OXFORD UNIVERSITY PRESS | Bioinformatics

MuClone: Detection and classification of somatic mutations through probabilistic integration of clonal population structure

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2016-0542
Category:	Original Paper
Date Submitted by the Author:	05-Apr-2016
Complete List of Authors:	Dorri, Fatemeh; University of British Columbia, Computer Science Jewell, Sean; University of Washington, Statistics Bouchard-Cote, Alexandre; University of British Columbia, Statistics Shah, Sohrab; BC Cancer Agency, CTAG
Keywords:	HiTSeq, Single nuclotide variants, Tumour clones, Mutation detection, Mutation classification, Cancer evolution

BIOINFORMATICS

Vol. 00 no. 00 2016
Pages 1–7

MuClone: Detection and classification of somatic mutations through probabilistic integration of clonal population structure

Fatemeh Dorri [1,*], Sean Jewell [2], Alexandre Bouchard-Côté [3] and Sohrab P. Shah [4,5,*]

[1]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada.
[2]Department of Statistics, University of Washington, Seattle, WA, USA.
[3]Department of Statistics, University of British Columbia, Vancouver, BC, Canada.
[4]Department of Molecular Oncology, British Columbia Cancer Research Center, Vancouver, BC, Canada.
[5]Department of Pathology, University of British Columbia, Vancouver, BC, Canada.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Accurate detection and classification of somatic single nucleotide variants (SNVs) is important in defining the clonal composition of human cancers. Existing tools are prone to missing low prevalence mutations and methods for classification of mutations into clonal groups across the whole genome are underdeveloped. Increasing interest in deciphering clonal population dynamics over multiple samples in time or anatomic space from the same patient is resulting in whole genome sequence (WGS) data from phylogenetically related samples. We posited that injecting clonal structure information into the inference of mutations from multiple samples would improve mutation detection.

Results: We developed MuClone: a novel statistical framework for simultaneous detection and classification of mutations across multiple samples of a patient from whole genome or exome sequencing data. The key advance lies in incorporating prior knowledge about the cellular prevalences of clones to improve the performance of detecting mutations, particularly low prevalence mutations. We evaluated MuClone through synthetic and real data from spatially sampled ovarian cancers. The results support the hypothesis that clonal information improves the sensitivity without compromising the specificity. In addition, MuClone classifies mutations across whole genomes of multiple samples into biologically meaningful groups that can provide additional phylogenetic insights and permits studying clonal dynamics from WGS data.

Availability: MuClone is available from http://compbio.bccrc.ca/.

Contact: sshah@bccrc.ca

1 INTRODUCTION

Somatic single nucleotide variants (SNVs) in cancer tissues are point mutations resulting from the substitutions of a single nucleotide in the genomes of tumour cells. Accumulation of somatic SNVs can disrupt the regular activity of cells and may result in cancer initiation and progression. High-throughput sequencing technologies provide the opportunity to systematically profile SNVs at modest cost from excised patient tumour material. Collectively, the complete set of SNVs across the genome (numbering in the thousands) form a statistically robust marker for inferring clonal populations and studying tumour evolution. As such, accurate detection of SNVs, including low prevalence ones, is vital as they can define clones with phenotypic properties such as treatment resistance or metastatic potential and signal clones that may expand under therapeutic selective pressures.

Phylogenetic analysis can encode the evolutionary lineage of tumour cells across time and anatomic space (Yachida *et al.*, 2010; Govindan *et al.*, 2012; Shah *et al.*, 2009; Nik-Zainal *et al.*, 2012). Gerlinger *et al.* (2012) sequenced multiple spatially separated samples from renal cell carcinomas and related metastatic sites to reveal the evolutionary patterns. They related samples through phylogenetic analysis and distinguished at a coarse level mutations that were shared and ancestral from those that occured in subsets of cells. In a lung cancer study, 25 regions from seven non-small sections of malignant patients were sequenced. Evidence of branched evolution was characterized at the level of single nucleotide variations, copy number alteration, and structural variations (De Bruin *et al.*, 2014). Ding *et al.* (2012b) investigated the genetic changes associated with acute myeloid leukaemia (AML) relapse and established clonal evolution patterns in AML relapse specimens. Our recent work has determined clonal population dynamics over time in breast cancer xenografts (Eirew *et al.*, 2015), and anatomic space in intraperitoneal sites of primary high grade serous ovarian cancer (McPherson *et al.*, 2016), showing that the relative composition of constituent clones in multi-sample studies provides major insight into disease spread.

In the limit case, all cells likely harbour unique genomes. However due to the nature of branched evolutionary processes, clones could be thought of as major clades in the cell lineage phylogeny of a cancer. These clades share the majority of mutations, leaving open the possibility of defining first approximations to the

[*]to whom correspondence should be addressed

genotypes of clones. Clonal genotypes and their relative abundances in the cancer cell population can be approximated by clustering mutations measured in bulk tissues and estimating the cellular prevalences (the variant fraction of tumour cells) (Roth *et al.*, 2014). Alternatively, single cell sequencing (though noise properties interfere) can be used to measure prevalences (Roth *et al.*, 2016).

Phylogenetic algorithms mostly use mutations (represented as binary genetic markers), as inputs to infer the branched evolutionary lineages of tumour cells (Popic *et al.*, 2015; Deshwar *et al.*, 2015). Thus, accuracy of mutation detection will impact the performance of phylogenetic inference algorithms. Moreover, identifying low prevalence mutations brings us closer to identifying the complete repertoire of mutations in a cancer. Detection of low prevalence mutations from whole genome shotgun sequencing of bulk DNA is a major challenge due to their weak signal to noise ratio. The weak signal is because of (i) impure samples which are contaminated by normal cells, (ii) copy number alteration of the genome, and (iii) the presence of mutations in only a fraction of tumour cells (intra-tumour heterogeneity). We assert in this work that prior knowledge of clonal population structure will improve detection of mutations defining low prevalence clonal genotypes.

1.1 Previous work

There has been progress in developing SNV calling algorithms in recent years, but the problem remains challenging particularly for detecting low allelic ratio mutations. Algorithms have been developed for calling mutations from a single sample (Goya *et al.*, 2010; Koboldt *et al.*, 2009), paired samples (matched normal and tumour) (Kim *et al.*, 2013; Roth *et al.*, 2012; Ding *et al.*, 2012a; Cibulskis *et al.*, 2013), or multiple samples (Josephidou *et al.*, 2015; van Rens *et al.*, 2015). Cibulskis *et al.* (2013) proposed Mutect which uses a Bayesian classifier to detect mutations from matched normal and tumour samples. It uses various filters to ensure high specificity. Ding *et al.* (2012a) uses a feature based classifier for calling mutations. The features are constructed from matched paired normal and tumour samples. Josephidou *et al.* (2015) proposed multiSNV which jointly considers all available samples under a Bayesian framework to improve the performance of calling shared mutations.

1.2 Our contribution

Existing tools incorporate different information to detect mutations. However, none of them use the cellular prevalences of existing clones across samples of a patient to detect mutations. In MuClone, we exploit prior knowledge of the cellular prevalences of clones together with copy number information to improve the performance of detecting mutations, particularly the ones forming low prevalence clones. In addition, Muclone classifies mutations into clusters sharing similar cellular prevalence adding a rich layer of interpretation into the detection process. We tested MuClone through simulation studies and an application to real, multiple sample patient data. The findings in synthetic studies reveal that incorporating the cellular prevalences of different clones improves the performance of detecting mutations. In real data, the analysis show MuClone has higher sensitivity without compromising specificity compared with other methods.

2 NOTATION

We begin by introducing notation used in the MuClone model. We consider bulk tumour DNA as representing three populations: (i) normal cells; (ii) reference cells (malignant cells without a mutation of interest); and (iii) variant population (malignant cells with the mutation of interest). These are coded respectively through the genotype variables g_N^n, g_R^n, g_V^n. Human genomes normally have two copy of each locus. For variant loci, letter A denotes the allele that matches the reference genome and letter B denotes the one that does not match. Therefore, the genotype of a diploid locus can be AA, AB or BB. If there is a coincident copy number alteration, the possible genotypes will change accordingly. Each of the genotype variables of g_N^n, g_R^n, g_V^n take values in $\mathcal{G} = \{-, A, B, AA, AB, BB, AAA, AAB, \ldots\}$, e.g., the genotype AAB refers to a genotype with two reference alleles and one variant allele. $\psi_m^n \in \mathcal{G}^3$ is the genotype state, (g_N^n, g_R^n, g_V^n), for the n^{th} locus from m^{th} sample, and π is the prior over genotype states. The samples are indexed from $m = 1 \ldots M$.

Let t_m denote the fraction of cancer cells in m^{th} sample, called the tumour content. This implies that the fraction of normal cells is $1 - t_m$. Within the fraction of cells which are cancerous, we further divide into the fraction of cells from the variant population, ϕ_m^n, and the fraction which are from the reference population, $1 - \phi_m^n$. To recover copy number and number of variant alleles at a given locus, consider the mappings, for $g \in \mathcal{G}$:

$$\begin{aligned} c(g) \quad &: \quad \mathcal{G} \rightarrow \mathbb{N} \quad \text{returns the copy number,} \\ b(g) \quad &: \quad \mathcal{G} \rightarrow \mathbb{N} \quad \text{returns the number of variant loci,} \end{aligned}$$

so that, for example, $c(AAB) = 3$ and $b(AAB) = 1$. With these notations, it is now relatively easy to define a probability measure following binomial distribution for the variant allele from genotype g:

$$\mu(g) \quad = \quad \begin{cases} \frac{b(g)}{c(g)} & b(g) \neq 0, b(g) \neq c(g) \\ \epsilon & b(g) = 0 \\ 1 - \epsilon & b(g) = c(g), \end{cases} \tag{1}$$

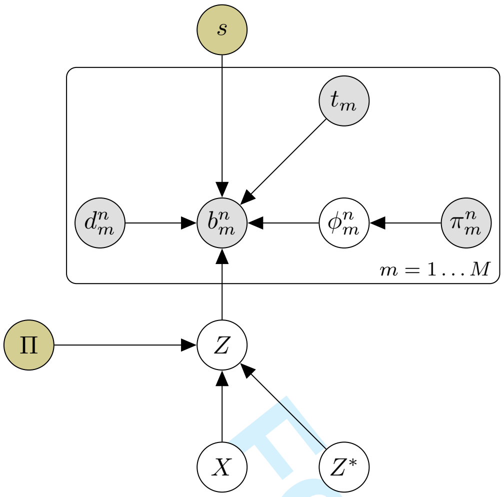
where ϵ accounts for sequencing error.

3 MUCLONE

The main premise of MuClone is to use tumour clones' prevalence information as a key prior knowledge in order to improve mutation detection and classification. In each sample from a patient, MuClone calls mutations from joint analysis of multiple samples. We encode this process in a generative probabilistic framework to perform joint statistical inference of multiple observations (from multiple samples) of the variant allele counts of a mutation of interest. The probabilistic graphical model of MuClone is depicted in Figure 1.

As such, MuClone assigns each locus of the genome to either (i) a wildtype (non-variant) cluster corresponding to cellular prevalence ϵ accounting for sequencing errors; or (ii) mutation clusters with given cellular prevalences. A locus is in the wildtype cluster, if it is wildtype across all samples, and it is in the mutation cluster, if it is mutated in at least one of the M samples.

Given a locus n, and tumour clone indicator $Z \in \{1, \ldots, k\}$, the following relationships are defined for the existing clones plus



$$b_m^n | d_m^n, \phi_m^n, z, s \propto \text{BetaBinomial}\left(d_m^n, \xi(\psi_m^n, \phi_m^n(z), t_m)(z)\right)$$

$$X|Z \propto \text{Bernoulli}\left(p^n = \textstyle\sum_{j \in \{1,\ldots,k\}} P^n(Z=j)\right)$$

$$\phi_m^n | \pi_m^n \propto \text{Categorical}\left(\pi_m^n\right)$$

$$Z^*|Z \propto \text{Categorical}\left({p_j}_{j \in \{1,\ldots,k\}} = P^n(Z=j)\right).$$

Fig. 1. Probabilistic graphical model of MuClone: white blank nodes are unobserved variables; grey shaded nodes are observed variables; golden nodes are external information.

the inclusion of a clone $Z = 0$ meant to capture the wildtype population,

$$\begin{aligned} b_m^n | Z = 0 &\sim \text{BetaBinomial}\left(d_m^n, \epsilon, s\right) \\ b_m^n | Z = z &\sim \text{BetaBinomial}\left(d_m^n, \kappa_m^n(z), s\right), \end{aligned} \tag{2}$$

where $\kappa_m^n(z) = \xi(\psi_m^n, \phi_m^n(z), t_m)(z)$ and ϵ are the means of the BetaBinomial distributions. Let ξ be the probability of sampling a read containing the variant allele overlapping the locus with genotype state of ψ_m^n. $\phi_m^n(z)$ is the cellular prevalence of clone z. Therefore:

$$\begin{aligned} W \cdot \xi(\psi_m^n, \phi_m^n, t) &= (1-t)c(g_N)\mu(g_N) + \\ &\quad t(1-\phi_m^n)c(g_R)\mu(g_R) + \\ &\quad t\phi_m^n c(g_V)\mu(g_V), \end{aligned} \tag{3}$$

where $W = (1-t)c(g_N) + t(1-\phi_m^n)c(g_R) + t\phi_m^n c(g_V)$ is the normalizing constant.

The probability of a locus n in the genome belonging to the clone, $Z = j$, is calculated as:

$$P^n(Z=j) \propto \tau_j \left(\prod_{m=1}^{M} \sum_{i \in I} \pi_i l(d_m^n, b_m^n, \kappa_m^n(z)) \right), \tag{4}$$

where the tumour clone indicator $Z \in \{1, \ldots, k\}$ represents mutation clusters and the wildtype cluster. i indexes π over the genotype states, $I = \{1 \ldots |\mathcal{G}|\}$, and $|\mathcal{G}|$ is the number of possible genotype states for n^{th} locus from m^{th} sample. τ_j is the prior over the potential clones. The likelihood function, $l\big(d_m^n, b_m^n, \kappa_m^n(z)\big)$, equals to $p\big(d_m^n, b_m^n | \kappa_m^n(z)\big)$, where the number of variants b_m^n follows the BetaBinomial distribution defined in Equation 2.

MuClone defines X as a cluster assignment variable, following a Bernoulli probability distribution with parameter p^n as Equation 5.

$$X|Z \propto \text{Bernoulli}\left(p^n = \sum_{j \in \{1,\ldots,k\}} P^n(Z=j)\right) \tag{5}$$

Based on basic decision theory, a decision can be extracted from a posterior distribution given a loss function. Therefore, $0-1$ loss function is defined:

$$L = \begin{cases} 1 & a \neq y \\ 0 & \text{elsewise} \end{cases} \tag{6}$$

where $a \in \mathcal{A}$ is an action, i.e., the new loci is a mutation or wildtype and $y \in \mathcal{Y}$ is the true state (here, $\mathcal{A} \equiv \mathcal{X}$, the space of values X can take). Then an estimator $\delta(D)$ that minimizes the expected loss is estimated as follows:

$$\begin{aligned} \delta(D) &= \underset{a \in \mathcal{A}}{\text{argmin}}\left(\mathbb{E}[\mathcal{L}(a, Y)|D]\right) && (7) \\ &= \underset{a \in \mathcal{A}}{\text{argmin}}\left(\left[\sum_{y \in \mathcal{Y}} p_y(y|D)\right] - p_y(y|D)\right) \\ &= \underset{y \in \mathcal{Y}}{\text{argmin}}\left(1 - p_y(y|D)\right) \\ &= \underset{y \in \mathcal{Y}}{\text{argmax}}\left(p_y(y|D)\right) \equiv \delta_{MAP}(D), && (8) \end{aligned}$$

where D is the data. Therefore, a locus belongs to the mutation cluster if p^n is greater than $P^n(Z=0)$.

MuClone defines Z^* as clone assignment for the locus which belongs to mutation cluster, following a categorical probability distribution:

$$Z^*|Z \propto \text{Categorical}\left({p_j}_{j \in \{1,\ldots,k\}} = P^n(Z=j)\right). \tag{9}$$

Based on decision theory, MuClone assigns locus n to clone Z^* which maximized the following:

$$Z^* = \underset{j \in \{1,\ldots,k\}}{\text{argmax}}\left(P^n(Z=j)\right). \tag{10}$$

Then, the probability of locus n from sample m, being a mutation is estimated as:

$$P_m^n(\text{mutant}) = \sum_{j \in J_m^*} P(Z^* = j). \tag{11}$$

where, J_m^* is the set of clones of sample m whose cellular prevalences are greater than a fixed threshold called Phi_threshold:

$$J_m^* = \{j \mid \phi_m(j) > \text{Phi\_threshold}\}.$$

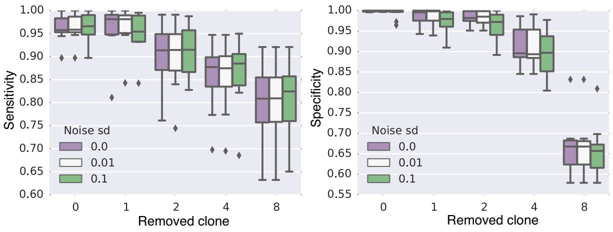


Fig. 2. MuClone's sensitivity and specificity with inaccurate clonal information. The synthetic data is generated for 200 positions from 4 samples of a patient, with 10 underlying clones. Maximum copy number is 5, and error rate equals 0.01. The average depth of sequencing is about 100. Noise sd is the standard deviation of noise added to (or subtracted from) the clones cellular prevalences. Removed clone is the number of clones that their clonal information is not accessible to MuClone. MuClone's parameters: Wildtype prior is 0.5, Phi_threshold is 0.03, error rate is 0.01, and precision parameter equals 1000.

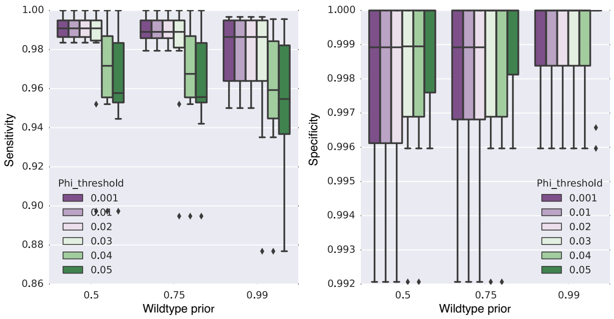


Fig. 3. MuClone's sensitivity and specificity at three different wildtype prior and six different Phi_threshold values. The synthetic data is generated for 200 positions from 4 samples of a patient, with 10 underlying clones. Maximum copy number is 5, and error rate equals 0.01. MuClone's parameters: error rate is 0.01, and precision parameter equals 1000.

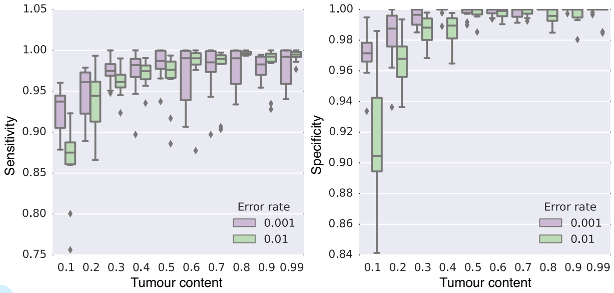


Fig. 4. MuClone's sensitivity and specificity with two different error rate values and tumour content values from 0.1 to 0.99. The synthetic data is generated for 200 positions from 4 samples of a patient, with 10 underlying clones. Maximum copy number is 5. MuClone's parameters: Wildtype prior is 0.5, Phi_threshold is 0.03, and precision parameter equals 1000.

4 EXPERIMENTAL RESULT

In order to evaluate the performance of MuClone and to show that exploiting clonal information improves the sensitivity of calling mutations, we tested MuClone on both synthetic and real data.

4.1 Synthetic data

4.1.1 Data generation Synthetic data was generated for the given number of positions, n, and samples, m. The other given parameters included: the number of tumour clones, k, sequencing error rate, ϵ, maximum copy number, C_{max}, and the mean coverage of the data, d_m, and tumour content, t_m, for each sample.

The cellular prevalences of tumour clones were sampled from a Uniform distribution over the closed interval $[0, 1]$. The positions were assigned to different clones. Then for each position in each sample, the coverage was sampled from a Poisson distribution with mean d_m. Wildtype copy number was deterministically set to 2 and a copy number profile (major and minor copy number) was generated by the following steps: First, total copy number was sampled from integer numbers between 1 and Cmax. C^b is a random integer between 1 and C^t, and C^a is defined as $C^a = C^t - C^b$, with the major copy number as the maximum of C^b and C^a and minor copy number as the minimum of those two values.

Then, corresponding to each clone, the number of variant reads were sampled from the Beta-binomial distribution described in Equation 2.

4.1.2 Synthetic data evaluation We simulated synthetic data for 200 positions from 4 samples of a patient, with 10 underlying clones, including ancestral clone. The maximum copy number was 5, and error rate was 0.01. The average depth of sequencing was assumed to be 100. This process was repeated 10 times in order to compute distributions over accuracy metrics.

We began evaluation by investigating the effect of systematically 'shielding' MuClone from clonal information during inference (Figure 2). Clonal information was perturbed by (i) adding noise to the cellular prevalences of tumour clones, or (ii) removing clonal information. The noise was generated from a normal distribution with different standard deviations (sd) equals 0, 0.01, 0.1, or 0.25. Then it is randomly either added to or subtracted from the cellular prevalence of clone, unless it got larger than 1 or smaller than 0. The clones were also systematically removed starting from the one with the smallest cellular prevalence to removing all except the ancestral clone (the number of removed clone is 9, when only the ancestral clone remained). As expected, both sensitivity and specificity are highest when clonal information is most complete and most accurate (Figure 2). This suggests that clonal information can indeed improve the accuracy of detecting mutations and establishes the theory of the MuClone approach. Furthermore, sensitivity and specificity were only marginally impacted by added noise in the clonal information, suggesting MuClone should be able to cope with modest levels of erroneous information in the prior. Naturally, accuracy was most severely impacted when the least amount of clonal information and the highest noise levels were given to the model (Figure 2). For different noise sd value, the sensitivity and specificity of removing various number of clones are compared through Kruskal-Wallis test (the p values are between $3e{-}7$ and $1e{-}5$) which shows the sweep of performance is significant with the changes in the amount of clonal information.

We next explored how sensitivity and specificity change as a function of the wildtype prior and Phi_threshold values (Figure

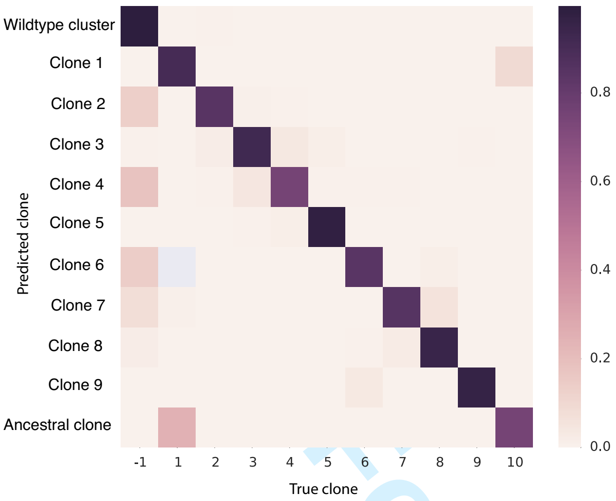


Fig. 5. Classification of mutations in different clones by MuClone. The synthetic data is generated for 200 positions from 4 samples of a patient, with 10 underlying clones. Maximum copy number is 5, error rate is 0.01, and noise sd is 0.01. MuClone's parameters: Wildtype prior is 0.5, Phi_threshold is 0.03, error rate is 0.01, and precision parameter equals 1000. Bin (i, j) shows the normalized number of mutations in clone i but classified in clone j. Diagonal elements show 89% of the mutations are classified correctly.

3). We generated data setting the wildtype prior at 0.5, 0.75, and 0.99; and Phi_threshold at 0.001, 0.01, 0.02, 0.03, 0.04, and 0.05. MuClone's sensitivity and specificity are near 1 for Phi_threshold less than 0.03. Sensitivity dropped in higher Phi_threshold, because mutations are miscalled with wildtype. The optimal Phi_threshold was about 0.03 when wildtype prior $= 0.5$. These parameters were used in the following experimental results.

The performance of MuClone was tested with different tumour content (range from 0.1 to 0.99) and error rate (0.01 or 0.001) (Figure 4). For samples with tumour content greater than 0.5, sensitivity remained slightly less than 1 and specificity near 1. Sensitivity and specificity dropped to only about 0.9 when tumour content in the sample was as low as 0.1, establishing promising performance over different ranges of tumour content with different error rates (likely scenarios in real data).

Figure 5 demonstrates how mutations are classified by MuClone. The input clonal information has been perturbed by adding noise with standard deviation of 0.01 to simulate a more realistic scenario. Each bin in row i and column j shows the number of mutations belonging to clone i and MuClone classifies them in clone j, divided by the total number of mutations. Figure 5 shows 89% of mutations are classified to the right clone as the diagonal elements are larger than the other ones.

4.2 Real data

We next tested MuClone's performance on whole genome sequencing data ($30X$) from multiple tumour samples surgically resected 7 high grade serous ovarian cancer patients (McPherson *et al.*, 2016). The samples were obtained from different spatially distributed metastatic sites. Brief details about the number of samples for each patient and sample sites are shown in Table 4.2.

Patient	samples	#validated positions	Anatomic samples
1	6	413	Right Ovary Site 1-4; Omentum 1; Small Bowel Site l
2	4	346	Omentum 1,2;Right Ovary Site1,2
3	4	373	Right Ovary Site 1,2;Omentum 1; Left Ovary Site 2
4	5	343	Right Ovary Site 1-4;Right Pelvic Mass
5	3	453	Left Ovary Site 1; Brain Metastasis; Right Pelvic Mass
6	5	435	Right Ovary Site 1; Left Ovary Site 1; Omentum Site 1,2
7	4	385	Right Ovary Site 1-4

Table 1. Description of the real data set (McPherson *et al.*, 2016).

The clonal information and experimentally re-validated mutations status were taken from McPherson *et al.* (2016) (accepted manuscript, to appear), estimated from running PyClone on the deep targeted sequencing data from the same samples and in three patients with accompanying single cell sequencing data (McPherson *et al.*, 2016). Performance was benchmarked against Strelka, MutationSeq, Mutect and MultiSNV.

The number of validated positions for each patient is shown in Table 4.2. The sensitivity and specificity for each sample is estimated separately. The distribution of performance across different samples of 7 patients are shown in Figure 6. It shows the sensitivity, specificity and Youden's index of different methods across all patients. Youden's index is defined as:

$$\text{Youden's index} = \text{Sensitivity} + \text{Specificity - 1}$$

In aggregate, MuClone outperforms other methods by improving the sensitivity without compromising the specificity (Figure 6). Strelka, MutationSeq and Mutect have lower performance as they do not incorporate the information across multiple samples for calling mutations. We calculated Welch's t-test for unequal population variances on MuClone and MultiSNV Youden's index results. The performance of MuClone was statistically higher than MultiSNV (p-value $= 0.01$). Importantly, MuClone improves sensitivity, enabling detection of more mutations across whole genome.

In addition MuClone, classifies mutations into different clones. Figure 7 depicts classification of mutations into clones relative to ground truth as established in (McPherson *et al.*, 2016). Each bin shows the normalized number of mutations. 88% of the elements in Figure 7 are diagonal which means MuClone classifies them correctly.

5 DISCUSSION AND CONCLUSION

We studied the use of clonal information for the purpose of somatic mutation detection and classification in multi-sample whole genome sequencing data. The proposed statistical framework uses the clones' cellular prevalences and copy number information for detection and classification of mutations. MuClone outperformed other popular mutation detection tools while having the benefit of classifying whole genome sequencing mutations into biologically relevant groups. Both simulation and real data results showed the cellular prevalences of tumour clones are beneficial information for improving the sensitivity. Importantly, our results suggest this improvement in sensitivity can be achieved without loss the specificity. As the accuracy of detecting mutations can affect the performance of phylogenetic analysis, we suggest this improvement will impact the field of multi-region sequencing for cancer evolution studies. As the field matures, we expect the method presented

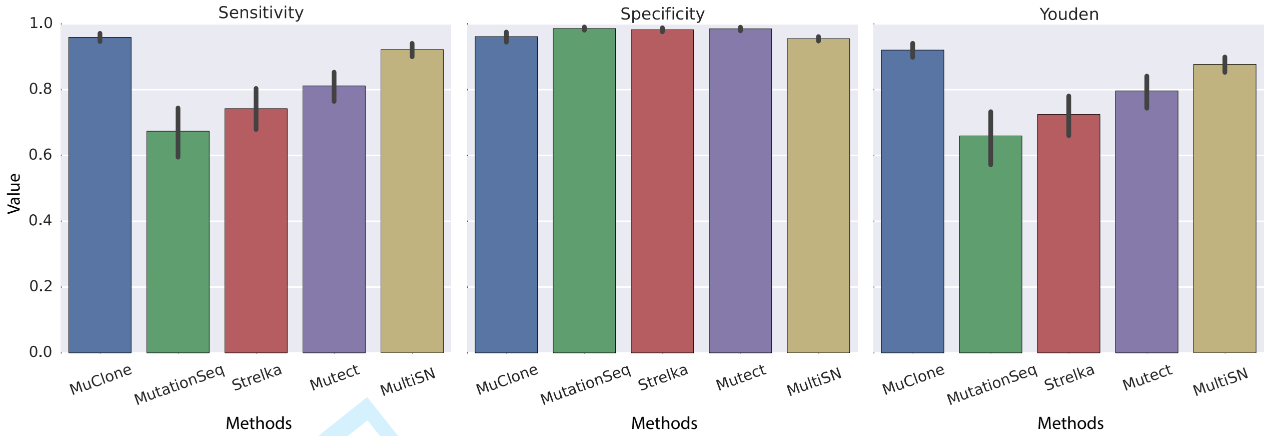


Fig. 6. The comparison of the sensitivity, specificity and Youden's index of different mutation detection methods. MuClone's parameters: Wildtype prior is 0.5, Phi_threshold is 0.03, error rate is 0.01, and precision parameter equals 1000. 88% of the elements are diagonal.

here will be incorporated into more analytically comprehensive modelling of whole genome sequencing data when multiple samples are used to infer properties of clonal dynamics. We suggest the next steps are a unified, iterative algorithm that alternates between identifying phylogenetic structure of the constituent clones comprising each tumour sample, and detection of mutations leveraging the new phylogenetic structure. As sequencing costs continue to decrease (e.g. with Illumina's X platform), multi-sample whole genome sequencing of tumours will continue to proliferate as a viable experimental design. Thus, MuClone's model will be an asset in the battery of analytical methods deployed to interpret evolutionary properties of cancer and to gain insights into clonal dynamics in time and space.

ACKNOWLEDGEMENTS

This work is supported by long-term funding support from the British Columbia Cancer Foundation. SPS and ABC received funding from a Natural Sciences and Research Council of Canada Frontiers grant #448167-13. SPS received a Canadian Institutes for Health Research Foundation award #FDN 143246 to support this work. SPS holds a Canada Research Chair and is a Michael Smith Foundation for Health Research scholar.

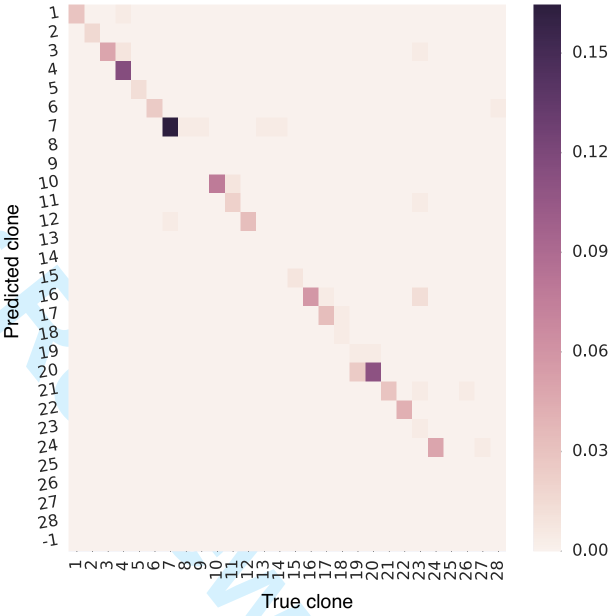


Fig. 7. Classification of 413 mutations of patient 1 across 6 samples. Bin (i, j) shows the number of mutations in clone i which were classified in class j by MuClone, divided by the total number of mutations. MuClone's parameters: Wildtype prior is 0.5, Phi_threshold is 0.03, error rate is 0.01, and precision parameter equals 1000.

REFERENCES

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, **31**(3), 213–9.

De Bruin, E. C., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A. J., Grnroos, E., Muhammad, M. A., Horswell, S., Gerlinger, M., Varela, I., Jones, D., Marshall, J., Voet, T., Van Loo, P., Rassl, D. M., Rintoul, R. C., Janes, S. M., Lee, S.-M., Forster, M., Ahmad, T., Lawrence, D., Falzon, M., Capitanio, A., Harkins, T. T., Lee, C. C., Tom, W., Teefe, E., Chen, S.-C., Begum, S., Rabinowitz, A., Phillimore, B., Spencer-Dene, B., Stamp, G., Szallasi, Z., Matthews, N., Stewart, A., Campbell, P., and Swanton, C. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, **346**(6206), 251–256.

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. (2015). Phylowgs: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, **16**(1), 1–20.

Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M. A., Condon, A., Aparicio, S., and Shah, S. P. (2012a). Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics*, **28**(2), 167–75.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., McLellan, M. D., McMichael, J. F., Wallis, J. W., Lu, C., Shen, D., Harris, C. C., Dooling, D. J., Fulton, R. S., Fulton, L. L., Chen, K., Schmidt, H., Kalicki-Veizer, J., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Wendl, M. C., Heath, S., Watson, M. A., Link, D. C., Tomasson, M. H., Shannon, W. D., Payton, J. E., Kulkarni, S., Westervelt, P., Walter, M. J., Graubert, T. A., Mardis, E. R., Wilson, R. K., and DiPersio, J. F. (2012b). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**(7382), 506–10.

Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., Laks, E., Biele, J., Shumansky, K., Rosner, J., McPherson, A., Nielsen, C., Roth, A. J. L., Lefebvre, C., Bashashati, A., de Souza, C., Siu, C., Aniba, R., Brimhall, J., Oloumi, A., Osako, T., Bruna, A., Sandoval, J. L., Algara, T., Greenwood, W., Leung, K., Cheng, H., Xue, H., Wang, Y., Lin, D., Mungall, A. J., Moore, R., Zhao, Y., Lorette, J., Nguyen, L., Huntsman, D., Eaves, C. J., Hansen, C., Marra, M. A., Caldas, C., Shah, S. P., and Aparicio, S. (2015). Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, **518**(7539), 422–426.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A., and Swanton, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, **366**(10), 883–892.

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N. D., Kanchi, K. L., Maher, C. A., Fulton, R., Fulton, L., Wallis, J., Chen, K., Walker, J., McDonald, S., Bose, R., Ornitz, D., Xiong, D., You, M., Dooling, D. J., Watson, M., Mardis, E. R., and Wilson, R. K. (2012). Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*, **150**(6), 1121–34.

Goya, R., Sun, M. G. F., Morin, R. D., Leung, G., Ha, G., Wiegand, K. C., Senz, J., Crisan, A., Marra, M. A., Hirst, M., Huntsman, D., Murphy, K. P., Aparicio, S., and Shah, S. P. (2010). Snvmix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**(6), 730–6.

Josephidou, M., Lynch, A. G., and Tavaré, S. (2015). multisnv: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. *Nucleic Acids Research*, **43**(9), e61.

Kim, S., Jeong, K., Bhutani, K., Lee, J., Patel, A., Scott, E., Nam, H., Lee, H., Gleeson, J. G., and Bafna, V. (2013). Virmid: accurate detection of somatic mutations with sample impurity inference. *Genome biology*, **14**(8), R90.

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., and Ding, L. (2009). Varscan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**(17), 2283–5.

McPherson, A., Roth, A., Laks, E., Masud, T., Bashashati, A., W. Zhang, A., Ha, G., Biele, J., Yap, D., Wan, A., M. Prentice, L., Khattra, J., Smith, M., Nielsen, C., Mullaly, S. C., Kalloger, S., Karnezis, A., Shumansky, K., Siu, C., Rosner, J., Li Chan, H., Ho, J., Melnyk, N., Senz, J., Yang, W., Moore, R., Mungall, A., Marra, M. A., Bouchard-Co te, A., Gilks, C. B., Huntsman, D. G., McAlpine, J. N., Aparicio, S., and Shah, S. P. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*. (accepted manuscript).

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

Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R. B., and Batzoglou, S. (2015). Fast and scalable inference of multi-sample cancer lineages. *Genome Biology*, **16**(1), 1–17.

Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., Bashashati, A., Hirst, M., Turashvili, G., Oloumi, A., Marra, M. A., Aparicio, S., and Shah, S. P. (2012). Jointsnvmix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, **28**(7), 907–13.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2014). Pyclone: statistical inference of clonal population structure in cancer. *Nature Methods*, **11**(4), 396–8.

Roth, A., McPherson, A., Laks, E., Biele, J., Yap, D., Wan, A., McAlpine, J. N., Aparicio, S., Bouchard-Co te, A., and Shah, S. P. (2016). Simultaneous inference of clonal genotypes and population structure from single cell tumour sequencing. *Nature Methods*. (accepted manuscript).

Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., Steidl, C., Holt, R. A., Jones, S., Sun, M., Leung, G., Moore, R., Severson, T., Taylor, G. A., Teschendorff, A. E., Tse, K., Turashvili, G., Varhol, R., Warren, R. L., Watson, P., Zhao, Y., Caldas, C., Huntsman, D., Hirst, M., Marra, M. A., and Aparicio, S. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**(7265), 809–13.

van Rens, K. E., Mäkinen, V., and Tomescu, A. I. (2015). Snv-ppilp: refined snv calling for tumor data using perfect phylogenies and ilp. *Bioinformatics*, **31**(7), 1133–5.

Yachida, S., Jones, S., Ivana, B., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., and Iacobuzio-Donahue, C. A. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, **467**(7319), 1114–7.