Ccube: A fast and robust method for estimating cancer cell fractions

Ke Yuan[a,b,e], Geoff Macintyre[b,c], Wei Liu[b,d], PCAWG-11 working group, Florian Markowetz[b,e]

[a] *School of Computing Science, University of Glasgow, Sir Alwyn Williams Building, Glasgow, G12 8RZ, UK*
[b] *Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, UK*
[c] *Department of Computing and Information Systems, University of Melbourne, Parkville 3010, Victoria, Australia*
[d] *Current address: Institute of Biodiversity Animal Health and Comparative Medicine, Graham Kerr Building, University of Glasgow, Glasgow, G12 8QQ, UK*
[e] *Correspondence to: Ke Yuan ke.yuan@glasgow.ac.uk and Florian Markowetz florian.markowetz@cruk.cam.ac.uk*

Abstract

Estimating and clustering cancer cell fractions of genomic alterations are central tasks for studying intratumour heterogeneity. We present Ccube, a probabilistic framework for inferring the cancer cell fraction of somatic point mutations and the subclonal composition from whole-genome sequencing data. We develop a variational inference method for model fitting, which allows us to handle samples with large number of the variants (more than 2 million) while quantifying uncertainty in a Bayesian fashion. Ccube is available at `https://github.com/keyuan/ccube`.

Keywords: intratumour heterogeneity, cancer cell fraction, variant allele frequency, variational inference

1. Introduction

A fundamental problem when studying intratumour heterogeneity is to estimate the cancer cell fraction (CCF) of a single nucleotide variants (SNVs). The key difficulty is that CCF is proportional to the number of mutated chromosomal copies, known as the multiplicity of a mutation, which is also unknown. For any given mutation, the observed variant allele frequency can be modelled by CCF times multiplicity. As a result, it is impossible to estimate CCF without making strong assumptions about what multiplicity is. Several methods choose to prefix multiplicities, for example, DPClust [1] and PyClone [2]. PhyloSub [3] and PhyloWGS [4] use phylogenetic trees to estimate the multiplicities. This approach significantly increase the complexity of the model making model inference difficult to scale.

We develop Ccube, a method using clustering (i.e. assuming multiple mutations share the same CCF) to determine what the appropriate multiplicities are. The method takes sequencing reads profiles of SNVs, corrects them for copy number alterations and purity, and produces CCF estimates for all mutations within the sample.

2. Results

2.1. Mapping between variant allele frequency and cancer cell fraction

Given the purity of the sample and copy number profile at the locus of a SNV of interest, there is a mapping between VAF and the CCF. Following [2], we formulate relationship between VAF and CCF from a probability stand point. Generally speaking, we are dealing with two questions here: first, are we observing a variant allele on the a sequencing read covering the locus of the mutation i.e. read variable. Second, where is the read comes from i.e. population variable. The key for the probabilistic view point is to consider VAF as the marginal probability of observing a variant read, where the population variable is integrated out:

$$p(\text{read} = \text{variant allele}) = \sum_s p(\text{read} = \text{variant allele}, \text{population} = s) \qquad (1)$$

Following the definition in [2], the population variable has three possible states: normal cell population, cancer cells that don't bear the SNV, defined as the reference population and cancer cells that carry the mutation, defined as the variant population. Based these three possible populations. For each population, we have the probability of observing a read coming from the population, and the conditional probability of the population contributing a variant read. The marginalisation can be written as,

$$\begin{aligned}&\sum_s p(\text{read} = \text{variant allele}, \text{population} = s) \qquad (2)\\ &= p(\text{population} = \text{normal})p(\text{read} = \text{variant allele}|\text{population} = \text{normal})\\ &\quad + p(\text{population} = \text{reference})p(\text{read} = \text{variant allele}|\text{population} = \text{reference})\\ &\quad + p(\text{population} = \text{variant})p(\text{read} = \text{variant allele}|\text{population} = \text{variant})\end{aligned}$$

The probability of observing a read coming from a population is proportional to the prevalence of the population in the sample times its total copy number. Specifically, the prevalence the variant population is CCF, ϕ. Therefore, we have

$$p(\text{population} = \text{normal}) = \frac{(1-t)n_{tot_n}}{C} \qquad (3)$$

$$p(\text{population} = \text{reference}) = \frac{t(1-\phi)n_{tot_{ref}}}{C} \qquad (4)$$

$$p(\text{population} = \text{variant}) = \frac{t\phi n_{tot_{var}}}{C} \qquad (5)$$

where $n_{tot_{ref}}$, $n_{tot_{var}}$, n_{tot_n} are the total copy numbers of cells in the reference, variant, and normal populations, respectively. The normalising constant, $C = (1-t)n_{tot_n} + t(1-\phi)n_{tot_{ref}} + t\phi n_{tot_{var}}$, make sure the probabilities add up to one.

We further assume the reference shares the same total copy number with the variant population, $n_{tot_{ref}} = n_{tot_{var}}$. This corresponds to assuming all the cancer cells in the sample share the same total copy number at the site of SNV, i.e. the copy number is clonal [1]. Using n_{tot_t} to represent the clonal total copy number. We have:

$$p(\text{population} = \text{normal}) = \frac{(1-t)n_{tot_n}}{C} \qquad (6)$$

$$p(\text{population} = \text{reference}) = \frac{t(1-\phi)n_{tot_t}}{C} \qquad (7)$$

$$p(\text{population} = \text{variant}) = \frac{t\phi n_{tot_t}}{C} \qquad (8)$$

$$C = tn_{tot_t} + (1-t)n_{tot_n} \qquad (9)$$

The conditional probabilities for populations to produce a variant read are:

$$p(\text{read} = \text{variant allele}|\text{population} = \text{normal}) = \epsilon \tag{10}$$

$$p(\text{read} = \text{variant allele}|\text{population} = \text{reference}) = \epsilon \tag{11}$$

$$p(\text{read} = \text{variant allele}|\text{population} = \text{variant}) = \frac{m}{n_{tot_t}}(1-\epsilon) \tag{12}$$

where ϵ is a uniform sequencing error. m is the number of mutated chromosomal copy, the multiplicity of the mutations.

Taken together, we obtain a linear mapping between the probability of observing a variant read at a mutated locus, f, and the CCF of the mutation ϕ:

$$f = w\phi + \epsilon \tag{13}$$

$$w = \frac{t(m(1-\epsilon) - n_{tot_t}\epsilon)}{(1-t)n_{tot_n} + tn_{tot_t}} \tag{14}$$

Assuming the number of read carrying the variant allele follows a binomial distribution, the VAF is an unbiased estimator of f.

2.2. The Ccube model for estimating and clustering cancer cell fractions

Let $i \in \{1, ..., N\}$ denotes the index for each variant considered, and $k \in 1, ..., K$ denote the index for the number of CCF cluster identified in the mixture. For the ith variant, b_i and d_i denote the number of reads reporting variant allele and the total number of reads. The copy number profile at the ith mutation includes the total copy number of the tumour n_{i,tot_t}, the total copy number of the normal population n_{i,tot_n}, the copy number of the major allele n_{i,maj_t}, the copy number of the minor allele n_{i,min_t}. The number of mutated chromosomal copies is denoted by m_i.

The basis of Ccube is a Binomial mixture model. We assume the ith variant allele read count b_i follows a Binomial Distribution with total read count d_i and expected VAF f_i as its parameter

$$b_i \sim \text{Binomial}(b_i|d_i, f_{i,k}) \tag{15}$$

where

$$f_{i,k} = w_i \phi_k + \epsilon \tag{16}$$

. where $w_i = \frac{t(m_i(1-\epsilon)-n_{i,tot_t}\epsilon)}{(1-t)n_{i,tot_n}+\rho n_{i,tot_t}}$.

The Ccube model can be shown as the following:

$$\begin{aligned}
b_i|z_{i,k} &\sim \text{Binomial}(b_i|d_i, f_{i,k})^{z_{i,k}} &(17)\\
f_{i,k} &= w_i \phi_k + \epsilon &(18)\\
z_{i,k} &\sim \pi_k^{z_{i,k}} &(19)\\
\pi_1, ...\pi_k &\sim \text{Dir}(\alpha, ..., \alpha) &(20)\\
\phi_k &\sim \mathcal{N}(\phi_k|\mu_0, \sigma_0^2) &(21)
\end{aligned}$$

2.3. Variational inference for Ccube

The variational inference maximises the evidence lower bound (ELBO) of the marginal likelihood of the model:

$$\begin{aligned}
\log p(\mathbf{b}|\mathbf{d}, \mathbf{m}) &= \log \int p(\mathbf{b}|\mathbf{d}, \mathbf{Z}, \boldsymbol{\phi}, \mathbf{m}) p(\mathbf{Z}|\boldsymbol{\pi}) p(\boldsymbol{\phi}) p(\boldsymbol{\pi}) d\mathbf{Z} d\boldsymbol{\phi} d\boldsymbol{\pi} &(22)\\
&= \log \int p(\mathbf{b}, \mathbf{Z}, \boldsymbol{\phi}, \boldsymbol{\pi}|\mathbf{d}, \mathbf{m}) d\mathbf{Z} d\boldsymbol{\phi} d\boldsymbol{\pi} &(23)\\
&\geq \mathbb{E}_{q(\mathbf{Z},\boldsymbol{\phi},\boldsymbol{\pi})} [\log p(\mathbf{b}, \mathbf{Z}, \boldsymbol{\phi}, \boldsymbol{\pi}|\mathbf{d}, \mathbf{m})] - \mathbb{E}_{q(\mathbf{Z},\boldsymbol{\phi},\boldsymbol{\pi})} [\log q(\mathbf{Z}, \boldsymbol{\phi}, \boldsymbol{\pi})] &(24)
\end{aligned}$$

where $\mathbf{b} = \{b_i\}$, $\mathbf{d} = \{d_i\}$, $\mathbf{Z} = \{z_{i,k}\}$, $\boldsymbol{\phi} = \{\phi_k\}$, $\boldsymbol{\pi} = \{\pi_k\}$, $\mathbf{m} = \{m_i\}$.

We adopt the common fixed-form mean field approximation, in which $\mathbf{Z}, \boldsymbol{\phi}, \boldsymbol{\pi}$ are independent:

$$q(\mathbf{Z}, \boldsymbol{\phi}, \boldsymbol{\pi}) = q(\mathbf{Z}) q(\boldsymbol{\phi}) q(\boldsymbol{\pi}) \tag{25}$$

Maximising the ELBO with respect to the above $q(\mathbf{Z}, \boldsymbol{\phi}, \boldsymbol{\pi})$ yields the following forms:

$$\begin{aligned}
q(\mathbf{Z}) &\propto \exp(\mathbb{E}_{q(\boldsymbol{\phi},\boldsymbol{\pi})} [\log p(\mathbf{b}, \mathbf{Z}, \boldsymbol{\phi}, \boldsymbol{\pi}|\mathbf{d}, \mathbf{m})]) &(26)\\
q(\boldsymbol{\phi}) &\propto \exp(\mathbb{E}_{q(\mathbf{Z},\boldsymbol{\pi})} [\log p(\mathbf{b}, \mathbf{Z}, \boldsymbol{\phi}, \boldsymbol{\pi}|\mathbf{d}, \mathbf{m})]) &(27)\\
q(\boldsymbol{\pi}) &\propto \exp(\mathbb{E}_{q(\mathbf{Z},\boldsymbol{\phi})} [\log p(\mathbf{b}, \mathbf{Z}, \boldsymbol{\phi}, \boldsymbol{\pi}|\mathbf{d}, \mathbf{m})]) &(28)
\end{aligned}$$

These distributions are fitted to the data with a variational E-step and variational M-step. In the E-step, we update the approximate posterior of the assignment.

$$\log q(z_{i,k}=1) \propto \mathbb{E}_{q(\boldsymbol{\phi})}[\log p(b_i|d_i, z_{i,k}=1, \phi_k, m_i)] + \mathbb{E}_{q(\boldsymbol{\pi})}[\log p(z_{i,k}=1|\pi_k)] \tag{29}$$

$$= \log \gamma_{i,k} \tag{30}$$

$$q(z_{i,k}=1) = \frac{\gamma_{i,k}}{\sum_j \gamma_{i,j}} \tag{31}$$

In the variational M-step, we update the approximate posteriors on parameters $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$.

$$q(\phi_k) = \mathcal{N}(\mu_k, \sigma_k^2) \tag{32}$$

where,

$$\mu_k = \underset{\phi_k}{\operatorname{argmax}}\ g(\mathbf{b}, \mathbf{d}, \mathbf{m}, \mathbf{z}_k, \phi_k) \tag{33}$$

$$\sigma_k^2 = -\left(\frac{\partial^2 g(\mathbf{b}, \mathbf{d}, \mathbf{m}, \mathbf{z}_k, \phi_k)}{\partial \phi_k^2}\right)^{-1} \tag{34}$$

$$g(\mathbf{b}, \mathbf{d}, \mathbf{m}, \mathbf{z}_k, \phi_k) = \log p(\phi_k|\mu_0, \sigma_0^2) + \sum_{i=1}^{N} \mathbb{E}_{q(z_{i,k})}[\log p(b_i|d_i, f_{i,k}, z_{i,k}=1)] \tag{35}$$

The multiplicities are estimated as the following

$$\hat{m}_i = \underset{m_i \in \mathcal{M}_{clonal}}{\operatorname{argmax}} \sum_{k}^{K} \mathbb{E}_{q(z_{i,k}, \phi_k)}[\log p(b_i|d_i, f_{i,k}, z_{i,k}=1)] \tag{36}$$

where $\mathcal{M}_{clonal} = \{1, ...n_{i,maj_t}\}$.

Finally, $q(\boldsymbol{\pi})$ is obtained as standard variational approximation for mixing weight in mixture models [5].

3. Ccube pipeline

Preprocessing: The Ccube pipeline uses the clonal copy number, consensus purity and variant and reference allele read counts from somatic point mutation calls.

Postprocessing steps: The core variational inference is ran with a range of possible number of clusters. The solution with the best ELBO is selected. The solution is consisted of 1) posterior distributions of in terms of means and variances; 2) posterior probabilities of each mutation to be assigned to all clusters and final assignments; 3) multiplicities and observed CCFs based multiplicities. Clusters with less than 1% of mutation assigned are removed. The mutations are re-assigned with variational expectation step. Clusters with mean CCF closer than 10% are merged by re-running the inference with merged cluster configuration. A typical graphical summary can be found in figure 1.

3.1. Estimating purity

Ccube pipeline also produces an independent purity estimate using mutations from balanced copy number regions. For samples without whole-genome duplication, only mutations in normal copy number regions are included. For samples with whole-genome duplication, all mutations in balanced copy number regions are included. We then convert the allele frequencies to cellular prevalence using the following equation:

$$f_i = \frac{\eta_i n_{i,tot_t}}{2((1-\eta_i)n_{i,tot_n} + \eta_i n_{i,tot_t})} \tag{37}$$

Where is the VAF of the ith mutation obtained as the ratio between variant and wild-type allele counts, η is the cellular prevalence of the th mutation, are the total copy number of normal and tumour populations respectively. We then cluster the using students-t mixture model. The model is fitted with the variational Bayes approach described in [6]. The purity corresponds to the component with the largest mean, in additional the eligible component must have at least more than 1.5% of mutation assigned to it.

4. Reference

[1] 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. Jönsson, O. Mariani, S. Boyault, P. Miron, A. Fatima, A. Langerod, S. A. Aparicio, A. Tutt, A. M. Sieuwerts, . Borg, G. Thomas, A. V. Salomon, A. L. Richardson, A. L. Borresen-Dale, P. A. Futreal, M. R. Stratton, P. J. Campbell, The life history of 21 breast cancers, Cell 149 (2012) 994–1007.

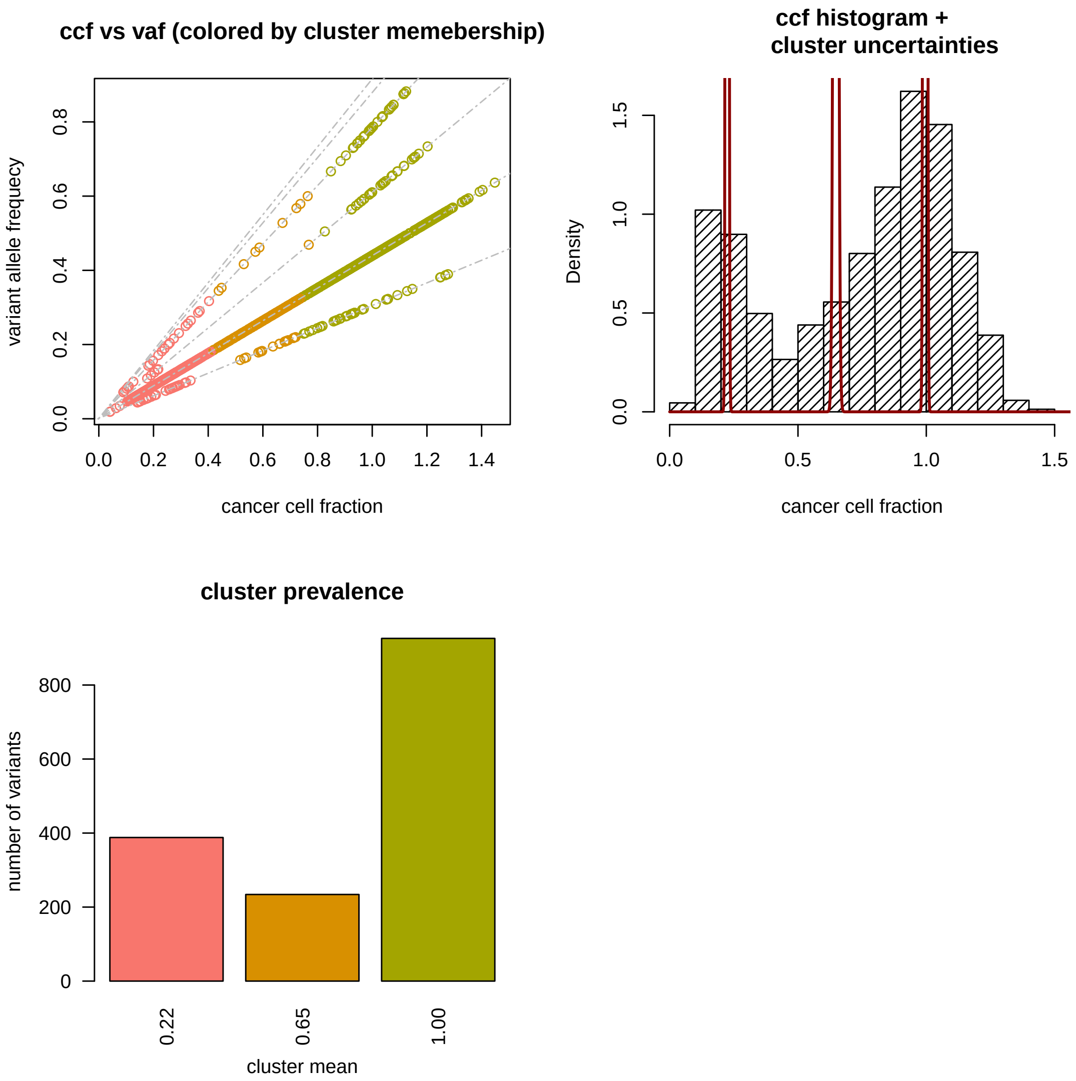

Figure 1: An example of Ccube sample results summary. A: Scatter plot of VAF and CCF. Each point in the figure is a mutation color coded by its cluster membership. The gray dashed line are all possible linear mappings (eq. 1) determined by copy number and multiplicity configurations in the sample. B: Histogram of observed CCFs. The red solid line shows the approximated posterior distribution of CCF cluster centers. The peak at CCF=1 corresponds to the clonal cluster. C: The number of variants assigned each CCF cluster. Each CCF cluster is labelled by it cluster center.

[2] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, S. P. Shah, PyClone: statistical inference of clonal population structure in cancer., Nature methods 11 (2014) 396–8.

[3] W. Jiao, S. Vembu, A. G. Deshwar, L. Stein, Q. Morris, Inferring clonal evolution of tumors from single nucleotide somatic mutations, in: J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, A. Culotta (Eds.), BMC Bioinformatics, volume 15, Curran Associates, Inc., 2014, p. 35.

[4] A. G. Deshwar, S. Vembu, C. K. Yung, G. Jang, L. Stein, Q. Morris, PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors, Genome Biology 16 (2015) 35.

[5] C. M. Bishop, Patterns Recognition and Machine Learning, 2006.

[6] C. Archambeau, M. Verleysen, Robust Bayesian clustering., Neural networks : the official journal of the International Neural Network Society 20 (2007) 129–38.