A PROBABILISTIC MODEL FOR RECONSTRUCTING INTRA-TUMOR PHYLOGENIES

NB-LAB + FM-LAB

Abstract. We present an intra-tumor oncogenetic tree (ITOT) model for inferring tumor evolutionary histories from NGS data obtained from bulk sequencing of mixed tumor samples. Our methods not only infer the number of clones per sample but also the most likely life history of the tumor. Our methods allow us to (i) infer early events in tumour development, (ii) link cancer heterogeneity to clinical outcome, (iii) compare clonal evolution in metastasis to evolution in the primary tumour.

1. Introduction

Cancers are heterogenous, the usual bla bla Aparicio and Caldas (2013); Nik-Zainal et al. (2012a,b); Shah et al. (2009)

Progress has been made on clustering sequencing data into clonal subpopulations (Shah et al. (2012), Ben Raphael's TheTA, Perou+Mardis paper soon), but the evolutionary relationships between these clones, the so called life history of a cancer, has so far only been estimated by eye (Nik-Zainal et al., 2012b).

To change this sad situation, we need rigorous and accurate phylogenetic methods to automate the inference of the evolutionary history of a tumor. These methods will enable us to infer early driver events on a large scale, test whether evolutionary trajectories are predictive of outcome, and compare clonal evolution between pimary and metastatic tumors. Better phylogenies will improve the usefulness of all cancer heterogeneity studies, which so far are limited to enumerating longer and longer lists of mutations and aberrations. Our work extends methods proposed in Schwarz and many others (2013a,b) from multiple-sampled copy-number profiles to SNVs from deep-sequencing data.

We assume that the mixed tumor cell population has been bulk sequenced, such that the resulting reads provide a statistical sample of the underlying population. Ignoring structural variation and copy number alterations, we aim at inferring a phylogenetic tree model that represents the evolutionary history of the tumor clones from the observed short read data focusing only on single-nucleotide variations (SNVs). This task is different from classical phylogenetic approaches in at least three ways: (i) The observed data are short reads covering only a small fraction of the cancer genotypes that define the subclones. (ii) Due to the bulk sequencing approach of mixed samples, it is unknown which read originated from which cancer genotype. (iii) Read counts allow for inference of relative frequencies of mutations, which are informative about the order in which mutations have occurred.

The starting point of our method development will be existing methods for estimating clonal populations. A prominent example is pyclone (Shah et al., 2012), which corrects mutation frequencies for copy number alterations and loss of heterozygosity and clusters the frequency distribution using a Dirichlet process mixture model. The outcome of this analysis is the number of clones and their frequencies; what is still unknown is the evolutionary history of the clones and previous approaches have relied on visual analysis to place clones in a phylogenetic tree (Nik-Zainal et al., 2012b).

Date: July 10, 2013.

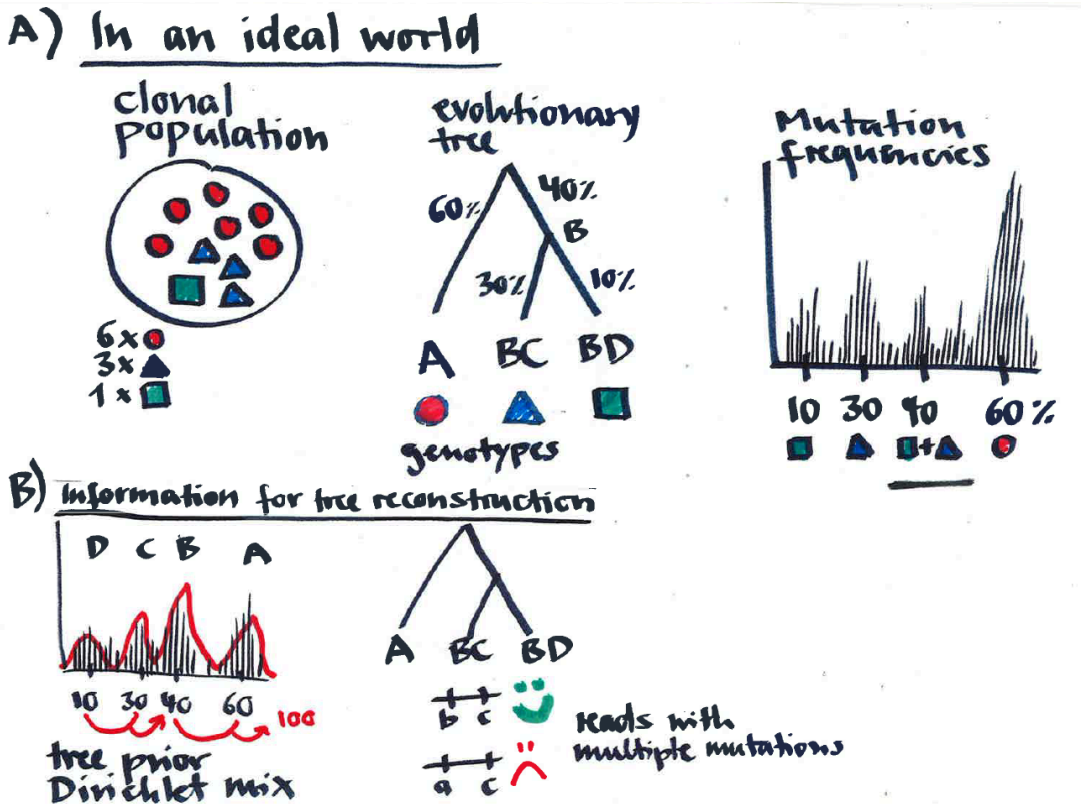

FIGURE 1. **A)** Overview how heterogeneity in a tumor relates to clonal evolution and is reflected in mutation frequencies. **B)** The sources of information we have available for inference.

We will automate this process by implementing the key ideas of Nik-Zainal et al. (2012b) in a rigorous statistical model. The central insight is that the way clonal frequencies add up hints at their evolutionary history. Figure 1A illustrates this idea in an idealized scenario by showing how a population of three clones (with genotypes A, BC, BD) results in a mutation frequency distribution with four peaks (centred around the frequencies of A,B,C,D). The tree structure is reflected in the frequency distribution: frequencies of A and B add up to 100% and frequencies of C and D add up to the frequency of B, their ancestor in the tree. However, in noisy data these relationships can be blurred, which makes tree construction difficult. In particular, the clustering could result in spurious clusters or misplaced cluster centers that cannot be fit in a tree. This is the reason why in the manual reconstruction in Nik-Zainal et al. (2012b) one of the big clusters had to be split by hand into three smaller ones that were attached to different parts of the tree.

might be too confined

To address these problems, we propose the following approach. We will start by calling mutations from deep-sequencing data and correcting their frequencies for copy-number and loss of heterozygosity in the same way as pyClone. To infer clonal evolutionary trees we will use two complementary pieces of information (Figure 1B): First, we will constrain the clustering of mutation frequencies to ensure that clonal frequencies are tree-like. Instead of sequentially clustering and tree-building we will combine both steps into a single model, following ideas of Adams et al. (2010); Williams (1999). As a result our mixture model will automatically (1) estimate the number of clones and (2) place them in an evolutionary tree.

A second source of information to validate this tree are the reads that contain more than one mutation. In the example of Figure 2 reads with mutations from both B and C would agree with the tree, but all reads with mutations from A and C would be evidence against it. Our final model will integrate the tree-constrained clustering with the information in multiply mutated reads into a joint likelihood to estimate the most likely tree topology.

2. Our goals

Our method will allow us to investigate three pivotal questions:

1. In which order did genomic events occur during tumor development? Earlier driver events will be more important drug targets than newer ones. To time events, our first goal will be to automate the phylogenetic inference of clonal evolution, as exemplified in Nik-Zainal et al (2012).

We will validate our approach using simulations of tumour evolution driven by basic biological principles, following the ideas developed in Schwarz and many others (2013a).

This project will be our main focus in the beginning. Once the model is established, the next project is basically for free (except for the blood, sweat and tears of applying it).

Data: We will use the data of Nik-Zainal et al. (2012a) to compare our automatically reconstructed trees to their visual analysis.

Are there a other clona trees out there that could comp against?

2. How predictive are life histories of tumours for clinical endpoints? In esophageal adenocarcinoma measures of clonal heterogeneity (entropy: number of clones weighted by their frequency) predict disease progression (Maley et al., 2006). This is a landmark study, because the link between heterogeneity and clinical variables is still widely unexplored. Maley et al. (2006) identified clones by any difference in flow cytometric DNA content (for differences 40.2N), LOH, microsatellite shifts (new alleles) and CDKN2A or TP53 sequence mutations.

We will extend the analysis of Maley et al. (2006) to deep sequencing data. Using the trees from our method we will compute summary statistics that measure genomic heterogeneity of the tumor and quantify features of its evolutionary history. The simplest summary statistics (and baseline of our analysis) are the number of clones and entropy of their distribution (Maley et al., 2006), which do not take any features of the tree

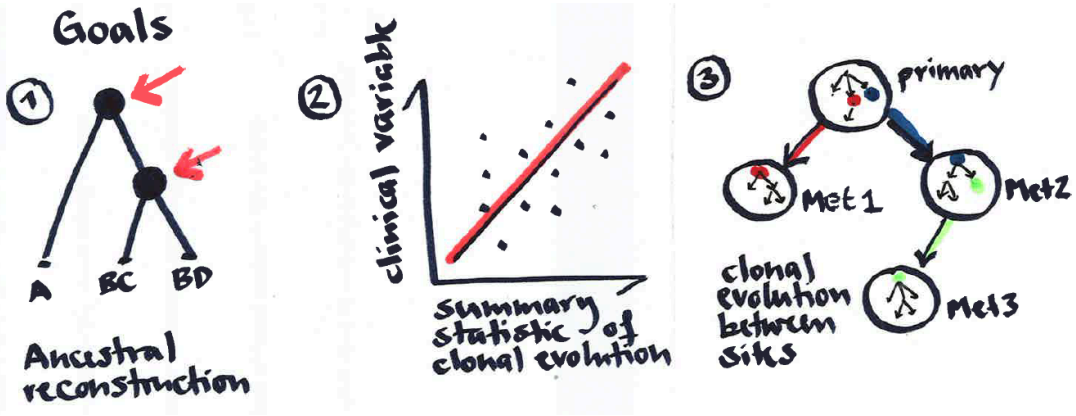

Figure 2. **Goals of our analysis are** (i) to infer early events in tumour development, (ii) to link cancer heterogeneity to clinical outcome, (iii) to compare clonal evolution in metastasis to evolution in the primary tumour.

an already
conpute
l distribu-
+ entropy
ll theses
oles (as
et al have
already)
correlate
outcome.

re do we
his?

ect all
!

into account. Our hypothesis is that measures of heterogeneity, which explicitly rely on features of the tree structure, are more informative than the number of clones.

Data: We will apply our method to the 65 triple-negative breast cancers sequenced in Shah et al. (2012), for which clinical data is available. We will also need a validation set of at least equal size.We can use all available data from all cancers. Being able to compare the predictive power of trees between cancer types would make us stronger.

3. How does clonal evolution in the patient span primary tumour and metastases? In a (so far) unpublished project Elaine Mardis and Chuck Perou work on data of sequencing samples of a primary breast tumour and five metastases at different anatomical sites. They clustered sequencing data into clones individually for each sample and ordered the samples into a tree using Phylip on the average genomic profile.

We will algorithmically improve this analysis by developing methods for joint inference of sample-specific trees and a global tree, which together will show how clones evolve within a tumor and spread to other anatomical sites to start a metastasis. This approach will borrow information across samples to identify small clonal subpopulations that might be missed if samples are treated independently.

The goal is to infer clonal evolution in the patient, spanning different anatomic sites and showing how clones move from one site to the next and populate a metastases.

Data: These data are not yet published, but Mardis is giving talks – can't be long. It would be good to have a basic method ready when the data come out.

3. How to get started

- First step: during Thomas' visit we will start merging our methodological ideas (see below) and derive a battle plan for method development. DONE!
- Repository for code, manuscripts and data(?). Any objections to using bitbucket? DONE!
- Data collection: these genetic data take a while to collect because of all the bureaucratic hoops to jump through. We need to start now basically even if we don't plan to touch them for months.
- Regular meetings. Every two weeks? MOSTLY EMAIL. SKYPE MEETINGS WHEN MILESTONES ARE REACHED. NEXT MILESTONE: RECONSRUCT NIK-ZAINAL TREE.

4. Methods

We assume the sequencing data is restricted to loci with clonal monosomies. Furthermore, each read is overlapping with two SNV positions, but does not necessarily contain the two or one SNV. Considering the Bayesian nonparametric approach introduced in Adams et al. (2010) we construct two (alternative) models:

4.1. **Pairing models.**

Factored Bernoulli likelihood-based approach. The data consists of two-dimensional binary vectors indicating the statuses of the SNVs covering this read and two genomic positions indicating the positions of those two SNVs: $X_n \in (\{0,1\}^2, position1, position2)$

The node wise likelihood is parametrized by a latent two-dimensional real vector which is then component-wise logit transformed: $\theta_\epsilon \in \mathbb{R}^N$

N is the length of the genomic loci of interest or the cardinality of $\mathcal{S}$.

$$f(x_n \mid \theta_\epsilon) = (1 + exp(-\theta_\epsilon)^{(position1)})^{-x_n^{(1)}}) * (1 + exp(\theta_\epsilon)^{(position1)})^{1-x_n^{(1)}}) * \tag{1}$$

$$(1 + exp(-\theta_\epsilon)^{(position2)})^{-x_n^{(2)}}) * (1 + exp(\theta_\epsilon)^{(position2)})^{1-x_n^{(2)}}) * \tag{2}$$

$$\frac{1}{N * (N-1)} \tag{3}$$

The transition kernel consists of a product of independent Gamma distributions:

$$T(\theta_{\epsilon\epsilon_i} \leftarrow \theta_\epsilon) = \prod_{j=1}^{N} Gamma(\theta_{\epsilon\epsilon_i}^{(j)} - \theta_\epsilon^{(j)} \mid \alpha, \beta) \tag{4}$$

The root level prior:
$p(\theta_0)$ =distribution that fits the clonal SNVs of interest
Discrete SNV-pool-based approach. $\mathcal{S}$ contains all identified clonal and subclonal SNVs in the loci of interest.

ε is the global error parameter.

The data is structure similar to the Factored Bernoulli likelihood-based approach: $X_n \in (\{0,1\}^2, position1, position2)$

The node wise likelihood is parametrized by a latent SNV set containing a subset of $\mathcal{S}$: $\theta_\epsilon \in \mathcal{S}$

The node wise data likelihood is define as:

$$f(x_n \mid \theta_\epsilon) = \begin{cases} \varepsilon^2 & \text{if } |\{s,t\} \cap \theta_\epsilon| = 0 \\ 2\varepsilon(1-\varepsilon) & \text{if } |\{s,t\} \cap \theta_\epsilon| = 1 \\ (1-\varepsilon)^2 & \text{if } |\{s,t\} \cap \theta_\epsilon| = 2 \end{cases} \tag{5}$$

The transition kernel adds a random subset of $\mathcal{S}$ without the parent SNVs θ_ϵ and adds them to the child SNV set $\theta_{\epsilon\epsilon_i}$:

$$T(\theta_{\epsilon\epsilon_i} \leftarrow \theta_\epsilon) = MUT(\theta_{\epsilon\epsilon_i} \mid \theta_\epsilon, \mathcal{S}) \tag{6}$$

MUT: a random number of elements of the set $\mathcal{S}$
θ_ϵ is chosen. PMF: combinatorial
The root level prior: $p(\theta_0)$ =distribution that fits the clonal SNVs of interest

4.2. **The frequency models.** The frequency data could use either Binomial (including PyClone) or Gaussian variables.

For the Binomial model, the data takes the following format, $X_n \in \{y_n, q_n \in \mathbb{N}, s \in \mathcal{S}\}$. The likelihood is:

$$f(x_n|\theta_\epsilon) = \text{Binomial}(y_n|q_n, \theta_\epsilon) \tag{7}$$

The cellularity can be used to scale θ_ϵ. For PyClone, the parameter θ_ϵ becomes a function of copy number state and others.

For the Gaussian model, the data becomes $X_n \in \{x_n \in [0,1], s \in \mathcal{S}\}$. The likelihood is:

$$f(x_n|\theta_\epsilon) = \text{Normal}(x_n|\theta_\epsilon^{(1)}, \theta_\epsilon^{(2)}) \tag{8}$$

Similarly, the cellularity can also be use to scale θ_ϵ.

Priors. The frequency models share the same transition kernel is the same as Eq.

$$T(\theta_{\epsilon\epsilon_i} \leftarrow \theta_\epsilon) = \text{Beta}\left(\frac{\theta_{\epsilon\epsilon_i}}{\theta_\epsilon}\Big|\alpha, \beta\right) \tag{9}$$

The root level prior

$$P(\theta_0) = \text{Normal}(\theta_0|\mu, \sigma^2) \tag{10}$$

where, μ is close to 1, σ^2 is small.

4.3. **Construct a prior from the frequency model for the pairing model.** Form the results from a frequency model, based on the MAP solution, we have a label c_n for each SNV. Let the depth of the node in pairing model tree be $d = |\epsilon|$, we can get all the SNVs which are at corresponding location in the frequency model tree. $n \in \{|c_n| = \{d, d \pm 1\}\}$

For the transition kernel,

$$T(\theta_{\epsilon\epsilon_i} \leftarrow \theta_\epsilon) = \prod_{j=1}^{N} \text{Gamma}(\theta^{(j)}_{\epsilon\epsilon_i} - \theta^{(j)}_{\epsilon} \mid \alpha_j, \beta_j) \tag{11}$$

where

$$\alpha_j, \beta_j = \begin{cases} 1.5, 0.05 & \text{if } j \in \{|c_j| = d - 1\} \\ 1.5, 0.1 & \text{if } j \in \{|c_j| = d\} \\ 1.5, 0, 05 & \text{if } j \in \{|c_j| = d + 1\} \\ 1, 0.05 & \text{otherwise} \end{cases} \tag{12}$$

$p(\theta_0)$ =distribution that fits the clonal SNVs of interest

5. Results

6. Discussion

References

R. P. Adams, Z. Ghahramani, and M. I. Jordan. Tree-structured stick breaking processes for hierarchical data. June 2010.

S. Aparicio and C. Caldas. The implications of clonal genome evolution for cancer medicine. *N Engl J Med*, 368(9):842–851, Feb 2013. doi: 10.1056/NEJMra1204892. URL `http://dx.doi.org/10.1056/NEJMra1204892`.

C. C. Maley, P. C. Galipeau, J. C. Finley, V. J. Wongsurawat, X. Li, C. A. Sanchez, T. G. Paulson, P. L. Blount, R.-A. Risques, P. S. Rabinovitch, and B. J. Reid. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet*, 38 (4):468–473, Apr 2006. doi: 10.1038/ng1768. URL `http://dx.doi.org/10.1038/ng1768`.

S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, A. Menzies, S. Martin, K. Leung, L. Chen, C. Leroy, M. Ramakrishna, R. Rance, K. W. Lau, L. J. Mudie, I. Varela, D. J. McBride, G. R. Bignell, S. L. Cooke, A. Shlien, J. Gamble, I. Whitmore, M. Maddison, P. S. Tarpey, H. R. Davies, E. Papaemmanuil, P. J. Stephens, S. McLaren, A. P. Butler, J. W. Teague, G. Jnsson, J. E. Garber, D. Silver, P. Miron, A. Fatima, S. Boyault, A. Langerd, A. Tutt, J. W. M. Martens, S. A. J. R. Aparicio, . Borg, A. V. Salomon, G. Thomas, A.-L. Brresen-Dale, A. L. Richardson, M. S. Neuberger, P. A. Futreal, P. J. Campbell, M. R. Stratton, and B. C. W. G. o. t. I.

C. G. C. . Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, May 2012a.

S. Nik-Zainal, P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, A. Shlien, S. L. Cooke, J. Hinton, A. Menzies, L. A. Stebbings, C. Leroy, M. Jia, R. Rance, L. J. Mudie, S. J. Gamble, P. J. Stephens, S. McLaren, P. S. Tarpey, E. Papaemmanuil, H. R. Davies, I. Varela, D. J. McBride, G. R. Bignell, K. Leung, A. P. Butler, J. W. Teague, S. Martin, G. Jnsson, O. Mariani, S. Boyault, P. Miron, A. Fatima, A. Langerd, S. A. J. R. Aparicio, A. Tutt, A. M. Sieuwerts, . Borg, G. Thomas, A. V. Salomon, A. L. Richardson, A.-L. Brresen-Dale, P. A. Futreal, M. R. Stratton, P. J. Campbell, and B. C. W. G. o. t. I. C. G. C. . The life history of 21 breast cancers. *Cell*, 149(5): 994–1007, May 2012b.

R. F. Schwarz and many others. Phylogenetic quantification of intra-tumor heterogeneity. *yes*, yes:yes, 2013a.

R. F. Schwarz and many others. Phylogenetic quantification of intra-tumor heterogeneity predicts time to relapse in high-grade serous ovarian cancer. *yes*, yes:yes, 2013b.

S. P. Shah, R. D. Morin, J. Khattra, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany, J. Senz, C. Steidl, R. A. Holt, S. Jones, M. Sun, G. Leung, R. Moore, T. Severson, G. A. Taylor, A. E. Teschendorff, K. Tse, G. Turashvili, R. Varhol, R. L. Warren, P. Watson, Y. Zhao, C. Caldas, D. Huntsman, M. Hirst, M. A. Marra, and S. Aparicio. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 461(7265):809–813, Oct 2009. doi: 10.1038/nature08489. URL `http://dx.doi.org/10.1038/nature08489`.

S. P. Shah, A. Roth, R. Goya, A. Oloumi, G. Ha, Y. Zhao, G. Turashvili, J. Ding, K. Tse, G. Haffari, A. Bashashati, L. M. Prentice, J. Khattra, A. Burleigh, D. Yap, V. Bernard, A. McPherson, K. Shumansky, A. Crisan, R. Giuliany, A. Heravi-Moussavi, J. Rosner, D. Lai, I. Birol, R. Varhol, A. Tam, N. Dhalla, T. Zeng, K. Ma, S. K. Chan, M. Griffith, A. Moradian, S.-W. G. Cheng, G. B. Morin, P. Watson, K. Gelmon, S. Chia, S.-F. Chin, C. Curtis, O. M. Rueda, P. D. Pharoah, S. Damaraju, J. Mackey, K. Hoon, T. Harkins, V. Tadigotla, M. Sigaroudinia, P. Gascard, T. Tlsty, J. F. Costello, I. M. Meyer, C. J. Eaves, W. W. Wasserman, S. Jones, D. Huntsman, M. Hirst, C. Caldas, M. A. Marra, and S. Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–399, Jun 2012. doi: 10.1038/nature10933. URL `http://dx.doi.org/10.1038/nature10933`.

C. K. I. Williams. A mcmc approach to hierarchical mixture modelling. In *NIPS*, 1999.