples in preference to bulk genotyping. Single-crypt samples from our simulation saw a molecular clock even though all larger samples did not. Small sample genotyping (single crypts or cells) is technically challenging but possible in some systems [9, 29]. This approach completely avoids the hazards of bulk genotyping. However, a pure small-sample genotyping strategy may be inefficient, as little information is gained about mutation frequency unless a large number of small samples are genotyped. A single small-sample genotype is probably less useful than a bulk genotype for detecting biologically critical mutations such as putative “drivers” as it will also contain many biologically irrelevant private polymorphisms.
- (2) Combine small-sample and bulk genotyping. For example, a sample could consist of a bulk genotype of the whole biopsy and individual genotypes of a few cells or crypts from it, as done for example by [11]. The single crypts provide information about the mutation content of individual crypts, while the bulk samples provide frequency information [30].

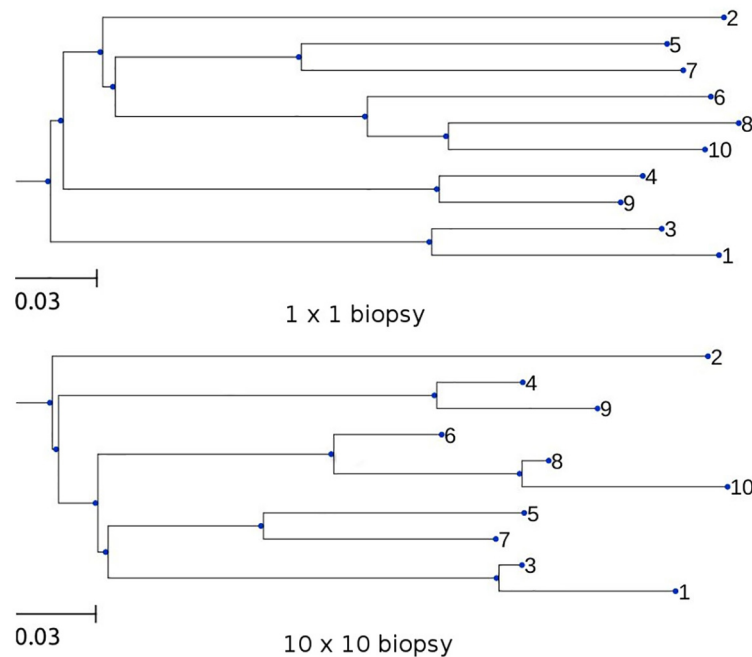


Fig 4. Single-crypt and bulk-biopsy inferred trees. Trees were inferred, without assuming a clock, from the simulation of [S6 Table](#). Upper panel, sample of size 1 crypt. Lower panel, bulk sample of size 10x10 crypts using a 40% cutoff.

doi:10.1371/journal.pcbi.1004413.g004

A biopsy in which the individual crypts show far more mutations than the bulk sample is revealed to be heterogeneous, and frequencies inferred from it can be viewed in this light.

- (3) Determine genetic diversity within a biopsy by computational means. Considerable recent work has focused on determining the number and relative frequencies of different lineages in a biopsy [31–33]. The problem is extremely challenging because different cells may vary in underlying ploidy as well as in mutation content. Progress in this area may require both computational improvements in deconvolution algorithms and experimental improvements in data collection.

Once subclones within a biopsy have been detected via approaches (2) or (3), this information needs to be incorporated into the analysis. For analytic methods involving phylogenetics, mixed samples are particularly challenging because when two variants are found at similar frequency in a sample, there is no easy way to determine whether they represent one lineage with two mutations or two lineages with one mutation each [33]. In principle it would be possible for a statistical analysis to sum over these possibilities using an approach analogous to that of [34]. The high computational burden of this approach will have to be compared with the experimental burden of small-sample typing. Alternatively, one could use the minority alleles to estimate biopsy diversity, without attempting to reconstruct minority genotypes.

One potential use of diversity estimations would be as the basis for corrected mutational distances to be used in phylogeny inference. Simulations or heuristics could be used to establish the relationship between observed mutational distance between two biopsies, internal diversity of each biopsy, and the true mutational distance. Distances corrected according to this relationship could then be used in a distance-based phylogeny algorithm to produce trees whose branch lengths more accurately represented the underlying mutational frequencies. We are currently developing such an algorithm.

Accurate inference of branch lengths is important in distinguishing, for example, a mutation rate increase in a specific lineage (presumably due to a mutator mutation or epigenetic change) from a mutation rate increase at a given time across the entire tissue (presumably due to an environmental change, since it manifests in unrelated lineages). Naive tree-drawing based on uncorrected data, as shown by the results in this paper, cannot answer such questions as the branch lengths of its trees are not proportional to time. This is shown dramatically in real BE data, where no separation is seen between data collected from time points many years apart [1]. This situation makes it difficult to draw conclusions about changes in rate over time, though in some cases coherent patterns have been detected [1]. Genetic distances corrected for the bias inherent in bulk-biopsy sampling could allow much more accurate separation of neutral from non-neutral processes in the development of tissues and cancers.

One further positive finding from this study is that the spread of a growing tissue tends to produce a characteristic fan-shaped pattern, as seen in Fig 1. As dense sampling of cancer and pre-cancer tissues becomes more feasible, it will become possible to detect this pattern or deviations from it which may indicate selection. An example is shown in Fig 5, which shows four typical results from a simulation with selected as well as neutral mutations. Note the disruption

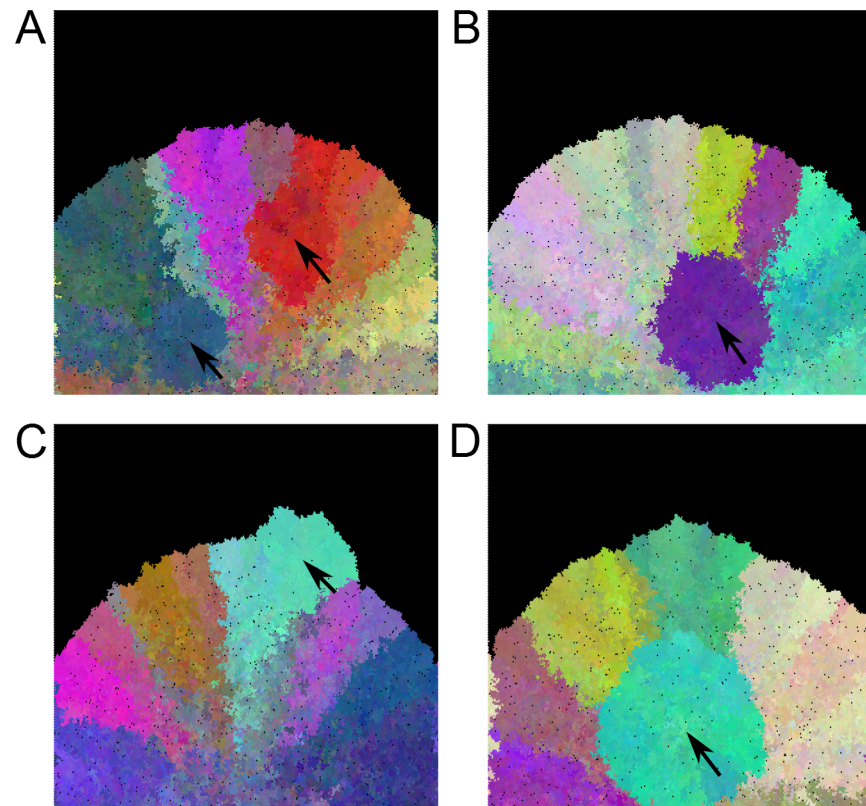


Fig 5. Simulation including a positively selected locus. Panels A-D represent the results of four independent simulations with selection. Black regions represent not-yet-colonized areas of the esophagus, and small black dots represent crypts which have died and not yet been replaced. Colors indicate the first three principal components of a PCA analysis of crypt genotypes at 100 neutral loci, essentially labeling distinct clones or cell lineages. Initial seeding of the esophagus was in the bottom-center of the figure. The left and right sides of the image wrap around. Arrows indicate clones carrying a selectively advantageous mutation. These images were chosen from a larger collection to illustrate typical selection signatures and do not represent a random sample. Selected clones appear more homogeneous and are round in shape whereas clones which expanded neutrally are wedge-shaped.

doi:10.1371/journal.pcbi.1004413.g005

of the fan pattern by lateral growth of selected clones. The spatial distribution of clones within an expanding tumor or neoplasm may therefore reveal selective processes, as has been explored by [11] in colorectal cancer. In other words, these simulations provide predictions for the nascent field of tumor phylogeography. More work is needed both to detect these patterns and to assess their significance.

Supporting Information

S1 Fig. Rejection of the clock with 1000 neutral loci, $\mu = 0.001$, equal allele frequencies.

Heat-map showing the percentage of cases rejecting the molecular clock at the 5% level; lighter colors indicate lower rejection of the clock. These data correspond to [S3 Table](#).

(TIFF)

S2 Fig. Rejection of the clock with 1000 neutral loci, $\mu = 0.002$, equal allele frequencies.

Heat-map showing the percentage of cases rejecting the molecular clock at the 5% level; lighter colors indicate lower rejection of the clock. These data correspond to [S4 Table](#).

(TIFF)

S3 Fig. Rejection of the clock with 100 neutral loci, $\mu = 0.001$, inferred allele frequencies.

Heat-map showing the percentage of cases rejecting the molecular clock at the 5% level; lighter colors indicate lower rejection of the clock. These data correspond to [S5 Table](#).

(TIFF)

S4 Fig. Rejection of the clock with 100 neutral loci, $\mu = 0.001$, equal allele frequencies.

Heat-map showing the percentage of cases rejecting the molecular clock at the 5% level; lighter colors indicate lower rejection of the clock. These data correspond to [S6 Table](#).

(TIFF)

S5 Fig. Rejection of the clock with 100 neutral loci, $\mu = 0.002$, equal allele frequencies.

Heat-map showing the percentage of cases rejecting the molecular clock at the 5% level; lighter colors indicate lower rejection of the clock. These data correspond to [S7 Table](#).

(TIFF)

S6 Fig. Rejection of the clock with 100 neutral loci, $\mu = 0.004$, equal allele frequencies.

Heat-map showing the percentage of cases rejecting the molecular clock at the 5% level; lighter colors indicate lower rejection of the clock. These data correspond to [S8 Table](#).

(TIFF)

S1 Table. Rejection of the clock with 1000 neutral loci, $\mu = 0.001$, inferred allele frequencies. μ , mutation rate per locus per generation. These data correspond to [Fig 2](#) in the main manuscript.

(PDF)

S2 Table. Rejection of the clock with 1000 neutral loci, $\mu = 0.002$, inferred allele frequencies. μ , mutation rate per locus per generation. These data correspond to [Fig 3](#) in the main manuscript.

(PDF)

S3 Table. Rejection of the clock with 1000 neutral loci, $\mu = 0.001$, equal allele frequencies. μ , mutation rate per locus per generation. These data correspond to [S1 Fig](#).

(PDF)

S4 Table. Rejection of the clock with 1000 neutral loci, $\mu = 0.002$, equal allele frequencies. μ , mutation rate per locus per generation. These data correspond to [S2 Fig](#).
(PDF)

S5 Table. Rejection of the clock with 100 neutral loci, $\mu = 0.001$, inferred allele frequencies. μ , mutation rate per locus per generation. These data correspond to [S3 Fig](#). A small proportion of runs could not be completed due to too many invariant biopsies, leading to sample sizes less than 500. For the last three entries in cutoff 90%, sample sizes were 499, 498, 497. For the last six entries in cutoff 100%, sample sizes were 497, 496, 494, 492, 492, 487.
(PDF)

S6 Table. Rejection of the clock with 100 neutral loci, $\mu = 0.001$, equal allele frequencies. μ , mutation rate per locus per generation. These data correspond to [S4 Fig](#). A small proportion of runs could not be completed due to too many invariant biopsies, leading to sample sizes less than 500. For the last three entries in cutoff 90%, sample sizes were 499, 498, 497. For the last six entries in cutoff 100%, sample sizes were 497, 496, 494, 492, 492, 487.
(PDF)

S7 Table. Rejection of the clock with 100 neutral loci, $\mu = 0.002$, equal allele frequencies. μ , mutation rate per locus per generation. These data correspond to [S5 Fig](#).
(PDF)

S8 Table. Rejection of the clock with 100 neutral loci, $\mu = 0.004$, equal allele frequencies. μ , mutation rate per locus per generation. These data correspond to [S6 Fig](#).
(PDF)

Acknowledgments

We thank Jon Yamato, Brian Reid, Carissa Sanchez, Patty Galipeau, Thomas Paulson, Ruchira Datta, Trevor Graham, and Joseph Felsenstein for helpful discussions, and Jon Yamato for assistance in preparing the paper.

Author Contributions

Conceived and designed the experiments: RK CCM MKK. Performed the experiments: RK MKK. Analyzed the data: RK MKK. Contributed reagents/materials/analysis tools: RK MKK. Wrote the paper: RK CCM MKK.

References

1. Kostadinov RL, Kuhner MK, Li X, Sanchez CA, Galipeau PC, et al. (2013) NSAIDs modulate clonal evolution in Barrett's esophagus. *PLoS Genetics* 9: e1003553. doi: [10.1371/journal.pgen.1003553](https://doi.org/10.1371/journal.pgen.1003553) PMID: [23785299](https://pubmed.ncbi.nlm.nih.gov/23785299/)
2. Campbell P, Pleasance E, Stephens P, Dicks E, Rance R, et al. (2008) Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *PNAS* 105: 13081–13086. doi: [10.1073/pnas.0801523105](https://doi.org/10.1073/pnas.0801523105) PMID: [18723673](https://pubmed.ncbi.nlm.nih.gov/18723673/)
3. Reid B, Kostadinov R, Maley C (2011) New strategies in barrett's esophagus: Integrating clonal evolutionary theory with clinical management. *Clinical cancer research* 17: 3512–3519. doi: [10.1158/1078-0432.CCR-09-2358](https://doi.org/10.1158/1078-0432.CCR-09-2358) PMID: [21498395](https://pubmed.ncbi.nlm.nih.gov/21498395/)
4. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumor evolution inferred by single-cell sequencing. *Nature*: 1–6.
5. Anderson K, Lutz C, van Delft F, Bateman C, Guo Y, et al. (2011) Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* 469: 356–361. doi: [10.1038/nature09650](https://doi.org/10.1038/nature09650) PMID: [21160474](https://pubmed.ncbi.nlm.nih.gov/21160474/)

6. Gerlinger M, Rowan A, Horswell S, Larkin J, Endesfelder D, et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England journal of medicine* 366: 883–892. doi: [10.1056/NEJMoa1113205](https://doi.org/10.1056/NEJMoa1113205) PMID: [22397650](https://pubmed.ncbi.nlm.nih.gov/22397650/)
7. Nik-Zainal S, Loo PV, Wedge D, Alexandrov L, Greenman C, et al. (2012) The life history of 21 breast cancers. *Cell* 149: 994–1007. doi: [10.1016/j.cell.2012.04.023](https://doi.org/10.1016/j.cell.2012.04.023) PMID: [22608083](https://pubmed.ncbi.nlm.nih.gov/22608083/)
8. Potter N, Ermini L, Papaemmanuil E, Cazzangia G, Vijayaraghavan G, et al. (2013) Single cell mutational profiling and clonal phylogeny in cancer. *Genome research*. doi: [10.1101/gr.159913.113](https://doi.org/10.1101/gr.159913.113) PMID: [24056532](https://pubmed.ncbi.nlm.nih.gov/24056532/)
9. Wang Y, Waters J, Leung M, Unruh A, Roh W, et al. (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512: 155–160. doi: [10.1038/nature13600](https://doi.org/10.1038/nature13600) PMID: [25079324](https://pubmed.ncbi.nlm.nih.gov/25079324/)
10. Johnson B, Mazor T, Hong C, Barnes M, Aihara K, et al. (2014) Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* 343: 189–193. doi: [10.1126/science.1239947](https://doi.org/10.1126/science.1239947) PMID: [24336570](https://pubmed.ncbi.nlm.nih.gov/24336570/)
11. Sottoriva A, Kang H, Ma Z, Graham T, Salaman M, et al. (2015) A big bang model of human colorectal cancer growth. *Nature Genetics* 47: 209–218. doi: [10.1038/ng.3214](https://doi.org/10.1038/ng.3214) PMID: [25665006](https://pubmed.ncbi.nlm.nih.gov/25665006/)
12. Yachida S, Jones S, Bozic I, Antal T, Leary R, et al. (2010) Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* 467: 1114–1117. doi: [10.1038/nature09515](https://doi.org/10.1038/nature09515) PMID: [20981102](https://pubmed.ncbi.nlm.nih.gov/20981102/)
13. van Delft F, Horsely S, Colman S, Anderson K, Bateman C, et al. (2011) Clonal origins of relapse in etv6-runx1 acute lymphoblastic leukemia. *Blood* 117: 6247–6254. doi: [10.1182/blood-2010-10-314674](https://doi.org/10.1182/blood-2010-10-314674) PMID: [21482711](https://pubmed.ncbi.nlm.nih.gov/21482711/)
14. Walter M, Shen D, Ding L, Shao J, Koboldt D, et al. (2012) Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med* 366: 1090–1098. doi: [10.1056/NEJMoa1106968](https://doi.org/10.1056/NEJMoa1106968) PMID: [22417201](https://pubmed.ncbi.nlm.nih.gov/22417201/)
15. Ding L, Ley T, Larson D, Miller C, Koboldt D, et al. (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481: 506–510. doi: [10.1038/nature10738](https://doi.org/10.1038/nature10738) PMID: [22237025](https://pubmed.ncbi.nlm.nih.gov/22237025/)
16. Li X, G P, Paulson T, Sanchez C, Arnaudo J, Liu K, et al. (2014) Temporal and spatial evolution of somatic chromosomal alterations: A case-cohort study of barrett's esophagus. *Cancer Prev Res*: 114–127.
17. Peiffer D, Le J, Steemers F, Chang W, Jenniges T, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 16: 1136–1148. doi: [10.1101/gr.5402306](https://doi.org/10.1101/gr.5402306) PMID: [16899659](https://pubmed.ncbi.nlm.nih.gov/16899659/)
18. Shah S, Morin R, Khattri J, Prentice L, Pugh T, et al. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* 461: 809–813. doi: [10.1038/nature08489](https://doi.org/10.1038/nature08489) PMID: [19812674](https://pubmed.ncbi.nlm.nih.gov/19812674/)
19. Cheng D, Cheng J, Mitchell T, Syed A, Zehir A, et al. (2014) Detection of mutations in myeloid malignancies through paired-sample analysis of microdroplet-pcr deep sequencing data. *J Mol Diagn* 16: 504–518. doi: [10.1016/j.jmoldx.2014.05.006](https://doi.org/10.1016/j.jmoldx.2014.05.006) PMID: [25017477](https://pubmed.ncbi.nlm.nih.gov/25017477/)
20. Totafurno J, Bjerknes M, Cheng H (1987) The crypt cycle. Crypt and villus production in the adult intestinal epithelium. *Biophysical Journal* 52: 279–294. doi: [10.1016/S0006-3495\(87\)83215-0](https://doi.org/10.1016/S0006-3495(87)83215-0) PMID: [3663832](https://pubmed.ncbi.nlm.nih.gov/3663832/)
21. Leedham S, Preston S, McDonald S, Elia G, Bhandari P, et al. (2008) Individual crypt genetic heterogeneity and the origin of metaplastic glandular epithelium in human barrett's oesophagus. *Gut* 57: 1041–1048. doi: [10.1136/gut.2007.143339](https://doi.org/10.1136/gut.2007.143339) PMID: [18305067](https://pubmed.ncbi.nlm.nih.gov/18305067/)
22. Kostadinov R (2012) Evolutionary dynamics of neoplastic cell populations in Barrett's esophagus. Dissertation, University of Pennsylvania, Philadelphia. URL <http://repository.upenn.edu/dissertations/AAI3509163>.
23. Cheng H, Matthew Bjerknes, Jack Amar, Geoffrey Gardiner (1986) Crypt production in normal and diseased human colonic epithelium. *The Anatomical Record* 216: 44–48. doi: [10.1002/ar.1092160108](https://doi.org/10.1002/ar.1092160108) PMID: [3094402](https://pubmed.ncbi.nlm.nih.gov/3094402/)
24. Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0. Sinauer Associates, Sunderland, Massachusetts.
25. Felsenstein J (1981) Evolutionary trees from dna sequences: a maximum likelihood approach. *J Molecular Evolution* 17: 368–376. doi: [10.1007/BF01734359](https://doi.org/10.1007/BF01734359)
26. Felsenstein J (1988) Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* 22: 521–565. doi: [10.1146/annurev.ge.22.120188.002513](https://doi.org/10.1146/annurev.ge.22.120188.002513) PMID: [3071258](https://pubmed.ncbi.nlm.nih.gov/3071258/)
27. Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions. *Trends Ecol Evol* 23: 347–351. doi: [10.1016/j.tree.2008.04.004](https://doi.org/10.1016/j.tree.2008.04.004) PMID: [18502536](https://pubmed.ncbi.nlm.nih.gov/18502536/)

28. Korolev K, Müller M, Karahan N, Murray A, Hallatschek O, et al. (2012) Selective sweeps in growing microbial colonies. *Phys Biol* 9: 026008. doi: [10.1088/1478-3975/9/2/026008](https://doi.org/10.1088/1478-3975/9/2/026008) PMID: [22476106](https://pubmed.ncbi.nlm.nih.gov/22476106/)
29. Xu X, Hou Y, Yin X, Bao L, Tang A, et al. (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 148: 886–895. doi: [10.1016/j.cell.2012.02.025](https://doi.org/10.1016/j.cell.2012.02.025) PMID: [22385958](https://pubmed.ncbi.nlm.nih.gov/22385958/)
30. Eirew P, Steif A, Khattra J, Ha G, Yap D, et al. (2014) Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*. doi: [10.1038/nature13952](https://doi.org/10.1038/nature13952) PMID: [25470049](https://pubmed.ncbi.nlm.nih.gov/25470049/)
31. Roth A, Khattra J, Yap D, Wan A, Laks E, et al. (2014) Pyclone: statistical inference of clonal population structure in cancer. *Nature Methods* 11: 396–398. doi: [10.1038/nmeth.2883](https://doi.org/10.1038/nmeth.2883) PMID: [24633410](https://pubmed.ncbi.nlm.nih.gov/24633410/)
32. Ha G, Roth A, Khattra J, Ho J, Yap D, et al. (2013) Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* 24: 1881–1893. doi: [10.1101/gr.180281.114](https://doi.org/10.1101/gr.180281.114)
33. Andor N, Harness JV, Muller S, Mewes HW, Petritsch C (2014) Expands: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* 30: 50–60. doi: [10.1093/bioinformatics/btt622](https://doi.org/10.1093/bioinformatics/btt622) PMID: [24177718](https://pubmed.ncbi.nlm.nih.gov/24177718/)
34. Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156: 439–447. PMID: [10978306](https://pubmed.ncbi.nlm.nih.gov/10978306/)