Clinical utility of subclonal evolution inferred using liquidCNA in metastatic breast cancer

Dohun Lee 이도헌

University of Cambridge

Thesis submitted for MPhil in Computational Biology

Word count: 9590

Acknowledgements

I would to like to thank Solon Karapanagiotis and Oscar Rueda. They showed me great kindness and their guidance helped me throughout the duration of the project.

Abstract

Background: Liquid biopsy and the analysis of cell free DNA (cfDNA) in the blood offer a minimally invasive method to monitor cancer evolution. Recently, liquidCNA, a computational algorithm that infers subclonal architecture has been released. Unlike current methods which uses mutation profiles, liquidCNA utilises information on somatic copy number aberrations from low-pass whole genome sequencing (lpWGS) 0.1x. LiquidCNA offers an economical method to characterise subclonal events. The purpose of this study was to assess the clinical utility of liquidCNA's output in monitoring response to therapy.

Methods: Data consisted cfDNA from 283 plasma samples from 80 advanced metastatic breast cancer patients. Using liquidCNA, tumour-fraction and the size of emergent subclones were estimated for each of the samples. These estimations were analysed with radiographic response as evaluated by Response Evaluation Criteria In Solid Tumours (RECIST).

Results: The size of emergent subclones were non-significantly associated with progression (p-value = 0.883) and metastasis (p-value = 0.302). Inferred subclonal events were also weak predictors of recurrent metastasis (p-value = 0.334, HR = 1.243, CI = 0.799–1.935). Our study showed the non-significant results to arise from limitations of the algorithm regarding tumour-fraction estimation. Additionally, the study highlighted prominent limitations of liquidCNA and provided modifications to the algorithm that can be implemented by future users.

Contents

Chapter 1

Introduction

In recent years, liquid biopsy has emerged as a promising alternative to traditional tissue biopsy. In particular, the analysis of cell free DNA (cfDNA) in the blood provides a minimally invasive method for cancer diagnosis, prognosis and treatment guidance. The quick, inexpensive and non-invasive nature of liquid biopsy makes cfDNA analysis particularly useful for clinical settings.

Cancer is an evolutionary process driven by clonal expansions of subclones. A tumour's subclonal architecture and the resulting intratumour heterogeneity (ITH) underlie the processes of disease progression, metastasis and therapy resistance. The advantages of liquid biopsy allows for routine serial sampling to infer subclonal architecture and track cancer evolution throughout the course of the disease. This is clinically informative and critical in controlling therapy-induced resistance and predicting disease outcome.

Current methods characterise subclonal events through mutation analysis of cfDNA. This requires deep sequencing with depth-of-coverage $>$ 30x, raising the price of the analysis. In this study, subclonal events will be inferred from a cohort of advanced metastatic breast cancer patients using a recently released algorithm, liquidCNA. LiquidCNA is a novel method utilising low-pass whole genome sequencing (lpWGS) data and copy number (CN) information. If shown insightful, it will provide an economical method to track cancer evolution and subclonal events. This study investigates the clinical utility of the method in monitoring and predicting disease outcomes. For the remaining sections of this

chapter, current literature and the motivations for the study will be introduced.

1.1 ctDNA as a cancer marker

Cells release fragmented nucleic acid during apoptosis, necrosis and active secretion [1]. These cell-free DNA are found in various biofluids including blood plasma, saliva and urine [2]. In 1989, Stroun et al. showed a subset of cfDNA in cancer patients to be tumour derived [3]. This subset was later termed circulating tumour DNA (ctDNA). In the following years, various genetic tumour markers including KRAS mutations were detected in ctDNA then matched to tumour tissue for a wide range of cancers [4]. In addition to mutations, cancer-associated genetic alterations including copy number aberrations (CNAs), aneuploidy, epigenetic alterations and rearrangement were also detected in ctDNA [1]. This has lead ctDNA to become a highly specific cancer marker.

Liquid biopsy involves the detection and molecular profiling of ctDNA to characterise cancer non-invasively. Characterising the above mentioned cancer-associated genetic alterations for diagnosis, cancer staging, prognosis and therapy guidance [2]. Over the years, there has been increasing confidence for liquid biopsy as an effective alternative to tissue biopsy. In principle, liquid biopsy captures the same genetic information whilst providing numerous advantages over the traditional tissue biopsy [5]. In the next section (1.2), the most prominent advantages of liquid biopsy will be discussed.

1.2 Liquid biopsy versus tissue biopsy

Minimally invasive. Liquid biopsy is minimally invasive compared to tissue biopsy [5]. This allows for serial sampling and routine monitoring for therapy response, metastasis, relapse and therapy-induced resistance [2]. Liquid biopsy is particularly useful for cancers with low accessibility; cancer such as non-small-cell lung cancer (NSCLC) which has up to 31% of its patients

with inaccessible tumour [6].

Global picture of cancer. Capturing tumour heterogeneity has been a consistent problem in tissue biopsy [7]. Multiple tissue biopsy conducted on the same patient showed both intratumour (i.e., across different regions of the same tumour) and intertumour (i.e., between primary tumour and a distant tumour) heterogeneity to be present in cancer [8]. This poses a problem as a single tissue biopsy will provide a limited representation of the cancer's genomic landscape; introducing a genomic bias when designing personalised-genotype-matched therapy [2]. Furthermore, multiple tissue biopsy is not only invasive but also typically infeasible due to time, cost and tissue accessibility.

Liquid biopsy better captures tumour heterogeneity [9]. This is because sampled ctDNA can originate from any tumour cell in the body. Thus, the genomic landscape from ctDNA analysis will reflect contributions from ITH and distant sites [10, 11]. For instance, ctDNA analysis from synchronous breast and ovarian cancer captured CNAs unique to both cancers [12].

Real-time picture of cancer. ctDNA sampled from the blood are fresh tumour-derived material which provides a real-time image of cancer's genetic landscape. This is the case as cfDNA are quickly removed from circulation [13]. They have a short half-life of 16 minutes to two hours, and are actively degraded by the body through DNase I activity and renal excretion into the urine [14, 15]. Furthermore, unlike tissue biopsy, the tumour material is not disturbed by preservatives.

Quick. Liquid biopsy has a substantially shorter turnover speed compared to tissue biopsy. Median time for assay is 11 days versus the 33 days for tissue biopsy [16, 17]. Such gain in speed is particularly important in clinical settings, allowing for quicker implementation of genomic-matched therapy. Shortening the time for implementation by 83%, from 5.9 months to just one month [16, 17].

Sensitivity and concordance. Liquid biopsy provides these advantages whilst having high sensitivity and concordance with respect to tissue biopsy. For instance, a blinded study comparing liquid and tissue biopsy in metastatic colorectal cancer showed the former to have 92% and 100% sensitivity for KRAS and BRAF mutations, respectively [18]. These mutations had 96% and 100% concordance with respect to paired tissue analysis.

In prostate cancer, 100% of the somatic mutations identified in matched metastatic tissue were concurrently identified in ctDNA analysis [19]. The shared mutations also showed similar hierarchy of variant allele fractions across the two biopsies. Additionally, CNAs also showed high correlation, with 88.9% concordance in clinically actionable genes. Such high sensitivity and concordance of ctDNA analysis has been demonstrated for a wide range of cancers.

With the advantages it provide and the high concordance it has, cfDNA is the ideal biomarker to assess cancer genomics. Importantly, with it being non-invasive, quick and inexpensive, cfDNA analysis can be easily implemented into clinical settings.

1.3 Current status of liquid biopsy

cfDNA analysis has been demonstrated to be useful across different areas of clinical settings. From initial diagnosis to monitoring of residual disease after therapy, ctDNA analysis is highly applicable throughout cancer's disease course. In this section, the current status of ctDNA analysis in research and application will be discussed.

Diagnosis. ctDNA analysis has been demonstrated to have high potential as a diagnostic tool [20, 21]. The assay has diagnostic power comparable to tissue biopsy and even allows for earlier diagnosis by up to two years [22]. However, cfDNA are yet to be used for diagnosis clinically. Currently, the main challenge is the detection of cancer-associated alterations in cfDNA sample. On one side, research has shown ctDNA to be detectable from tumours 50 million cells small [23]. However, on the other hand, ctDNA availability is also highly variable across different cancer types, amongst patients and disease stages [2]. Furthermore, there may be biological noise arising from mutated cfDNA of non-tumourigenic cells as well. For instance, even though leukaemia-associated mutations were found in 10% of individuals aged over 65, the absolute risk of cancer development was only 1% [24]. Similarly, cancer-associated genomic alterations were detected in plasma of healthy individuals as well [25]. Advancements in sequencing technology are thought to improve cfDNA analyses' sensitivity.

Cancer localisation. Encoded into cfDNA are tissue-specific nucleosome and methylation patterns [26]. Thus, relative contributions of cfDNA from different tissue types in the total sample can be quantified [26]. This is useful for tumour localisation, particularly when the primary site is unknown. Though it has not yet been demonstrated, this method may also allow for metastatic site localisation.

Disease burden and staging. The level of ctDNA correlates with the burden and progression of the disease. For instance, ctDNA levels significantly correlated with disease volume in patients of relapsed high-grade serous ovarian cancer and NSCLC [27, 28]. Similarly, plasma ctDNA levels were higher in advanced and metastatic cancer than in local diseases. In fact, ctDNA levels correlated with different stages of cancer. A 100-fold increase in median ctDNA concentration was observed between stage I and stage IV across 640 patients of diverse types and progressions of the disease [29]. This translates to an increase from $\sim$10 copies per 5ml of plasma to 100$\sim$1000 copies. Thus, ctDNA can be utilised for cancer staging.

Prognosis. Once ctDNA was established as a marker for assessing disease burden, ctDNA's utility for prognosis was studied soon after. Studies demonstrated patients with detectable ctDNA to have worst survival outcome relative to those without. For example, in a two year study done by Lecomte et al., patients of colorectal cancer with detectable ctDNA had 48% overall survival rate compared to the 100% of patients without any detected ctDNA [30]. As such, there is an inverse correlation between ctDNA levels and clinical outcome. In metastatic breast cancer, ctDNA concentration show a significant inverse correlation with overall survival up to 2000 copies per millimetre, with universally poor prognosis beyond this level [31].

Furthermore, ctDNA levels show correlation with therapy response and radiographic outcomes as well. Patients without clinical benefits show increased ctDNA levels, whilst the levels are stable or reduced for patients with amiable therapy response [32, 33]. Of note, ctDNA clearance precedes radiographic response, highlighting ctDNA as a real-time cancer marker [34].

Treatment guidance. Treatment guidance is one of the areas in which ctDNA analysis is actively used in the clinic. Currently, European Medicines Agency (EMA) and the US Food and Drug

Administration (FDA) approved usage of ctDNA analysis to select for NSCLC treatment when tissues are inaccessible and/or tissues biopsies are not evaluable [35, 36]. Furthermore, under the current guidance, liquid biopsy may negate the need for tissue biopsy if ctDNA are detected. However, tissue biopsy may still be required even if ctDNA are undetected. As such, though ctDNA are not fully replacing tissue biopsy, the confidence towards ctDNA analysis is increasing.

Surveillance. The non-invasive nature of liquid biopsy allows for serial sampling. This allows routine monitoring for residual disease, metastasis and resistance emergence during and after therapy [1]. For instance, recurrence through residuals can be detected through 'ctDNA relapse'. This is where ctDNA is detected after therapy completion and clearance. CtDNA relapse precedes imaging response and allows for earlier assessment of response [37]. As such, because ctDNA show a real-time snapshot of cancer, it is a prospective tool to combat resistance. It will allow for admitted therapy to be dynamically adapted in response to resistance emergence.

From diagnosis to surveillance, ctDNA analysis has been demonstrated to be a tool with wide applications. A key clinical application yet discussed is the use of ctDNA to infer clonal evolution and characterise tumour's subclonal architecture. In the following sections, the importance of understanding subclonal architecture and the opportunities provided by liquid biopsy regarding it will be discussed.

1.4 Cancer as an evolutionary process and subclonal architecture

The evolutionary process of cancer is parallel to Darwinian natural selection [38]. Cancer is initially founded by a single cell that grows and proliferates while accumulating mutations. Subset of the tumour cells with evolutionary advantageous 'driver' mutations will proliferate more than others. This evolutionary process known as clonal expansion gives rise to subclones - genetically related but distinct subpopulations of cancer cells in a tumour [38]. Clonal expansion results a tumour to be comprised

of multiple subclones and gives rise to intratumour heterogeneity [39].

Under the selective pressure of therapy, subclones are key ingredients for tumour evolution. A recent study showed chemotherapy to drive phenotypic changes in metastatic breast cancer [40]. The altered phenotypes were associated with enhanced drug resistance and immune system avoidance. Importantly, these 'novel' phenotypes were traced back to pre-treatment subclones that were now dominant in the tumour. This highlights the clinical importance of understanding tumour evolutionary history and clonal architecture: therapy-driven subclone emergence must be addressed to combat resistance emergence and treatment failure.

1.5 Benefits of deriving subclonal architecture from liquid biopsy

Liquid biopsy may be the optimal tool to track clonal evolution and infer subclonal architecture. With the advantages of liquid biopsy, longitudinal ctDNA data can be easily obtained across patients' disease course. If subclonal architecture can be accurately characterised from these ctDNA, emerging subclones and their putative resistance will be detected. In addition to tackling therapy resistance, understanding cancer phylogeny will be informative when predicting and managing metastasis and disease progression. As such, the inference of clonal evolution from ctDNA will be a quick and inexpensive approach that will provide a wide host of clinically actionable information.

Cancer phylogeny and subclonal architecture has been successfully inferred from ctDNA. In a study by Murtaza and colleagues, a patient of metastatic breast cancer was tracked over a period of three years. From the patient, samples were rountinely obtained from both liquid and tissue biopsy. The study showed mutation levels analysed from ctDNA to reflect subclonal architecture characterised from multiregional tumour biopsies [41]. As was the case for this study, current analysis of ctDNA infers subclonal architecture from mutation groups and their mutation allele frequency. This methodology requires deep amplicon sequencing with $>$ 30x depth of coverage. This results the analysis

to be too expensive for wide clinical application. Additionally, the use of selective few mutations may introduce bias in the selection and efficacy of genotype-matched therpay [2].

LiquidCNA, a computational algorithm, has recently been released to infer subclonal architecture from lpWGS [42]. Unlike current methods, liquidCNA uses CN information to estimate the size of the most dominant emerging subclone. SCNAs can be profiled to high precision from lpWGS data with a shallow depth-of-coverage of 0.1x. With the shallow depth, sequencing becomes considerably cheaper than the current methods. Furthermore, by using CN profiles, it evaluates subclones from the whole genome and is less prone to bias as with mutation. LiquidCNA is a novel tool that, if shown insightful, will provide an inexpensive method to analyse subclonal architecture from liquid biopsy samples.

Aim of the project: In this project, we have longitudinal ctDNA samples from metastatic breast cancer patients. We will utilise liquidCNA to infer the size of the emerging subclone in each sample. With the outputs from the algorithm, we will be able to track the dynamics of subclonal architecture and infer cancer evolution. We will test this information's prowess as a prognostic tool, evaluating their effectiveness in predicting disease progression and metastasis.

Chapter 2

Materials and methods

2.1 Data set

Data set consisted of aligned lpWGS (0.1x) data from 80 metastatic breast cancer patients [43]. For each patient, samples were collected across multiple timepoints — this amounted to 283 time samples in total. Patients had mean and median time samples of 3.5 and 3.0, respectively; they had a maximum and minimum of 11 and 2 time samples, respectively. The samples had read counts with mean of 3.3 million reads and median of 3.0 million reads. Reads aligned to the reference human genome GRCh38/hg38 were provided for this project. Additionally, each patient had complementary progression evaluation based on radiological assessments according to Response Evaluation Criteria In Solid Tumours (RECIST) guidelines.

2.2 QDNAseq: computing copy number profile

Depth of coverage CN profiles were computed using the R package QDNAseq [44]. In QDNAseq, the human reference genome was binned into non-overlapping bins of 500kb. Bins with less than 75% mappability (defined as average mappability of 50-mers within a bin) and those in the ENCODE blacklisted regions were filtered out. Then, for each bin, sequence read counts were

computed. Additionally, QDNAseq uses R package DNAcopy for segmentation [45]. Here, bins with contiguous copy number values were grouped together. Thus, QDNAseq outputted two files for each sample: one raw bin-wise CN values and one segmented CN values.

2.3 Pre-processing the CN profile for liquidCNA

For each patient, computed raw and segmented CN profiles for each of their time samples are inputted into liquidCNA. Prior to estimating subclonality, the CN profiles are filtered, renomarlised and recentred following the liquidCNA pipeline. Of note, preceding the estimations, segments are redefined such that their boundaries are constant across all time samples. Thus, for each patient, all time samples have equal number of these 'ensemble segments' — being equal in length and in the same genomic loci, but differing in their respective CN values. Ensemble segments are used by liquidCNA to estimate subclonality.

2.4 liquidCNA: Estimation of subclonality

R package liquidCNA is a computational algorithm that infers and tracks subclonal dynamics from longitudinally collected cfDNA samples [42]. LiquidCNA considers SCNA to be the dominant evolutionary force driving cancer evolution. Thus, it estimates the abundance of the most dominant emerging subclone (from now on defined as subclonality) from the CN profile. In this section, steps of liquidCNA subclonality estimation will be discussed.

2.4.1 cfDNA originates from three cell populations

LiquidCNA assumes that cfDNA are released from three distinct populations: normal cells (N), ancestral tumour cells (A) and emerging subclonal cells (S). The proportion of ctDNA (released from populations A&S) in the cfDNA pool is defined to be the tumour fraction (i.e., purity). For each time sample i, the proportion of these three populations will depend on purity (p_i) and subclonality (r_i):

$$N_i = 1 - p_i; \qquad A_i = p_i \cdot (1 - r_i); \qquad S_i = p_i \cdot r_i \tag{2.1}$$

As SCNA are considered to be the dominant evolutionary force, the three population are assumed to have distinct segment CN values. Thus, for a given segment j, CN state will be defined by each population's CN value for segment j and their population proportion. As normal cells are in diploid state, $C(N)^j = 2$ for all j. Then, the CN value of segment j in time sample i (i.e., C_i^j) will be defined as:

$$C_i^j = 2 + p_i((1 - r_i)C(A)^j + r_i C(S)^j - 2) \tag{2.2}$$

2.4.2 Segment classification

In liquidCNA, each segment is classified into one of three categories: clonal, subclonal or unstable. Segment classification depends on segments' CN values in the two tumour populations (A and S). Clonal segments are segments shared across the tumour populations; they have the same copy number in ancestral and subclonal populations, $C(A)^j = C(S)^j$. For clonal segments, CN will change across time samples following:

$$C_i^j = 2 + p_i(C(A)^j - 2) \tag{2.3}$$

Subclonal segments are segments specifically associated with the emerging subclone. These segments would have CNA driving the population's evolution; for instance, having CNA associated with therapy resistance. Thus, subclonal segments differ in their CN

between the two populations, $C(A)^j \neq C(S)^j$. For subclonal segments, CN will change across time samples following:

$$C_i^j = 2 + p_i(C(A)^j - 2 + r_i(C(S)^j - C(A)^j)) \tag{2.4}$$

Finally, unstable segments represent genomic regions with unreliable measurement or instability arising from ongoing CNA. Unstable segments does not depend on subclonality r_i but is instead represented by a time sample dependent tumour wide CN value, $\zeta(T)_i^j$. For unstable segments, CN will change across time samples following:

$$C_i^j = 2 + p_i(\zeta(T)_i^j) \tag{2.5}$$

Using Equation 2.3 and 2.4, subclonality is computed in four steps.

Step 1: Purity estimation. The first step in calculating subclonality is purity estimation. In a tumour with newly emerging subclone, majority of the segments are expected to be clonal and follow Equation 2.3. In the equation, $C(A)^j$ can only be integer values. Thus, the CN distribution are expected to peak at regular intervals of purity. Following this, liquidCNA estimates purity as the value which minimises the squared distance between the observed peaks and the expected integer peaks. Once purity values are estimated, samples' CN values are corrected by their respective purity (see Equation 2.6). This corrects for the contamination of normal cells.

$$\hat{C_i^j} = (\frac{1}{p_i} \cdot (C_i^j - 2) + 2) \tag{2.6}$$

Step 2: deltaCN. Once CN values are corrected by purity, the difference in CN (dCN) with respect to a designated baseline sample is calculated. Baseline samples are those assumed to have minimal and negligible proportion of the emerging subclone. This, for example, could be a sample collected at the point of diagnosis or prior to commencing therapy. dCN of purity-corrected CN values clearly shows genomic regions of CN gains and losses, thereby highlighting subclone associated segments. In our analysis, for each patient, time sample with the lowest mean purity-corrected segment CN value was chosen as baseline. Biologically, it would be more appropriate to choose the sample

from patients' first blood draw. However, as it will be discussed in Step 3, the algorithm does not consider chronological ordering of samples when analysing across longitudinal data. By choosing the sample with the smallest mean CN, the algorithm is able to identify subclonal segments more robustly.

Step 3: Segment classification. Using the computed dCNs, segments are classified into one of the three segment categories. Clonal segments are expected to have dCN values of 0. To account for measurement noise, segments with a standard deviation below a threshold ($\eta = 0.05$ for our analysis) are classified to be clonal. From the remaining non-clonal segments, subclone-specific segments are segregated from the unstable segments. For this, liquidCNA assumes that subclonal population increases over time points due to an evolutionary advantage over the ancestral cells. Under this assumption, subclone-associated segments will display either a monotonic increase or decrease over time. A segment following a monotonic pattern is classified subclonal, and those that don't are classified unstable.

However, subclones do not always follow this assumption of increasing population. Due to therapy or other evolutionary forces dictating over the tumour, the population size will fluctuate. Thus, liquidCNA rearranges the time samples into an ordering that follows this assumption. All ordering permutations are exhaustively computed, with each segment classified either subclonal or unstable for each sample ordering. Then, the sample order with the maximum number of subclonal segments is chosen to be the optimal sample order.

Step 4: Subclonal estimation. Segments classified subclonal are used to estimate the proportion of subclonal population within the tumour. With these segments, r_i is calculated following:

$$\hat{\Delta C_i^j} = r_i(C(S)^j - C(A)^j) \tag{2.7}$$

To account for measurement noise in dCN values, mixture of Gaussian distributions is fitted over the values. Here, each mixture represents a CN state and their means are constrained by r_i. Thus, the constrained mean of the best fitting Gaussian mixture model is defined as the size of the subclonal population.

2.5 Analysis of recurrent metastasis

Survival analysis was conducted to statistically analyse the occurrence of new metastasis. Unlike death, many clinical events (including metastasis) can occur numerous times for a patient. Despite this, most methods of survival analysis only consider the time to the first event and discard further recurrences. This makes popular methods such as the Kaplan-Meier method inappropriate for our data of metastasis. Thus, Andersen-Gill (AG) model was used for our analysis. AG model is an extension of the Cox proportional hazards model optimised for recurrent time-to-event data.

In our data, event of interest is metastasis. More specifically, as all our patients have metastatic breast cancer, an event was defined as non-primary metastasis detected at least 14 days after the initial metastasis. RECIST evaluation of metastasis was available for 73 of the 80 patients. Across these patients, there were 39 new metastatic events in total. Six of these events were recurrent. For the analysis, patients were right censored at their last CT scan date.

Measured time-varying covariates are included in the AG model. Time-varying covariates are predictor variables that could change between follow ups (i.e., over time). AG model considers recurrent events to be mediated by these covariates. Our study has two time-varying covariates: subclonality and purity estimated from liquidCNA. Prior to model fitting, both estimates were categorised into two groups. Subclonality estimates were grouped into either high or low subclonality. For the grouping, mean of all estimates were defined as threshold (0.3657). Subclonalities larger than this threshold were categorised 'high-subclonality' and smaller estimates were categorised 'low-subclonality'. For purity, mean of all estimates were again defined to be the threshold (0.13). Any purity estimates larger than this threshold were defined 'high' and smaller were defined 'low'. With the two time-varying covariates categorised, AG model was used for survival analysis of our recurrent time-to-metastasis data.

AG model was fitted to our data using the pipeline developed by Amorim and Cai [46]. R library 'survival' was used for the analysis [47].

2.6 Data and Code availability

The data for this project are confidential and could not be provided publicly. Code used for this project are available without restrictions at https://github.com/dhldhldhl/subclonality.

Chapter 3

Results

3.1 Quality check liquidCNA output

3.1.1 LiquidCNA detects clearly marked regions of CNA

Initial quality check of liquidCNA outputs were performed. LiquidCNA estimates subclonality exclusively from CN data, identifying subclone-associated segments and utilising them for the estimation. Thus, it was first checked whether the algorithm detects clearly marked regions of amplifications and deletions across the genome, classifying them to be subclone-associated.

Figure 3.1 shows the CN profiles of the three time samples of Patient 1005. The figure shows liquidCNA to accurately classify segments. For example, the subclonal segment near bin number 300 (shown along the dotted red line) is observed to amplify along the sample order — the segment's CN state increasing from 2.1 to 2.4, then to 3.3 by Sample3.

3.1.2 Purity and subclonality have weak correlation

LiquidCNA uses initial estimations of purity when computing subclonality. Thus, the second quality check asked whether the two values showed any correlation. Square of the correlation co-

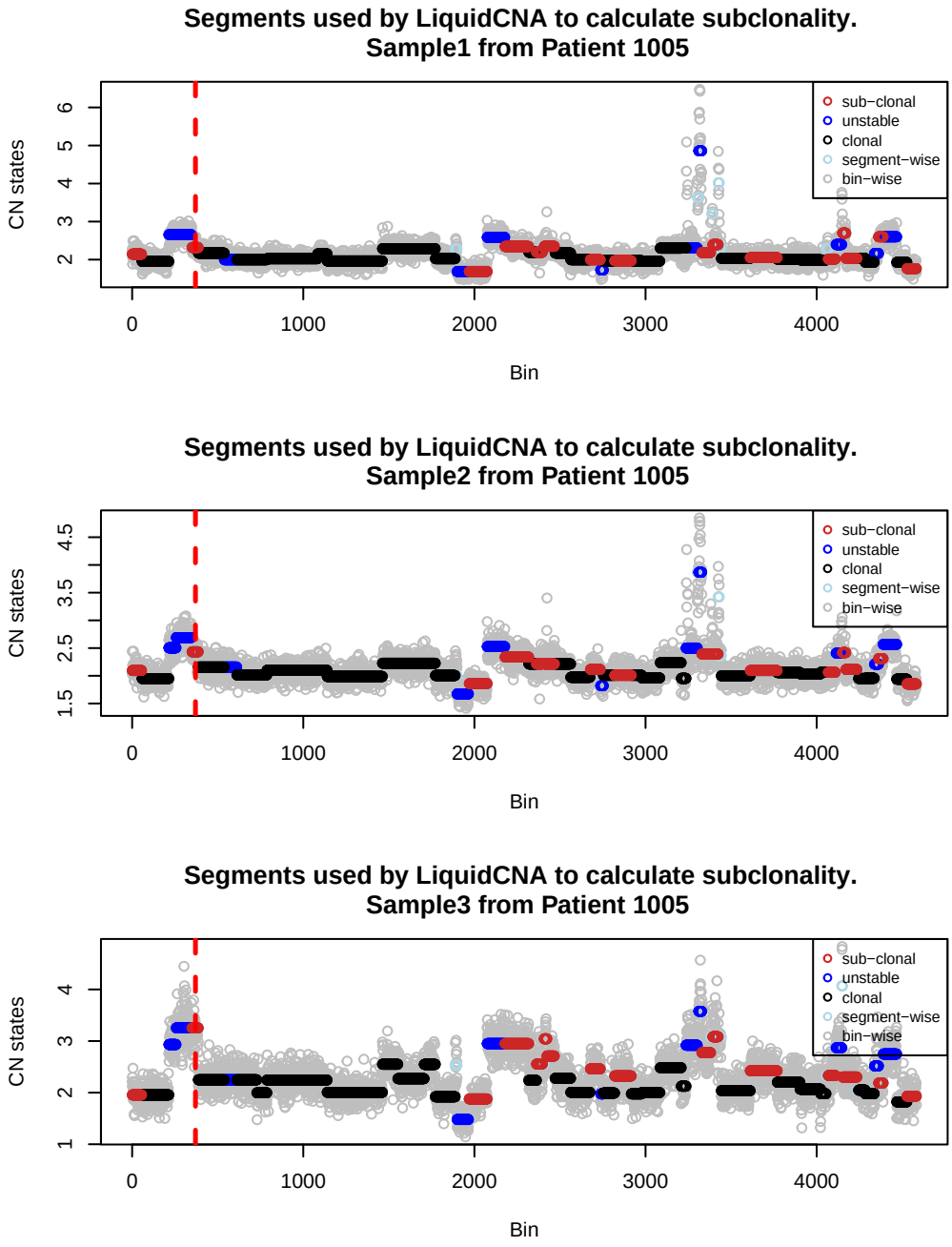


Figure 3.1: CN profiles for time samples of Patient 1005. LiquidCNA chose the following sample ordering for Patient 1005: Sample1 (as baseline), Sample2 then Sample3. Segments are coloured labelled by the algorithm's classification: clonal (black), unstable (blue) and subclonal (red). Points labelled segment-wise (in light blue) and bin-wise (in grey) are the two CN profiles outputted from QDNAseq.

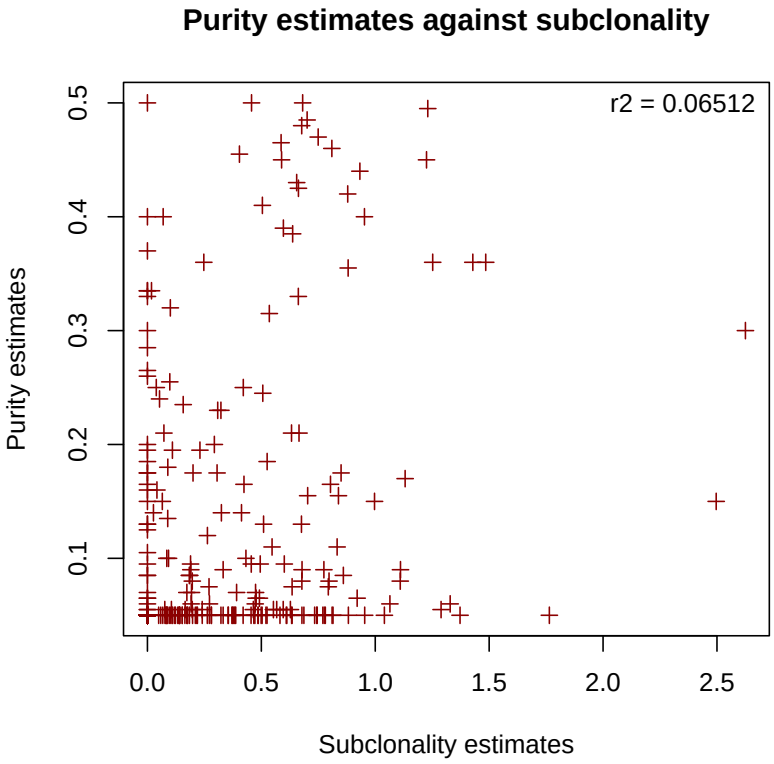


Figure 3.2: LiquidCNA estimates of purity and subclonality plotted against each other. With a r^2 value of 0.065, the two show low correlation.

efficient (r^2) was calculated and a scatterplot was created (Figure 3.2). The figure, alongside the low r^2 value of 0.065, shows the two estimates to have low correlation. This indicates subclonality (r_i) to be a tumour characteristic distinct from purity (p_i).

Two additional points are noted from Figure 3.2. First, there are subclonality estimates greater than one. Second, there is a clustering of samples with small purity estimates ($p_i < 0.1$). These two points will be discussed and addressed in the following sections.

3.2 Practical limitations of liquidCNA

Running the algorithm for our 80 patients, two practical limitations of liquidCNA became apparent. First, liquidCNA was incompatible for patients with two time samples. This was because

the algorithm classifies segments based on changes in dCN across time samples. For patients with two samples, their segments had only one dCN value. This meant that non-clonal segments could not be further classified subclonal or unstable. Second, in the other extreme, liquidCNA was incompatible for patients with more than six time samples. This was for two reasons, with the first being computational complexity. During the segment classification step, the algorithm exhaustively computes through all permutations of sample order. For patients with many time samples, the time complexity of this step became too large. For instance, segment classification is computed for nearly 40 million permutations for Patient 840 with 11 time samples. Second, and more importantly, even if time complexity was not an issue, rarely any of the segments are classified subclonal. This is because with a large number of dCN, segments do not follow the monotonic pattern of subclones across all time samples. This resulted none of the segments to be classified subclonal and subclonality could not be estimated. Of our 80 patients, 22 patients had two time samples and seven patients had over six samples. That is over a third of the patients for which the algorithm was incompatible. As it can be seen, these two short comings severely limits the use-case scenario for the algorithm. In the next section, modifications made to optimise the algorithm for these patients will be discussed.

3.2.1 Modifying liquidCNA for two time sample patients

The main problem for two sample patients was the inability to classify non-clonal segments. The best option was to omit the classification step and estimate subclonality using all non-clonal segments. Thereby, an estimation for the one of the two time samples (i.e., the non-baseline sample) can be made. Whilst this is an oversimplification of the algorithm, we believe there is still value in the output. First, as the estimation uses non-clonal segments, the estimation is based on genomic regions that differ from ancestral tumour cells. This estimation, however, will be of lower confidence and precision as it encompasses measurement error and genomic instability represented by unstable segments. However, despite the lower precision, the estimation will still be able to depict the trend of change in subclonality over the two time points.

3.2.2 liquidCNA for 6+ time sample patients

Subclonality was estimated for patients with 6+ time samples by running liquidCNA in batches. Time samples were chronologically ordered then equally divided into two batches. For patients with odd number of time samples, samples were batched such that the first batch had one more sample than the second. As the maximum number of time sample was 11, each batch had less than or equal to six time samples. Three options for stitching the two batches were explored:

Option 1 runs batch1 for each patient, then sets the 'top sample' as baseline sample for batch2. Here, top sample is the last sample in the sample ordering chosen by liquidCNA for segment classification. This sample represents the final point in the dynamics of subclonality across the first batch. Therefore, it was likely to be an appropriate baseline sample for batch2. After running the second batch, the two batches were stitched together with reference to the top sample.

Option 2 uses time sample with the highest estimated subclonality from batch1 as the baseline for batch2. This option supports liquidCNA's assumption that subclonality increases due to its evolutionary advantage over ancestral cells. Like with Option 1, the two batches were appended together with reference to the overlapping time sample.

Option 3 re-uses the same baseline sample from batch1 for batch2. As liquidCNA identifies subclone segments using dCN, this option allows all time samples across both batches to compute dCN against the same baseline. Furthermore, of the three options, only Option 3 holds liquidCNA's assumption that baseline is of minimum and negligible subclonality.

Options 1 and 2 resulted identical results. For all patients the two options used the same time samples to stitch the two batches together. Comparing options 1&2 against Option 3, Option 3 yielded the best results. Looking at Figure 3.3, for Option 1&2, an abrupt change in subclonality dynamics is observed between Sample1 and Sample5 as the batch changes. This change is less prominent in Option 3. It is less prominent because Option 3 has a consistent baseline sample; which allows dCNs computed between the two batches to be in a more similar range than in Option 1&2. This subsequently makes the stitching more regular.

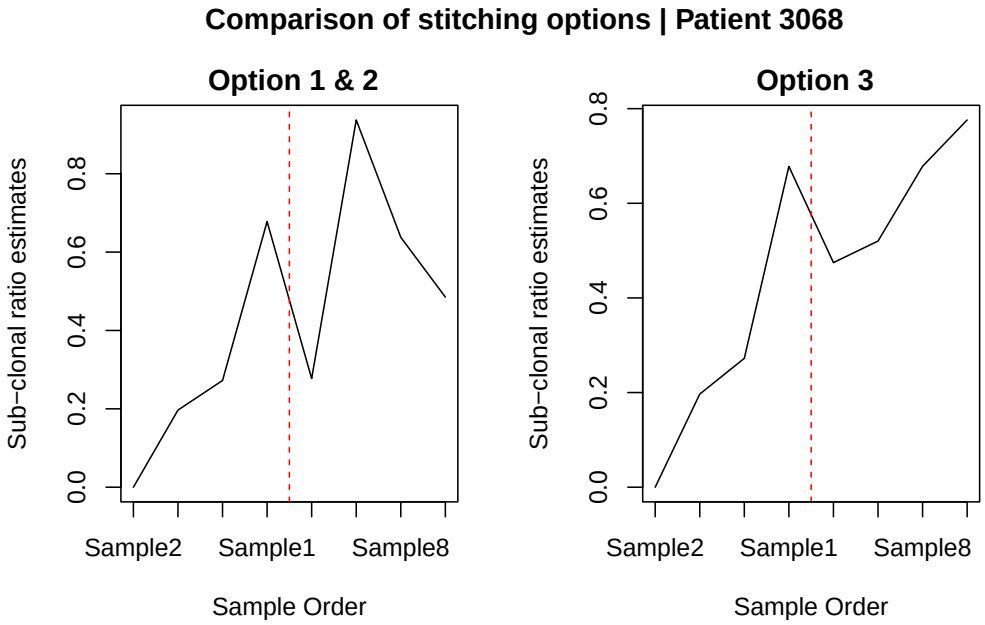


Figure 3.3: Comparing batching options explored for patients with 6+ time samples. Here, the results of the options are showed for Patient 3068. The dotted red line indicates the point where the batches change. For all options, liquidCNA chose the following sample order: Sample2, Sample3, Sample4, Sample1, Sample5, Sample7, Sample8 then Sample6.

With these results, Option 3 was chosen to batch time samples for our analysis.

Interestingly, despite having different baselines, the options have identical sample ordering. This shows how robust and confident liquidCNA is in identifying subclonal segments. However, despite the same permutation, the estimated subclonalities for batch2 are drastically different between the options. This highlights the algorithm's sensitivity and emphasises the care required when users choose their baseline samples. It also shows the limitations of using dCNs to calculate subclonality.

It must be noted that an assumption is made by batching the samples then re-stitching them. By batching the samples, sample ordering cannot be mixed across batches and time samples in batch2 are always ordered after the first batch. Thereby, we are explicitly assuming that the samples in batch2 have greater subclonality than those in batch1. This is an extension of the assumption that subclonality increases over time.

3.3 liquidCNA estimates subclonal ratio greater 1

Another prominent limitation of liquidCNA was that its subclonality estimations included values outside of the range [0,1]. In 25 out of the 283 time samples, subclonality r_i was greater than 1. These 25 samples came from 19 of the 80 patients. Estimations outside of the range are erroneous and uninterpretable as r_i was defined to be the proportion of subclonal cells within a tumour population.

Purity was likely to be one of the main causes for the erroneous estimation. This is because, aside from subclonality, purity is the major estimation computed by the algorithm. Purity is estimated upstream of subclonality and is used to correct each time sample's CN values for contamination from normal cells. These purity-corrected CN values are used to estimate subclonality. Thus, the error could have originated from either the purity-estimation step or the purity-correction step. First, to visualise any apparent relationship between purity and the erroneous values, subclonality estimates outside of [0,1] were plotted against their respective purity values (Figure 3.4).

Two points are observed from Figure 3.4. First, most erroneous estimates are between [1,2]. Second, erroneous estimates arise from two groups of purity: low purity group (with $p_i < 0.2$) and high purity group ($p_i = 0.4 \sim 0.5$). In the following subsections, reasons for liquidCNA's incorrect estimations will be explored.

3.3.1 Subclonality estimation failed at the purity-correction step due to low purity

Of the two groups of purity, low purity samples posed a larger problem. Looking at Figure 3.4, low purity samples make greater error, estimating subclonal proportions with values greater than six.

A patient with low purity and erroneous subclonality estimates was chosen. We then processed this patient through the steps of the algorithm to identify where the error was coming from. Pa-

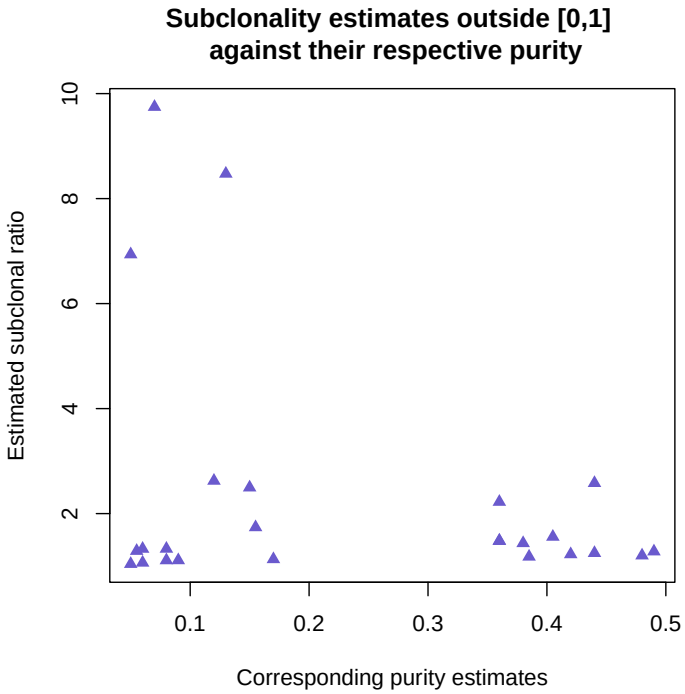


Figure 3.4: Erroneous subclonality estimates ($r_i > 1$) plotted against their respective purity estimates.

tient 3080 have two of the largest subclonality estimated (Figure 3.5). As previously mentioned, this error could have arisen from either the purity-estimation step or the purity-correction step. The latter was explored first.

In the algorithm, purity is corrected following Equation 2.6. In the equation, segment CN values are divided by purity. When the purity value is small and close to 0, the 'corrected-CN values' become large due to the division. The steps of the computation are illustrated in Figure 3.6 for six segments of Patient 3080. As seen from Fig 3.6A and Fig 3.6B, the low purity of Sample1 ($p_i = 0.05$) resulted its CN values to be corrected to drastically larger values compared to the other samples. Furthermore, as Sample1 was chosen to be the baseline for this patient, all dCNs were computed with reference to these erroneous CN values. This subsequently resulted the dCN values to be large (Fig 3.6C). As subclonality is estimated by fitting a GMM over these dCN values, the estimated subclonality was also large. This shows the purity-correction step of the algorithm to be the cause of erroneous subclonality estimations, highlighting the limitations liquidCNA face for low purity samples.

Patient 3080 estimates

	Sample order	Subclonality	Purity
(baseline)	Sample1	0.0000	0.05
	Sample3	8.4736	0.13
	Sample2	9.7463	0.07

Figure 3.5: Subclonality and purity estimated for the three time samples of Patient 3080.

(A) Segment CN values

Sample1	Sample2	Sample3
1.371	2.004	1.73
1.367	1.989	1.734
2.02	2.063	2.034
2.742	2.112	2.366
1.952	2.005	1.964
2.746	2.15	2.402

(B) Purity−corrected segment CN

Sample1	Sample2	Sample3
−10.577	2.058	−0.073
−10.666	1.847	−0.043
2.398	2.901	2.26
16.843	3.601	4.817
1.047	2.072	1.725
16.917	4.15	5.093

(C) Segment dCN values

Sample1	Sample2	Sample3
0	12.635	10.504
0	12.513	10.624
0	0.503	−0.138
0	−13.243	−12.027
0	1.025	0.678
0	−12.768	−11.825

Figure 3.6: Three steps of liquidCNA's algorithm illustrated for Patient 3080. Subfigure A shows six example segments chosen from the patient. For each sample, segments were corrected by their respective purity (subfigure B). Once purity-corrected, dCNs were computed with reference to the baseline sample (subfigure C).

3.3.2 Subclonality estimation failed in high purity samples due to low purity baseline samples.

Subclonality estimation failed for time samples with high purity estimates as well (Figure 3.4). Interestingly, these erroneous estimates were caused by the same reason, low purity. Despite the individual time samples having high purity values, estimations were erroneous when the baseline samples had low purity. As it was illustrated in Figure 3.6, low purity of the baseline resulted dCNs to be extreme. Extreme dCNs resulted erroneous subclonality estimates regardless of a sample's own purity. Following these results, a strict quality control for high purity baseline samples may be required for liquidCNA.

3.3.3 Low correlation between liquidCNA and ichorCNA estimations of purity.

Above, we illustrated how dependent the computations of subclonality are to the prior estimations of purity. This highlights the importance of having accurate purity estimates. However, without true purity values to compare to, it is difficult to test the accuracy and robustness of liquidCNA's estimations.

Figure 3.7 shows the CN profiles for two time samples of Patient 3269. These CN profiles are outputs from QDNAseq that are yet corrected by purity. Sample1 has high purity estimation of 0.48, whilst Sample3 has low purity estimation of 0.13. From looking at the figures, Sample1 with the higher purity has a more distinct CN profile with clearer signal. Conversely, Sample3 has less signal and higher measurement noise. This suggests that there may be a correlation between liquidCNA's estimation of purity and the quality of CN profile inputted into the algorithm.

The robustness of liquidCNA's purity estimation was tested by comparing it to another algorithm, ichorCNA. IchorCNA is an algorithm developed specifically to compute purity from low-coverage cfDNA samples (unlike liquidCNA which was developed for subclonality). It has been previously noted that ichorCNA makes more accurate estimation for samples with lower read count (i.e., higher measurement noise) and lower purity

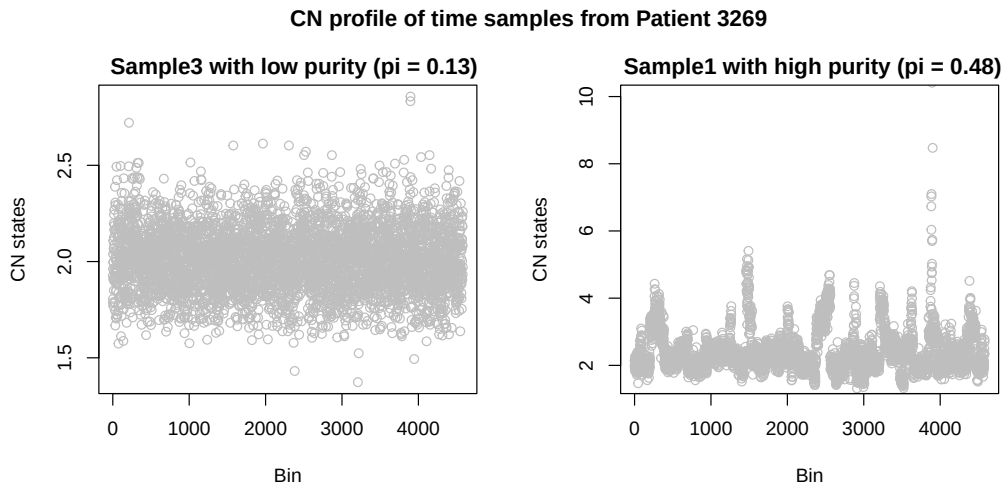


Figure 3.7: CN profiles of Sample3 ($p_i = 0.13$) and Sample1 ($p_i = 0.48$) from Patient 3269.

(for instance, for synthetic data where true purity was known). Comparing the estimates from the two algorithms, estimations show low consistency and low correlation (Figure 3.8). Looking at the figure, a vertical clustering is present along x = 0.05. In the cluster, whilst liquidCNA estimates all purity values to be 0.05, ichorCNA is observed to estimate p_i values ranging from $0 \sim 0.15$. Incidentally, purity of 0.05 is liquidCNA's lower limit of purity estimation. This further highlights the limitations of our package with regards to low purity samples. Whilst we could not evaluate which of the two algorithms were more accurate, the low correlation between the two makes it difficult to argue for liquidCNA's robustness.

3.3.4 Modifying liquidCNA to reduce erroneous estimations.

The algorithm was tweaked such that subclonality estimation is less influenced by low purity samples and the resulting extreme dCNs. In the algorithm, after dCN is computed and subclonal segments are identified, identified segments' dCN values are used to estimate subclonality. With the modification, segments with extreme dCN are filtered before the estimation. More specifically, segments with dCN > 5 in at least one of the time samples were removed. This is not an obtrusive modification to the algorithm because liquidCNA already consider only a limited number of

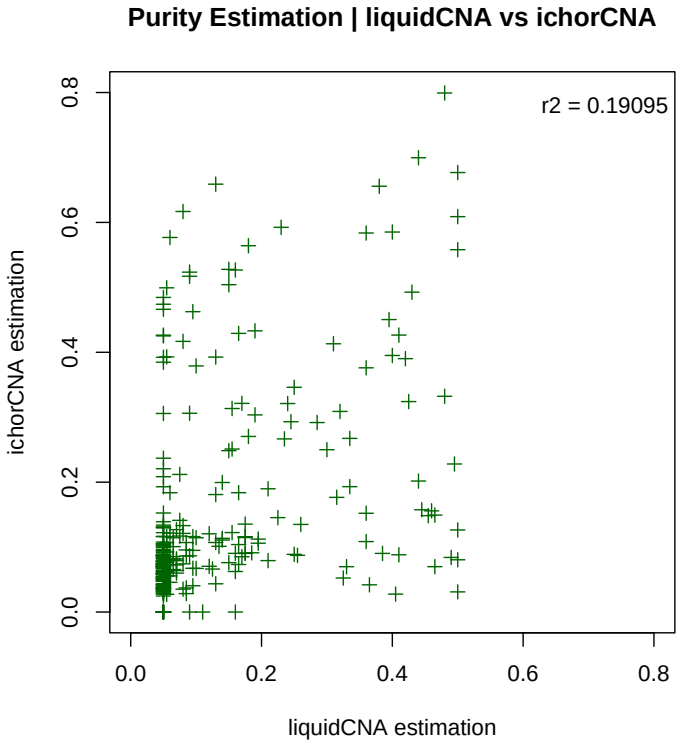


Figure 3.8: Purity estimations of liquidCNA plotted against their respetive estimates from ichorCNA. With r^2 value of 0.19, the two algorithms show low correlation.

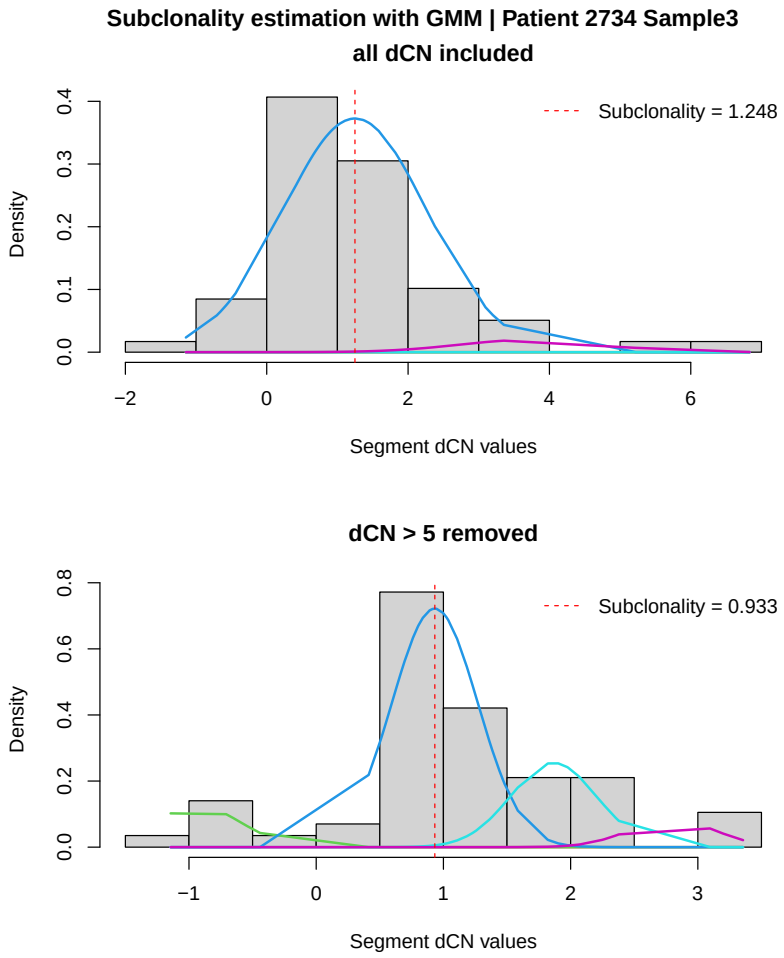


Figure 3.9: GMM fitted to dCN values of Patient 2734 Sample3. Without any filtering of dCNs, subclonality was estimated to be 1.248 (TOP). With the filtering of dCNs > 5, subclonality estimate is now within the range [0,1] (BOTTOM).

dCNs when estimating subclonality. During the GMM fitting step, only five Gaussians (each for dCN values -2, -1, 1, 2 and 3) are fitted. With the modification, we are now explicitly assuming that dCN states > 5 are not possible. These dCNs are regarded as artefacts of measurement noise and are removed. As seen in Figure 3.9, the filtering allows better fitting of Gaussian models. The modification 'saves' the subclonality estimation for this sample, reducing the initial estimate of 1.25 down to 0.93.

After the modification, number of erroneous estimates reduced from 25 to 16. Looking at Figure 3.10, the modification noticeably removed the extreme subclonality estimates > 6.0. All erroneous estimates are now smaller than 3.0, with the majority being in the range [1,2].

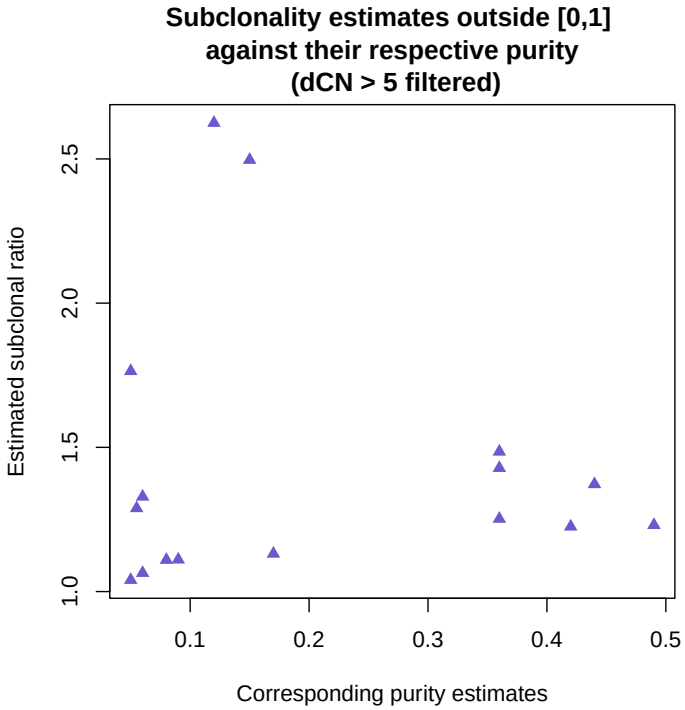


Figure 3.10: The number of erroneous subclonalities estimated are reduced to 16 after the filtering of dCNs > 5. These 16 estimates are plotted against their respective purities.

3.4 Subclonality is not a prognostic marker for progression and metastasis.

Subclonality is a poor prognostic marker for progression. For each of the time sample we have RECIST radiological evaluation for, subclonality estimates were grouped into either Progression evaluation of YES or NO. A statistical test was conducted to see whether the mean of the two groups' distributions were significantly different. T-test was inappropriate for the data as normality could not be assumed for either of the distributions (Shapiro–Wilk test p-value $< 1e-09$). Wilcoxon test was used instead. With a p-value of 0.883, the test showed the subclonality between the two groups of progression to not differ significantly. This indicates subclonality to not provide any prognostic information for progression (Figure 3.11).

The same analysis was done with time samples' purity. Purity estimates were grouped by progression and the difference in dis-

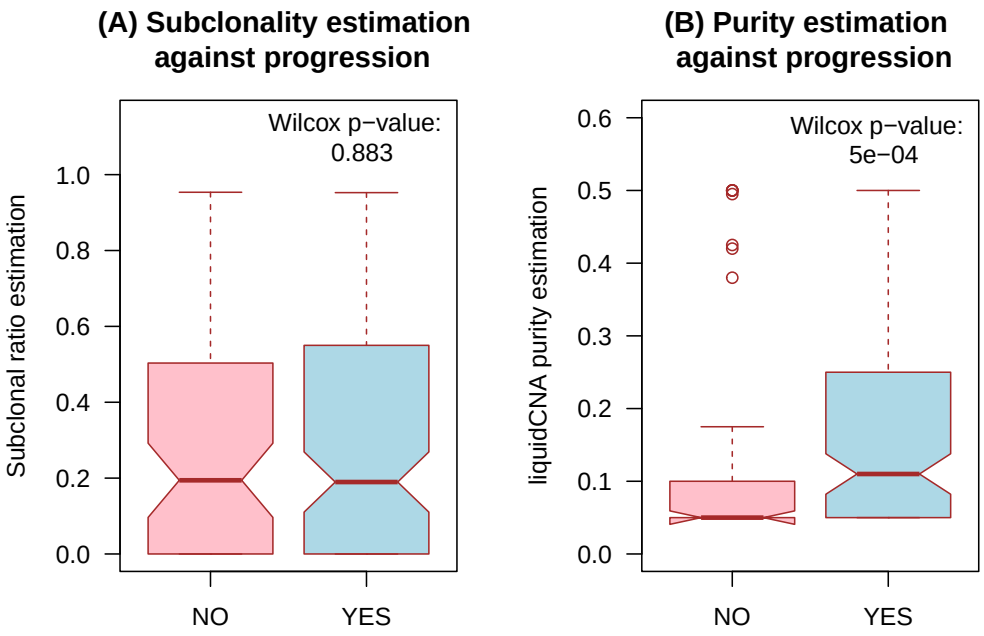


Figure 3.11: Boxplots of subclonality (LEFT) and purity (RIGHT) grouped by time samples' RECIST evaluation of progression.

tribution of the two groups were compared (Figure 3.11). Unlike with subclonality, the two distributions differed significantly with a p-value of 0.000466. This shows purity to be positively associated with disease progression and indicates its role as a potential prognostic biomarker.

Subclonality is a poor prognostic marker for metastasis as well. For each of the 73 patients for which we have RECIST data, patients were categorised into groups with or without new metastasis. As all patients were metastatic breast cancer patients, 'new metastasis' was defined as metastasis detected after at least 14 days from the primary metastasis. Wilcoxon test showed both purity and subclonality to be insignificantly different between the two groups of metastasis (Figure 3.12). Interestingly, though both were insignificant, the distribution of subclonality was observed to differ more than purity. This may indicate subclonality to have more prognostic power for metastasis compared to purity. This contrasts with the results for Progression.

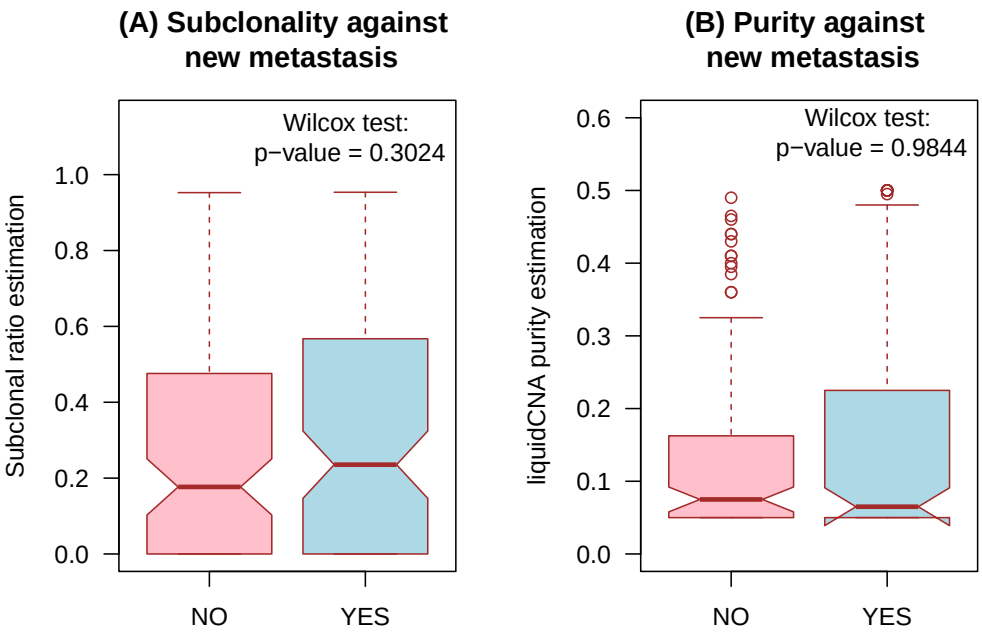
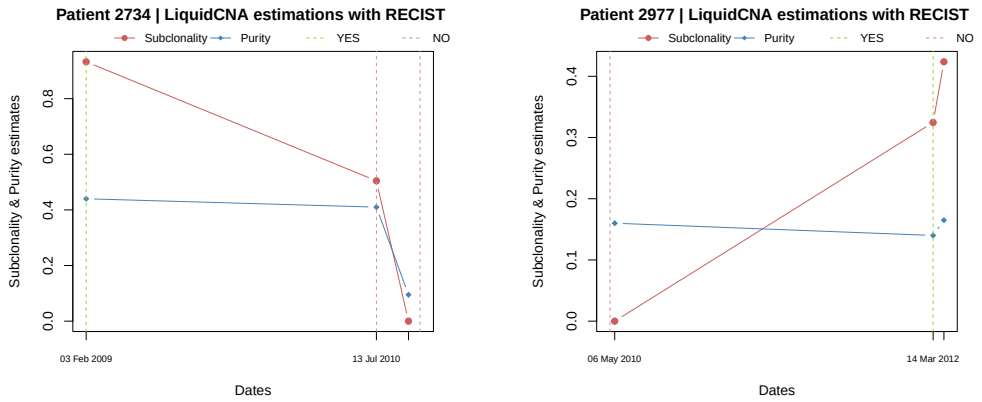


Figure 3.12: Boxplots of subclonality (LEFT) and purity (RIGHT) grouped by patients' RECIST evaluation of metastasis. To note, as all our patients have metastatic breast cancer, 'new metastasis' are non-primary recurrent metastatic events.

3.5 Subclonality is a better indicator of metastasis than purity.

For each patient, subclonality and purity estimates were plotted against time. The dynamics of the two estimates were analysed alongside RECIST evaluation of metastasis. Of the 73 patients, 33 patients had new non-primary metastasis. For a subset of these 33 patients, subclonality and purity was observed to provide prognostic information (Figure 3.13).

In the figure, subclonality was seen to be a better indicator of metastasis than purity. Overall, the two estimates followed similar dynamics. Subclonality and purity values were reduced when there were no metastasis, and were increased with metastatic events. Though similar, the magnitude of the dynamics was larger for subclonality and provided a clearer signal. For instance, in Patient 2734, the absolute range of change in subclonality was 0.93, whilst it was only 0.35 for purity - an approximately three times larger magnitude of response (Figure 3.13a). Additionally,



(a) (b)

Figure 3.13: LiquidCNA estimations of subclonality and purity plotted against time for Patient 2734 (LEFT) and 2977 (RIGHT). Dotted line labelled YES and NO indicate RECIST evaluation of new non-primary metastasis.

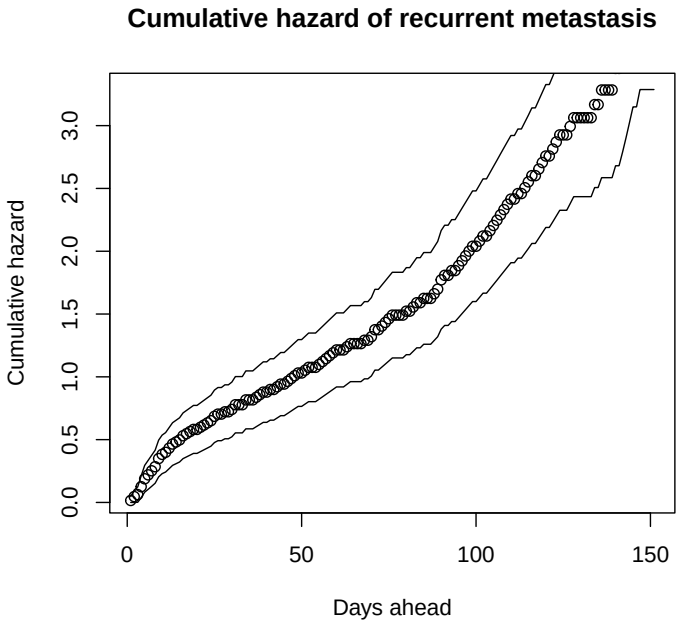
for both Patient 2734 and 2977, purity values changed only by the third sample and showed negligible response when evaluation of metastasis first changed. Thus, in addition to a clearer signal, subclonality provided earlier response of metastasis compared to purity.

3.6 Analysis of recurrent metastasis

In the previous sections, both subclonality and purity has been shown to be statistically insignificant prognostic markers for metastasis. However, from Figure 3.12 and 3.13, subclonality levels in particular were observed to respond to new metastatic events. Thus, survival analysis was conducted to assess and quantify the relationship purity and subclonality has with recurrent metastasis.

The key interpretable output from AG model is the hazard ratio (HR). For our data, HR can be interpreted as the instantaneous rate of a new metastasis occurring. A HR > 1 will be interpreted as increased risk of new metastasis whilst HR < 1 indicates reduced risk. For each of the time-varying covariates, AG model outputs HR alongside its 95% confidence interval (CI) and the statistical significance (Figure 3.14b).

Looking at the significances of the survival analysis, both purity



(a) Cumulative hazard of recurrent metastasis.

	Hazard_ratio	lower_95	upper_95	p_value
Subclonality	1.2434	0.7992	1.935	0.334
Purity	0.8246	0.5083	1.338	0.435

(b) AG model output for the two covariates, subclonality and purity.

Figure 3.14

and subclonality were not significantly associated with recurrent metastasis. Though not significant, the HR for the two covariates were interesting. For subclonality, samples in the high subclonality group had 24% higher risk of metastasis (HR = 1.243, CI = 0.799–1.935, p-value = 0.334). This aligns with our biological expectations, as tumours with higher subclonal ratio are expected to have poor prognosis. For purity, samples with high purity exhibited a reduction in risk by 18% (HR = 0.825, CI = 0.508–1.338, p-value = 0.435). This result is unexpected as it goes against studies that showed tumour fraction to be positively associated with metastasis [48].

Cumulative hazard is used to assess the risk of an event occurring over time. For our data, with purity and subclonality as covariates, a two times risk of a new metastatic event is predicted in 110 days (Figure 3.14a).

Chapter 4

Discussion

4.1 Low purity estimates may be the cause of non-significant results.

Looking at Figure 3.11(A) and 3.12(A), subclonality does not provide any prognostic information for progression (p-value = 0.883) and for metastasis (p-value = 0.302). These results directly disprove the aim of our study, that subclonal architecture derived from lpWGS CN profile will be informative. However, this is not a conclusive result due to liquidCNA's limitations with regards to purity.

Coming back to the two figures, subclonality is observed to have a bigger difference between the groups of Metastasis than it does for Progression; subclonality estimates are increased with metastasis (Figure 3.12a), whilst being very similar for progression (Figure 3.11a). Purity, interestingly, shows the opposite response. Purity values are significantly higher for patients with progression (p-value < 0.0005) (Figure 3.11b) - whilst being highly insignificant for metastasis (p-value = 0.98) (Figure 3.12b).

This inverse response between the two is very interesting because the two estimations are not independent. Rather, purity is estimated first, then liquidCNA uses this estimate to compute subclonality. In Section 3.3.1, we have shown the computational steps of subclonality estimation to be highly sensitive to purity. In particular, the accuracy of subclonality fell when the purity values were too low. This means that it would be inappropriate

to compare the distributions of subclonality estimates when the respective distributions of purities have been shown to be significantly different — e.g., for the two patient groups of Progression.

To illustrate, patient group without progression has a mean purity estimate of 0.114 whilst the group with progression has a mean of 0.171. The former group exhibits significantly lower purity estimates, with over half of it samples with purities smaller than 0.1. Consequently, subclonality estimated for this group will be more inaccurate than the group with progression. This means that had the purity estimates been larger and more uniform across the two groups, subclonality may have shown a more significant difference across progression. In fact, when samples with low purity estimates were filtered out, remaining subclonalities did indeed show a larger difference across progression (p-value = 0.37, Figure A.1). However, such filtering severely reduced the size of our data, making it impractical to implement for our study.

This problem is not prevalent for Metastasis as the difference in purity values between the two groups are highly insignificant (p-value = 0.984). Consequently, subclonality estimates across the groups will have equal accuracy. This may be the reason why subclonality is more different across metastasis (p-value = 0.302) than across progression (p-value = 0.883). To note, purity estimates across Metastasis is also low (with mean values for both groups being smaller than 0.1). Had the purity estimates been larger — thus, the subclonality estimates more accurate — the difference in subclonality across metastasis may have been statistically different.

As such, the limitations of the algorithm are obscuring the biological differences our study aimed to uncover. Thus, it seems premature to reject our hypothesis that subclonality derived from lpWGS will provide prognostic information. Our study highlights this limitation and provides an important insight that future users of liquidCNA must take into account: when comparing subclonality estimates across different groups, it must be first ensured that purity estimates are not significantly different. This may be a challenge as purity itself is a dynamic characteristic of the tumour. In most scenarios, purity will change throughout the course of the disease and will be influenced by factors such as treatment and progression.

A potential method to overcome this limitation is by filtering low purity time samples (of $p_i < 0.1 \sim 0.2$). Whilst this will improve the over all accuracy of the output, it will severely limit the size of the data. For our data set of 283 time samples, filtering by $p_i = 0.1$, 0.15 or 0.2 reduces the number of samples to 95, 78 and 53, respectively. The filtering results many of the patients to have only $0 \sim 2$ time samples remaining, rendering them to be either suboptimal or inappropriate for liquidCNA.

4.2 Concluding remarks

In our study, liquid biopsy was used to infer subclonal evolution in patients of metastatic breast cancer. The algorithm is a novel method that utilises SCNAs from lpWGS data of longitudinal liquid biopsy samples. Our study investigated whether liquidCNA's estimation of subclonality provided any clinical utility in predicting disease progression and metastasis. Unfortunately, the results of the study were not significant. Subclonality was non-significantly associated with progression and metastasis, and was a weak predictor of recurrent metastasis.

Despite the results of our study, we believe liquidCNA (and more specifically, the inference of subclonal events through lpWGS data) still has potential for clinical utility. First, liquidCNA is still a novel method with several limitations. Our study addressed some of these limitations and provided modifications that can be implemented by future users. Further refinements will improve the algorithm. Second, our data was not perfectly curated for the algorithm. For instance, there were only two time samples available for a quarter of the patients - a number that was insufficient to run liquidCNA before our modifications. This raises the question, would the results be different if the study was repeated with data specifically collected with liquidCNA in mind? A data set with more patients, each with $3 \sim 6$ time samples, and with each sample having matching radiological evaluation and high purity estimates. As such, the results of this study are far from conclusive, but rather calls the need for future work.

Future directions include tackling questions such as: if liquidCNA is sensitive to low purity, could purity estimates from ichorCNA be used in the liquidCNA pipeline? If so, how would the results be affected? An additional question would be: why is

there low correlation between ichorCNA and liquidCNA? Furthermore, our study only used a small fraction of the data available. Thus, could the data regarding the specific therapy each of the patients received be used to link, subclonality, metastasis and therapy resistance together? In this new and exciting field, our study is still preliminary. Further work will be required to refine methods and better determine the clinical utility of inferring subclonal evolution.

Bibliography

[1] Abel Jacobus Bronkhorst, Vida Ungerer, and Stefan Holdenrieder. "The emerging role of cell-free DNA as a molecular marker for cancer management". In: *Biomolecular Detection and Quantification* 17 (Mar. 2019). ISSN: 22147535. DOI: 10.1016/j.bdq.2019.100087.

[2] Jonathan C.M. Wan et al. "Liquid biopsies come of age: Towards implementation of circulating tumour DNA". In: *Nature Reviews Cancer* 17 (4 Apr. 2017), pp. 223–238. ISSN: 14741768. DOI: 10.1038/nrc.2017.7.

[3] M Stroun et al. "Neoplastic Characteristics of the DNA Found in the Plasma of Cancer Patients". In: *Oncology* 46 (5 1989), pp. 318–322. ISSN: 0030-2414. DOI: 10.1159/000226740. URL: https://www.karger.com/DOI/10.1159/000226740.

[4] George D Sorenson et al. *Soluble Normal and Mutated DNA Sequences from Single-Copy Genes in Human Blood*. 1994, pp. 67–71. URL: http://aacrjournals.org/cebp/article-pdf/3/1/67/2287024/67.pdf.

[5] Marius Ilié and Paul Hofman. "Pros: Can tissue biopsy be replaced by liquid biopsy?" In: *Translational Lung Cancer Research* 5 (4 Aug. 2016), pp. 420–423. ISSN: 22264477. DOI: 10.21037/tlcr.2016.08.06.

[6] Stephen Q. Wong et al. "Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing". In: *BMC Medical Genomics* 7 (1 May 2014). ISSN: 17558794. DOI: 10.1186/1755-8794-7-23.

[7] Helmut H. Popper. "Commentary on tumor heterogeneity". In: *Translational Lung Cancer Research* 5 (4 Aug. 2016), pp. 433–435. ISSN: 22264477. DOI: 10.21037/tlcr.2016.08.07.

[8] Elza C de Bruin et al. "Spatial and temporal diversity in genomic instability processes defines lung cancer evolution". In: *Science* 346 (6206 Oct. 2014), pp. 251–256. DOI: 10.1126/science.1253462. URL: https://doi.org/10.1126/science.1253462.

[9] L. De Mattos-Arruda et al. "Capturing intra-tumor genetic heterogeneity by de novo mutation profiling of circulating cell-free tumor DNA: A proof-of-principle". In: *Annals of Oncology* 25 (9 2014), pp. 1729–1735. ISSN: 15698041. DOI: 10.1093/annonc/mdu239.

[10] M. Jamal-Hanjani et al. "Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer". In: *Annals of Oncology* 27 (5 May 2016), pp. 862–867. ISSN: 15698041. DOI: 10.1093/annonc/mdw037.

[11] Ronald Lebofsky et al. "Circulating tumor DNA as a non-invasive substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types". In: *Molecular Oncology* 9 (4 Apr. 2015), pp. 783–790. ISSN: 18780261. DOI: 10.1016/j.molonc.2014.12.003.

[12] K. C.Allen Chan et al. "Cancer genome scanning in plasma: Detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing". In: *Clinical Chemistry* 59 (1 Jan. 2013), pp. 211–224. ISSN: 00099147. DOI: 10.1373/clinchem.2012.196014.

[13] Frank Diehl et al. "Circulating mutant DNA to assess tumor dynamics". In: *Nature Medicine* 14 (9 Sept. 2008), pp. 985–990. ISSN: 10788956. DOI: 10.1038/nm.1789.

[14] Svetlana N. Tamkovich et al. "Circulating DNA and DNase activity in human blood". In: vol. 1075. Blackwell Publishing Inc., 2006, pp. 191–196. ISBN: 157331627X. DOI: 10.1196/annals.1368.026.

[15] Nancy B.Y. Tsui et al. "High Resolution Size Analysis of Fetal DNA in the Urine of Pregnant Women by Paired-End Massively Parallel Sequencing". In: *PLoS ONE* 7 (10 Oct. 2012). ISSN: 19326203. DOI: 10.1371/journal.pone.0048319.

[16] Yoshiaki Nakamura et al. "Clinical utility of circulating tumor DNA sequencing in advanced gastrointestinal cancer: SCRUM-Japan GI-SCREEN and GOZILA studies". In: *Nature Medicine* 26 (12 Dec. 2020), pp. 1859–1864. ISSN: 1546170X. DOI: 10.1038/s41591-020-1063-5.

[17] Ryan B. Corcoran. "Liquid biopsy versus tumor biopsy for clinical-trial recruitment". In: *Nature Medicine* 26 (12 Dec. 2020), pp. 1815–1816. ISSN: 1546170X. DOI: 10.1038/s41591-020-01169-6.

[18] Alain R. Thierry et al. "Clinical validation of the detection of KRAS and BRAF mutations from circulating tumor DNA". In: *Nature Medicine* 20 (4 2014), pp. 430–435. ISSN: 1546170X. DOI: 10.1038/nm.3511.

[19] Alexander W. Wyatt et al. "Concordance of Circulating Tumor DNA and Matched Metastatic Tissue Biopsy in Prostate Cancer". In: *Journal of the National Cancer Institute* 109 (12 Dec. 2017). ISSN: 14602105. DOI: 10.1093/jnci/djx118.

[20] Yung Bin Kuo et al. "Comparison of KRAS mutation analysis of primary tumors and matched circulating cell-free DNA in plasmas of patients with colorectal cancer". In: *Clinica Chimica Acta* 433 (June 2014), pp. 284–289. ISSN: 18733492. DOI: 10.1016/j.cca.2014.03.024.

[21] Samanta Salvi et al. "Cell-free DNA as a diagnostic marker for cancer: Current insights". In: *OncoTargets and Therapy* 9 (Oct. 2016), pp. 6549–6559. ISSN: 11786930. DOI: 10.2147/OTT.S100901.

[22] Li Mao et al. "Detection of oncogene mutations in sputum precedes diagnosis of lung cancer". In: *Cancer research* 54.7 (1994), pp. 1634–1637.

[23] Luis A. Diaz et al. "The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers". In: *Nature* 486 (7404 June 2012), pp. 537–540. ISSN: 00280836. DOI: 10.1038/nature11219.

[24] Giulio Genovese et al. "Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence". In: *New England Journal of Medicine* 371.26 (2014), pp. 2477–2487.

[25] Emmanuelle Gormally et al. "TP53 and KRAS2 mutations in plasma DNA of healthy subjects and subsequent cancer occurrence: a prospective study". In: *Cancer research* 66.13 (2006), pp. 6871–6876.

[26] Kun Sun et al. "Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments". In: *Proceedings of the National Academy of Sciences of the United States of America* 112 (40 Oct. 2015), E5503–E5512. ISSN: 10916490. DOI: 10.1073/pnas.1508736112.

[27] Christopher Abbosh et al. "Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution". In: *Nature* 545 (7655 Apr. 2017), pp. 446–451. ISSN: 14764687. DOI: 10.1038/nature22364.

[28] Christine A Parkinson et al. "Exploratory analysis of TP53 mutations in circulating tumour DNA as biomarkers of treatment response for patients with relapsed high-grade serous ovarian carcinoma: a retrospective study". In: *PLoS medicine* 13.12 (2016), e1002198.

[29] Chetan Bettegowda et al. "Detection of circulating tumor DNA in early-and late-stage human malignancies". In: *Science translational medicine* 6.224 (2014), 224ra24–224ra24.

[30] Thierry Lecomte et al. "Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosis". In: *International journal of cancer* 100.5 (2002), pp. 542–548.

[31] Sarah-Jane Dawson et al. "Analysis of circulating tumor DNA to monitor metastatic breast cancer". In: *New England Journal of Medicine* 368.13 (2013), pp. 1199–1209.

[32] Rongyuan Zhuang et al. "The prognostic value of KRAS mutation by cell-free DNA in cancer patients: A systematic review and meta-analysis". In: *PLOS ONE* 12 (8 Aug. 2017), e0182562–. URL: https://doi.org/10.1371/journal.pone.0182562.

[33] Elin S Gray et al. *Circulating tumor DNA to monitor treatment response and detect acquired resistance in patients with metastatic melanoma*. 2015, 6:42008–42018. DOI: https://doi.org/10.18632/oncotarget.5788. URL: www.impactjournals.com/oncotarget.

[34] Sarah B. Goldberg et al. "Early assessment of lung cancer immunotherapy response via circulating tumor DNA". In: *Clinical Cancer Research* 24 (8 Apr. 2018), pp. 1872–1880. ISSN: 15573265. DOI: 10.1158/1078-0432.CCR-17-1341.

[35] US Food Drug Administration. *Premarket approval P150044 — Cobas EGFR MUTATION TEST V2. FDA*. 2016. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpma/pma.cfm?id=P150044.

[36] European Medicines Agency. *Tagrisso: public assessment report — product information*. 2016. URL: https://www.ema.europa.eu/en/documents/product-information/tagrisso-epar-product-information_en.pdf.

[37] Dhruvajyoti Roy et al. *Cell-free circulating tumor DNA profiling in cancer management*. Oct. 2021, pp. 1014–1015. DOI: 10.1016/j.molmed.2021.07.001.

[38] Mel Greaves and Carlo C. Maley. "Clonal evolution in cancer". In: *Nature* 481 (7381 Jan. 2012), pp. 306–313. ISSN: 00280836. DOI: 10.1038/nature10762.

[39] Noemi Andor et al. "Pan-cancer analysis of the extent and consequences of intratumor heterogeneity". In: *Nature Medicine* 22 (1 Jan. 2016), pp. 105–113. ISSN: 1546170X. DOI: 10.1038/nm.3984.

[40] Samuel W Brady et al. "Combating subclonal evolution of resistant cancer phenotypes". In: *Nature Communications* 8 (1 2017), p. 1231. ISSN: 2041-1723. DOI: 10.1038/s41467-017-01174-3. URL: https://doi.org/10.1038/s41467-017-01174-3.

[41] Muhammed Murtaza et al. "Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer". In: *Nature Communications* 6 (Nov. 2015). ISSN: 20411723. DOI: 10.1038/ncomms9760.

[42] Eszter Lakatos et al. "LiquidCNA: Tracking subclonal evolution from longitudinal liquid biopsies using somatic copy number alterations". In: *iScience* 24 (8 Aug. 2021). ISSN: 25890042. DOI: 10.1016/j.isci.2021.102889.

[43] Anja Brouwer et al. "HER-2 status of circulating tumor cells in a metastatic breast cancer cohort: A comparative study on characterization techniques". In: *PLOS ONE* 14 (9 Sept. 2019), e0220906–. URL: https://doi.org/10.1371/journal.pone.0220906.

[44] Ilari Scheinin et al. "DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly". In: *Genome Research* 24

(12 Dec. 2014), pp. 2022–2032. ISSN: 15495469. DOI: 10.1101/gr.175141.114.

[45] Venkatraman E. Seshan and Adam Olshen. *DNAcopy: DNA copy number data analysis*. R package version 1.68.0. 2021.

[46] Leila D.A.F. Amorim and Jianwen Cai. "Modelling recurrent events: A tutorial for analysis in epidemiology". In: *International Journal of Epidemiology* 44 (1 Feb. 2015), pp. 324–333. ISSN: 14643685. DOI: 10.1093/ije/dyu222.

[47] Terry M Therneau. *A Package for Survival Analysis in R*. R package version 3.3-1. 2022. URL: https://CRAN.R-project.org/package=survival.

[48] Atish D. Choudhury et al. "Tumor fraction in cell-free DNA as a biomarker in prostate cancer". In: *JCI insight* 3 (21 Nov. 2018). ISSN: 23793708. DOI: 10.1172/jci.insight.122109.

Appendix A

Supplementary Figures:

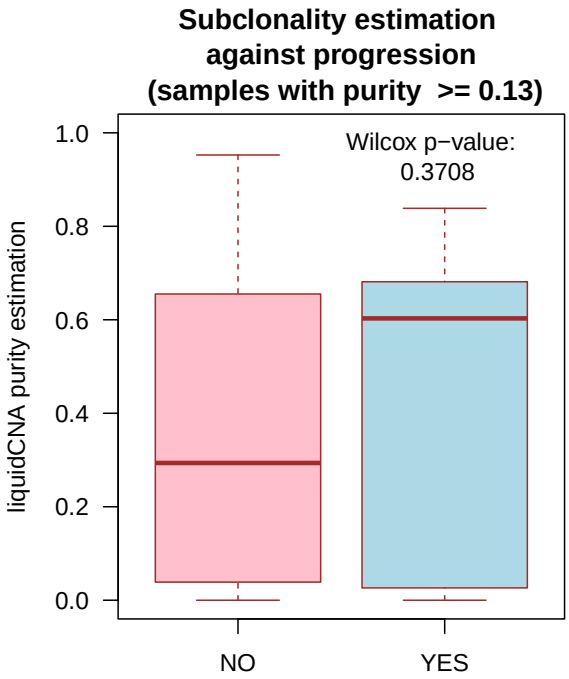


Figure A.1: Supplementary Figure. Boxplot of subclonality grouped by time samples' RECIST evaluation of progression. Time samples with $p_i < 0.13$ are filtered out.