Journal of Theoretical Biology ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Ancestral inference in tumors: How much can we know?

Q1 Junsong Zhao [a], Kimberly D. Siegmund [b], Darryl Shibata [c], Paul Marjoram [a,b,*]

[a] Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[b] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA
[c] Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA

HIGHLIGHTS

- The ability to retrieve ancestral information depends on the parameter being inferred.
- Ancestral state and other parameters are more accurately estimated for younger tumors.
- Methylation/demethylation rate ratio can be estimated in tumors in stationary phase.
- Number of cancer stem cells can be inferred in most tumors and varies significantly.

ARTICLE INFO

Article history:
Received 30 October 2013
Received in revised form
22 February 2014
Accepted 20 May 2014

Q2 *Keywords:*
Ancestry
Approximate Bayesian computation
Methylation
Phylogeny
Methylation error rate
Number of cancer stem cells

ABSTRACT

A tumor is thought to start from a single cell and genome. Yet genomes in the final tumor are typically heterogeneous. The mystery of this intratumoral heterogeneity (ITH) has not yet been uncovered, but much of this ITH may be secondary to replication errors. Methylation of cytosine bases often exhibits ITH and therefore may encode the ancestry of the tumor. In this study, we measure the passenger methylation patterns of a specific CpG region in 9 colorectal tumors by bisulfite sequencing and apply a tumor development model. Based on our model, we are able to retrieve information regarding the ancestry of each tumor using approximate Bayesian computation. With a large simulation study we explore the conditions under which we can estimate the model parameters, and the initial state of the first transformed cell. Finally we apply our analysis to clinical data to gain insight into the dynamics of tumor formation.

1. Introduction

The mechanisms by which tumors grow remain poorly understood. Various models have been proposed to study tumor initiation, growth and progression. An early study (Laird, 1964) showed that the Gompertzian model fitted experimental data remarkably, although later research indicated that a Gompertzian model will fail when the tumor is small or when the interaction between the tumor and the host immune system is included in the model (d'Onofrio, 2005). Tumor growth can also be modeled by partial differential equations and mixture theory (Ambrosi and Preziosi, 2002; Byrne and Preziosi, 2003) with an emphasis on mass build-up and the geometry of the tumor. Some later tumor models (Anderson et al., 2008; Klein and Hölzel, 2006) focus on single-cell level behavior. Technologic advances such as single-cell tumor sequencing (Navin et al., 2011) will increasingly provide more experimental data for inferring tumor population structure.

Fitting models of tumor growth is problematic because we do not typically observe that growth. Rather, we observe an end point of that growth. Furthermore, we are not able to observe the clonal expansion of a single cell that is thought to initiate tumor growth (Hong et al., 2010; Siegmund et al., 2009). Since the parameters of tumor growth, or state of initial single cell before clonal expansion, might contain important prognostic flags for future tumor behavior, it is vital to explore how well they might be inferred from data collected from the final tumor. In this paper we explore this issue using approximate Bayesian computation (ABC), a method that allows principled analysis in contexts such as ours where models are of sufficient complexity to make more traditional analysis methods intractable.

The key intuition that we exploit is that ancestry can be inferred from the variation between genomes (cf., inference of mtEVE, or Y-chromosome Adam, from human genotype data (Marjoram and Donnelly, 1997; Pritchard et al., 1999)). The greater the differences between genomes, on average the greater the time

* Corresponding author at: Department of Molecular and Computational Biology, Q3 University of Southern California, Los Angeles, CA 90089, USA. Tel.: +1 323 442 0111.
E-mail addresses: junsongz@usc.edu (J. Zhao), kims@usc.edu (K.D. Siegmund), dshibata@usc.edu (D. Shibata), pmarjora@usc.edu (P. Marjoram).

http://dx.doi.org/10.1016/j.jtbi.2014.05.027
0022-5193/

since a common ancestor (the *molecular clock* hypothesis (Bromham and Penny, 2003)). Molecular phylogeny is usually employed to reconstruct the pasts of macroscopic populations such as individuals or species, but it can also be used to infer the fates of somatic cells within an individual. Accurate inference of somatic cell phylogenies would be extremely valuable, especially for human tissues, because more direct experimental observations are often impractical. However, a problem with comparing somatic cell genomes within an individual is that few somatic mutations are expected to accumulate within a lifetime (Shibata and Lieber, 2010). To overcome this practical shortcoming, recent studies have employed epigenetic measurements such as DNA methylation patterns DNA methylation is a covalent modification at CpG dinucleotides that is also copied after DNA replication. However, unlike base replication, epigenetic replication fidelity is markedly lower at certain CpG rich regions. Therefore, DNA methylation patterns measurably change during normal human aging and are often highly polymorphic within an individual (Shibata, 2009). Consequently, the 5′ to 3′ order of DNA methylation can be used to infer the history of a tumor in a way that is directly analogous to the use of nucleotide variation to infer history of individuals (Shibata and Tavaré, 2006).

DNA methylation patterns at non-expressed CpG rich regions ("passenger methylation") have been used to reconstruct the past of human tissues such as colon crypts and tumors (Yatabe et al., 2001). However, it is uncertain with how much precision the pasts of somatic cells can be inferred from methylation patterns. Complicating factors include uncertainties imposed by rapid replication errors, stepwise changes (both methylation and demethylation are possible), and possible variations in error rates between neighboring CpG sites that may depend on the methylation status of neighboring sites. Potentially, certain aspects of ancestry are more recoverable from passenger methylation patterns.

Specifically for human tumorigenesis, simple unknowns are the ancestral state of the first tumor cell, how fast a tumor grows, and its mitotic age (numbers of divisions between the first tumor cell and tumor removal). To further explore the utility of passenger methylation patterns for the reconstruction of human tumorigenesis, we simulate data under a variety of tumor growth models, and evaluate our ability to estimate parameters capturing tumor growth behavior, extending earlier work (Hong et al., 2010; Siegmund et al., 2009) in which we focused on estimation of three parameters: the total number of cell divisions (tumor age), the number of cancer stem cells per gland, and the probability of asymmetric stem cell division.

2. Data, model and methods

2.1. Experimental data and model

We applied our analysis methodology to a data set that consists of information from 9 colorectal tumors. The methylation patterns of a short CpG-rich region (LOC, 14 CpG sites) were measured using bisulfite sequencing. We sampled eight cells per gland, and eight glands per half, in each tumor.

We model actual physical tumor growth, beginning with the clonal expansion of a single cell (Hong et al., 2010; Siegmund et al., 2009), applying a biological constraint on the total number of tumor cells (e.g. assuming 1 billion cells/1 cm^3), and making use of clinical data on tumor size to inform our model. Tumors arising from glandular tissues such as the colon, with cells organized into small tubular units, are typically adenocarcinomas which are composed of many neoplastic glands. Adenocarcinomas are also common in the breast, prostate, lung, pancreas, and stomach. As such, dividing cancer cells in our model are geographically confined to cancer glands, which also divide, with constraints on the total number of cells based on the size of the tumor (see Fig. 1). Our model directly reflects this glandular structure.

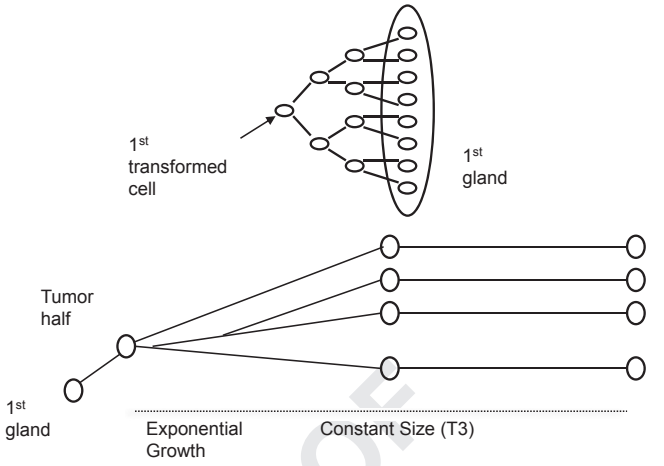

Fig. 1. The tumor growth model. Top graph shows the division of the 1st transformed cell into a gland. The bottom graph shows the exponential growth and the constant-size growth of the glands in one tumor half. See text for more details.

A tumor is simulated as the clonal expansion of a single transformed cell. A 4 cm^3 tumor contains approximately 4 billion cells, which is impossible to simulate at the single-cell level by forward simulation. However, the organization of tumor cells within glands allows for a flexible growth modeling across two different scales, cell level and gland level. Since one gland contains approximately 8000 cells, a 4 cm^3 tumor can be approximated by only 500,000 glands. This size is achieved after only 19 generations of exponential growth. We mimic the structure of our sampled data by sampling only eight glands from the ~500,000, and storing their ancestral tree. This is followed by the simulation of single-cells along the ancestral tree for the sampled glands. This approach allows us to simulate for each tumor a sample of ~33 K cells (=4096 cells/gland × 8 sampled glands) instead of a total of ~4 billion. This ensures computational tractability.

The cells and glands follow separate models for growth. We model gland growth as exponential growth followed by a period of constant size (see Fig. 1) At the cell-level, the single transformed cell undergoes exponential growth (cell doubling) until it attains the number required of the first cancer gland (see Fig. 1). In subsequent generations, the cells in the gland divide until they double in number, and then the gland divides. Both the cells and glands continue to divide, forming a second period of exponential growth (phase one for gland tree growth), until the tumor reaches its fixed biological size. The tumor then enters the second phase of the gland tree growth, in which the gland number remains constant, but the cells within glands divide and die, allowing for continued 'aging' in a tumor of fixed size (no growth). Cell division and death occurs via symmetric and asymmetric division. We refer to long-lived dividing cells lines as cancer stem cell lines. The model for cancer stem cell division is as follows. Under asymmetric division, a cancer stem cell differentiates into one cancer stem cell and one normal cancer cell, while under symmetric division, a cancer stem cell have 0.5 probability to give birth to two cancer stem cells and 0.5 probability to divide into two normal cancer cells. This is parameterized by probability of asymmetric division (PAD) that controls the proportion of cancer stem cells having asymmetric division. Finally, the DNA methylation patterns are sampled from approximately 16 glands per tumor, eight per tumor half. For a detailed mathematical description of the model, see (Siegmund et al., 2009), in which the same parameterization is used.

Please cite this article as: Zhao, J., et al., Ancestral inference in tumors: How much can we know? J. Theor. Biol. (2014), http://dx.doi.org/10.1016/j.jtbi.2014.05.027