A new method to accurately identify single nucleotide variants using small FFPE breast samples.

Angelo Fortunato[1,2,*], Diego Mallo[1,2, *], Shawn M. Rupp[1], Lorraine King[3], Timothy Hardman[3], Joseph Lo[3], Allison Hall[3], Jeffrey R. Marks[3], E. Shelley Hwang[3] and Carlo C. Maley[1,2].

[*] These authors contributed equally to this work.

1. Arizona Cancer Evolution Center, Biodesign Center for Biocomputing, Security and Society, Arizona State University, 727 E. Tyler St.,Tempe, AZ 85281, USA;

2. School of Life Sciences, Arizona State University, 427 East Tyler Mall, Tempe, AZ 85287, USA.

3. Duke University, Durham, NC, USA.

Keywords:

DCIS, NGS, exome, heterogeneity

Abstract

Most tissue collections of neoplasms are composed of formalin-fixed and paraffin-embedded (FFPE) excised tumor samples, and routine diagnosis in oncology relies on histopathological analysis of those samples. Genomic sequencing is becoming increasingly important in the clinical management as well as the basic science of cancer. Unfortunately, genomic sequencing of FFPE samples is difficult due to the small amounts of DNA available particularly from early cancers, as well as degradation of that DNA. We developed a new bioinformatic algorithm to robustly identify somatic mutations using small amounts of DNA extracted from archival FFPE samples of breast ductal carcinoma *in situ*, a preinvasive form of breast cancer. We optimized this strategy using 28 pairs of technical replicates, in which the same DNA sample was sequenced twice independently. After optimization, the mean similarity between replicates was 88.3%, range 66.7-100%, and we were able to detect an average of 19.9 (range 1-61) single nucleotide variants in each sample. We found that the accuracy of identifying SNVs severely declined when there was less than 40ng of DNA available. High depth resequencing also showed that insertion-deletion (indel) variants are an unreliable subset of mutations, using current methods. This new algorithm was empirically optimized and validated. It provides a significant improvement in detecting somatic single nucleotide variants in FFPE samples that can be used to accurately profile the genomes of neoplasms.

Introduction

Tumors are characterized by a high genetic heterogeneity both within the same tumor type and in different parts of the same neoplasm (Marusyk and Polyak 2010). Genetic heterogeneity determines the capacity of the neoplastic cell population to adapt to new microenvironments and to develop resistance to therapeutic treatments (McGranahan and Swanton 2015; Andor et al. 2016; Morris et al. 2016). We and others have hypothesized that the quantification of genetic heterogeneity will be generally useful for risk stratification of patients (Bedard et al. 2013; C. C. Maley et al. 2017). But to do so, we need accurate methods for identifying somatic genomic alterations in neoplasms.

Cancers can develop from different combinations of genetic mutations and each patient typically has a unique mutational profile, distributed among a mosaic of subclones (Dash et al. 2019). This makes it difficult to develop universal biomarkers to predict cancer progression based on specific mutations and a single sample from a neoplasm. Alternatively, measures that characterize the underlying evolutionary process do not focus on specific progression mechanisms or the particular mutations that occur, making them more generalizable (Maley et al. 2017). Intratumor heterogeneity is one such measure, and we have successfully used it in the past to predict cancer progression of pre-malignant diseases (Maley et al. 2006; Merlo et al. 2010; Martinez et al. 2016).

Routine diagnosis in oncology relies on histopathological analysis of formalin-fixed and paraffin-embedded (FFPE) excised tumor samples. Using these samples for genetic analysis has numerous advantages: histopathological analyses are already available for them; specific areas can be selected with precision, eliminating the need to take additional samples dedicated to genetic analysis, and they are archived in large numbers, readily available to carry out retrospective studies. On the other hand, these samples have several technical limitations when used for genetic analyses. Histological fixation and embedding partially degrades and binds amino acids to the DNA, which continues to deteriorate over time (Carrick et al. 2015).

Deamination of cytosine residues leading to apparent C to T transitions is also a common artefact in FFPE derived DNA (Chen et al. 2014). These problems are exacerbated when the amount of available DNA is limited, because DNA artifacts are not compensated by the abundance of intact molecules, leading to sequencing errors (Do and Dobrovic 2015; Sah et al. 2013). This is particularly relevant when studying early or pre cancerous conditions where the lesion can be very small. In order to study genomic intratumor heterogeneity using FFPE samples, we must often sequence the degraded and imperfectly purified DNA extracted from small focal areas of the tumor or pre cancer. Furthermore, estimates of intratumor heterogeneity, as well as other precision medicine efforts, are confounded by both false positives and false negatives in the detection of mutations. Precision medicine requires avoiding false positives and negatives which would potentially expose patients to the wrong therapeutic interventions. Thus, there is a need for robust and accurate methods for sequencing and detecting mutations in small amounts of DNA extracted from FFPE samples. We have developed a new bioinformatic method that reduces these obstacles for the estimation of genetic intratumor heterogeneity using paired FFPE samples. We developed this somatic-variant post-processing pipeline by empirical optimization using 28 technical replicates—DNA samples sequenced twice independently, and validated the results using a different, high depth, sequencing technique.

We selected a precursor of breast cancer, ductal carcinoma *in situ* (DCIS), to develop and optimize our pipeline because most of those tumors are detected in the early phase of their development and there is an urgent clinical need to be able to estimate the risk level of the tumor in this phase. Improved risk stratification in DCIS could guide improvements in management of the condition and therapeutic intervention. The majority of breast tumors develop in the terminal duct lobular unit, mainly starting among duct cells (Pandya and Moore 2011; Sims et al. 2007) (Fig. 1). The cancer cells proliferate within the ducts and deform their anatomical structure. Despite the ducts' growth in volume their walls remain intact, confining the tumor cells in the lumen, separating them from nearby tissues and limiting their dissemination. In this phase, the tumor is defined as ductal carcinoma *in situ* (DCIS). Subsequently, the cells may evolve to invasive disease, crossing the duct wall's boundaries, invading the surrounding tissue and potentially metastasizing. DCIS tumors can remain non-invasive but there is substantial evidence that a subset will invade and, in some cases, metastasize. We included 7 samples from invasive breast tumors as representatives of the endpoint of breast cancer progression. The integration of genetic heterogeneity analyses in the patient clinical evaluation could provide a significant contribution to the estimation of DCIS patients' risk for progressing to breast cancer.

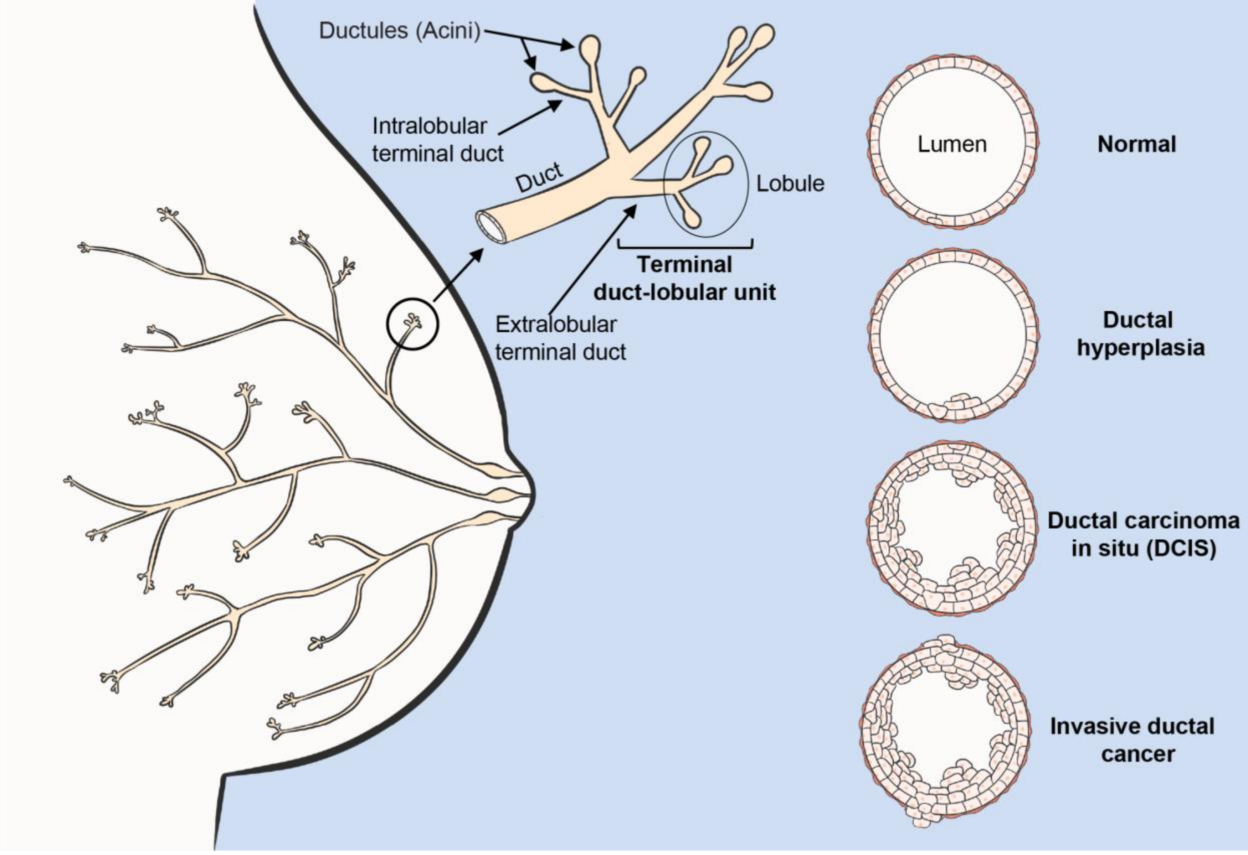
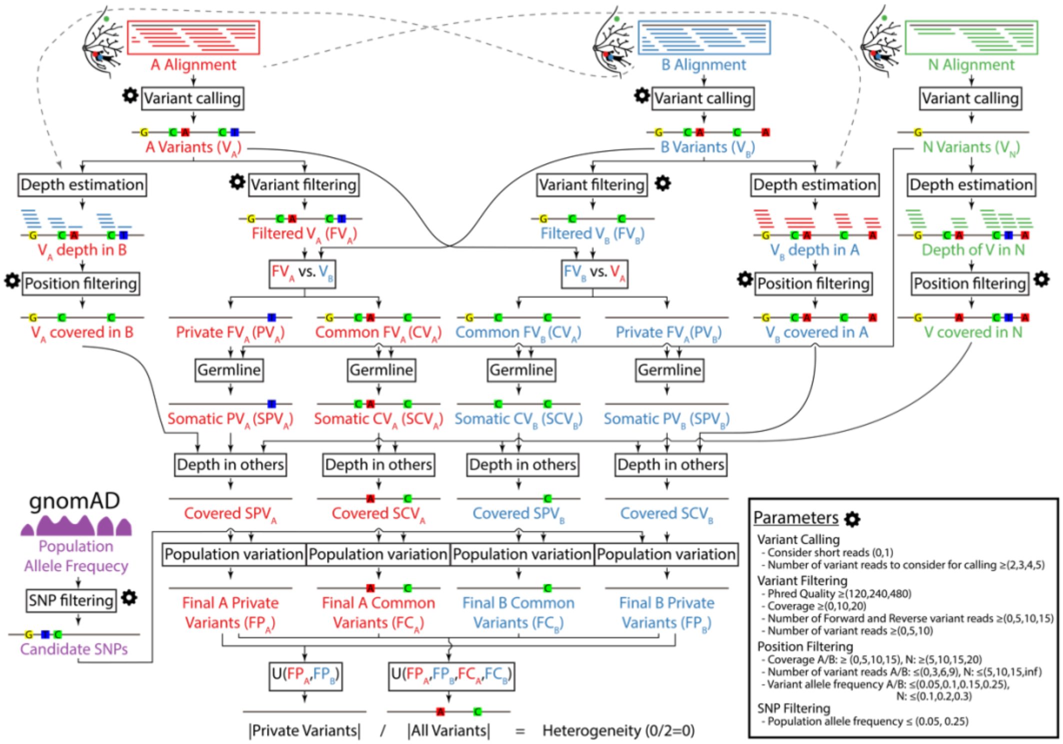

Ductules (Acini)
Intralobular terminal duct
Duct
Lobule
Terminal duct-lobular unit
Extralobular terminal duct
Lumen
Normal
Ductal hyperplasia
Ductal carcinoma in situ (DCIS)
Invasive ductal cancer

Figure 1: Breast cancer anatomy. Schematic representation of mammary gland anatomy and cancer development. The majority of breast tumors develop in the terminal duct lobular unit, 80% starting among ductal cells. Initially, the duct suffers a benign hypertrophic growth of cells that can progress into ductal carcinoma *in situ* (DCIS). In this phase the neoplasm is confined

within the duct's lumen and it is still clinically benign. Cancer cells can cross the duct wall's boundaries, invading nearby tissues (IDC) and metastasizing.

Results

Ideally, the same sample of tumor DNA, when sequenced twice with the same methodology, should give the same results (detect the same mutations). We developed our mutation detection pipeline (Fig. 2), optimized it using duplicate sequencing (technical replicate) assays of the same samples, and validated our results using deep sequencing.



Parameters

Variant Calling
- Consider short reads (0,1)
- Number of variant reads to consider for calling ≥(2,3,4,5)

Variant Filtering
- Phred Quality ≥(120,240,480)
- Coverage ≥(0,10,20)
- Number of Forward and Reverse variant reads ≥(0,5,10,15)
- Number of variant reads ≥(0,5,10)

Position Filtering
- Coverage A/B: ≥ (0,5,10,15), N: ≥(5,10,15,20)
- Number of variant reads A/B: ≤(0,3,6,9), N: ≤(5,10,15,inf)
- Variant allele frequency A/B: ≤(0.05,0.1,0.15,0.25), N: ≤(0.1,0.2,0.3)

SNP Filtering
- Population allele frequency ≤ (0.05, 0.25)

Figure 2: Flowchart of the algorithm to estimate the genetic heterogeneity between two samples. Inputs: aligned sequences (BAM files) of the two samples (A, in red; and B, in blue) and their healthy tissue control (N, in green), population allele frequency data from the gnomAD

database (single nucleotide polymorphisms, SNPs, in purple), and user-specified configuration parameters (gear icon). Outputs: estimate of the genetic heterogeneity between samples A and B, and set of variants (level of detail user-specified). The key step of this algorithm is the generation of two sets of private and common variants by comparing the variants in the two samples twice, alternatively filtering one of the sets and using all variants from the other. The parameters that control this pipeline, and the values assayed in our bioinformatic optimization, are detailed in the Parameters box.

Pipeline optimization

We used an empirical method for optimizing the analysis algorithm through the comparison of technical replicates of whole exome sequences. Any variant detected only in one sample but not in the other is likely the result of a sequencing or data processing error. This approach allowed us to systematically and objectively compare alternative parameterizations of the estimation pipeline to single out the best overall, and to find the most generalizable parameter values using cross-validation.

In order to optimize our pipeline, we assigned a range of values to explore for each of the 13 parameters that control its execution (Fig. 2) and explored every possible combination of them, scoring each using a statistic that integrates the central tendency and dispersion of the heterogeneity across the 28 technical replicates. Furthermore, we used DNA quantity (from 20 ng to >100 ng) in order to evaluate the efficiency of the method on different quantities of input DNA, in order to determine the limits of the method on small amounts of DNA (Suppl. table 1S).

The resulting algorithm yielded a mean similarity across the 28 technical replicates of 88.3% (range 66.7-100%) (Fig. 3), which constitutes a 5-fold improvement over using the same variant caller–Platypus without any post-processing of the results (Rimmer et al. 2014), 17.8%, range: 0.1-61.8%). We identified a mean of 19.9 (range 1-61) single nucleotide variants per sample (Table 1), which are distributed throughout the entire exome (Fig. 3S).

We also assayed an alternative implementation of our algorithm that uses Mutect2 to call variants, but it achieved a much lower accuracy, with a mean similarity (including indels) across the 28 technical replicates of only 2.4%, range 0.4-6.9%. Overall, we found 25 (14.9%) variants (SNVs) overlap between our main pipeline and this alternative implementation using Mutect2.

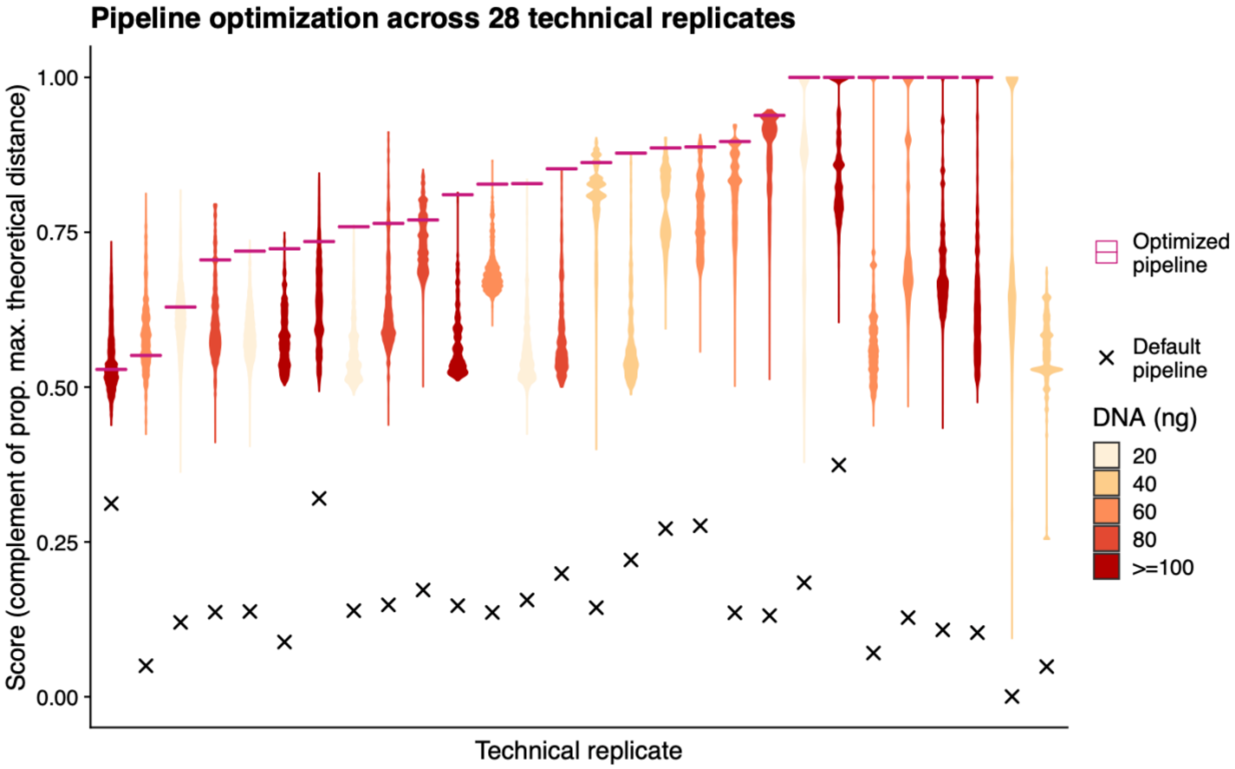
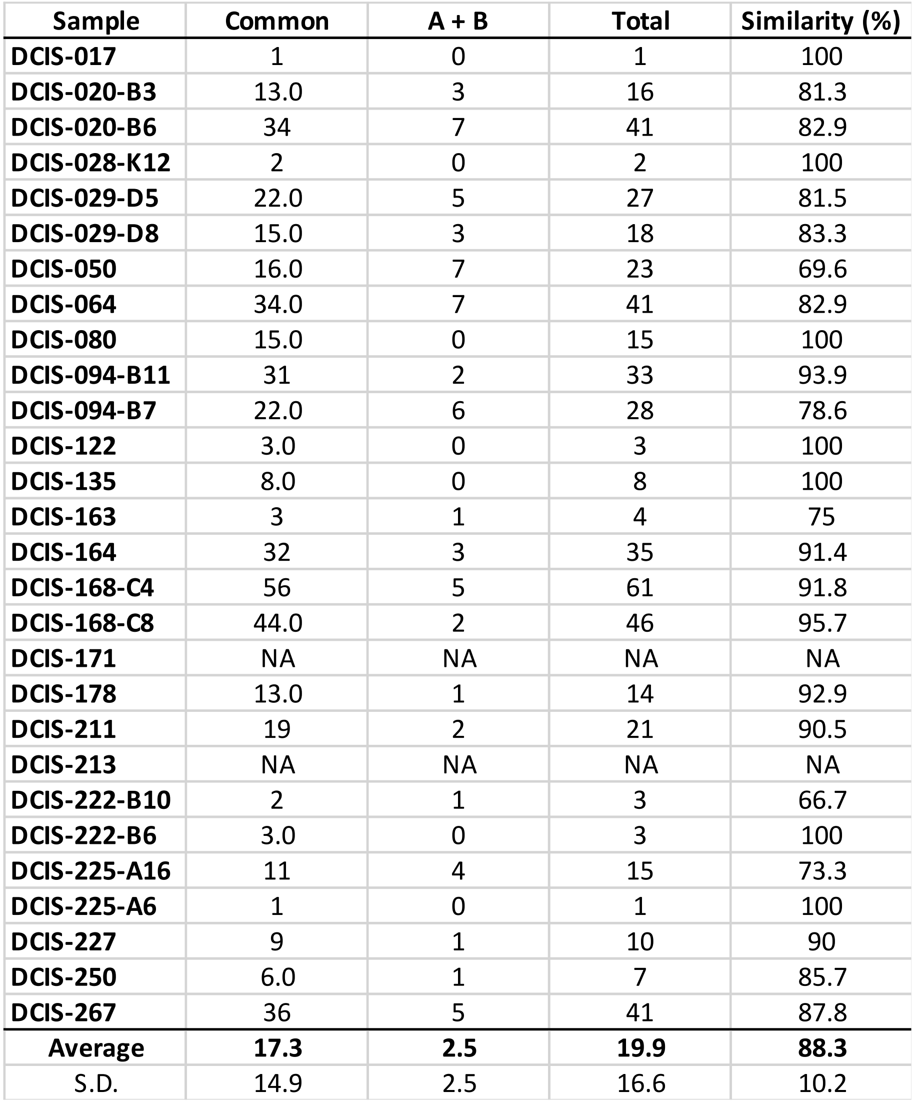


Figure 3: Empirical optimization of the variant post-processing algorithm. Each violin plot summarizes the distribution of optimization scores of 5,308,416 combinations of values of the 13 parameters that control the pipeline for one of the 28 technical replicates (same DNA sample processed twice independently). The optimization score indicates the two-dimensional euclidean distance to the theoretical optimum value of similarity between technical replicates (1) and proportion of final common variants that have a population allele frequency below 0.05 (1) relative to the maximum possible distance. After parameter optimization the similarity between the technical replicates was 88.3 %, range 66.7-100% on average (x= score before optimization; — : score after optimization; colors indicate the amount (ng) of DNA used as template).



Sample	Common	A + B	Total	Similarity (%)
DCIS-017	1	0	1	100
DCIS-020-B3	13.0	3	16	81.3
DCIS-020-B6	34	7	41	82.9
DCIS-028-K12	2	0	2	100
DCIS-029-D5	22.0	5	27	81.5
DCIS-029-D8	15.0	3	18	83.3
DCIS-050	16.0	7	23	69.6
DCIS-064	34.0	7	41	82.9
DCIS-080	15.0	0	15	100
DCIS-094-B11	31	2	33	93.9
DCIS-094-B7	22.0	6	28	78.6
DCIS-122	3.0	0	3	100
DCIS-135	8.0	0	8	100
DCIS-163	3	1	4	75
DCIS-164	32	3	35	91.4
DCIS-168-C4	56	5	61	91.8
DCIS-168-C8	44.0	2	46	95.7
DCIS-171	NA	NA	NA	NA
DCIS-178	13.0	1	14	92.9
DCIS-211	19	2	21	90.5
DCIS-213	NA	NA	NA	NA
DCIS-222-B10	2	1	3	66.7
DCIS-222-B6	3.0	0	3	100
DCIS-225-A16	11	4	15	73.3
DCIS-225-A6	1	0	1	100
DCIS-227	9	1	10	90
DCIS-250	6.0	1	7	85.7
DCIS-267	36	5	41	87.8
Average	**17.3**	**2.5**	**19.9**	**88.3**
S.D.	14.9	2.5	16.6	10.2

Table 1: Similarity between technical replicates and number of variants.

The similarity between technical replicates on average is 88.3%, range 66.7-100%. Number of total, common and private SNVs. Common SNVs: SNVs detected in both replicas of the same DNA samples; Private SNVs: SNVs detected only in one of the two DNA sequences of the same DNA.

Intratumor genetic heterogeneity estimation pipeline

In order to estimate the genetic heterogeneity between two samples (A, B), we applied the concept that the presence of a high confidence variant in one sample should increase the confidence of that variant in the other sample. This concept could also be applied to multi-region sequencing projects. We implemented this in a crossed unequal comparison scheme (Fig. 2), by which the set of filtered variants detected in a sample is compared against all variants estimated in the other sample. This comparison is then reversed, to finally integrate the result of the two comparisons by considering any variant found common in either comparison as common, and private otherwise. Thus, if a variant has been detected with high confidence in one sample and has also been detected in the other sample–even if with low confidence–the variant is considered present in both samples. However, if a variant is detected with low confidence in both samples the variant is discarded, preventing an artificial increase in the confidence of shared variants. Finally, variants that are detected with high confidence in only one sample and not detected even at low confidence in the other sample, are considered private. Before the integration step, the algorithm refines the variants removing detected germline variants, known germline variants in human populations, and variants with insufficient coverage in either the normal sample (all variants) or the other sample (private variants) (see Methods for additional details).

Validation of filtering parameters

We performed a 5-fold cross-validation study to assess the sensitivity of the optimization strategy to input data, and how well the algorithm generalizes to independent datasets. The optimization strategy is extremely robust to the input data, returning a mean evaluation score (empirical cumulative distribution of test score) of 0.98, range 0.95 - 1 and being equal to the optimal overall in 4 out of 5 folds (Suppl. fig. 1S). Additionally, this experiment shows the robustness of the overall optimal model across different cross-validation folds, being the model with the highest mean test and training score in this cross-validation analysis. The test score of the overall optimal model is always as good or better than the model selected based on the training score for each fold.

Sensitivity analysis of the number of technical replicates

We saw a fast increase in the relative score, reaching a plateau with just 6 technical replicates and only obtaining diminishing returns when going over 10 technical replicates (Suppl. fig. 2S). With 6 technical replicates the results are very close to the ones obtained using the whole dataset, with a mean empirical distribution of the optimization score of 0.97.

Validation of somatic variants

In order to validate the identified mutations with our new method, we analyzed the same DNA used for the exome sequences using targeted primers and the AmpliSeq^{TM} technology. We achieved an average of 18,311 (tumor) and 14,906 (control) read coverage for each single nucleotide variant in the validation set. We found that insertion-deletion variants are an unreliable sub-set of mutations (58 indels tested: 19 (32.7%) indels fully validated, 16 (27%) indels partially validated, in which not all nucleotides have been confirmed). The comparison of the data confirmed 83.9% (with stringent parameters, S) and 81.2% (with relaxed parameters, P) single nucleotide variants identified by applying our pipeline to the exome sequence (Table 2). We found 12 (S) or 15 (P) of the unconfirmed variants belong to the same gene MUC6 characterized by highly repetitive sequences, thus subject to read alignment errors and known to have an unreliable reference sequence (Svensson et al. 2018). Excluding MUC6 (16 (S) or 17 (P) variants), we validated 90.1% (S) or 85.2% (P) of the remaining variants. We found that 17.6% (S) and 16.6 % (P) of the confirmed variants are also present in the control samples with a frequency >10%; thus, these could be SNPs and not somatic mutations (Table 2). However, the expected frequency (50%) of the two alternative alleles of a germline SNP only occurs in 7 (S) or 16 (P) cases, if we include alleles with frequency >40% (Suppl. table 2S, Fig. 4S). Importantly, we found a strong negative correlation between the amount of input DNA used (20, 40, 60, 80, >100) for the NGS libraries and the inability to identify correctly the SNPs including all variants detected with a frequency >1% in the germ line DNA (Spearman correlation r = -1, p<0.01 (S and

R); Suppl. table 2S). Excluding MUC6 variants and DNA samples with less than 40 ng, we validated 94% (S) or 91% (P) of the variants, however, 7 (6%) (S) or 13 (6.3%) (P) variants were detected only in one of the two technical replicates.

Stringent filter (S)	Variants	Common (A and B)	Private (A or B)	Variants in controls (>10%)
Total number of SNVs	168	151 (89.9%)	17 (10.1%)	31 (18.4%)
Validated variants	141 (83.9%)	128 (85%)	13 (76.5%)	30 (96.8%)
Non-validated variants	27 (16.1%)	23 (15%)	4 (23.5%)	1 (3.2%)
MUC6-excluded SNVs	152	140 (92.1%)	12 (7.9%)	30 (19.3%)
Validated variants	140 (90.1%)	127 (90.7%)	10 (83.3%)	30 (100%)
Non-validated variants	15 (9.9%)	13 (9.3%)	2 (16.7%)	0 (0%)
MUC6-excluded, DNA ≧ 40ng SNVs	116	109 (94%)	7 (6%)	14 (12%)
Validated variants	109 (94%)	102 (93.6%)	7 (100%)	14 (100%)
Non-validated variants	7 (6%)	7 (6.4%)	0 (0%)	0 (0%)
Permissive filter (P)	**Variants**	**Common (A and B)**	**Private (A or B)**	**Variants in controls (>10%)**
Total number of SNVs	308	279 (90.6%)	32 (10.4%)	51 (16.6%)
Validated variants	250 (81.2%)	236 (84.6%)	19 (59.4%)	49 (96.1%)
Non-validated variants	58 (18.8%)	43 (15.4%)	13 (40.63%)	2 (3.9%)
MUC6-excluded SNVs	291	195 (93.8%)	13 (6.3%)	50 (17.2%)
Validated variants	248 (85.2%)	172 (88.2%)	10 (76.9%)	49(98%)
Non-validated variants	43 (14.8%)	23 (11.8%)	3 (23.1%)	2 (0%)
MUC6-excluded, DNA ≧ 40ng SNVs	208	188 (90.4%)	12 (6%)	23 (11.1%)
Validated variants	190 (91%)	172 (91.5)	10 (83.3%)	21 (91.3%)
Non-validated variants	18 (9%)	16 (8.5%)	2 (16.7%)	2 (8.7%)

Table 2: Validation.

Targeted sequencing confirmed that 83.9% (Stringent filtering pipeline) and 81.2% (Relaxed filtering pipeline) of single nucleotide variants identified using our algorithm. Excluding MUC6, subject to sequencing errors we validated 90.1% (S) or 85.2% (P) of the remaining variants. Excluding MUC6 and low input amounts of DNA we validated 94% (S) or 91% (P) of variants. We found that the 12% (S) or 11.1% (P) of the confirmed variants are also present in the control samples with a frequency >10%. These variants may be SNPs.

Breast cancer genetic and functional divergence

Notably, the IDC samples have a significantly higher number of variants (on average) than pure DCIS samples ((S) one-way ANOVA (*F*(2,17) = 5.228 p= 0.017, post-hoc Tukey test: pure DCIS vs synchronous DCIS=NS, pure DCIS vs IDC= p=0.013, synchronous DCIS vs IDC=NS; (P) one-way ANOVA (*F*(2,18) = 5.406 p= 0.014, post-hoc Tukey test: pure DCIS vs synchronous

DCIS=NS, pure DCIS vs IDC= p=0.011, synchronous DCIS vs IDC=NS). Moreover, our method allowed us to measure the genetic divergence (heterogeneity) between pairs of samples taken from different locations in 6 different tumors. Our preliminary analysis based on 6 patients resulted in (S) 77.3% ± 15.9 S.D. and (P) 78.7% ± 10.5 S.D. genetic divergence.

The analysis of the variants allowed us to identify mutated genes that are typical of breast cancer such as *MDM2*, *TP53*, *NCOR1*, *PIK3CA*, *PIK3CA* (Suppl. table 3S). There is a marginal overlap of mutated genes among the 3 different tumors analyzed: pure DCIS, synchronous DCIS, IDC = 1% (S), 1.6% (P); pure DCIS, IDC = 2.3% (S), 3.9% (P); pure DCIS, synchronous DCIS = 7.6% (S), 5.8% (P); synchronous DCIS, IDC = 3.2% (S), 3.5% (P).

Discussion

Cancer is a disease of clonal evolution, and intra-tumor heterogeneity is its fuel. Unfortunately, this heterogeneity poses a challenge for traditional sampling and prognosis, as different biopsies may sample different clones with variable relevance to the future behavior of the tumor. However, because heterogeneity itself helps to drive clonal evolution, measurements of heterogeneity itself may be prognostic. Our previous studies of metrics of intratumor heterogeneity (which we also refer to as genetic diversity), showed that one robust measure is the degree to which two samples from the same tumor have genetically diverged (Merlo et al. 2010). This measure has the useful property that the more of the genome that is sequenced, the more accurate it gets. We hypothesized that ductal carcinoma *in situ* (DCIS) with more clonal heterogeneity would be more likely to progress to invasion and metastasis. But before we can test that hypothesis, we need a reliable method to measure clonal heterogeneity in this experimental system. Here we have developed, characterized, and validated a reliable method to measure genetic divergence from two FFPE derived DNA samples from the same tumor, solving this limitation.

The sequencing of reduced quantities of DNA extracted from FFPE samples leads to substantial sequencing errors that require correction in order to obtain accurate detection of somatic mutations. Variant calling software packages need to be optimized to reduce the impact of sequencing errors. This is particularly important in the study of heterogeneity, as well as precision medicine, as both false positive and false negative detection of mutations are important for clinical care decisions, and diminish the predictive power of heterogeneity as a risk marker .

Any study of tumor heterogeneity using comparable DNA samples needs to account for and minimize technical variation. We found 88.3% of the variants were detected in both duplicated sequences and 94% excluding the MUC6 gene and those samples with <40ng input DNA. This despite the fact that we included old FFPE samples and that we integrated in the analysis sequences with small inserts that are the main source of sequence errors. We found that keeping reads with small inserts in the analysis improved the quality of the results possibly because the DNA extracted from FFPE is fragmented and their exclusion diminishes the ability to detect variants. Both levels of filtering stringency tested (Stringent and Permissive) have proven successful. As expected, the relaxed version of the algorithm allows the detection of a higher number of variants in exchange for a small reduction of accuracy. It is remarkable that, when not using a post-processing pipeline such as the one presented here, variant callers like Platypus and Mutect2 generate very inaccurate results in this system, with similarities between the technical replicates of only 17.8% and 2.4% respectively.

We have validated the bioinformatic algorithm by re-sequencing the regions containing the variants using a different sequencing technique: AmpliSeq^{TM}. This technology allows for a deep re-sequencing of the regions of interest, improving our ability to identify mutations correctly. The comparison between the data obtained with these two techniques allowed us to validate the new algorithm. Among these, some are presumably SNPs and not somatic variants. However, the frequency of the two alternative alleles is often far from the expected frequency of 50%. This could be because of difficulties encountered when sequencing with AmpliSeq™ to analyze DNA extracted from FFPE, or biological signals of neoplastic DNA present in the control samples. The fact that there is a strong statistically significant negative correlation between the amount of DNA used for the preparation of the libraries and the presence of SNPs detected as SNVs suggests that it is recommended to use DNA quantities at least 40 ng. In particular, this result indicates that the quality and quantity of control DNA is a key factor in the ability to correctly identify somatic mutations in tumors. In many instances, control DNA is not a limiting factor and higher amounts can be used for the preparation of the NGS libraries. Moreover, control samples could be collected during surgery or from blood cells, obtaining DNA from specimens that have not undergone the effect of fixation and DNA deterioration. Our algorithm allows us to modulate the stringency of SNP filtering parameters and to obtain the frequency of each potential SNP in the population.

The variants detected using our algorithm were distributed over the entire exome and we have cataloged numerous mutations in well-known breast cancer genes. The marginal overlap between the genes mutated in the different tumor groups suggests a functional differentiation between the groups that should be investigated with a higher number of patients. Despite the small number of samples analyzed, we found a statistically significant difference in the number of mutations between the invasive ductal carcinoma samples and the DCIS samples. This result encourages the application of this method to a larger cohort of patients focusing on the study of DCIS heterogeneity to identify tumors that may be at elevated risk of progressing to invasive cancer. We are in the process of applying this approach to a larger cohort of patients.

The current version of our algorithm has been developed and implemented to analyze only two samples per patient, which fits our needs but is too restrictive for many multi-region datasets. The generalization of this algorithm to use any number of samples per patient is not complicated but has some nuances that may need to be adjusted depending on the final purpose of the called SNVs. The removal of variants with insufficient coverage in other samples is the main focus of these decisions. For example, for a downstream analysis that does not integrate uncertainty easily, the algorithm could require enough coverage in most (or all) samples, discarding variants with a lot of missing data, while for other applications those SNVs could be kept if they are at least present in another sample, assigning missing values or a measure of uncertainty to samples with insufficient coverage. The core step of the algorithm—comparison of filtered and unfiltered sets of variants—could be kept as it is. However, we also envision more stringent alternatives in which a variant need to be present in more than one non-filtered sample to be kept in the final set. The removal of germline variants and SNPs would remain, since it does not depend on the number of samples.

We developed a bioinformatics pipeline to analyze pairs of DCIS samples taken from the same neoplasm. We identified the mutations present in each sample and we showed that this method is capable of identifying different levels of genetic heterogeneity. This algorithm is easily modifiable and can be integrated with additional parameters, allowing investigators to choose different levels of filtering stringency. These parameters can be re-optimized for a different experimental system with as few as six sets of technical replicates, and the optimized set of parameters is robust to changes in the input data. These characteristics make our algorithm readily translatable to large tissue banks of FFPE samples of any neoplasm.

Methods

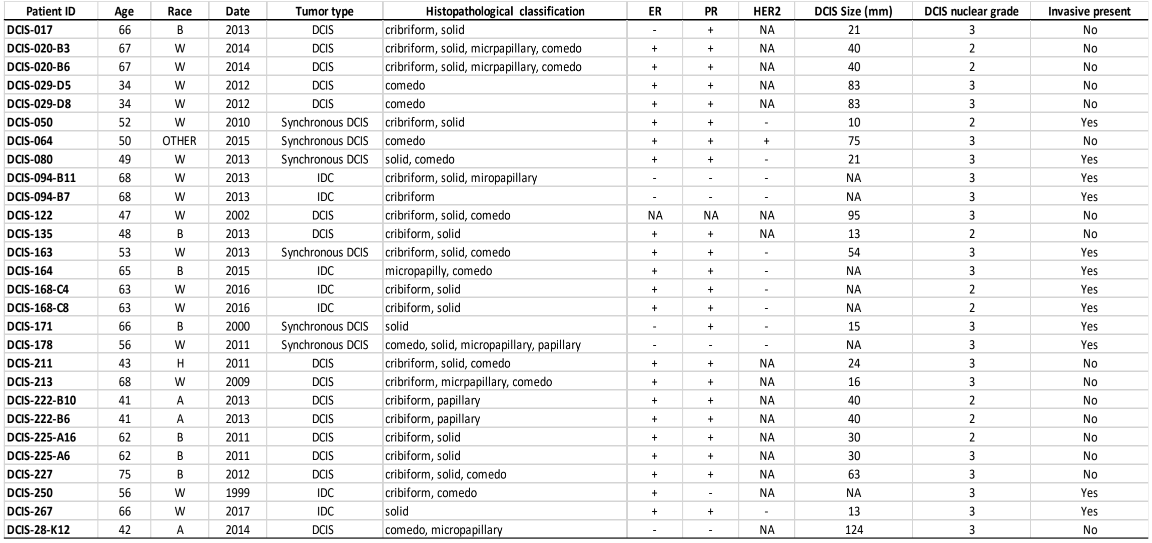
Patients clinical data and biological samples.

This study was approved by the Institutional Review Board (IRB) of Duke University Medical Center, and a waiver of consent was obtained, according to the approved protocol. Formalin-fixed paraffin embedded (FFPE) breast tissue blocks were retrieved from Duke Pathology archives. All cases underwent pathology review (AH) for tissue diagnosis and case eligibility.

A total of 22 separate patients are included in this study (Table 3). All DNA was extracted from archival formalin fixed paraffin embedded thin sections stained with hematoxylin.

For tumors, the study pathologist identified areas of DCIS or invasive cancer that were macrodissected to enrich for tumor epithelial cells. Control DNA was extracted from either distant benign areas of the breast or a benign lymph node using the same procedure employed for the tumor containing areas. These benign areas were confirmed to be devoid of tumor by the study pathologist. A total of 28 breast tumor DNA samples included six patients where two tumor samples located at least 8 mm apart were macrodissected and extracted separately.

Breast tumors were classified using the World Health Organization (WHO) criteria (Tan et al. 2020). Following pathology review, pure DCIS (DCIS not associated with invasion; n=15 tumors, from 11 patients), synchronous DCIS (DCIS identified concurrently with invasive cancer; n=6 tumors, from 6 patients) and invasive =ductal carcinoma (IDC; n=7 tumors, from 5 patients), were included in this study (Table 3). IDCs and DCIS were graded according to the Nottingham grading system (Elston and Ellis 2002) or recommendations from the Consensus conference on DCIS classification ("Consensus Conference on the Classification of Ductal Carcinoma in Situ" 1997), respectively.



Patient ID	Age	Race	Date	Tumor type	Histopathological classification	ER	PR	HER2	DCIS Size (mm)	DCIS nuclear grade	Invasive present
DCIS-017	66	B	2013	DCIS	cribriform, solid	-	+	NA	21	3	No
DCIS-020-B3	67	W	2014	DCIS	cribriform, solid, micrpapillary, comedo	+	+	NA	40	2	No
DCIS-020-B6	67	W	2014	DCIS	cribriform, solid, micrpapillary, comedo	+	+	NA	40	2	No
DCIS-029-D5	34	W	2012	DCIS	comedo	+	+	NA	83	3	No
DCIS-029-D8	34	W	2012	DCIS	comedo	+	+	NA	83	3	No
DCIS-050	52	W	2010	Synchronous DCIS	cribriform, solid	+	+	-	10	2	Yes
DCIS-064	50	OTHER	2015	Synchronous DCIS	comedo	+	+	+	75	3	No
DCIS-080	49	W	2013	Synchronous DCIS	solid, comedo	+	+	-	21	3	Yes
DCIS-094-B11	68	W	2013	IDC	cribriform, solid, miropapillary	-	-	-	NA	3	Yes
DCIS-094-B7	68	W	2013	IDC	cribriform	-	-	-	NA	3	Yes
DCIS-122	47	W	2002	DCIS	cribriform, solid, comedo	NA	NA	NA	95	3	No
DCIS-135	48	B	2013	DCIS	cribiform, solid	+	+	NA	13	2	No
DCIS-163	53	W	2013	Synchronous DCIS	cribriform, solid, comedo	+	+	-	54	3	Yes
DCIS-164	65	B	2015	IDC	micropapilly, comedo	+	+	-	NA	3	Yes
DCIS-168-C4	63	W	2016	IDC	cribiform, solid	+	+	-	NA	2	Yes
DCIS-168-C8	63	W	2016	IDC	cribiform, solid	+	+	-	NA	2	Yes
DCIS-171	66	B	2000	Synchronous DCIS	solid	-	+	-	15	3	Yes
DCIS-178	56	W	2011	Synchronous DCIS	comedo, solid, micropapillary, papillary	-	-	-	NA	3	Yes
DCIS-211	43	H	2011	DCIS	cribriform, solid, comedo	+	+	NA	24	3	No
DCIS-213	68	W	2009	DCIS	cribriform, micrpapillary, comedo	+	+	NA	16	3	No
DCIS-222-B10	41	A	2013	DCIS	cribiform, papillary	+	+	NA	40	2	No
DCIS-222-B6	41	A	2013	DCIS	cribiform, papillary	+	+	NA	40	2	No
DCIS-225-A16	62	B	2011	DCIS	cribiform, solid	+	+	NA	30	2	No
DCIS-225-A6	62	B	2011	DCIS	cribiform, solid	+	+	NA	30	3	No
DCIS-227	75	B	2012	DCIS	cribiform, solid, comedo	+	+	NA	63	3	No
DCIS-250	56	W	1999	IDC	cribiform, comedo	+	-	NA	NA	3	Yes
DCIS-267	66	W	2017	IDC	solid	+	+	-	13	3	Yes
DCIS-28-K12	42	A	2014	DCIS	comedo, micropapillary	-	-	NA	124	3	No

Table 3: Patients clinical data.

Clinical data of the 22 patients included in the study. The histopathological analysis showed that 11 patients are DCIS while 6 are synchronous DCIS and 5 have invasive features (IDC). We selected FFPE samples of different ages (1999-2017). ER: estrogen receptors, PR: progesterone receptors, HER2: human epidermal growth factor receptor 2 expression is qualitatively estimated (positive (+), negative (-), non-present (NP)) using histochemistry stains.

DNA extraction

The DCIS component of all cases as well as IDC from synchronous DCIS cases were macrodissected separately, following hematoxylin staining, of between 10 and 25 five-micron-thick histological sections. The first and last slides were stained with hematoxylin-eosin (H&E) staining, and reviewed by a pathologist to confirm the presence of >=70% of neoplastic cells.

DNA was extracted using the FFPE GeneRead DNA Kit which incorporates enzymatic cleavage of DNA at uracil residues via uracil DNA glycosylase reducing the problem of cytosine deamination (Qiagen, cat n. 180134) according to manufacturers' instructions. DNA quantification was performed using a Qubit™ 1X dsDNA HS Assay Kits (ThermoFisher, cat. n. Q33230), and DNA quality assessed with an Agilent 2100 Bioanalyzer.

DNA sequencing

We sequenced different quantities of genomic DNA (20, 40, 60, 80, 100, >100 ng) to estimate the effects of DNA quantity on the estimation of intratumor genomic heterogeneity. All technical replicates were separated into two aliquots from the same tube of DNA sample before all subsequent steps. Each aliquot was sheared to a mean fragment length of 250 bp using the Covaris LE200 instrument, and Illumina sequencing libraries were generated as dual-indexed, with unique bar-code identifiers, using the Accel-NGS 2S PCR-Free library kit (Swift Biosciences, cat. n. 20096). We pooled groups of 96 equimolar libraries (100 ng/library) for hybrid capture using two target panels, the human exome and a panel containing all exons of the

83 genes in the breast cancer gene panel (BRC83, suppl. table 4S). To capture BRC83 we used biotinylated "ultramer" oligonucleotides synthesized by Integrated DNA Technologies (Coralville, Iowa), and to capture the human exome we used IDT's xGen Exome Research Panel v1.0. After hybridization, capture pools were quantitated via qPCR (KAPA Biosystems kit). We sequenced the final product using an Illumina HiSeq 2500 1T instrument multiplexing nine tumor samples per lane.

After binning the sample data according to its index identifier, we aligned it to the Genome Reference Consortium Human Build 37 using the BWA-MEM (Li, 2013) algorithm, and marked sequencing duplicates with Picard's MarkDuplicates. The resulting BAM files are the input data for our pipeline for intratumor genetic heterogeneity calculation. We discarded samples with less than 40% of the target covered at 40X (Suppl. table 1S). This sequencing protocol was performed at the McDonnell Genome Institute at Washington University School of Medicine in St. Louis.

Intratumor genetic heterogeneity estimation pipeline

We implemented our heterogeneity estimation pipeline (Fig. 2) in a series of Perl scripts, tailored to be run at Arizona State University's research computing high performance computing clusters. Variants are first called using Platypus 0.8.1 (Rimmer et al. 2014) against the Genome Reference Consortium Human Build 37 reference genome using the default settings except for the parameters regulated during pipeline optimization (Fig. 2): The inclusion of reads with small inserts (--filterReadPairsWithSmallInserts), and the minimum number of reads supporting a variant to consider it for calling (--minReads). Before downstream analyses, our pipeline splits multiallelic sites into biallelic sites, and clusters of variants into individual SNVs. The variant filtering step uses SnpSift 4.2 (Ruden et al. 2012) (Phred Quality: QUAL, Coverage: GEN[*].NR[*], Forward and Reverse variant reads: NF & NR, Variant reads: GEN[*].NV[*]). The depth estimation step, which estimates the coverage of the position of a variant in the other samples (and the proportion of reads supporting that specific allele) is carried out by first generating a bed file integrating deletions, insertions, and SNVs using BEDOPS (Neph et al. 2012), and then using it as intervals input for GATK 3.5.0's UnifiedGenotyper, executed to output data for all sites (--output_mode EMIT_ALL_SITES, -glm BOTH). The position filtering step is carried out in the inhouse pipeline with these results. This step differs slightly in the comparison between tumor samples and the comparison against the normal. In the first case, a variant is discarded if any of the conditions is not met, while in the second both the allele frequency and the number of variants need not be met for them to trigger the discard of a variant while the coverage filter acts independently. Importantly, while the steps of variant removal are generally applied to all sets of variants (e.g., removal of germline variants, candidate SNPs, and positions with lack of support in the normal), the removal of variants based on insufficient coverage in the other tumor samples only applies to private variants.

Population allele frequency estimates are obtained from the gnomAD 2.1.1 genomic database (Karczewski et al. 2020), which spans 15,708 whole-genome sequences, and filtering using this information is carried out within our pipeline. All variant comparisons within our pipeline are genotype specific.

We also implemented an alternative version of this pipeline identifying somatic mutations using Mutect2 (McKenna et al. 2010) version 4.0.5.0 for comparison purposes against a developing version of our pipeline, both lacking the population allele frequency step (Fig. 2), and using slightly different parameter values, which were optimal at that stage of development (Suppl. table 5S). To use this variant caller, first we generated a panel of normals using all control tissue samples and the CreateSomaticPanelOfNormals GATK command. Then, we called variants on all paired tumor files using the panel of normals, IDT's xGen Exome Research Panel v1.0, and the AllowAllReadsReadFilter. We filtered the resulting variants with an equivalent re-implementation of our post-processing pipeline that uses Bcftools isec to perform comparisons between sets of variants and ran FilterMutectCalls to obtain the final calls.

Optimization of the intratumor genetic heterogeneity pipeline

We assigned a range of values to explore for each of the 13 parameters that control the genetic heterogeneity estimation pipeline (Fig. 2) and explored every possible combination of them with the data from all 28 technical replicates, assessing a total of 5,308,416 parameter combinations. We calculated the score of a condition (set of parameter values) as the minimum value of the 90% confidence interval of the mean (p=0.9) of the scores of that condition across the 28 technical replicates. We used this statistic to integrate central tendency and dispersion in the same measure. The score of each technical replicate was calculated as the two-dimensional euclidean distance to the theoretical optimum value of similarity between technical replicates (1) and proportion of final common variants that have a population allele frequency below 0.05 (1) relative to the maximum possible distance. This score ranging from 0 to 1, allowed us to co-optimize the similarity between technical replicates and the sets of variants with the least chance of being dominated by germline variants not detected in the normal and detected as somatic common variants. We performed a 5-fold cross-validation study stratified by amount of DNA, in which patients were partitioned randomly into 5 subsets, with at least 1 patient from each DNA category 20, 40, 60, 80, ≥100 ng. In each of the 5 interactions, one of the subsets (testing set) was held out of the parameter optimization and then evaluated based on the optimal parameter values obtained from the training set. We implemented the optimization and cross-validation steps in R (Team and Others 2013), using the LSR (Navarro 2015), and cowplot (Wilke, n.d.) packages.

Sensitivity analysis on the number of technical replicates

We subsampled our dataset to create smaller technical replicate datasets of k={2,...,28} sizes. For each k, we generated all combinations of size k with our 28 technical replicates, and took a random sample of 104 of them (or all if ≤104) without replacement. We optimized the pipeline using each of these resampled subsets and reported the empirical cumulative probability of its optimization score using all samples. This statistic indicates how this resulting pipeline compares with the overall optimal pipeline in the complete dataset.

Validation of somatic variants

In order to validate the robustness of the method we used both the optimized stringent (S) parameter values and a permissive (P) version of the algorithm (minimum number of forward and reverse reads supporting the variant=10 instead of 15). The permissive version allowed us to increase the number of the variants selected. We randomly selected for validation a subset of single nucleotide variants (S=168 out of 517, P=308 out of 1047) and insertion-deletion mutations (S= 22 out of 265, P=57 out of 512) sequencing DNA amplicons containing the variants detected with our bioinformatic algorithm by targeted re-sequencing using AmpliSeq^{TM} technology (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's specification. The AmpliSeq^{TM} technology allows for a deep re-sequencing of the regions of interest, improving our ability to identify mutations correctly. We re-sequenced both tumor and control samples. Alternative alleles were validated if their frequency was ≥1%.

Variants annotation and functional analysis

We annotated the variants using Annovar 20140714 (Wang, Li, and Hakonarson 2010), Ensembl (Yates et al. 2020) and we investigated their function using DAVID (Huang, Sherman, and Lempicki 2009a, [b] 2009).

Software availability

All software developed to carry out this study is distributed under the GPLv3 license. The implementation of the intratumor heterogeneity estimation pipeline—*ITHE*, can be found at https://github.com/adamallo/ITHE, scripts to carry out the cross-validation study and data analysis can be found at https://github.com/adamallo/ITHE_analyses, and the alternative implementation of our intratumor genetic heterogeneity pipeline using Mutect2 to call variants can be found at https://github.com/icwells/mutect2Parallel.

Acknowledgments

This work was supported primarily by a CDMRP Breast Cancer Research Program Award BC132057, as well as NIH grants U54 CA217376, U2C CA233254, P01 CA91955, R01 CA170595, R01 CA185138 and R01 CA140657, and the Arizona Biomedical Research Commission grant ADHS18-198847. The findings, opinions and recommendations expressed here are those of the authors and not necessarily those of the universities where the research was performed or the funding bodies.

Author Contributions

A.F. and D.M. developed the method and analyzed the data. S.M.R. contributed to the data analysis and implemented the alternative version of the bioinformatic pipeline that uses Mutect2. L.K. and T.H. collected patient clinical data and extracted the DNA from FFPE samples. A.H. review the histopathological status of the samples. A.F. and D.M. wrote the manuscript with support from C.C.M., J.R.M. and E.S.H. All authors discussed the results. C.C.M., J.R.M. and E.S.H supervised the project.

Competing Interests statement

The authors declare that they have no competing interests.

References

Andor, Noemi, Trevor A. Graham, Marnix Jansen, Li C. Xia, C. Athena Aktipis, Claudia Petritsch, Hanlee P. Ji, and Carlo C. Maley. 2016. “Pan-Cancer Analysis of the Extent and Consequences of Intratumor Heterogeneity.” *Nature Medicine* 22 (1): 105–13.

Bedard, Philippe L., Aaron R. Hansen, Mark J. Ratain, and Lillian L. Siu. 2013. “Tumour Heterogeneity in the Clinic.” *Nature* 501 (7467): 355–64.

Carrick, Danielle Mercatante, Michele G. Mehaffey, Michael C. Sachs, Sean Altekruse, Corinne Camalier, Rodrigo Chuaqui, Wendy Cozen, et al. 2015. “Robustness of Next Generation Sequencing on Older Formalin-Fixed Paraffin-Embedded Tissue.” *PloS One* 10 (7): e0127353.

Chen, Guoli, Stacy Mosier, Christopher D. Gocke, Ming-Tseh Lin, and James R. Eshleman. 2014. “Cytosine Deamination Is a Major Cause of Baseline Noise in next-Generation Sequencing.” *Molecular Diagnosis & Therapy* 18 (5): 587–93.

“Consensus Conference on the Classification of Ductal Carcinoma in Situ.” 1997. *Human Pathology* 28 (11): 1221–25.

Dash, Sajal, Nicholas A. Kinney, Robin T. Varghese, Harold R. Garner, Wu-Chun Feng, and Ramu Anandakrishnan. 2019. “Differentiating between Cancer and Normal Tissue Samples Using Multi-Hit Combinations of Genetic Mutations.” *Scientific Reports* 9 (1): 1005.

Do, Hongdo, and Alexander Dobrovic. 2015. “Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization.” *Clinical Chemistry* 61 (1): 64–71.

Elston, C. W., and I. O. Ellis. 2002. “Pathological Prognostic Factors in Breast Cancer. I. The Value of Histological Grade in Breast Cancer: Experience from a Large Study with Long-Term Follow-Up. C. W. Elston & I. O. Ellis. Histopathology 1991; 19; 403-410. AUTHOR COMMENTARY.” *Histopathology*. https://doi.org/10.1046/j.1365-2559.2002.14691.x.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009a. “Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources.” *Nature Protocols*.

https://doi.org/10.1038/nprot.2008.211.

———. 2009b. “Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists.” *Nucleic Acids Research* 37 (1): 1–13.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

Maley, Carlo C., Patricia C. Galipeau, Jennifer C. Finley, V. Jon Wongsurawat, Xiaohong Li, Carissa A. Sanchez, Thomas G. Paulson, et al. 2006. “Genetic Clonal Diversity Predicts Progression to Esophageal Adenocarcinoma.” *Nature Genetics* 38 (4): 468–73.

Maley, C. C., A. Aktipis, T. A. Graham, A. Sottoriva, A. M. Boddy, M. Janiszewska, A. S. Silva, et al. 2017. “Classifying the Evolutionary and Ecological Features of Neoplasms.” *Nature Reviews. Cancer* 17 (10): 605–19.

Martinez, P., M. R. Timmer, C. T. Lau, S. Calpe, C. Sancho-Serra Mdel, D. Straub, A. M. Baker, et al. 2016. “Dynamic Clonal Equilibrium and Predetermined Cancer Risk in Barrett’s Oesophagus.” *Nature Communications* 7: 12158.

Marusyk, Andriy, and Kornelia Polyak. 2010. “Tumor Heterogeneity: Causes and Consequences.” *Biochimica et Biophysica Acta* 1805 (1): 105–17.

McGranahan, Nicholas, and Charles Swanton. 2015. “Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution.” *Cancer Cell* 27 (1): 15–26.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303.

Merlo, Lauren M. F., Najaf A. Shah, Xiaohong Li, Patricia L. Blount, Thomas L. Vaughan, Brian J. Reid, and Carlo C. Maley. 2010. “A Comprehensive Survey of Clonal Diversity Measures in Barrett’s Esophagus as Biomarkers of Progression to Esophageal Adenocarcinoma.” *Cancer Prevention Research* 3 (11): 1388–97.

Morris, L. G., N. Riaz, A. Desrichard, Y. Senbabaoglu, A. A. Hakimi, V. Makarov, J. S. Reis-Filho, and T. A. Chan. 2016. "Pan-Cancer Analysis of Intratumor Heterogeneity as a Prognostic Determinant of Survival." *Oncotarget* 7 (9): 10051–63.

Neph, Shane, M. Scott Kuehn, Alex P. Reynolds, Eric Haugen, Robert E. Thurman, Audra K. Johnson, Eric Rynes, et al. 2012. "BEDOPS: High-Performance Genomic Feature Operations." *Bioinformatics* 28 (14): 1919–20.

Pandya, Sonali, and Richard G. Moore. 2011. "Breast Development and Anatomy." *Clinical Obstetrics and Gynecology* 54 (1): 91–95.

Rimmer, Andy, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R. F. Twigg, WGS500 Consortium, Andrew O. M. Wilkie, Gil McVean, and Gerton Lunter. 2014. "Integrating Mapping-, Assembly- and Haplotype-Based Approaches for Calling Variants in Clinical Sequencing Applications." *Nature Genetics* 46 (8): 912–18.

Ruden, Douglas, Pablo Cingolani, Viral Patel, Melissa Coon, Tung Nguyen, Susan Land, and Xiangyi Lu. 2012. "Using Drosophila Melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift." *Frontiers in Genetics* 3: 35.

Sah, Sachin, Liangjing Chen, Jeffrey Houghton, Jon Kemppainen, Adam C. Marko, Robert Zeigler, and Gary J. Latham. 2013. "Functional DNA Quantification Guides Accurate next-Generation Sequencing Mutation Detection in Formalin-Fixed, Paraffin-Embedded Tumor Biopsies." *Genome Medicine* 5 (8): 77.

Sims, Andrew H., Anthony Howell, Sacha J. Howell, and Robert B. Clarke. 2007. "Origins of Breast Cancer Subtypes and Therapeutic Implications." *Nature Clinical Practice. Oncology* 4 (9): 516–25.

Svensson, Frida, Tiange Lang, Malin E. V. Johansson, and Gunnar C. Hansson. 2018. "The Central Exons of the Human MUC2 and MUC6 Mucins Are Highly Repetitive and Variable in Sequence between Individuals." *Scientific Reports* 8 (1): 17503.

Tan, Puay Hoon, Ian Ellis, Kimberly Allison, Edi Brogi, Stephen B. Fox, Sunil Lakhani, Alexander J. Lazar, et al. 2020. "The 2019 WHO Classification of Tumours of the Breast." *Histopathology*.

https://doi.org/10.1111/his.14091.

Team, R. Core, and Others. 2013. “R: A Language and Environment for Statistical Computing.” Vienna, Austria. http://finzi.psych.upenn.edu/R/library/dplR/doc/intro-dplR.pdf.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. “ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data.” *Nucleic Acids Research* 38 (16): e164.

Wilke, C. O. n.d. “Cowplot: Streamlined Plot Theme and Plot Annotations for ‘ggplot2’. R Package Version 0.9. 2; 2017.” *URL https://CRAN. R-Project. Org/package= Cowplot*. https://CRAN.R-project.org/package=cowplot.

Yates, Andrew D., Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, et al. 2020. “Ensembl 2020.” *Nucleic Acids Research* 48 (D1): D682–88.