# 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 3, 2013.

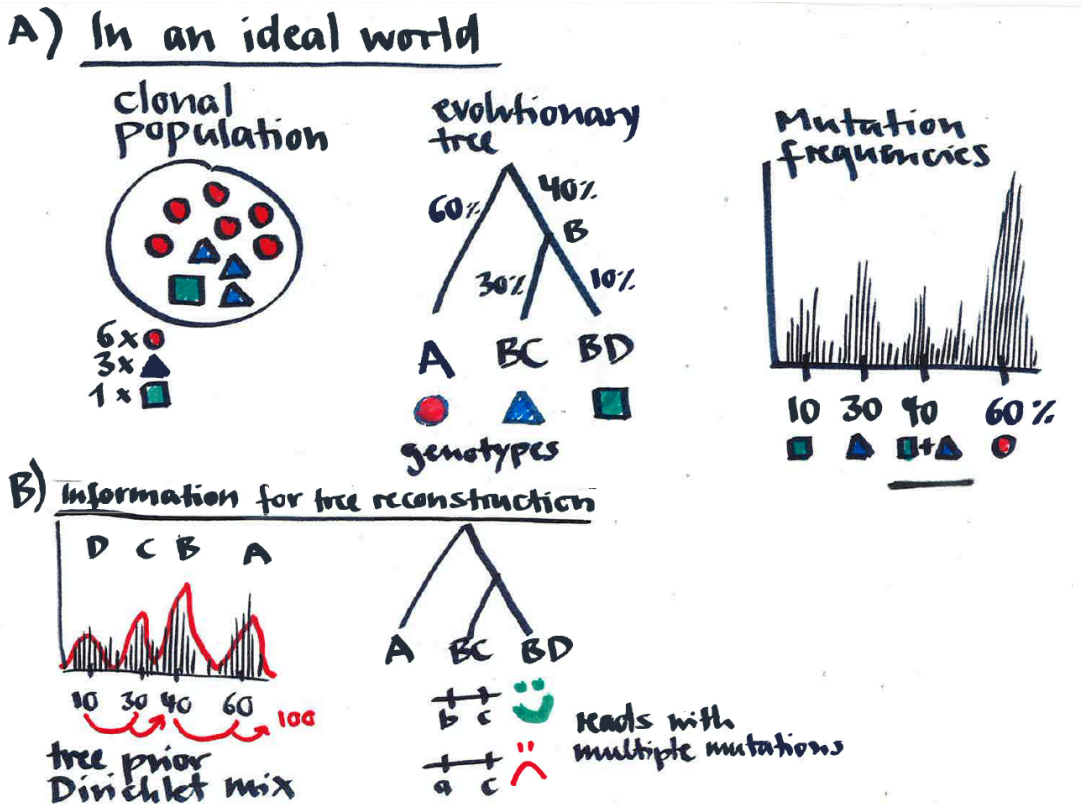


FIGURE 1. **A)** Overview how heterogeneity in a tumor relates to clonal evolution and is reflected in mutation frequencies. **B)** The sources of information we have available for inference.

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

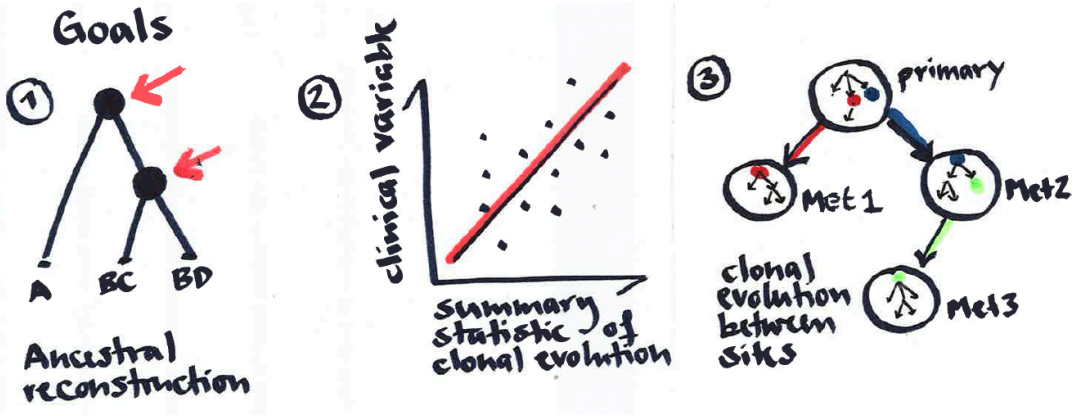


Figure 2. **Goals of our analysis are** (i) to infer early events in tumour development, (ii) to link cancer heterogeneity to clinical outcome, (iii) to compare clonal evolution in metastasis to evolution in the primary tumour.

can 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.
- Repository for code, manuscripts and data(?). Any objections to using bitbucket?
- 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?

## 4. Methods (Basel)

An SNV $x$ is defined by its genomic position $\mathrm{pos}(x)$ and its variant nucleotide $\mathrm{nt}(x) \in \mathcal{A} := \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Let $\mathcal{S}$ denote the set of all SNVs in a given tumor. NGS of the tumor produces short reads and we consider only those reads that overlap with at least one segregating site $j \in \{\mathrm{pos}(x) \mid x \in \mathcal{S}\}$. We assume that each read contains either one or two SNVs (relative to the most frequent nucleotide at the respective position). Reads without SNVs are not informative and those with more than two SNVs are so rare that we can ignore them. Then the data $X$ consists of $N_1$ reads carrying one SNV, $x_1^{(i)} \in \mathcal{S}$, $i = 1, \ldots, N_1$, and $N_2$ reads with pairs of SNVs, $x_2^{(i)} \in \mathcal{S}^2$, $i = 1, \ldots, N_2$.

An intra-tumor oncogenetic tree (ITOT) is a labeled binary rooted tree $T = (V, E, r, g, f, s)$ (Figure 3). The edges $e \in E$ are labeled with sets of SNVs, $s : E \to 2^{\mathcal{S}}$, that have occurred on this branch such that they define a partition of the set of all SNVs, i.e., $s(e_1) \cap s(e_2) = \emptyset$ for all edges $e_1 \neq e_2$ and $\cup_{e \in E} s(e) = \mathcal{S}$. The vertices $v \in V$ represent subclones and are labeled (uniquely) with genotypes. A genotype is a set of co-ocurring SNVs and each vertex is labeled by the set of all SNVs that have given rise to it. Formally, let $(e_1, \ldots, e_M)$ be the unique path (ordered set of edges) connecting the root vertex $r$ with vertex $v \in V$. Then the genotype label $g : V \to 2^{\mathcal{S}}$ of vertex $v$ is $g(v) = \cup_{m=1}^{M} s(e_m)$. It follows that $g(v) = g(w_1) \cap g(w_2)$ for all branching points $v$ into daughter clones $w_1$ and $w_2$. The root of the tree is the genotype after the last clonal expansion of the tumor, the leaves of the

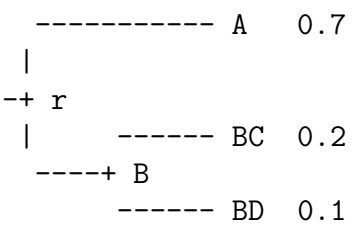


FIGURE 3. ITOT for three clones. Edges are labeled from top to bottom by SNV sets A, C, B, D. Vertices are labeled from top to bottom by genotypes A (leaf), r (root), BC (leaf), B (internal), BD (leaf), and by clone frequencies 0.7, 1.0, 0.2, 0.3, 0.1.

tree are the subclones at the time of observation (diagnosis), and the interior vertices are extinct common ancestor clones. We denote by $L \subset V$ the set of leaves (contemporary clones). In addition, vertices are labeled with their relative frequencies within the tumor, $f : V \to (0, 1]$, subject to the following constraints: (i) $f(r) = 1$, (ii) $\sum_{v \in L} f(v) = 1$, and (iii) $f(v) = f(w_1) + f(w_2)$ for all branching points $v$ into daughter clones $w_1$ and $w_2$. Note that if (ii) holds for a function $f : L \to (0, 1]$, then $f$ can always be extended in a unique way to $V$ such that (i) and (iii) also hold.

The ITOT model consists of an ITOT $T = (V, E, r, g, f, s)$ for generating genotypes $G$ and a probabilistic model for generating reads, i.e., single SNVs $X_1$ and SNV pairs $X_2$, from genotypes. Let $L = \{g_1, \ldots, g_K\}$ denote the genotypes of the contemporary clones, i.e., the leaves of the tree. Each leaf genotype $G$ is generated with probability

$$\Pr(G = g_k) = f_k \tag{1}$$

where $f_k \in (0, 1]$ are parameters such that $\sum_{k=1}^{K} f_k = 1$. The parameters $f_k$ define the clone frequencies of the ITOT via $f(g_k) = f_k$ and its unique recursive extension to $V$. The conditional probabilities of observing single and pair SNVs are, respectively,

$$\Pr(X_1 = s \mid G = g) = \begin{cases} \varepsilon & \text{if } |\{s\} \cap g| = 0 \\ 1 - \varepsilon & \text{if } |\{s\} \cap g| = 1 \end{cases} \tag{2}$$

$$\Pr(X_2 = (s, t) \mid G = g) = \begin{cases} \varepsilon^2 & \text{if } |\{s, t\} \cap g| = 0 \\ 2\varepsilon(1 - \varepsilon) & \text{if } |\{s, t\} \cap g| = 1 \\ (1 - \varepsilon)^2 & \text{if } |\{s, t\} \cap g| = 2 \end{cases} \tag{3}$$

where $\varepsilon$ is the error probability accounting for misassignment of SNVs to their clones. The ITOT model is the mixture model

$$\Pr(X_m \mid T) = \sum_{k=1}^{K} \Pr(X_m, G = g_k) = \sum_{k=1}^{K} \Pr(X_m \mid G = g_k) \Pr(G = g_k), \quad m = 1, 2$$

To learn the structure and parameters of the ITOT model for fixed $K$ from observed SNV data, we devise a structural EM algorithm. In the E step, we compute the responsibility of each leaf genotype for each observed read. In the M step, we re-estimate the ITOT.

4.1. **E step.** The responsibility of leaf genotype $g_k$ for observation $x_m^{(i)}$, $m = 1, 2$, $i = 1, \ldots, N_m$, is

$$\gamma_{mik} := \Pr(G = g_k \mid X_m = x_m^{(i)}) \tag{4}$$

$$= \frac{\Pr(X_m = x_x^{(i)} \mid G = g_k)\Pr(G = g_k)}{\sum_{k=1}^{K} \Pr(X_m = x_x^{(i)} \mid G = g_k)\Pr(G = g_k)} \tag{5}$$

4.2. **M step.** The ML step consists of maximization over the discrete structures tree topologies and SNV partitions, and over the continuous parameters $f$ (clone frequencies) and $\varepsilon$ (error probability). For given leaf genotypes $g_1, \ldots, g_K$, the MLEs of $f$ and $\varepsilon$ are, respectively,

$$\hat{f}_k = \frac{\sum_{m=1,2}\sum_{i=1}^{N_m} \gamma_{mik}}{N_1 + N_2}, \quad k = 1, \ldots, K \tag{6}$$

$$\hat{\varepsilon} = \frac{\sum_{m=1,2}\sum_{i=1}^{N_m} \gamma_{mik}(m - |x_m^{(i)} \cap g_k|)}{N_1 + 2N_2} \tag{7}$$

Estimation of the leaf genotypes involves reconstruction of the ITOT from the responsibilities $\gamma_{mik}$. Exact ML estimation seems difficult, but we can employ heuristics, such as distance-based phylogenetic methods. The distance between clone $k$ and clone $l$ may be defined as

$$d_{kl} = \sum_{m=1,2}\sum_{i=1}^{N_m}(1 - 2\min\{\gamma_{mik}, \gamma_{mil}\}) \tag{8}$$

The resulting tree $(V, E)$ can be labeled recursively from the leaves to the root by setting $g(g_k) = \cup_{m,i}\{x_m^{(i)} \mid \gamma_{mik} \geq \eta\}$ for all leaves $g_k$, and $g(v) = g(w_1) \cap g(w_2)$ for all branchings $v$ into daughters $w_1$, $w_2$. Here, the paramter $\eta$ controls the multiple, hard assignment of SNVs to genotypes. $\eta \approx \varepsilon$ for single and $\eta \approx \varepsilon^2$ for pairs of SNVs appears a reasonable choice, or else a preset fixed cutoff.

Alternatively, the triplet search may be employed here ...

From $g : V \to 2^{\mathcal{S}}$, edge labels $s : E \to 2^{\mathcal{S}}$ can be inferred. Clone frequencies $f$ can then be estimated for all leaves as described above and extended to $V$ as described earlier.

## 5. Methods (Cambridge)

to come

## 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.