

## Shallow whole genome sequencing for robust copy number profiling of formalin-fixed paraffin-embedded breast cancers



Suet-Feung Chin<sup>a,b,i,\*</sup>, Angela Santonja<sup>c,l</sup>, Marta Grzelak<sup>a</sup>, Soomin Ahn<sup>d,m</sup>, Stephen-John Sammut<sup>a,b,i</sup>, Harry Clifford<sup>a</sup>, Oscar M. Rueda<sup>a,i</sup>, Michelle Pugh<sup>e</sup>, Mae A. Goldgraben<sup>f</sup>, Helen A. Bardwell<sup>a</sup>, Eun Yoon Cho<sup>g</sup>, Elena Provenzano<sup>h,i</sup>, Federico Rojo<sup>j,k</sup>, Emilio Alba<sup>e,k,l</sup>, Carlos Caldas<sup>a,b,h,i,\*</sup>

<sup>a</sup> Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK

<sup>b</sup> Department of Oncology, University of Cambridge, Cambridge CB2 2QQ, UK

<sup>c</sup> Medical Oncology Service, Hospital Universitario Regional y Virgen de la Victoria, Instituto de Investigación Biomédica de Málaga (IBIMA), Málaga, Spain

<sup>d</sup> Department of Pathology, Seoul National University Bundang Hospital, 82, Gumi-ro 173 Beon-gil, Bundang-gu, Seongnam, Gyeonggi 13620, Republic of Korea

<sup>e</sup> Invivata UK, The Portway Building, Granta Park, Cambridge CB21 6GS, UK

<sup>f</sup> Department of Medical Genetics, University of Cambridge, Cambridge CB2 0QQ, UK

<sup>g</sup> Department of Pathology and Translational Genomics, Samsung Medical Center, Sungkyunkwan University School of Medicine, 50 Irwon-dong, Gangnam-gu, Seoul 135-710, Republic of Korea

<sup>h</sup> Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust, NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK

<sup>i</sup> Cancer Research UK Cambridge Cancer Centre, Cambridge CB2 0QQ, UK

<sup>j</sup> Pathology Department, Instituto de Investigación Sanitaria Fundación Jiménez Díaz (IIS-FJD), Madrid, Spain

<sup>k</sup> GEICAM-Spanish Breast Cancer Research Group, Madrid, Spain

<sup>l</sup> Laboratorio de Biología Molecular del Cáncer, Centro de Investigaciones Médico-Sanitarias (CIMES), Universidad de Málaga, Málaga, Spain

<sup>m</sup> Invivata, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

### ARTICLE INFO

#### Keywords:

Formalin-fixed paraffin-embedded (FFPE)  
Shallow whole genome sequencing (sWGS)  
Copy number (CN) and breast cancer

### ABSTRACT

Pathology archives with linked clinical data are an invaluable resource for translational research, with the limitation that most cancer samples are formalin-fixed paraffin-embedded (FFPE) tissues. Therefore, FFPE tissues are an important resource for genomic profiling studies but are under-utilised due to the low amount and quality of extracted nucleic acids. We profiled the copy number landscape of 356 breast cancer patients using DNA extracted FFPE tissues by shallow whole genome sequencing. We generated a total of 491 sequencing libraries from 2 kits and obtained data from 98.4% of libraries with 86.4% being of good quality. We generated libraries from as low as 3.8 ng of input DNA and found that the success was independent of input DNA amount and quality, processing site and age of the fixed tissues. Since copy number alterations (CNA) play a major role in breast cancer, it is imperative that we are able to use FFPE archives and we have shown in this study that sWGS is a robust method to do such profiling.

### 1. Introduction

Comparative Genomic Hybridisation (CGH) (Kallioniemi et al., 1992) has had a significant impact in the study of cancer genomes. Chromosomal regions gained or lost in the tumor could be easily visualised by hybridization onto normal human metaphase spreads, allowing characterisation of genome-wide copy number alterations (CNA) in tumours (Kallioniemi et al., 1992). Microarrays with DNA probes (cloned DNA or oligonucleotides) spotted onto glass slides representing the entire genome soon replaced normal chromosomes

(Pinkel et al., 1998) making it faster and easier to profile. The importance of characterizing somatic CNAs in cancer is now well established, with a recent TCGA pan-cancer analysis showing that human tumours can be classified into mutation driven (M-class) or copy-number driven (C-class) subtypes. Breast cancer is a C-class cancer type (Ciriello et al., 2013) and we have previously shown that CNAs are the main determinants of the expression architecture of breast cancers. Using gene expression driven in *cis* by CNAs, we have generated a new molecular taxonomy of breast cancer with 10 genomic driver-based subtypes termed Integrative Clusters. The samples used in this analysis

\* Corresponding authors at: Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK.  
E-mail addresses: [suet-feung.chin@cruc.cam.ac.uk](mailto:suet-feung.chin@cruc.cam.ac.uk) (S.-F. Chin), [carlos.caldas@cruc.cam.ac.uk](mailto:carlos.caldas@cruc.cam.ac.uk) (C. Caldas).

were derived from the METABRIC cohort, which encompassed a large biobank of fresh frozen tumor samples collected across five major teaching hospitals in the UK and Canada (Curtis et al., 2012).

Formalin-fixed paraffin-embedded (FFPE) tissue samples are more routinely collected and hence more representative of cancer in the general population. These FFPE archives are a valuable resource for molecular profiling in cancer research. Whilst the fixation process is essential to protect cellular morphology and protein expression, it is detrimental to nucleic acids and results in their chemical modification and degradation. As a result, extraction of DNA from FFPE tissues results in lower yields when compared to extraction from fresh frozen tissues. DNA extracted from FFPE works well for downstream applications using polymerase chain reaction (PCR), particularly for small size amplicons (< 300 base pairs), but for other applications, including microarray based CGH, where efficient labelling of the DNA is dependent on its integrity, its use is more challenging. There have been several studies describing different methods for DNA extraction (Janecka et al., 2015), quality control (van Beers et al., 2006; Dang et al., 2016), labelling (Salawu et al., 2012) and other optimisation protocols (Hosein et al., 2013) to improve the performance of FFPE DNA on microarrays. In the past, we have tried to profile CNAs using FFPE DNA on microarrays with limited success. Only Illumina Infinium and Molecular Inversion Probe (MIP, Affymetrix) arrays yielded good results but these required good quality and at least 200 ng of DNA (Iddawela et al., 2017).

Next generation sequencing has revolutionised cancer genomics. It is now relatively easy and inexpensive to sequence an entire genome. However, as with microarrays, the robustness of the results obtained are dependent on the quality of the input DNA. Two recent studies have demonstrated the feasibility of doing shallow whole-genome sequencing (sWGS) for CNA profiling using DNA extracted from FFPE tissue material (Scheinin et al., 2014; Kader et al., 2016). The first report used 250 ng of DNA from FFPE tissues and a breast cancer cell line to produce libraries and developed an analytical method for sWGS. The second study compared several sequencing library production kits and reported generating successful sequencing libraries with low input DNA in a small number of FFPE samples.

Here we present extensive sWGS data generated from DNA extracted from FFPE breast cancer samples to describe steps to ensure successful libraries.

## 2. Materials

### 2.1. Specimen collection

FFPE tissue samples from invasive breast cancer patients diagnosed between 1997 and 2014 were obtained from several tumor repositories: Addenbrooke's Hospital in Cambridge (n = 62), a consortium of hospitals participating in clinical trials (GEICAM) in Spain (n = 172), and Samsung Medical Center in South Korea (n = 122). In some cases, we extracted DNA from adjacent normal (n = 15) and DCIS (n = 115) samples. Some of the clinical trials samples were biopsies taken at diagnosis (n = 107) and/or surgery (n = 106) where 41 are paired. All tumor samples were collected with informed patient consent and their use for genomics profiling had ethics approval from the institutional review board for each of the biobanks (Cambridge: REC ref. 07/H0308/161; South Korea: 2014-10-041; Spain: NCT00432172 & NCT00841828). Detailed information on the sample cohort is collated in Table 1.

### 2.2. DNA extraction and quality control

DNA was extracted from either one mm cores punched from tissue blocks or from 10 × 30 μm sections (Cambridge and Korea), or 4-6 × 10 μm sections (Spain) from FFPE blocks, using Qiagen QIAmp DNeasy Kits (Qiagen, Germany) according to the manufacturer's

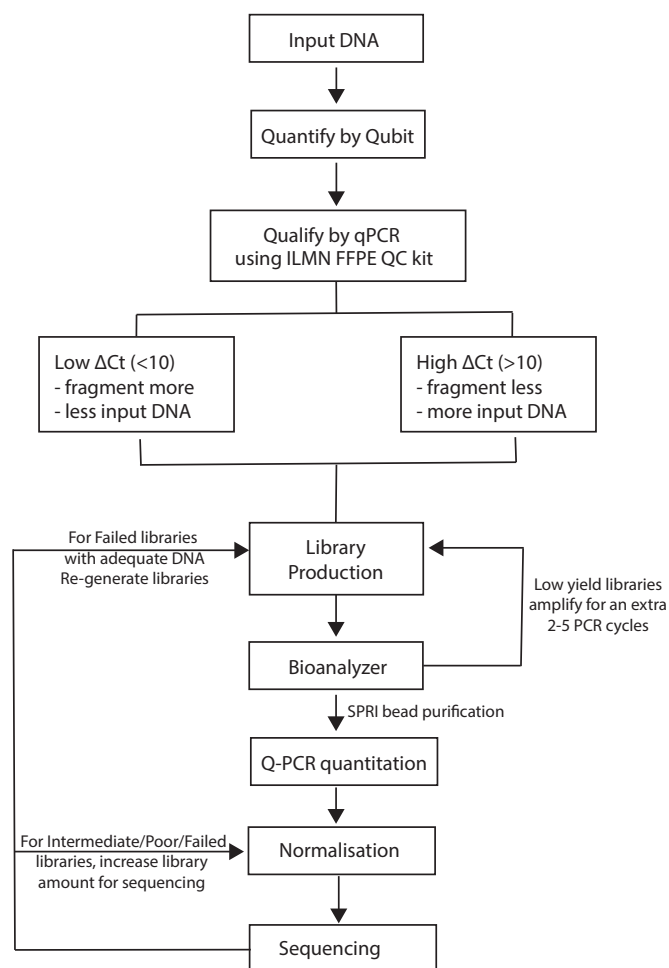
**Table 1**

Features of input DNA and libraries generated from FFPE blocks collected at three different sites. Data provided in minimum-maximum range and median in brackets.

Cohort	Cambridge	Korea	Spain
Patients	62	122	172
Age (years)	2–19 (12.4)	3–8 (6.4)	9–10(9)
DNA quality ( $\Delta$ Ct)	–0.3–8.1 (3.5)	1–17.1* (5.7)	–3.2–7.8 (3.5)
Fragment size (bp)	217–324 (247.5)	197–288* (236)	180–251* (225)
Library yield (nM)	0–77 (17.3)	0–925 (12.3)	0.18–278 (13.8)
Good quality libraries (%)	48 (87.3)	48 (97.1)	163 (74.8)

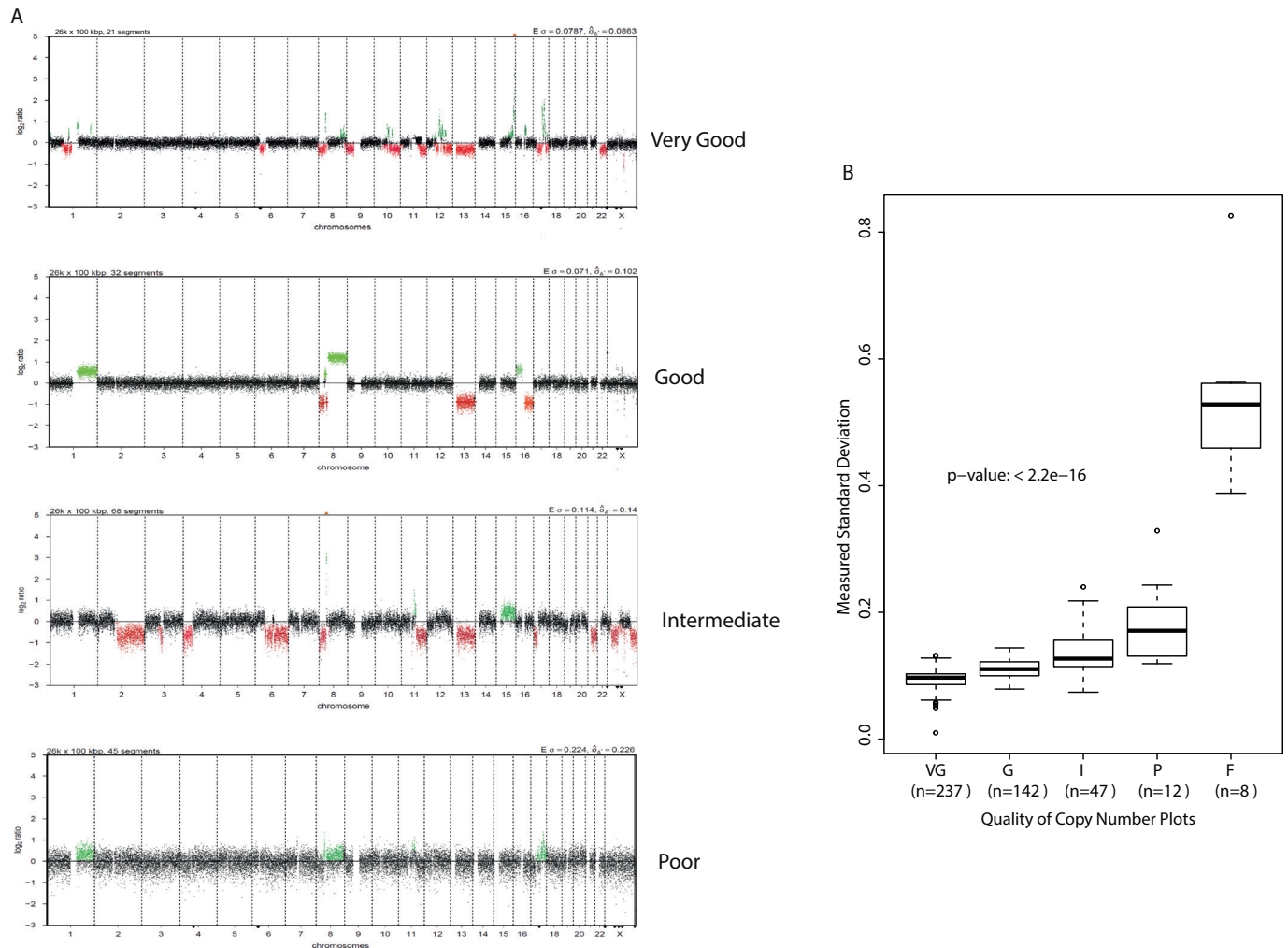
Age = years since blocks were generated.  $\Delta$ Ct = difference between the cycle threshold of test to the control template ACD1 provided in the kit. ng = nanogram, PCR = polymerase chain reaction, bp = base pairs, nM = nanomoles.

\* Denotes the site where there is a significant difference to the index group (ie Cambridge).



**Fig. 1.** Overall Design: Schematic showing the workflow to ensure successful shallow whole genome sequencing (sWGS) libraries.

instructions. All DNA samples were quantified fluorometrically using the Qubit dsDNA High Sensitivity Quantification Reagent (ThermoFisher, USA). The DNA quality was assessed using Illumina's FFPE QC kit, a quantitative PCR (q-PCR) assay. All test DNAs and the template control provided in the kit (ACD1) were diluted to 0.25 ng/μl and PCR reactions set up in triplicate as per manufacturer's instructions. DNA quality was quantified as the difference between the Ct (cycle threshold) value of the test FFPE-extracted DNA against the Ct value of the control DNA template.



**Fig. 2.** Categorisation of copy number profiles.

A. Examples of QDNaseq copy number plots scored as Very Good, Good, Intermediate and Poor. Failed libraries had very few reads and are not shown. Green dots represent regions of gains/amplifications and red dots represent regions of loss/deletion. Dots represent individual samples within each category. VG = very good, G = good, I = intermediate, P = poor, F = fail.

### 2.3. DNA fragmentation

DNA samples of different concentrations (4–500 ng) were diluted in water to a final volume of 15  $\mu$ l in Covaris microTUBE-15 8 strip tubes (Covaris, USA) and fragmented to an average size distribution of 150–180 bp with Covaris LE220 Focused Ultrasonicator with Adaptive Focused Acoustics technology. The following parameters were used for shearing: Peak Incident Power: 180 W; Duty Factor: 30%; Cycles per Burst: 50; with the fragmentation time: 250 s for DNA with  $\Delta$ Ct < 10, and 200 s for DNA with  $\Delta$ Ct  $\geq$  10.

### 2.4. Sequencing library generation

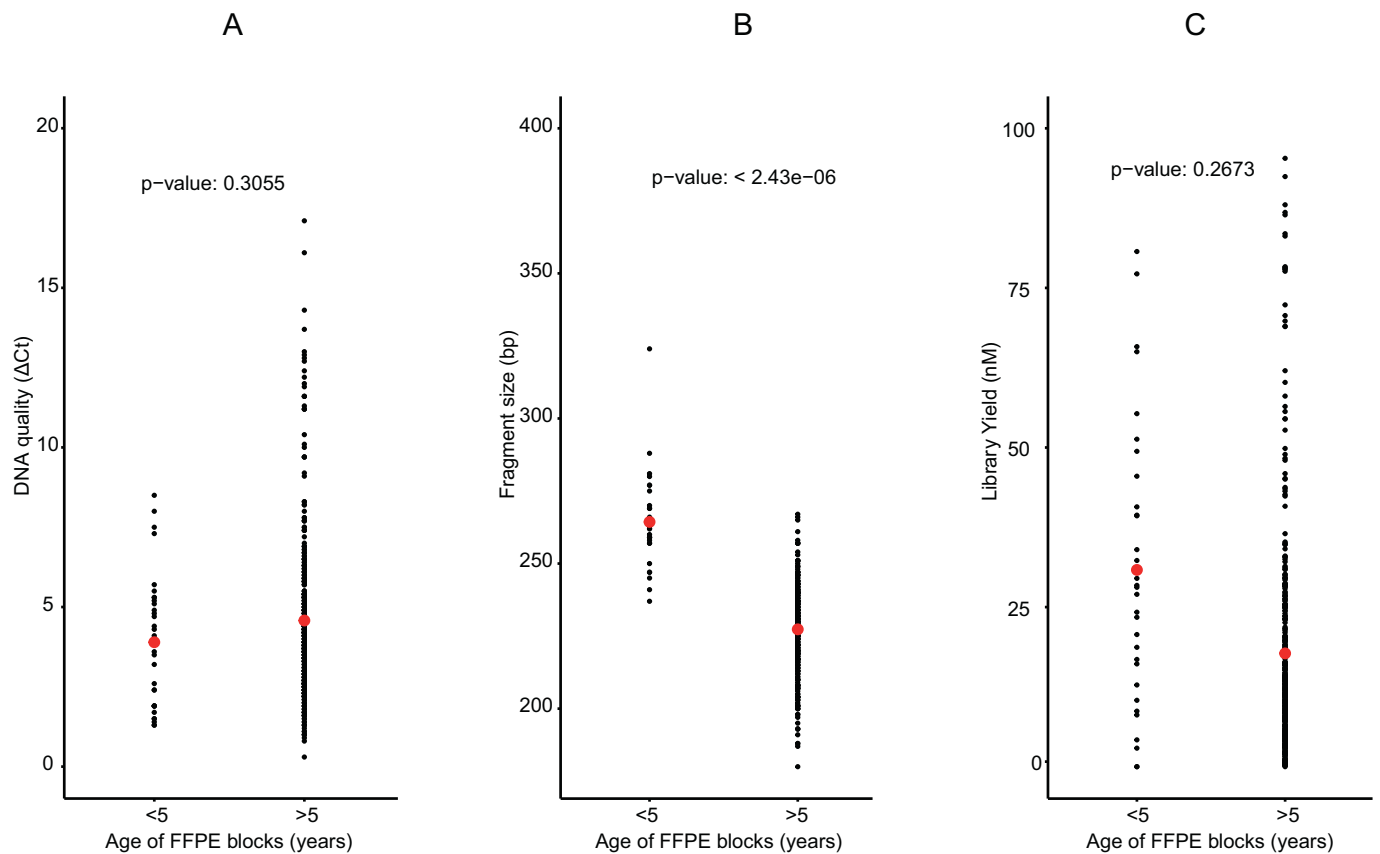
Sequencing libraries were generated using either the beta testing version of the Illumina FFPE TruSeq kit (ILMN, libraries = 45) or the Rubicon Genomics ThruPLEX DNaseq (RGT, libraries = 446), as per manufacturer's instructions. For four samples, we generated sequencing libraries using both kits to compare their performance (Supplementary Fig. 1a–b). The sample metrics for both kits are presented in Supplementary Table 1.

The ILMN libraries were generated manually whilst RGT libraries were generated either on the Agilent Bravo (n = 228) or manually (Spain, n = 218). Final libraries were purified using magnetic beads

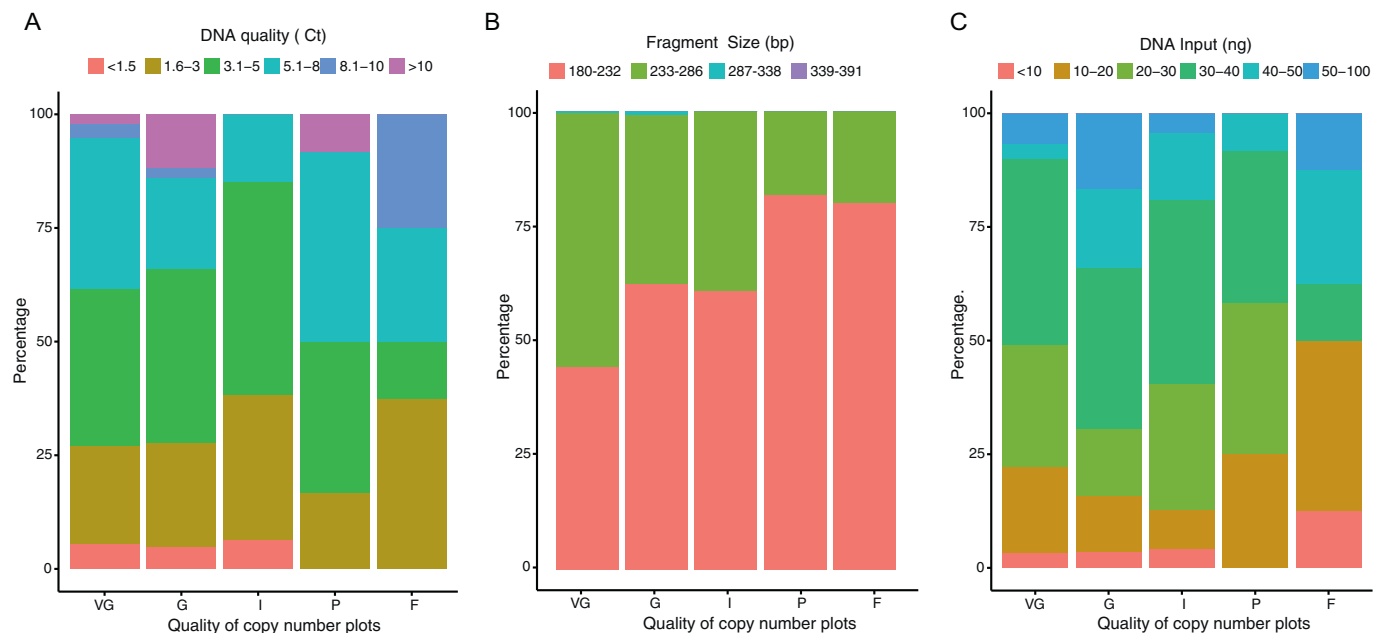
(Agencourt SPRI beads, Becton Dickinson, USA) and eluted libraries were quantified using Kapa Library Quantification kit (Roche Life Technologies, USA). Fragment size distributions were analysed utilising a 2100 Bioanalyzer with a DNA High Sensitivity kit (Agilent Technologies, USA). Two nanomoles (nM) of each library were prepared and 48 samples were pooled in one lane for sequencing on a HiSeq4000 (Illumina, USA). The pools were re-quantified and normalised to 10 nM. Single end sequencing was conducted for 50 cycles, generating on average  $4.3 \times 10^8$  reads per lane.

### 2.5. Bioinformatics

Alignment against the GRCh 37 assembly of the human genome was performed using BWA ver. 0.7.9 (Li and Durbin, 2009) or NovoAlign ver. 3.2.13 (NovoCraft, Malaysia). PCR and optical duplicates were identified using Picard tools (<https://broadinstitute.github.io/picard>) or Novosort (NovoCraft, Malaysia). Circular binary segmentation on the aligned files was performed in 100 kb windows using the QDNaseq R package available on Bioconductor, which corrects for mappability and GC content (Scheinin et al., 2014). All statistical analyses were performed in R using the functions lm() for fitting linear models and t-test() for Welch two-sample t-test.



**Fig. 3.** Features of input DNA and libraries generated from blocks less and more than five years. Dot plots represent the range (minimum-maximum) of observed values for each of the following categories and the red dot represents the median. A. The quality of input DNA inferred by  $\Delta Ct$ . B. Fragment sizes of the libraries in base pair. C. The library yield in nanomoles.



**Fig. 4.** Measured standard deviations from the QDNASEQ copy number plots and associations with the quality of sequencing libraries. A. Bar charts showing proportion of samples with different input DNA quality (based on  $\Delta Ct$ ) in each sequencing quality group. B. Bar charts showing proportion of samples from FFPE blocks of different fragment sizes in each sequencing quality group. C. Bar charts showing proportion of samples with different amount of input DNA in each sequencing quality group. VG = very good, G = good, I = intermediate, P = poor, F = fail.

**Table 2**

Features of input DNA and libraries for the different categories of copy number data. Data provided in minimum-maximum range and median values in brackets.

Quality	DNA input (ng)	DNA quality ( $\Delta$ Ct)	Library yield (nM)	Fragment size (bp)	Observed SD
VG	3.8–100 (30.3)	–0.3–12.7 (4.4)	0.28–245.1 (16.1)	180–288 (235)	0.01–0.132 (0.097)
G	4.3–107 (35.9)*	1–17.1 (4.2)	0.18–139.7 (11.3)	195–324 (229)*	0.0789–0.144 (0.1105)*
I	4–59 (33.1)	–3.2–6.3 (3.5)*	1.06–924.8 (16.0)*	191–258 (231)*	0.0739–0.24 (0.127)*
P	14.7–50 (28.7)	2–11.3 (4.9)	0–55.63 (8.7)	187–243 (219)*	0.119–0.329 (0.171)*
F	4.8–51.1 (25.2)	1.9–8.2 (4.9)	0–179.6 (8.2)	204–244 (225)	0.388–0.826 (0.528)*

$\Delta$ Ct = difference between the cycle threshold of test to the control template ACD1 provided in the kit, ng = nanogram, bp = base pairs, nM = nanomoles, SD = standard deviation.

\* Denotes the site where there is a significant difference to the index group (ie Cambridge).

### 3. Results

The majority of the FFPE samples available were core biopsies collected as part of a neoadjuvant clinical trial (GEICAM/2006-03, n = 107) yielding low amounts of DNA (range = 4–61 ng, median 30 ng). Therefore, to successfully generate libraries for CNA profiling using limited input DNA, we needed to understand how different variables could influence the quality of libraries and steps that can be taken to ensure good sequencing results (Fig. 1).

#### 3.1. Assessment of the copy number plots

We examined the copy number plots by manual inspection and categorised them based on the variance in the CN data for each case into categories: “Very Good”, “Good”, “Intermediate” and “Poor” (Fig. 2a). We also used QDNAseq (Scheinin et al., 2014) which calculates the expected (estimated from read depth) and measured (using read depth and influenced by DNA quality) standard deviation of the summarised reads, as a measure of variance. Both measures increased as the quality of library decreased and validated our categorisation of library quality (measured standard deviation shown in Fig. 2b).

#### 3.2. Assessment of different sequencing kits

We tested two kits (Illumina FFPE TruSEQ kit and Rubicon Genomics ThruPLEX DNaseq) using four FFPE samples to generate sequencing libraries and found comparable results (Supplementary Fig. 1a–b). The CNA profiles obtained using DNA processed with the ILMN kit had less variance (noise) than those processed using the RGT kit however the ILMN libraries were generated using more input DNA (200–500 ng (ILMN) versus 50 ng (RGT)) and were sequenced deeper (average coverage  $0.9 \times$  (ILMN) versus  $0.08 \times$  (RGT)). For a more comparable evaluation, we down-sampled ILMN sequencing data to a similar read depth as RGT; this showed comparable copy number profile qualities between the two library preparation technologies.

In theory, increasing the sequencing depth should improve the copy number results by reducing the variance. We examined this by increasing the sequencing depth of 23 RGT kit libraries which had less reads (from  $0.08 \times$  up to  $0.15 \times$ ) and found improvement in the data quality in 20 out of 23 libraries (examples shown in Supplementary Fig. 2a). To examine the association between sequencing depth and variance, we down-sampled the number of reads (in steps of  $1 \times 10^6$  reads) for six libraries with high read counts (up to  $24 \times 10^6$  reads). We found a significant improvement in the quality of copy number plots with increasing number of reads ( $p < 2.2e-16$ ; Supplementary Fig. 2b). It is interesting to note that the noise reduction levels off at approximately  $7 \times 10^6$  reads suggesting that increasing the read depth  $> 7 \times 10^6$  reads provides little benefit to variance reduction.

#### 3.3. Performance of sWGS for copy number profiling using the RGT kit

Due to the limited amount of DNA available for most samples, we chose the RGT kit as it required less input DNA due to fewer processing

steps, in particular purifications. Sequencing libraries were generated from as little as 3.8 ng of DNA, and out of 16 libraries prepared from  $< 10$  ng of DNA, only one failed, 13 generated good quality CNA plots, and 2 generated intermediate quality CNA plots. Information for all the libraries generated are summarised in Supplementary Table 3.

#### 3.4. Recovery of under-performing RGT libraries

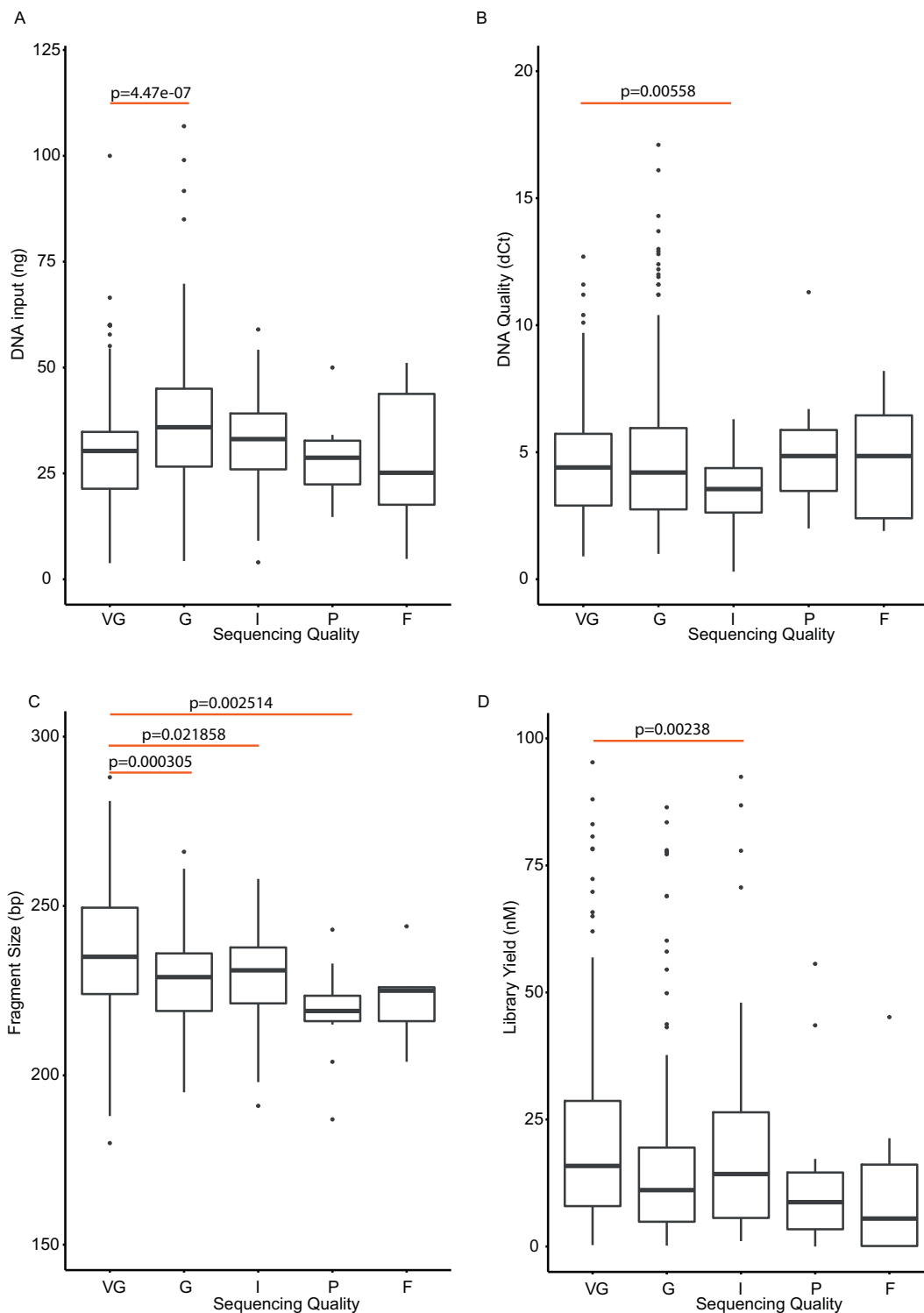
Eight (1.8%) libraries failed and 12 (2.7%) generated poor quality libraries out of 446 libraries. To recover some of these failed/poor samples, we prepared fresh libraries from samples with sufficient DNA (n = 6) or repeated the sequencing using three-fold more library material for samples with insufficient DNA to generate new libraries (n = 8). Thirteen of these new/re-sequenced libraries generated good quality data. The one repeat sample that failed was from the re-sequencing group. Consequently, only two out of 446 RGT libraries (taking into consideration the repeated libraries and re-sequencing) failed, resulting in a 99.5% success rate. Good sWGS data produced from 379/446 (84.9%) samples.

#### 3.5. Association between FFPE storage time, site, and sequencing quality

The FFPE samples were collected from three different tissue banks, spanning 20 years (Table 1). The effect of storage time on the DNA extracted was analysed (Fig. 3). DNA from older FFPE blocks ( $> 5$  years) was generally of poorer quality: higher  $\Delta$ Ct values, shorter fragment size, generating lower yield sequencing libraries. We compared the quality metrics for each banking site and found that overall FFPE samples from different sites were comparable (Table 1 and Supplementary Fig. 3).

#### 3.6. Association between input DNA characteristics and sequencing library yield

We used the Illumina FFPE QC kit, a quantitative-PCR assay to estimate the quality of FFPE-extracted DNA. This assay measures the difference in Ct (cycle threshold) value of the test FFPE-extracted DNA against the Ct value of the control DNA template provided in the kit. Increasing  $\Delta$ Ct values indicate decreasing DNA quality with Illumina quality thresholds set at:  $\Delta$ Ct  $< 1.5$  denotes high quality (HQ),  $\Delta$ Ct  $< 3.0$  denotes medium quality (MQ), and  $\Delta$ Ct  $> 3$  denotes low quality (LQ) DNA. The Illumina DNA-input recommendations for sWGS are 50 ng DNA with HQ DNA, 200 ng with MQ DNA, and exclusion of LQ DNA. Using the ILMN kit, we could generate good quality sWGS using 50 ng HQ and MQ DNA, and 200–500 ng of LQ DNA. Unsurprisingly, for eight samples with paired libraries generated from 50 ng and 200 or 500 ng of input DNA using the ILMN kit, we found that the sequencing library yields generated with more DNA was significantly higher than when using only 50 ng ( $p$ -value: 0.000265; Supplementary Fig. 3a). This is an important consideration if these libraries were destined for downstream target enrichment assays for mutation detection that require 500 ng of library material. Data from all the generated ILMN libraries (n = 45) showed a library yield that

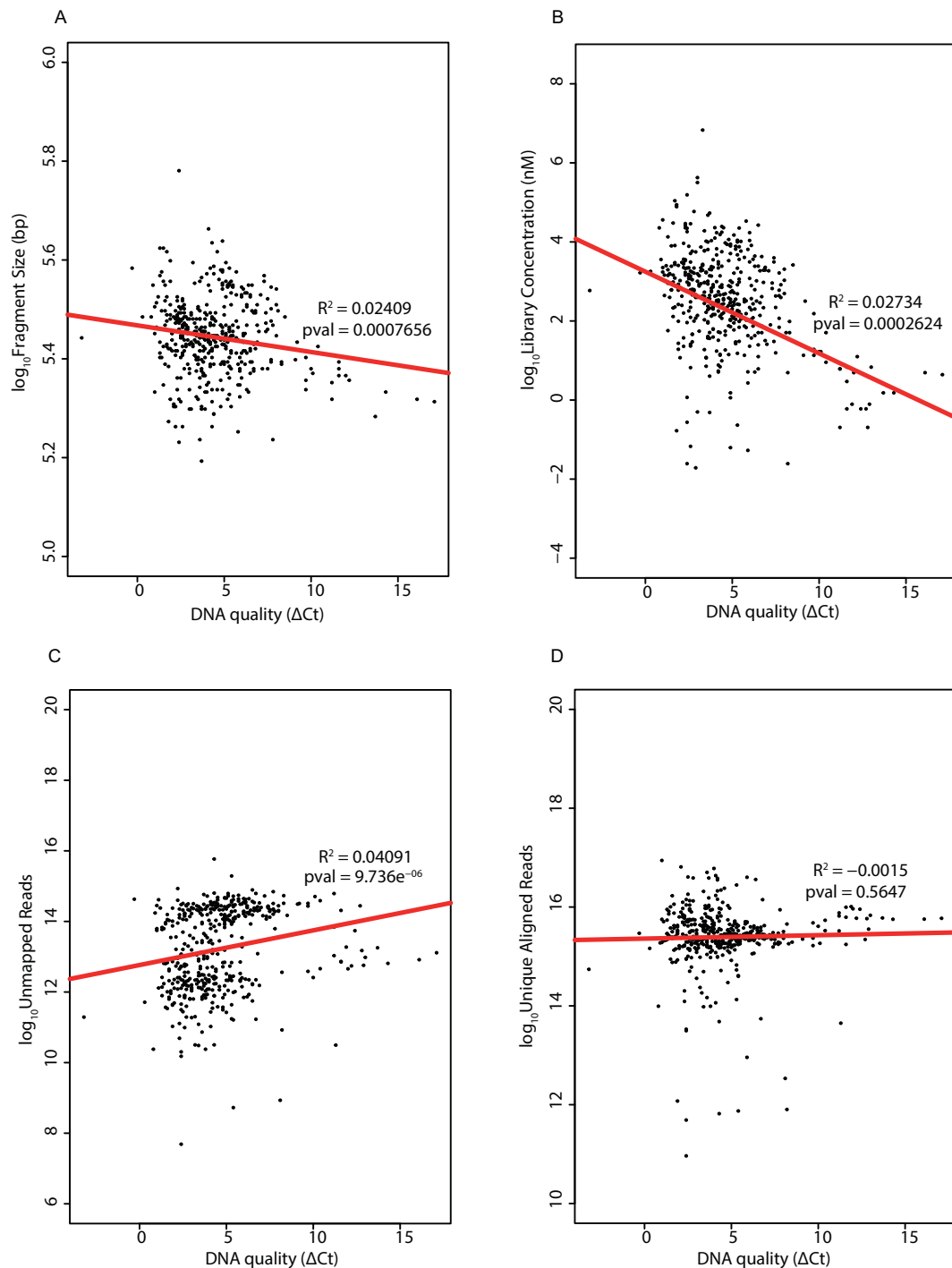


**Fig. 5.** Features of the sequencing libraries. Boxplots showing different features of input DNA and library yield relative to the different library qualities. A. Input DNA. B. Quality of input DNA inferred from  $\Delta Ct$ . C. Fragment size of libraries. D. Library yield. VG = very good, G = good, I = intermediate, P = poor, F = fail.

averaged 5.6 nM using 50 ng FFPE-extracted DNA, which was significantly less than with libraries made with more input DNA (200 ng, 23.6 nM, Welch Two Sample *t*-test,  $p = 5.24e^{-06}$ ; 500 ng: 23.0 nM, Welch Two Sample *t*-test,  $p = 0.0121$ ). There was no difference in library yield when using either 200 or 500 ng of DNA (Welch Two Sample *t*-test,  $p = 0.2401$ ). This is probably due to the quality of the

input DNA as libraries produced from 200 ng of DNA had lower  $\Delta Ct$  values (better quality) than those using 500 ng (Welch Two Sample *t*-test,  $p = 0.0179$ , Supplementary Fig. 4a–c).

Using the RGT kit, we found no correlation between amount of input DNA and sequencing library yield ( $r^2 = -0.002$ ,  $p = 0.81$ ). This is probably due to the fewer library-washing steps using the RGT kit (six



**Fig. 6.** Effect of input DNA quality.

Scatterplots showing the association between quality of input DNA with different features of the sequencing libraries.

A. Fragment size of libraries. B Library yield. C. Unmapped Reads. D. Unique aligned reads.

washing steps in the ILMN protocol versus one in RGT).

### 3.7. Association between input DNA characteristics and sequencing library quality

Next we sought to determine if sequencing quality was influenced by the nature of the input DNA by looking at the proportion of samples from all quality ( $\Delta Ct$ ) groups (Fig. 4a), fragment sizes (Fig. 4b) and different input groups (Fig. 4c) in each of the sequencing quality categories. Reassuringly, we found no biases in sampling that contributed

to the sequencing quality. In other words, each copy number plot quality group had samples from all DNA quality ( $\Delta Ct$ ) groups, fragment sizes and input quantity groups, suggesting that we could generate good quality libraries from most of our FFPE DNA regardless of these features.

Using our copy number output categorisation scoring, we examined if quality of the libraries (analysed as “all sequencing quality groups” versus “very good”) can be attributed to the different features of the input DNA and library yields (Table 2, Fig. 5). We found that the quantity of template was only significantly different in the good quality

libraries. Meanwhile, the quality of input DNA was significantly different in the intermediate libraries only when compared to the “very good” libraries. Therefore, the lesser quality sequencing libraries (I, P and F) cannot be attributed simply to either quantity or quality of the template DNA. The DNA fragment sizes, which should reflect the length of the template as the DNA was sheared under similar conditions, were found to be significantly different in all groups (progressively becoming shorter) except the failures. We found that low quality DNA was associated with shorter DNA fragments, lower library yield and higher number of unmapped reads but no association with the total number of unique reads aligned (Fig. 6a–d). The recovery of most of the poor/failed libraries described previously, was achieved by either repeating the library generation or re-sequencing to generate more reads. Consequently, we suspect the poor/failed libraries could be due to a loss of DNA during the purification steps or that the Q-PCR quantification of the libraries prior to normalisation, over-estimated the library concentration resulting in inadequate amount of library being used for sequencing. This would explain why by simply increasing the quantity of libraries for sequencing and reducing the number of samples in a single pool, ensured adequate read counts and successful sequencing.

#### 4. Discussion

In this study, we have looked at the effect that quantity and quality of DNA from FFPE tissues has on successful sWGS library preparation for CN profiling of human breast cancers. Both the quantity and quality of DNA have always been an important consideration for sample selection and in deciding which genomic application to use. For example, microarrays require 100 ng–2.5 µg of DNA depending on the resolution of the arrays whereas PCR based methods require only 10 ng of DNA. In our hands, we have not had much success in obtaining CN data with DNA extracted from FFPE DNA using microarrays, especially when the extracted DNAs are more fragmented and of lower quality (judging from absorbance ratios of 260 nm to 280 nm and multiplex PCR for quality control).

Here we have robustly shown that we can generate CN data from virtually all archived FFPE samples using sWGS. We show good CN profiling data irrespective of the quality of input DNA, as inferred by whether it can be amplified with Q-PCR ( $\Delta C_t$ ). Previous work has extensively tested the utility of FFPE DNA for mutation analysis (Astolfi et al., 2015; De Paoli-Iseppi et al., 2016; Dumur et al., 2015; Holley et al., 2012; Kerick et al., 2011) but to date no comprehensive study has shown its use for CN profiling. Since many human cancer types, including breast and ovarian cancers, are driven mostly by CNA (C-class) rather than point mutations or indels (M-class), we believe more effort should be focussed on characterizing the copy number landscapes of these cancers (Ciriello et al., 2013). We found sWGS to be very robust in generating these CN profiles, independently of the kits used, quantity and quality of DNA. sWGS is also significantly cheaper (~50%) than microarray-based methods (Supplementary Table 2).

Another advantage of generating sWGS libraries is the ability to use the same library for targeted sequence enrichment to identify mutations. There have been other methods reported for CN profiling using DNA extracted from FFPE samples but these methods do not generate sequencing libraries that can then be used for target enrichment and sequencing (Hughesman et al., 2016) or if they do, are expensive (Singh et al., 2016). In addition, sWGS will also serve as a quality control for the libraries, given its relative low cost when compared to that of generating targeted sequencing libraries. Only libraries that generate good CN profiles should be used for target enrichment and mutation detection (Kinde et al., 2012). Whilst we haven't performed target enrichment on our FFPE libraries, we expect the performance of these FFPE libraries for mutation analysis to be similar to that of published data, including known artefacts caused by formalin-based fixation effects on the DNA template (De Paoli-Iseppi et al., 2016; Holley et al., 2012; Carrick et al., 2015; Fassunke et al., 2015; Hadd et al., 2013).

#### 5. Conclusions

We have shown that sWGS is a robust and cost-effective method for obtaining good quality CN data from FFPE cancer samples, irrespective of the DNA quality and quantity used. In the case of breast cancer, CN profiles can be used to stratify breast cancers into one of the 10 Integrative Clusters (Ali et al., 2014), reiterating the importance of FFPE tumor archives. The methods described here are also of relevance to other cancers, e.g. ovarian cancers where CN profiling is essential to characterise their genomic landscapes.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.yexmp.2018.03.006>.

#### Conflicts of interest

We disclose no conflicts of interest.

#### Acknowledgements

This sequencing project was funded by Cancer Research UK program grant A16942. We thank the Genomics, Histopathology, and Biorepository Core Facilities at the Cancer Research UK Cambridge Institute; the Addenbrooke's Human Research Tissue Bank (supported by the National Institute for Health Research Cambridge Biomedical Research Centre). We thank all the patients who donated tissue and the associated pseudo-anonymized clinical data for this project.

#### References

- Ali, H.R., Rueda, O.M., Chin, S.F., et al., 2014. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol.* 15, 431.
- Astolfi, A., Urbini, M., Indio, V., et al., 2015. Whole exome sequencing (WES) on formalin-fixed, paraffin-embedded (FFPE) tumor tissue in gastrointestinal stromal tumors (GIST). *BMC Genomics* 16, 892.
- Carrick, D.M., Mehaffey, M.G., Sachs, M.C., et al., 2015. Robustness of next generation sequencing on older formalin-fixed paraffin-embedded tissue. *PLoS One* 10, e0127353.
- Ciriello, G., Miller, M.L., Aksoy, B.A., et al., 2013. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45, 1127–1133.
- Curtis, C., Shah, S.P., Chin, S.F., et al., 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352.
- Dang, J., Mendez, P., Lee, S., et al., 2016. Development of a robust DNA quality and quantity assessment qPCR assay for targeted next-generation sequencing library preparation. *BMC Cancer* 16, 1755–1765.
- De Paoli-Iseppi, R., Johansson, P.A., Menzies, A.M., et al., 2016. Comparison of whole-exome sequencing of matched fresh and formalin fixed paraffin embedded melanoma tumours: implications for clinical decision making. *Pathology* 48, 261–266.
- Dumur, C.L., Almenara, J.A., Powers, C.N., et al., 2015. Quality control material for the detection of somatic mutations in fixed clinical specimens by next-generation sequencing. *Diagn. Pathol.* 10, 169.
- Fassunke, J., Haller, F., Hebele, S., et al., 2015. Utility of different massive parallel sequencing platforms for mutation profiling in clinical samples and identification of pitfalls using FFPE tissue. *Int. J. Mol. Med.* 36, 1233–1243.
- Hadd, A.G., Houghton, J., Choudhary, A., et al., 2013. Targeted, high-depth, next-generation sequencing of cancer genes in formalin-fixed, paraffin-embedded and fine-needle aspiration tumor specimens. *J. Mol. Diagn.* 15, 234–247.
- Holley, T., Lenkiewicz, E., Evers, L., et al., 2012. Deep clonal profiling of formalin fixed paraffin embedded clinical samples. *PLoS One* 7, e50586.
- Hosein, A.N., Song, S., McCart Reed, A.E., et al., 2013. Evaluating the repair of DNA derived from formalin-fixed paraffin-embedded tissues prior to genomic profiling by SNP-CGH analysis. *Lab. Invest.* 93, 701–710.
- Hughesman, C.B., Lu, X.J., Liu, K.Y., et al., 2016. A robust protocol for using multiplexed droplet digital PCR to quantify somatic copy number alterations in clinical tissue specimens. *PLoS One* 11, e0161274.
- Iddawela, M., Rueda, O., Eremin, J., et al., 2017. Integrative analysis of copy number and gene expression in breast cancer using formalin-fixed paraffin-embedded core biopsy tissue: a feasibility study. *BMC Genomics* 18, 526.
- Janecka, A., Adamczyk, A., Gasinska, A., 2015. Comparison of eight commercially available kits for DNA extraction from formalin-fixed paraffin-embedded tissues. *Anal. Biochem.* 476, 8–10.
- Kader, T., Goode, D.L., Wong, S.Q., et al., 2016. Copy number analysis by low coverage whole genome sequencing using ultra low-input DNA from formalin-fixed paraffin embedded tumor tissue. *Genome Med.* 8, 121.
- Kallioniemi, A., Kallioniemi, O.P., Sudar, D., et al., 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (New York, N.Y.)* 258, 818–821.
- Kerick, M., Isau, M., Timmermann, B., et al., 2011. Targeted high throughput sequencing



- in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med. Genet.* 4, 68.
- Kinde, I., Papadopoulos, N., Kinzler, K.W., et al., 2012. FAST-SeqS: a simple and efficient method for the detection of aneuploidy by massively parallel sequencing. *PLoS One* 7, e41162.
- Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25, 1754–1760.
- Pinkel, D., Seagraves, R., Sudar, D., et al., 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 20, 207–211.
- Salawu, A., Ul-Hassan, A., Hammond, D., et al., 2012. High quality genomic copy number data from archival formalin-fixed paraffin-embedded leiomyosarcoma: optimisation of universal linkage system labelling. *PLoS One* 7, e50415.
- Scheinin, I., Sie, D., Bengtsson, H., et al., 2014. 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. *Genome Res.* 24, 2022–2032.
- Singh, R.R., Mehrotra, M., Chen, H., et al., 2016. Comprehensive screening of gene copy number aberrations in formalin-fixed, paraffin-embedded solid tumors using molecular inversion probe-based single-nucleotide polymorphism Array. *J. Mol. Diagnost.* 18, 676–687.
- van Beers, E.H., Joosse, S.A., Ligtenberg, M.J., et al., 2006. A multiplex PCR predictor for aCGH success of FFPE samples. *Br. J. Cancer* 94, 333–337.