On the Origin, Progression, and Evolution of Ovarian Cancer



Mohammad KaramiNejadRanjbar

St. Hugh's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2018

Declaration

I declare that this thesis, "On the origin, progression, and evolution of ovarian cancer", is entirely my own work, and except where otherwise stated, describes my own research. Any figures that had been taken from other resources are clearly indicated and the permission reference is provided when necessary. Any assistance that had been received in addition to what had been provided by my two supervisors, Prof. Ahmed Ahmed and Dr. Donatien Chedom Fotso, is declared below. All experiments were performed in the Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford, OX3 9DS, UK.

- Laser capture micro-dissected samples 1210-1218 were obtained from a sample collection previously prepared by Dr. Karin Hellner.
- I received advice for the pathology status of LCM samples from Dr. Sunanda Dhar.
- Library E011 was prepared by Ms. Salma El-Sahhar under my supervision as part of her training.
- The main body of the code for MutLearn algorithm was written by Mr. Sahand Sharifzadeh under my specific request. The initial training for MutL was performed by Mr. Sharifzadeh. The exact structure of the MLP was decided by Mr. Sharifzadeh, with features and specifications that were provided by me. I adopted the codes for MutL.2 and MutLX from the MutL algorithm.
- Mr. Bahram Kheradmand, assisted me in structuring the information and ordering the references for section 1.2.3 of chapter 1.

Mohammad KaramiNejadRanjbar

St. Hugh's College, Oxford

I dedicate this thesis to my wonderful sisters, Maryam and Mina.

In their presence I always feel joyful and happy.

On the Origin, Progression, and Evolution of Ovarian Cancer

Mohammad KaramiNejadRanjbar, St. Hugh's college

University of Oxford, Trinity 2018

Abstract

High-grade serous ovarian cancer (HGSOC) is amongst the deadliest of cancers. Despite decades of research on this disease, the exact mechanisms through which this carcinoma initiates, spreads, and relapses remain unclear. I believe the main reason behind this discrepancy between importance and available knowledge roots in the fact that currently available genomics methods are not sensitive enough for accurate study of limited amount of clinical materials.

To address this issue, I optimised a novel library preparation protocol called DigiPico to generate reliable next generation sequencing data from picogram quantities of DNA that can be obtained from 10-20 human cells. Moreover, I implemented DigiTitan and MutLX analysis algorithms to identify clone-specific copy number alterations and small nucleotide variations from DigiPico data. Validating MutLX algorithm using experimental and simulated data I demonstrated that DigiPico/MutLX is a uniquely sensitive, accurate, and reliable method for analysis of micro-heterogeneity in solid tumours.

Next, using DigiPico, I conducted the first accurate case study of genetic micro-heterogeneity on a collection of pre-chemotherapy tumour islets and post-chemotherapy microscopic residual disease (MRD) samples from one HGSOC patient. Analysing the results using DigiTitan and MutLX, firstly, I introduced tumour heterogeneity and genetic diversity as a hallmark of HGSOC. Secondly, based on new observations from this study, I challenged a pre-existing notion that mutations in TP53 gene are responsible for the initiation of HGSOC. I proposed that mutations in the regulatory regions that adversely affect the DNA repair machinery which can lead to subsequent large-scale genomic alterations precede the loss of TP53 function in the tumorigenesis pathway of HGSOC. Thirdly, comparing the somatic mutation load of pre-chemotherapy, MRD and recurrent tumours, I demonstrated that MRD harbours the least number of somatic mutations. Thereby, I hypothesized that MRD is likely to be composed of stem cell-like tumour cells that have survived the chemotherapy therefore suggesting that MRD is likely be the recurrence precursor. Lastly, I propose experimental approaches that are required to effectively validate or nullify these hypotheses.

Acknowledgments

First and foremost, I would like to thank my supervisor Prof. Ahmed Ahmed for accepting me in his lab and offering me his mentorship in life and research. I would also like to thank my second supervisor Dr. Donatien Chedom Fotso for advising me and supporting me on the bioinformatics aspects of my project.

I am thankful of St. Hugh's college, Nuffield Department of Women's & Reproductive Health, and the Weatherall Institute of Molecular Medicine for giving me the opportunity to conduct my research, in particular I am extremely grateful of the staff and scientists in the sorting facility and the computational unit in the WIMM who never even blinked in helping me with my complex requests.

With great pleasure, I would like to thank Helen Clarke Fund and more specifically the Clarke family who not only financially supported my DPhil but have always been extremely encouraging and supportive of my research.

These acknowledgments would be incomplete without specifically thanking my great friend, ex-lab mate, and coffee buddy, Dr. Toni A. Baeumler. Nothing but quoting this sentence about our friendship from his thesis can better complete this expression of gratitude: *Without our infamous dinner in Goettingen, both of us would have never applied to Oxford, and the rest is history*[1].

Lastly, I am extremely thankful of all my friends and my family, this work would never have been successful without your constant support.

Thank you all.
Mo,
Winter 2018

[1] Baeumler, T. A. (2018). Redirecting the cellular information flow with programmable dCas9-based chimeric receptors (DPhil), page VII. University of Oxford, UK.

Abbreviations

ANN	Artificial Neural Network
BAF	B Allele Fraction
bp	Base pair
C-FC	Control-FREE C
CGH	Comparative Genomic Hybridization
chr	Chromosome
CNA	Copy Number Alteration
CoREF	Controlled Enzymatic Fragmentation
DAC	Dual-Allelic Compartment
DigiPico	Digital Analysis of Picogram Quantities of DNA
DNA	Deoxyribonucleic Acid
DSB	Double Strand Break
DSS	Dual Size Selection
EndoIV	Endonuclease IV
FFPE	Formalin-Fixed Paraffin-Embedded
fg	Femto gram
FP	False Positive
FPR	False Positive Rate
HCC	Hexamine Cobalt (iii) Chloride
HDR	Homology Directed Repair
HGSOC	High-grade Serous Ovarian Cancer
IGV	Integrated Genome Viewer
Kb	Kilo base pair
LCM	Laser Capture Microdissection
LFR	Long Fragment Read Sequencing
LOH	Loss of Heterozygosity
LRR	Log Read Ratio
MDA	Multiple Displacement Amplification
MLP	Multilayer Perceptron
MRCA	Most Recent Common Ancestor

MRD	Microscopic Residual Disease
MutL	MutLearn
ng	Nanogram
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End Joining
OCT	Optimal Cutting Temperature Compound
OSE	Ovarian surface epithelium
PAGE	Polyacrylamide Gel Electrophoresis
PEG	Polyethylene Glycol
pg	Picogram
PP	Posterior Probability
QF	Quality Filter
qPCR	Quantitative Polymerase Chain Reaction
ROC	Receiver Operating Characteristic
rSAP	Shrimp Alkaline Phosphatase
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SPRI	Solid Phase Reversible Immobilization
ssDNA	Single Stranded DNA
STIC	Serous Tubal Intraepithelial Carcinoma
TP	True Positive
TPR	True Positive Rate
UDG	Uracil DNA Glycosylase
UTD	Unique to DigiPico
VAF	Variant Allele Fraction
VAWF	Variant Allele Well Fraction
WES	Whole Exome Sequencing
WGA	Whole Genome Amplification
WGS	Whole Genome Sequencing

Table of contents

Chapter 1 : Introduction **14**
1.1 Ovarian cancer, a deadly disease 14
1.2 Genetics of HGSOC 16
1.2.1 Importance of TP53 in ovarian cancer 17
1.2.2 Significance of BRCA1/BRCA2 mutations in ovarian cancer 18
1.2.3 Cell of origin in HGSOC 19
1.3 Chemotherapy resistance and tumour recurrence 20
1.3.1 Modes of chemotherapy resistance 22
1.3.2 Chemotherapy resistance in HGSOC 23
1.4 Current state of HGSOC studies 26
1.4.1 Limitations and obstacles 27
1.4.2 State of the art genomics methods for the study of HGSOC 28
1.4.3 Long Fragment Read sequencing 30
1.5 Thesis aims and scopes 31
Chapter 2 : DigiPico Sequencing **33**
2.1 Optimization of a streamlined protocol 35
2.1.1 Controlled Enzymatic Fragmentation 36
2.1.2 Adapter ligation optimization 38
2.1.3 WGA optimization and automated library preparation 45
2.2 Reaction miniaturization 55
2.2.1 WGA miniaturization 55
2.2.2 Reaction volume adjustment for Mosquito 57
2.2.3 DigiPico library preparation and sequencing: Library 1110 59
2.3 Direct application of DigiPico on LCM samples 63
2.3.1 DigiPico on a frozen tumour section: Library 1111 63
2.3.2 Quantification of crude DNA from LCM samples 67
2.4 Increasing the multiplexabilty of DigiPico libraries 69
2.5 Further advancements with DigiPico sequencing 70
2.5.1 Expanding the DigiPico datasets to three patients 70
2.5.2 Effect of sequencing depth on DigiPico data 70
2.5.3 DigiPicoExome 74

Chapter 3 : DigiPico Analysis Pipeline 76
3.1 CNA prediction from DigiPico data 78
3.1.1 QDNASeq 80
3.1.2 Control-FREEC 80
3.1.3 TitanCNA 90
3.2 De novo variant calling from DigiPico data 96
3.2.1 Evaluation of variant detection sensitivity of different algorithms 98
3.2.2 Evaluation of false positive rate of different algorithms 100
3.2.3 Filter optimisation for Platypus variant caller 101
3.2.4 MutSwipe and visual processing of UTD variant calls 107
3.2.5 Development of MutLearn algorithm for filtering low-quality DigiPico variants 107
3.2.6 Experimental validation of high-quality UTD variant calls 111
3.2.7 Dual-Allelic Compartment filter for further filtering of Platypus UTD calls 113
3.2.8 Experimental validation of DAC filtered UTD calls 114
3.2.9 Development of MutLearn2 algorithm for filtering high-quality FP UTDs 119
3.2.10 MutLearnX: Development of an unbiased, effective and customisable variant filtering algorithm for DigiPico data 123
3.2.11 Application of MutLX algorithm to DigiPicoExome data 127
3.2.12 Accurate prediction of the status of clone-specific SNVs using MutLX 130
Chapter 4 : Inferring the Evolutionary History of a HGSOC using DigiPico: A case study 134
4.1 Study of copy number alterations 135
4.1.1 Micro-heterogeneity of CNAs in the pre-chemotherapy HGSOC 137
4.1.2 CNA diversity in the recurrence HGSOC tumour 144
4.2 Study of single nucleotide variations 148
4.2.1 Inference of evolutionary relationship of tumour islets based on SNVs 148
4.2.2 17p LOH might precede the TP53 mutation in patient 11152 151
4.2.3 Presence of SNV micro-heterogeneity in the recurrence but not the pre-chemotherapy HGSOC tumour 153
4.2.4 Investigation of active mutational processes in recurrence tumours 155
4.3 Study of MRD could reveal early mutations in HGSOC 159
Chapter 5 : Discussion and Conclusions 165
5.1 Discussion 165
5.1.1 Optimisation of DigiPico sequencing and analysis pipeline 165
5.1.2 To better understand HGSOC using DigiPico 166

5.2 Conclusions 174
5.2.1 On the origin 174
5.2.2 On the progression 174
5.2.3 On the evolution 174
5.3 Future perspectives 175
5.3.1 DigiPico sequencing 175
5.3.2 DigiPico analysis 175
5.3.3 Study of HGSOC 176
Chapter 6 : Methods 178
6.1 Ethical approval, patient consent, and sample collection 178
6.2 Tissue sectioning and LCM 178
6.2.1 Sectioning 178
6.2.2 H&E staining 178
6.2.3 Cresyl Violet staining and LCM 179
6.3 Library preparation 179
6.3.1 Standard WGS library 179
6.3.2 WES library 180
6.3.3 WGA WGS library 180
6.3.4 DigiPico library preparation 180
6.4 Illumina sequencing and data analysis 183
6.4.1 Illumina sequencing 183
6.4.2 Read mapping 184
6.4.3 Copy number analysis 184
6.4.5 SNV calling and analysis 185
6.4.6 MutSwipe application 185
6.4.7 Machine learning 186
Chapter 7 : Bibliography 187
Appendix I 194
Appendix II 195

List of Figures

Figure 1.1. Histological subtypes of ovarian cancer. 15
Figure 1.2. Model of ovarian cancer origin from fallopian tube epithelium. 21
Figure 1.3. Library preparation procedure for Long Fragment Read (LFR) sequencing. 32
Figure 2.1. Streamlined CoREF process for DigiPico. 37
Figure 2.2. Anatomy of Illumina adapters used in DigiPico protocol. 39
Figure 2.3. Streamlined A-tailing and adapter ligation for DigiPico. 42
Figure 2.4. SPRI bead purification of ligation product to remove self-ligated adapters. 44
Figure 2.5. Optimsing WGA using REPLI-mini in 5 µl 47
Figure 2.6. Inconsistent adapter ligation in small scale DigiPico library preparation. 48
Figure 2.7. Addition of PEG 8000 can significantly enhance the adapter ligation efficiency 50
Figure 2.8. Heatmap presentation of a failed DigiPico library preparation. 52
Figure 2.9. Comparison of WGA in 5 µl using four different MDA kits. 54
Figure 2.10. Reaction miniaturization and CoREF optimisation for REPLI-g Single Cell kit. 56
Figure 2.11. Miniaturized DigiPico workflow. 58
Figure 2.12. Library 1110 – First successful DigiPico library. 60
Figure 2.13. Quality measures and sequencing statistics of library 1110. 62
Figure 2.14. Quality measures and sequencing statistics of library 1111. 66
Figure 2.15. Evaluation of quantification assays for detection of pg quantities of DNA from crude lysates. 68
Figure 2.16. Comparison of basic sequencing statistics of various DigiPico libraries. 72
Figure 2.17. Effect of sequencing depth on the per well breadth and depth of coverage in DigiPico. 73
Figure 2.18. Target enrichment efficiency in DigiPico Exome. 75
Figure 3.1. Results of QDNASeq analysis of various DigiPico runs. 81
Figure 3.2. Copy number profile estimate of standard WGS run 1101 using Control-FREEC. 83
Figure 3.3. BAF plots of standard WGS run 1101 from Control-FREEC. 84
Figure 3.4. Copy number profile estimate of standard WGS run E101 using Control-FREEC. 85
Figure 3.5. BAF plots of standard WGS run E101 from Control-FREEC. 86
Figure 3.6. Copy number Profile estimation from merged DigiPico runs of patient 11152 using Control-FREEC. 88
Figure 3.7. Copy number Profile estimation from WGA DNA and merged DigiPico runs of patient 11513 using Control-FREEC. 89

Figure 3.8. Copy number profile estimation from WGS and merged DigiPico data of patient 11152 using TitanCNA. 91
Figure 3.9. Copy number profile estimation from WGS, standard WGA WGS, and merged DigiPico data of patient 11513 using TitanCNA. 92
Figure 3.10. Effect of supplying digitised and phased DigiPico data on the accuracy of copy number estimation using Titan algorithm..... 95
Figure 3.11. De novo variant calling sensitivity from DigiPico data. 99
Figure 3.12. Analysis of mutation type distribution of 'Unique to DigiPico' variants from SCcaller and Platypus. 102
Figure 3.13. Comparing the distribution of values from various Platypus quality measure parameters between known SNVs and UTD SNVs. 103
Figure 3.14. Threshold optimisation for filtering Platypus variant calls from DigiPico data. 105
Figure 3.15. Visualisation of Platypus UTD variants from run E111..... 106
Figure 3.16. Using visual information and machine learning to filter FP calls from DigiPico data. 109
Figure 3.17. IGV visualisation of the variant position chr1:207222565 in sequencing data from patient 11152. 117
Figure 3.18. Effect of various filtering criteria on the mutation type distribution of Platypus UTD variants. 118
Figure 3.19. Development and application of MutLearn2 algorithm for improved variant filtering of DigiPico data. 121
Figure 3.20. Validation of MutLearnX Artificial Neural-Network pipeline using UTD* variants. ... 125
Figure 3.21. Comparison of TPR of UTD* and unaltered known variants..... 126
Figure 3.22. Application of MutLX to DigiPico data. 128
Figure 3.23. Application of MutLX algorithm to DigiPicoExome data. 129
Figure 3.24. Mutation type distribution of MutLX filtered UTD variants..... 131
Figure 3.25. Effect of MutLX threshold values on the frequency of different mutation types. 132
Figure 3.26. Comparison of C>A mutation level plots in various DigiPico libraries. 133
Figure 4.1. Laser-Capture Microdissection of tumour islets from a pre-chemotherapy mass of tumour..... 136
Figure 4.2. Copy number analysis of bulk sequencing data from pre-chemotherapy and recurrence tumour samples. 138
Figure 4.3. Analysis of CNAs across various pre-chemotherapy tumour islets using DigiTitan. 143
Figure 4.4. Genome-wide comparison of copy number alterations in recurrence tumours. 147

Figure 4.5. Inference of evolutionary relationship of tumour islets with recurrence based on SNVs. 150
Figure 4.6. Investigation of the TP53:C275F status in 1211, 1212, and 1511 samples. 152
Figure 4.7. MutLX analysis of clone-specific somatic mutations of pre-chemotherapy and recurrence DigiPico data. 154
Figure 4.8. Analysis of the mutational process signatures of the pre-chemotherapy and recurrence tumours. 157
Figure 4.9. Rainfall plots of mutation distances reveals the presence of clone specific kataegis events. 158
Figure 4.10. Study of somatic SNVs of MRD. 163
Figure 4.11. VAF of potential early mutations in all of the studied tumour samples. 164
Figure 5.1. Proposed tumorigenesis pathway of HGSOC. 170

List of Tables

Table 2.1. Relative activity of CoREF and adapter ligation enzymes in various commercially available buffers. 38
Table 3.1. Number of 'Unique To DigiPico' variants identified in various DigiPico runs using different variant calling methods. 101
Table 3.2. Effect of variant filtering on the average detection rate of known variants (top rows, highlighted) and number of UTDs (bottom rows). 110
Table 3.3. Results of targeted sequencing validation of high-quality UTD variants from run E111. 112
Table 3.4. Effect of the DAC filter on the variant detection sensitivity (top row, highlighted) and the number of Platypus UTD calls (bottom row). 114
Table 3.5. Results of targeted sequencing validation of high-quality, DAC filtered UTD variants from run 1111. 116
Table 3.6. Effect of application of MutLearn2 variant filtering model on the number of UTDs (bottom rows) and average detection rate of known variants from DigiPico data (top rows, highlighted). 120
Table 3.7. Effect of combining MutL.2 with various other variant filtering criteria on the average detection rate of known variants (top rows, highlighted) and number of UTDs (bottom rows). 122
Table 4.1. Variant re-calling rate of all of the available DigiPicoExome libraries from patient 11152. 160

Chapter 1 : Introduction

1.1 Ovarian cancer, a deadly disease

Ovarian cancer claims the lives of over 4,000 women each year in the UK (CRUK, 2015). Accounting for over 5% of all deaths caused by cancer, ovarian cancer is the deadliest gynaecological malignancy (Torre *et al.*, 2018). The main reason for the high mortality rate of ovarian cancer is often attributed to the lack of a reliable screening method which results in late presentation, leading to a 5-year survival rate of only about 30% (Siegel *et al.*, 2014). Over 90% of the ovarian cancer cases have been found to be from an epithelial origin. Among the five subtypes of epithelial ovarian cancer (Figure 1.1), high-grade serous ovarian cancer (HGSOC) is the most commonly observed malignancy which accounts for over 70% of the diagnosed cases (Matulonis *et al.*, 2016). In general, four main diagnostic stages had been defined for HGSOC by International Federation of Gynaecology and Obstetrics (FIGO). Stage I is identified by limited tumour presentation only in the ovaries. Stage II requires the extension of the disease into the pelvic or peritoneal area. At stage III the disease has metastasized to retroperitoneal lymph nodes and/or to the peritoneum outside the pelvis and at stage IV the disease is characterised by distant metastasis into other organs ('American Cancer Society', 2018). With a survival rate of about 90% when diagnosed at stage I but 15% when diagnosed at stage IV, it appears that the stage of diagnosis has a significant effect on the mortality rate of HGSOC ('American Cancer Society', 2018). Despite this, because of the lack of a reliable early diagnostic method, currently, over 60% of the HGSOC cases are diagnosed at stages III/IV which results in the high mortality rate associated with this malignancy (Bowtell *et al.*, 2015). Consequently, neoadjuvant chemotherapy with paclitaxel and carboplatin followed by interval surgical debulking of the tumour mass has become the standard method of treatment for HGSOC.

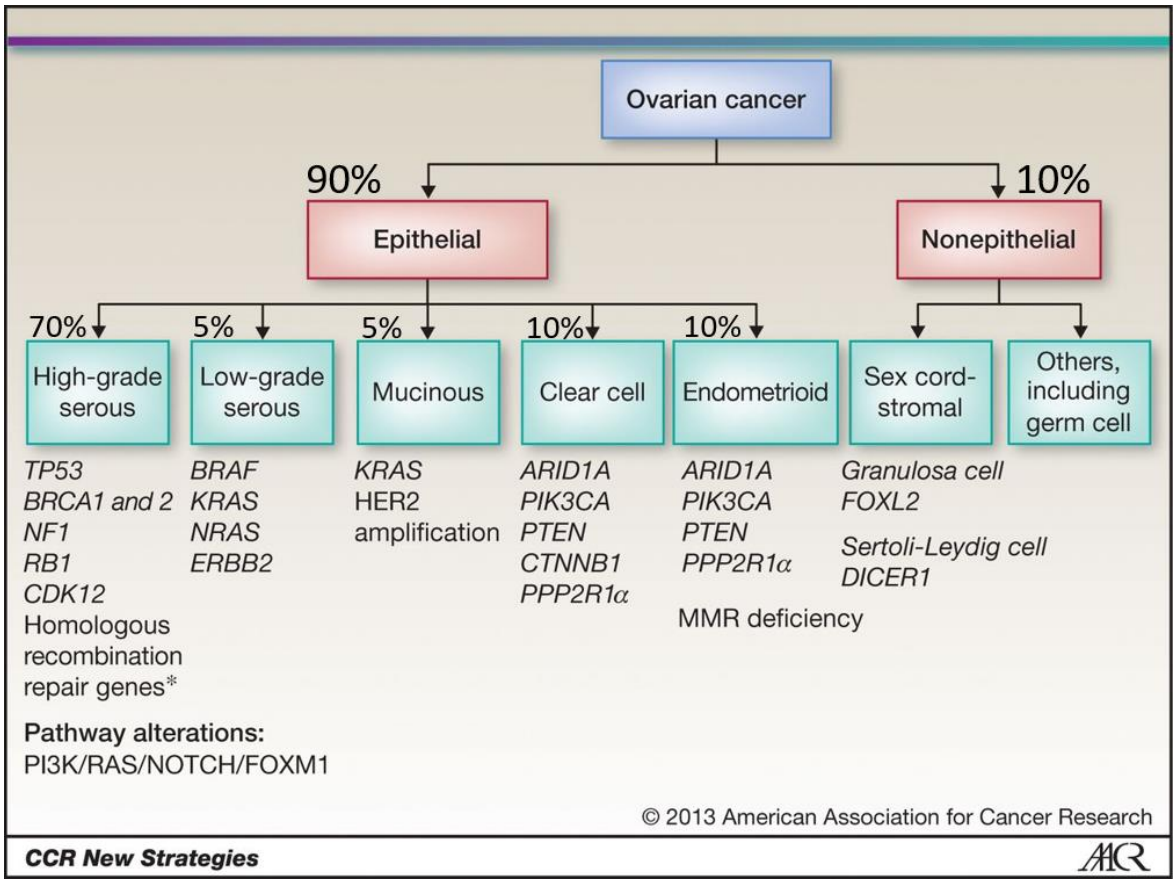


Figure 1.1. Histological subtypes of ovarian cancer. Over 90% of the incidences of ovarian cancer are of the epithelial type. Five subtypes of epithelial ovarian cancer, High-grade serous, Low-grade serous, Mucinous, Clear cell, and Endometroid, differ from each other based on pathological appearance, patterns of transcription and the underlying genetic mutations. Numbers on each box represent the relative incidence rate of each subtype. Gene names under each box show the mostly mutated/altered genes in the respective subtype. High-grade serous ovarian cancer (HGSOC) accounts for over 70% of the epithelial ovarian cancer cases and is characterised by deleterious TP53 mutations and extensive copy number alterations across the genome. Modified and reprinted by permission from American Association for Cancer Research, Clinical Cancer Research (Banerjee and Kaye, 2013). Licence #4441641479102.

However, in spite of an apparent initial response to chemotherapy with complete disappearance of all of the visible masses, over 70% of the cases suffer from tumour recurrence within two years after the completion of treatment (Jayson *et al.*, 2014). Subsequent treatment of patients with recurrent tumours often results in progressively shorter disease-free intervals with tumours showing a stronger resistance phenotype after each round of treatment, resulting in substantial reduction in the life quality of the patient and eventually death (Jayson *et al.*, 2014). These statistics indicate that there is an imminent need to address both the late presentation and high relapse rate for efficient management of HGSOC. These issues however, could only be addressed with an accurate and in depth understanding of the molecular mechanisms behind the origin, progression, and evolution of ovarian cancer.

1.2 Genetics of HGSOC

Cancer is often referred to as the disease of the genome and 'mutations and genomic instability' had been introduced as one of the main hallmarks of cancer (Hanahan and Weinberg, 2011). Mutations can disrupt or inactivate tumour suppressors genes or can activate or upregulate oncogenes and pro-survival factors. Such mutations that result in the initiation and progression of a tumour are termed *driver mutations*. As a complementary term, mutations that just happen to occur at the same time as the driver mutations and lack any functional significance in the evolution of the tumour are often called *passenger mutations*. Although all forms of genetic alterations, from small single-nucleotide variations (SNVs) to loss of an entire chromosome arm, can occur in cancers, different types of cancer are often enriched with specific types of mutations (Figure 1.1). Some of these diversities had been stratified by the introduction of signatures of mutational processes in various cancer types (Alexandrov *et al.*, 2013, 2018; Macintyre *et al.*, 2018). In this regard, HGSOCs are characterised by dominant presence of mutational signatures 1, 3, and 5 as well as a high number of large-scale genomic aberrations across the genome in various forms, from copy-number alterations (CNA) to deletions and duplications as well as various chromosomal translocations

(Alexandrov *et al.*, 2013; Macintyre *et al.*, 2018). Although SNVs are less frequently observed in HGSOC, several recurring small-scale mutations targeting specific genes have been identified in HGSOC. These genes include: TP53, BRCA1, BRCA2, RB1, NF1, FAT3, CSMD3, GABRA6, and CDK12 (TCGA, 2011). Most notably among these genes are the TP53 and BRCA1/BRCA2 genes which are involved in important cancer related cellular processes such as DNA repair and apoptosis.

1.2.1 Importance of TP53 in ovarian cancer

TP53 is the mostly mutated tumour suppressor gene in human cancers and its functional deficiency had been observed in nearly all types of cancer (Nigro *et al.*, 1989). Reviewing the diverse functions of TP53 in a mammalian cells I would like to refer to it as the *Master Regulator of Everything* that is related to a cell's division, survival, and differentiation. Regulation of apoptosis, DNA damage response, cell cycle arrest, autophagy, and senescence are amongst the most important cancer related activities of TP53 (Farnebo, Bykov and Wiman, 2010). Therefore, the exact functional importance of TP53 mutations can be at various levels of tumour evolution from initiation and progression to metastasis and aggressiveness depending on the cancer type (Nigro *et al.*, 1989; Rivlin *et al.*, 2011). In HGSOC it is often assumed that the deactivation of TP53 is amongst the earliest events that initiate the malignancy (Lee *et al.*, 2007). TP53 mutations or functional defects had been observed in nearly 100% of the HGSOC cases, over 96% of which had been found to be in the form of SNVs or Indels (Ahmed *et al.*, 2010). Among these, about 60% had been found to be in the form of missense mutations, with the remaining taking the form of Indels and nonsense (stop-gain) mutations that often result in the production of non-functional truncated proteins (Ahmed *et al.*, 2010; Rivlin *et al.*, 2011). This however, is unlike all of the other known tumour suppressor genes in which majority of the cancer related mutations are in the form of inactivating mutations. One possible explanation for this discrepancy relies on the way TP53 protein functions at the molecular level. Human TP53 protein is composed of 3 main domains: An N-terminal transactivation domain, a middle DNA-binding domain, and a C-terminal tetramerization domain.

Since the functional form of the TP53 transcription factor is in a tetrameric state, a missense mutation in the DNA-binding domain of one copy of the TP53 gene in the genome can result in the production of defective proteins that are able to participate in the formation of a tetramer with intact proteins. Such interactions can potentially impair the overall functionality of the entire complex. This phenomenon has been referred to as Dominant-Negative effect (DNE) and had been confirmed for some of the mostly recurring TP53 missense mutations such as R273C (Michalovitz, Halevy and Oren, 1991; Kern *et al.*, 1992; Willis *et al.*, 2004). Nevertheless, while with the DNE a single heterozygous missense mutation should be sufficient to transform a cell, in HGSOC, heterozygous TP53 mutations had rarely been observed. In nearly 87% of the HGSOC cases, a loss of heterozygosity event at chromosome 17p, where the TP53 locus resides, results in the complete removal of the intact copy of the TP53, leaving only the mutated version behind (TCGA, 2011). This indicates that DNE cannot solely explain the observed high missense mutation rate in TP53 gene and hence blurring the exact role of TP53 mutations in the initiation of HGSOC.

1.2.2 Significance of BRCA1/BRCA2 mutations in ovarian cancer

BRCA1 and BRCA2 proteins are two of the main components of the homologous recombination machinery in mammalian cells. Their main responsibility is to help maintain the genomic integrity by repairing double-strand DNA breaks (DSB) through 'error free' homologous recombination. DSB can occur in the genome for several natural and unnatural reasons such as exposure to radiation. These lesions are often dealt with via two main repair mechanisms, error prone non-homologous end joining (NHEJ) or more accurate homology directed repair (HDR). The latter uses the information from the homologous strand of the DNA to accurately repair the break. During the HDR, BRCA1 is initially recruited to the site of the damage and enables the start of a process referred to as strand recession which leads to the production of ssDNA at each end of the break. Next, through an interaction with a bridging protein called PALB2, it recruits BRCA2 which in turn helps RAD51 recombinase to load onto the ssDNA filaments and initiate the homologous

recombination process (Chen *et al.*, 1999; Prakash *et al.*, 2015). In situations where the genomic stress is too high BRCA1/2 had also been shown to be able to induce apoptosis (Yan *et al.*, 2002), hence, directly functioning as a tumour suppressor gene. Therefore, functional defects in these genes can result in genomic instability which can fuel the initiation, progression and overall evolution of the tumour. Women with deleterious germline mutations in the BRCA1/2 genes show a 40% lifetime risk for ovarian cancer which is 25-fold higher than women with two intact copies for these genes. Moreover, mutations in BRCA1/2 genes had been observed in over 20% of HGSOC cases (TCGA, 2011). Since mutations in BRCA1/2 genes can impair DNA damage response pathways in the cell, they can contribute to the high rate of CNAs observed in HGSOC, to the extent that nearly all of the ovarian cancers with defective BRCA1/2 genes are in the form of high-grade serous.

1.2.3 Cell of origin in HGSOC

The exact cell of origin in HGSOC had been a point of debate over the previous decades (Kurman and Shih, 2010). Although currently it is commonly accepted that majority of HGSOC cases originate from the fallopian tube (Labidi-Galy *et al.*, 2017) there still remain some evidence that at least in some cases ovarian surface epithelium (OSE) might be the source of the disease (Ng and Barker, 2015). Originally, OSE cells were proposed to be the cell of origin in HGSOC (Fathalla, 1971). It was hypothesised that the transformation of these cells is mainly a result of multiple rounds of ovulations without conception, which is unique to humans. After it was found that HGSOC is linked to tumour suppressor genes BRCA1 and BRCA2 (Miki *et al.*, 1994) bilateral salpingo-oophorectomy (complete removal of both fallopian tubes and ovaries) was advised for patients with inherited BRCA1 and BRCA2 mutations as an effective preventive measure (Kauff *et al.*, 2008). Detailed analysis of the fallopian tubes that were obtained from these preventive surgeries, resulted in the identification of secretory epithelial cells in the dysplastic lesions found in the fimbrial end of the fallopian tubes. The high proliferative index of these cells suggested that fallopian tube secretory cells might be the key cell type involved in the initiation of the HGSOC (Piek *et al.*, 2001).

Follow-up histological studies on these fallopian tubes helped further definition of the tumour development steps (Figure 1.2). It was proposed that mutations in TP53 gene result in the formation of *P53 signature* areas in the fallopian tube epithelium (Lee *et al.*, 2007). Further genetic changes in the genome of *P53 signature* cells can result in subsequent loss of polarity and formation of multi-layered serous epithelium, referred to as Serous Tubal Intraepithelial Carcinoma (STIC) (Carcangiu *et al.*, 2004). However, despite the observations that *P53 signature*, STICs, and the primary tumour can have identical TP53 mutations (Lee *et al.*, 2007; Labidi-Galy *et al.*, 2017), the exact transformation mechanism of STICs to HGSOC remains a rather controversial topic, with one study suggesting that at least some STICs might originate from metastasis of the tumour back into the fallopian tubes (Eckert *et al.*, 2016). Nevertheless, although it seems that not all HGSOC may have emerged from STICs, there is evidence to support that fallopian tube secretory cells are the main source of HGSOC (Perets and Drapkin, 2016).

1.3 Chemotherapy resistance and tumour recurrence

Neoadjuvant chemotherapy followed by interval tumour debulking is currently the main method of treatment for majority of HGSOC patients. The main chemotherapeutic agents that are used for treatment of HGSOC patients are Paclitaxel and Carboplatin. Moreover, Olaparib, a potent PARP inhibitor, may also be administered to some patients, especially in cases with BRCA1/2 mutant tumours. Paclitaxel, and in general Taxanes, exert their chemotherapeutic function by tightly binding to β-tubulin molecules only when they are in a microtubule structure. Thereby, stabilizing the microtubule structures to an extent that leads to mitotic defect, cell cycle arrest and apoptosis (Dumontet and Sikic, 1999). Carboplatin, and in general Platinum based chemotherapy agents, act by forming inter- and intra-strand DNA crosslinks. Such lesions activate two main DNA repair mechanisms in the cell, Nucleotide Excision Repair (NER) and Mismatch Repair (MMR). The exact mechanism of action of these platinum-DNA adducts that leads to apoptosis is not entirely known.

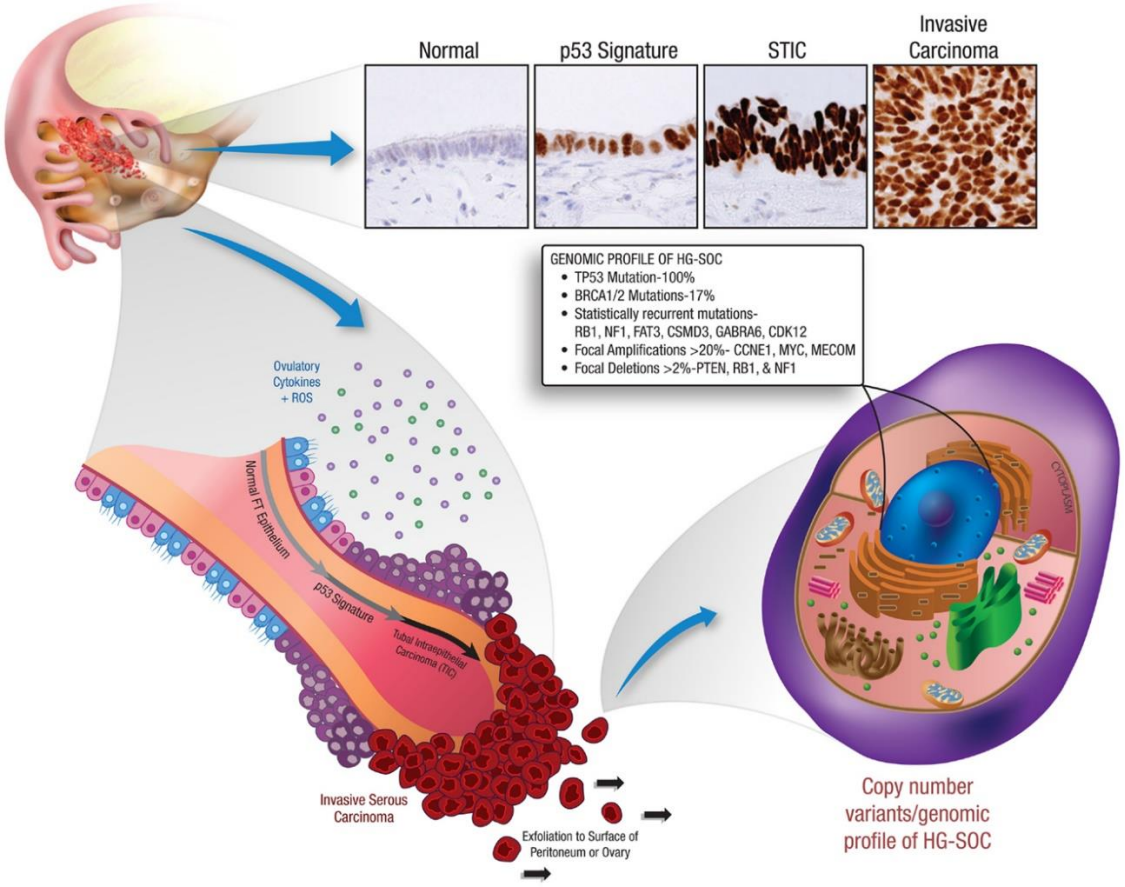


Figure 1.2. Model of ovarian cancer origin from fallopian tube epithelium. Based on this model, under the effect of ovulatory cytokines and harsh oxidative stress, deleterious mutations occur in TP53 gene which results in the overexpression of this protein and formation of *P53 signature*. Further mutations and genetic changes will lead to loss of polarity and formation of multi-layered STICs which will eventually evolve into invasive carcinoma through accumulation of new driver mutations. Most important frequently observed genetic changes in HGSOC are represented in the white box. Reprinted under the terms of the Creative Commons Attribution License (CC BY 3.0) from Frontiers in Oncology (Jones and Drapkin, 2013).

However, it had been proposed that in HGSOC, the DSB products of NER pathway stay unrepaired in the absence of an HDR pathway in HGSOC cells which leads to the activation of apoptotic pathways and selective removal of cancer cells (Sundar *et al.*, 2016). PARP inhibitors, on the other hand, perform their function through the synthetic lethality interaction between PARP1 and BRCA1/2. PARP1 is one of the main components of DNA repair mechanisms for ssDNA breaks. Such lesions may occur spontaneously or as a result of Base Excision Repair (BER) pathway. PARP1 identifies the ssDNA breaks (nicks) and recruits the remaining enzymes that are required for the repair of the damage. In the absence of PARP1 activity, these nicks accumulate in the genome and turn into DSBs. In normal cells, these, DSBs are often repaired by HDR with the help of BRCA1 and BRCA2. In BRCA1/2 mutant cells, however, inhibition of PARP1 results in the formation of a large number of unrepaired DSBs which eventually lead to apoptosis and cell death (Dziadkowiec *et al.*, 2016). Although all three of these agents are extremely good at exerting their intended function, but cancer cells, like any other living entity, are constantly evolving and adapting to their environment (Blackman and Brothers, 2004). More so, the uncontrolled cell division, lack of a fully functional repair pathway, as well as a potentially hostile metabolic environment further increases the rate of mutagenesis which leads to a high genetic diversity within the tumour. This in turn, fuels an accelerated evolutionary process which can lead to adaptation to chemotherapeutic agents.

1.3.1 Modes of chemotherapy resistance

Adaptation to chemotherapeutic agents can occur through two different mechanisms. In the *intrinsic resistance* mode, some cancer cells within the heterogenous population of the tumour intrinsically harbour the resistance phenotype from the beginning. Some have speculated that these resistant cells are stem cell-like cancer cells that can inherently tolerate the cytotoxic agents because of reduced apoptosis, better DNA repair mechanisms as well as ability to efflux cytotoxic agents (Hong, Lee and Nam, 2015; Prieto-Vila *et al.*, 2017). Others have suggested that the intrinsic

resistance is a result of an extreme genetic diversity within the tumour. Therefore, it is assumed that the mutations that may render the cells resistant to chemotherapy are already present in some of the cancer cells and the administration of chemotherapy simply applies a selection pressure that allows for clonal selection and expansion of these resistant cells (Greaves and Maley, 2012). Although these two models of intrinsic resistance seem to differ in subtle details, at core they both unify on one concept, there are some populations of cells that do not respond to chemotherapy. This lack of response, might have resulted from epigenetic or genetic adaptations. The former results in formation of stem cell-like resistant cells, the latter results in clonal evolution of resistant cancer cells (Holohan *et al.*, 2013; Prasetyanti and Medema, 2017). However, in the second mode of resistance, the *acquired resistance*, the resistance to chemotherapy occurs during the course of treatment. This can again either be in the form of epigenetic changes, which may result in the formation of stem cell-like cancer cells (Levina *et al.*, 2008), or it might occur by appearance of new mutations, which may or may not have been caused by the action of the chemotherapy agent itself (Cowin *et al.*, 2012). Although all of these suggested mechanisms had been observed in different cancer types, both in patients and model systems, the exact extent that each mechanism may apply to a certain type of cancer or even a specific patient may vary.

1.3.2 Chemotherapy resistance in HGSOC

In HGSOC, despite an apparent good response to chemotherapy over 70% of the patients present with recurrence. These recurrence tumours often exhibit a higher tolerance to further cycles of chemotherapy indicting their adaptation to cytotoxic agents (Jayson *et al.*, 2014). Previous studies have shown that HGSOCs are clonally heterogeneous (Bashashati *et al.*, 2013; McPherson *et al.*, 2016) and that the recurrent tumours emerge from selective persistence of pre-existing tumour clones (Castellarin *et al.*, 2013). Moreover, analysis of relapse specific mutations has failed to identify a dominant platinum-induced mutation signature, indicating that chemotherapy itself is unlikely to have contributed to the emergence of recurrent tumours (Lambrechts *et al.*, 2016).

Nevertheless, both the intrinsic and acquired modes of chemotherapy resistance had previously been reported in HGSOC. The main form of intrinsic resistance to platinum therapy had been associated with amplification of Cyclin E1 (CCNE1) on chromosome 19q which had been observed in 20% of the HGSOC cases (Etemadmoghadam *et al.*, 2010; Patch *et al.*, 2015). Nevertheless, since reactivation of HDR is the main component of resistance to platinum therapy, this observation might simply be due to the synthetic lethality interaction between the gain of CCNE1 and the loss of BRCA1/2 rather than a direct function of CCNE1 (Etemadmoghadam *et al.*, 2013). Furthermore, a significant upregulation of extracellular matrix related gene, had also been observed in patients with primary platinum-resistance that do not have CCNE1 amplification (Etemadmoghadam *et al.*, 2009). Moreover, it has been shown that loss or downregulation of LRP1B in HGSOC can result in primary resistance to liposomal doxorubicin treatment (Cowin *et al.*, 2012; Patch *et al.*, 2015).

Unlike intrinsic resistance that appears to be limited to a few mechanisms in HGSOC, acquired chemotherapy resistance had been observed in various different forms. BRCA1/2 reactivating changes, including reversion of deleterious mutations or removal of promoter methylations, had been reported to be one of the most common forms of acquired resistance to platinum therapy and PARP inhibitors (Etemadmoghadam *et al.*, 2009; Christie and Bowtell, 2017). Moreover, it has been shown that genomic rearrangements involving ABCB1 that result in gene/promoter fusions which lead to overexpression of the product of this gene, MDR1, also result in acquired resistance in HGSOC (Christie and Bowtell, 2017). MDR1 (Multidrug Resistance protein 1), is an ABC transporter that can actively increase the efflux of the cytotoxic molecules from the cytoplasm, thereby reducing the effective concentration of the chemotherapeutic agent. Furthermore, acquired resistance to chemotherapy may arise from epigenetic changes in the genome of the cancer cells. For instance, methylation and therefore, silencing of the MLH1 gene, that is involved in the MMR pathway, has also been shown to result in chemotherapy resistance in HGSOC (Zeller *et al.*, 2012).

Overall it is evident that in patients that show an initial good response to chemotherapy, acquired resistance is likely to be the main mode of chemotherapy resistance which leads to tumour relapse. However, considering that in majority of the patients, as part of the initial treatment, the fallopian tubes and ovaries have been removed, it is logical to assume that recurrence is initiated from metastatic cancer cells that have survived the chemotherapy. Although, some have already hypothesised the involvement of metastatic cancer stem cells in the relapse process (Giornelli and Mandó, 2017), there still is no conclusive evidence to show the exact source of tumour recurrence in HGSOC. Extrapolating from our knowledge in other types of cancer, the source of recurrence could either be from circulating tumour cells (Crosbie *et al.*, 2016), or from minimal residual disease in bone marrow/lymph nodes (Sai and Xiang, 2018). But, neither the circulating tumour cells nor the bone marrow/lymph node deposits have frequently been reported in HGSOC. However, another form of residual disease, referred to as Microscopic Residual Disease (MRD), is often identified in HGSOC patients with an apparent complete response to chemotherapy. Since the presence of MRD after chemotherapy in these patients had been shown to significantly reduce the survival (Narod, 2016), it is commonly believed that these tumour 'leftovers' can potentially be the source of chemotherapy resistance and tumour relapse in HGSOC. Nevertheless, to date, no experimental evidence for this hypothesis has been provided. This is mainly because of technical limitations in accurate and reliable genetic investigation of few numbers of cells that are present in the MRD. Therefore, to better understand the mechanisms of resistance to chemotherapy and the importance of MRD in this process, it is essential to improve our technical abilities to an extent that allows for reliable genomic studies on the few numbers of cells that can be obtained from the MRD of HGSOC patients.

1.4 Current state of HGSOC studies

Currently the majority of our understanding on the genomics of ovarian cancer arises from whole genome sequencing (WGS), whole exome sequencing (WES), and transcriptome analysis of bulk tumour samples from primary or recurrent disease (TCGA, 2011; Patch *et al.*, 2015; McPherson *et al.*, 2016). Such studies have generated a vast amount of knowledge in regards to the progression and evolution of HGSOC. Important driver genes, potential new drug targets, new mechanisms of chemotherapy resistance, and reliable predictive markers are amongst the most important findings from these studies. Most importantly, a 2016 study by McPherson et al. on 68 tumour samples from seven HGSOC patients, had been the most extensive genomics study on HGSOC to date (McPherson *et al.*, 2016). This study has clarified the modes of clonal divergence and clonal spread in HGSOC and has shown the importance of studying the clonal composition of tumours in inferring the evolutionary history of the disease. However, majority of the clonal evolution interpretations in this study had been obtained by computational inference of clonal compositions from bulk sequencing data and therefore are affected by the underlying assumptions of the mathematical models. Furthermore, in studies in regards to the origin of the ovarian cancer, bulk sequencing data of the tumour had often been combined with WES or targeted sequencing data of Formalin-Fixed Paraffin Embedded (FFPE) samples obtained from potentially premalignant lesions of fallopian tubes (Lee *et al.*, 2007; Labidi-Galy *et al.*, 2017). Despite the controversial nature of genetic studies on FFPE samples, these studies have led to the discovery of a clonal relationship between the primary tumour and potential premalignant lesions in the fallopian tubes. In contrary, another study with a similar structure have suggested that in at least some cases, the identified lesions in the fallopian tube are a result of metastasis from the primary tumour (Eckert *et al.*, 2016). Therefore, despite decades of research on the evolution of HGSOC, because of scarcity of high-quality samples and lack of a reliable genomics method for the study of limited available materials, various aspects of HGSOC evolution still remain unclear.

1.4.1 Limitations and obstacles

Although the aforementioned previous studies on HGSOC have had a profound impact on our knowledge about this deadly disease, two key questions in regards to origin, progression and evolution of HGSOC remains to be answered:

1) What mutation or mutations initiate the malignancy in HGSOC? All of the current studies assume that deleterious mutations in TP53 gene accompanied by loss of heterozygosity (LOH) on chromosome 17p are the first driver mutation in HGSOC. However, the discovery of SOX2 overexpression in premalignant lesions of fallopian tube that have lacked the TP53 mutation has raised the possibility of non-coding regulatory mutations that may precede the functional loss of TP53 in the initiation of HGSOC (Hellner *et al.*, 2016). Such mutations have been left unidentified mainly because nearly all of the extensive studies on HGSOC's origin had been performed on WES or targeted sequencing data obtained from FFPE samples.

2) Which cells are responsible for the tumour recurrence? Although the clonal evolution study by McPherson et al. had been able to identify a small population of cells in one of the primary tumour sites that is likely to have resulted in the emergence and evolution of the recurrence tumour in one of the patients, the exact mechanism through which these cells have survived the chemotherapy and their hiding place have not clearly been defined. Moreover, despite strongly correlating with poor survival, the relevance and importance of MRD in the relapse of HGSOC is still unclear, mainly because of lack of reliable genomics studies on the MRD.

Currently the main obstacle in answering these questions remains to be the technical difficulties in dealing with limited amount of material that can be obtained from potential premalignant lesions and microscopic residual disease (MRD). Moreover, the analysis of bulk sequencing data from primary tumours also results in the loss of a significant amount of information in regards to the intra-tumour heterogeneity. Additionally, normal cell contamination that commonly occurs in bulk WGS and WES data could significantly affect the result of the analysis.

The latter two issues can easily be addressed by using laser capture microdissection (LCM) to selectively isolate clonal *tumour islets* from cross-section of a tumour mass. Although studying such tumour islets will preserve the heterogeneous information that has resided in the tumour mass and will not suffer from normal cell contamination, but the use of LCM will inevitably limit the amount of available material for genomics studies. Therefore, a method that allows for accurate identification of SNVs and CNAs from the limited number of cells that can be obtained via laser-capture microdissection of potential premalignant lesions, MRD, and isolated clonal tumour islets can simultaneously help addressing all of the aforementioned issues.

1.4.2 State of the art genomics methods for the study of HGSOC

With currently available methods and equipment several strategies can be employed to achieve sequencing data from limited amounts of material. The simplest strategy to achieve data from small number of LCM tumour cells would be to use whole genome amplification (WGA) methods to increase the amount of available material. Several different methods of WGA have been used in cancer studies. Multiple displacement amplification (MDA), MALBAC, PicoPlex, and Ampli1 have all been previously used and shown to be able to result in reliable CNA predictions from limited amount of tumour DNA (Deleye *et al.*, 2015, 2017). However, extensive amplification, results in loss of information through allelic drop-outs and often produces a large number of artefact mutations across the amplified genomes, making SNV calling from such samples extremely unreliable. While single nucleus sequencing strategies require fewer cycles of amplification and have previously been shown to be applicable to HGSOC tumour samples (McPherson *et al.*, 2016), these methods however, suffer from two main issues: Firstly, the quality of single cell/nuclei data is often not sufficient for de novo discoveries and the low per cell breadth and depth of coverage significantly reduces the negative predictive value of the data. Secondly, dissociation of the frozen sample into single nuclei disrupts the spatial organisation of the tissue and therefore, makes this method inapplicable to MRD, where majority of the tissue is composed of normal cells.

On the other hand, droplet and compartmentalization-based WGA methods not only can result in reliable CNA predictions, but also have been shown to be able to reduce the number of false positive (FP) artefact mutations associated with WGA (Peters *et al.*, 2012). The basic idea behind all of such methods is to subdivide the already low amount of template into a large number of independent compartments to reduce the number of DNA molecules in each compartment as much as possible. Whole genome amplification, often in the form of MDA, is then performed on these spatially separated molecules of DNA. This compartmentalisation of WGA reactions minimises the competition between different DNA molecules and therefore, reduces the rate of allelic drop-outs. Furthermore, since the WGA artefacts are believed to occur randomly across the genome, the independent amplification of fragments allows for identification of these FP mutations by computational methods: As repeated occurrence of same random artefact mutation in multiple compartments is extremely unlikely such FP mutations can be identified by considering the number of compartments that harbour the specific mutation. In addition, since mutations that are on the same DNA fragment will always co-occur in the same compartments, the spatial segregation of large molecules of DNA would also allow for inference of phasing information of these mutations using mathematical models. Complete Genomics was one of the first who successfully implemented this idea in the form of Long Fragment Read (LFR) sequencing method in a 384-well plate format (Peters *et al.*, 2012). This was followed by the introduction of droplet based Linked-Read sequencing by 10x-Genomics few years later (Zheng *et al.*, 2016). While LFR sequencing is often performed on about 100-200 pg of template DNA, Linked-Read sequencing requires up to 10 folds more template DNA. Therefore, in this study, LFR technology was found to be more suitable for the analysis of the small number of cells that can be obtained from MRD and tumour islets.

1.4.3 Long Fragment Read sequencing

In its core, LFR relies on sub-haploid dilution of large fragments of DNA into a 384-well plate to a degree where each compartment receives no more than 10% of a haploid genome. This guarantees that in over 95% of the wells only one allele from each locus will be present and therefore, minimizing the chances of allelic competition (which could result in allelic drop-out) while allowing for accurate extraction of phase information. After compartmentalisation of fragments, in the LFR method, DNA template fragments in each well will first undergo MDA based WGA using Phi29 DNA Polymerase in the presence of a certain amount of dUTP nucleotides. The incorporation of 'U' nucleotides in the amplified genome through WGA reaction allows for a two-step enzymatic fragmentation strategy referred to as CoREF (Controlled Enzymatic fragmentation) (Figure 1.3). During CoREF, Initially the whole genome amplified DNA is treated with two DNA modifying enzymes, UDG and EndoIV, in the same reaction. UDG (Uracil-DNA Glycosylase) is an enzyme that in *E. coli* bacteria is responsible for the removal of 'U' nucleotides from the DNA. The product of this enzyme is an abasic site which, in an *in vivo* scenario is recognised by DNA repair mechanisms and gets repaired. In CoREF, however, EndoIV (Endonuclease IV) identifies this abasic site and introduces a nick upstream of this position. In the second step of CoREF, DNA polymerase I (PolI) from *E. coli* is supplied to the reaction. The enzyme identifies the nicked sites and starts a process known as *Nick Translation*, through which the PolI enzyme moves the position of the nick towards the 3' of the DNA strand. During CoREF, when two PolI enzymes that are operating on opposite strands collide with each other, a DSB forms and DNA fragments. Depending on the initial dUTP/dTTP ratio in the WGA reaction, the nicks will be spaced at different distances from each other and consequently, will result in different fragment sizes upon application of PolI enzyme. Therefore, by adjusting the dUTP concentration in the WGA reaction it would be possible to fragment the DNA molecules into any specific fragment range that is required depending on the Next Generation Sequencing (NGS) strategy. Following fragmentation using CoREF, the resulting DNA fragments in each compartment will be ligated to a compartment specific indexing adapter.

These adapters will mark all of the molecule in the same compartment with a well-specific index sequence. After this step, DNA molecules from all of the wells can be mixed together to finalise the library preparation. Since every compartment has received a unique index, after sequencing, all of the reads that have emerged from a specific well can be grouped together for further downstream analyses (Peters *et al.*, 2012).

1.5 Thesis aims and scopes

My main goal throughout my DPhil project and hence this thesis, had always been to answer the two main unanswered questions about ovarian cancer:

What exactly starts it and why does it come back so often?

However, evidently, there are several reasons why these fundamental questions have not yet been answered about this deadly disease. The issue, at its base, is the lack of a reliable genomics method to fundamentally address such questions. Therefore, as a step towards finding an answer for these questions, in this thesis, I

- Implemented, optimised, and validated DigiPico sequencing method for the analysis of picogram quantities of DNA obtained from frozen HGSOC clinical samples.
- Designed and optimised a robust and reliable analysis pipeline for the processing and interpretation of DigiPico sequencing data.
- Performed the case study of a HGSOC tumour using DigiPico and demonstrated the importance and power of this method.
- And lastly, based on the observations from my case study, I discussed three hypotheses on the origin, progression, and evolution of ovarian cancer, and suggested the experimental means that are required to confirm or reject these hypotheses.

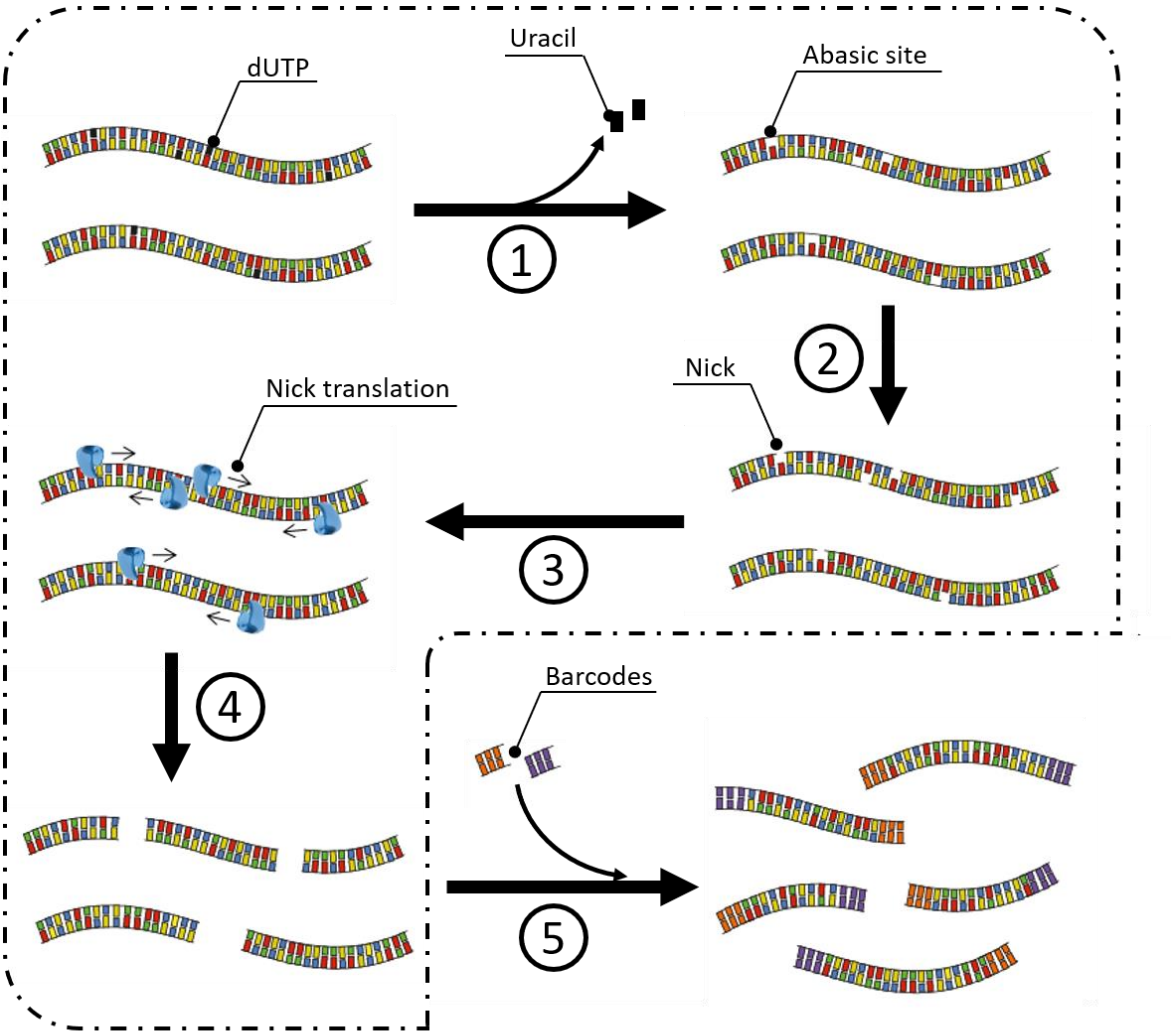


Figure 1.3. Library preparation procedure for Long Fragment Read (LFR) sequencing. 100-200 pg of template DNA is first distributed into a 384-well plate. Template DNA in each well is whole genome amplified in the presence of dUTP. Incorporation of dUTP in the amplified DNA allows for Controlled Enzymatic Fragmentation (CoREF) of the product. During CoREF: **(1)** UDG enzyme removes the Uracil bases from the DNA leaving abasic sites behind. **(2)** EndoIV enzyme, recognises the abasic sites and introduces a nick upstream of the lesion. **(3)** PolI enzyme recognises the nicked DNA and starts the Nick translation process. **(4)** When two opposing PolI enzymes meet each other the co-occurrence of nicks from opposite strands turns into a double-strand break which leads to the fragmentation of the DNA. The length distribution of final DNA fragments correlates with the initial concentration of the dUTP in the WGA reaction. The higher the concentration of the dUTP, the smaller the final products will be. **(5)** As part of the library preparation process, well-specific barcodes (indices) are ligated to the fragmented DNA molecules. After indexing the products of all wells can be combined to finalise the library preparation.

Chapter 2 : DigiPico Sequencing

As a result of the advantages of LFR sequencing, I decided to use LFR sequencing for a detailed study of tumour islets and MRD from multiple HGSOC patients. To begin with, as part of an ongoing collaboration, over 100 LCM samples from tumour islets and MRD of several patients were submitted to Complete Genomics for LFR sequencing and analysis. Only about 30% of the submitted samples were successfully processed and the sequencing data were analysed by Complete Genomics' experts. The final variant calls and phase information were then obtained from Complete Genomics. Further analysis and experimental validation of the variant calls, however, indicated that a large number of FP calls were present in the data. Application of more strict filtering criteria suggested by Complete Genomics did not appear to solve this issue. Since Complete Genomics uses its own proprietary human genome sequencing and analysis technology, the analysis of raw output data to identify the root issue was not possible. These complications as well as the low success rate in data generation soon indicated that LFR sequencing, as offered by Complete Genomics, does not seem to be applicable to clinical samples for the intended evolutionary analyses. Therefore, to enable the study of target samples for this study, I decided to implement an in-house compartmentalisation-based WGA and library preparation strategy and optimise it for clinical samples. This in-house library preparation strategy was named DigiPico sequencing, for 'Digital analysis of Picogram quantities of DNA'. DigiPico method is based on the main ideas from LFR technology, such as sub-haploid dilution of template in a 384-well plate and CoREF fragmentation. However, unlike LFR, the libraries prepared with DigiPico method is compatible with Illumina sequencing platforms and the analysis pipeline will be based on the available bioinformatic tools. In this chapter I will explain the step-by-step optimization approach which resulted in an optimal and streamlined library preparation protocol that consistently results in good quality libraries from minute amounts of template DNA or directly from LCM tissues.

The implementation of the DigiPico library preparation protocol consisted of multiple optimisation steps during which I had to achieve four main targets:

1. **Streamlined process with no purification step between the reactions.** As the DigiPico library preparation consists of 5 distinct enzymatic reactions (WGA, two steps of CoREF and two steps of adapter ligation) in a 384-well plate, purification of the products from each reaction step would render the method extremely expensive and time consuming. Moreover, multiple rounds of purification would increase the risk of cross contamination between the wells. Therefore, I decided that DigiPico library preparation pipeline should be designed in a way that enzymatic reactions could directly be performed on the unpurified products of the previous steps.
2. **Miniaturized reaction volumes.** Because of the previous criteria the total volume of the reactions will increase after each step. Therefore, it is important to ensure that the volume of the final product will not exceed the maximum capacity of 384-well reaction plates. Moreover, to reduce the overall cost of the enzymes and reaction mixtures for this process I decided to minimize the reaction volumes for DigiPico as much as possible, relying on the available automated liquid handling robots.
3. **Directly applicable to clinical samples obtained by LCM.** As the main goal of this project is to analyse rare clinical samples obtained from LCM of frozen tissues, it is of high importance for me to be able to apply the DigiPico method on minute amount of material that can be obtained by LCM. Direct application of the protocol on LCM tissues, without intermediate DNA purification steps, guarantees a minimal loss of material. This, therefore, lowers the threshold for the minimum number of cells that could be processed using DigiPico sequencing protocol.

4. **Highly multiplexed libraries.** Lastly, to enable independent analysis of each of the compartments in DigiPico and to allow for pooling of multiple DigiPico libraries on high throughput Illumina NGS platforms, such as NovaSeq, a highly multiplexable indexing strategy is required. Such multiplexing strategy would need to be custom-designed as currently, the diversity of commercially available indexing systems caps at 384, which, is the minimum number of required indices for only one DigiPico library.

2.1 Optimization of a streamlined protocol

DigiPico process, similar to LFR, starts with sub-haploid dilution of genomic DNA into a 384-well plate followed by WGA of 0.3 – 0.6 fg of denatured DNA in the presence of dUTP in each well. The presence of dUTP in the WGA reaction is essential as it allows for a time- and concentration-independent fragmentation of DNA through a two-step process referred to as CoREF (Figure 1.3). During CoREF, Uracil bases are first cleaved off the WGA DNA using UDG and the upstream phosphodiester bond of these abasic sites is hydrolysed by EndoIV to generate a nick. In the second step, DNA Polymerase I from *E. coli* (PolI) is used for *nick translation* which ultimately leads to the fragmentation of DNA (Figure 1.3). However, since the strategy for DigiPico protocol is to perform each reaction step by adding the components of the new reactions to the products of the previous steps, the excess dUTP nucleotides from the WGA reaction are likely to interfere with the second step of the CoREF reaction. Therefore, to ensure that no excess dUTPs are transferred to this step, an alkaline phosphatase, rSAP, is also required to be added to the first CoREF reaction mixture. After successful fragmentation of WGA products, fragmented DNA is A-tailed using Klenow fragment, and lastly, barcoded Illumina sequencing adapters are ligated to the end of the fragments using T4 DNA Ligase. As the Illumina sequencing adapters will contain an index sequence that uniquely identifies each well, after the ligation step all of the products can be pooled and be processed for sequencing on an Illumina NGS platform.

2.1.1 Controlled Enzymatic Fragmentation

Firstly, to be able to optimize a streamlined protocol, I found it essential to find a common buffer in which all of the enzymes that are required for the CoREF and adapter ligation stay fully functional. Table 2.1 shows the list of enzymes that are required for all of the library preparation steps and their relative activity in several commercially available buffers. It immediately becomes clear that NEBuffers 2 and 3 are the only choices that support the desired multistep enzymatic reactions. As EndoIV seems to be more sensitive to the reaction conditions compared to the other enzymes in this list, its recommended buffer, NEBuffer 3, was selected to ensure the highest activity for this enzyme. However, comparing the composition of buffers ('NEB', 2018) indicates that, while EndoIV requires a high concentration of salt (50 – 100 mM NaCl) for activity, the high concentration of monovalent salts may not be desired for T4 DNA Ligase enzyme (Hayashi *et al.*, 1985). Therefore, as the rest of the enzymes in the list do not seem to be sensitive to NaCl concentrations, a possible strategy would be to start the reactions in NEBuffer 3 where EndoIV functions optimally. It is then possible to only dilute out the NaCl in later steps by providing the reaction mixtures in an NaCl-free buffer with adequate amounts of $MgCl_2$ and DTT. To check the feasibility of this strategy, genomic DNA from a tumour sample was first whole genome amplified in the presence of various concentrations of dUTP (0, 2.5, 5, 10, 25, and 50 µM final concentrations) using REPLI-g mini MDA kit following manufacturer's instructions. The products of these WGA reactions were then directly used for a series of enzymatic reactions as depicted in Figure 2.1.

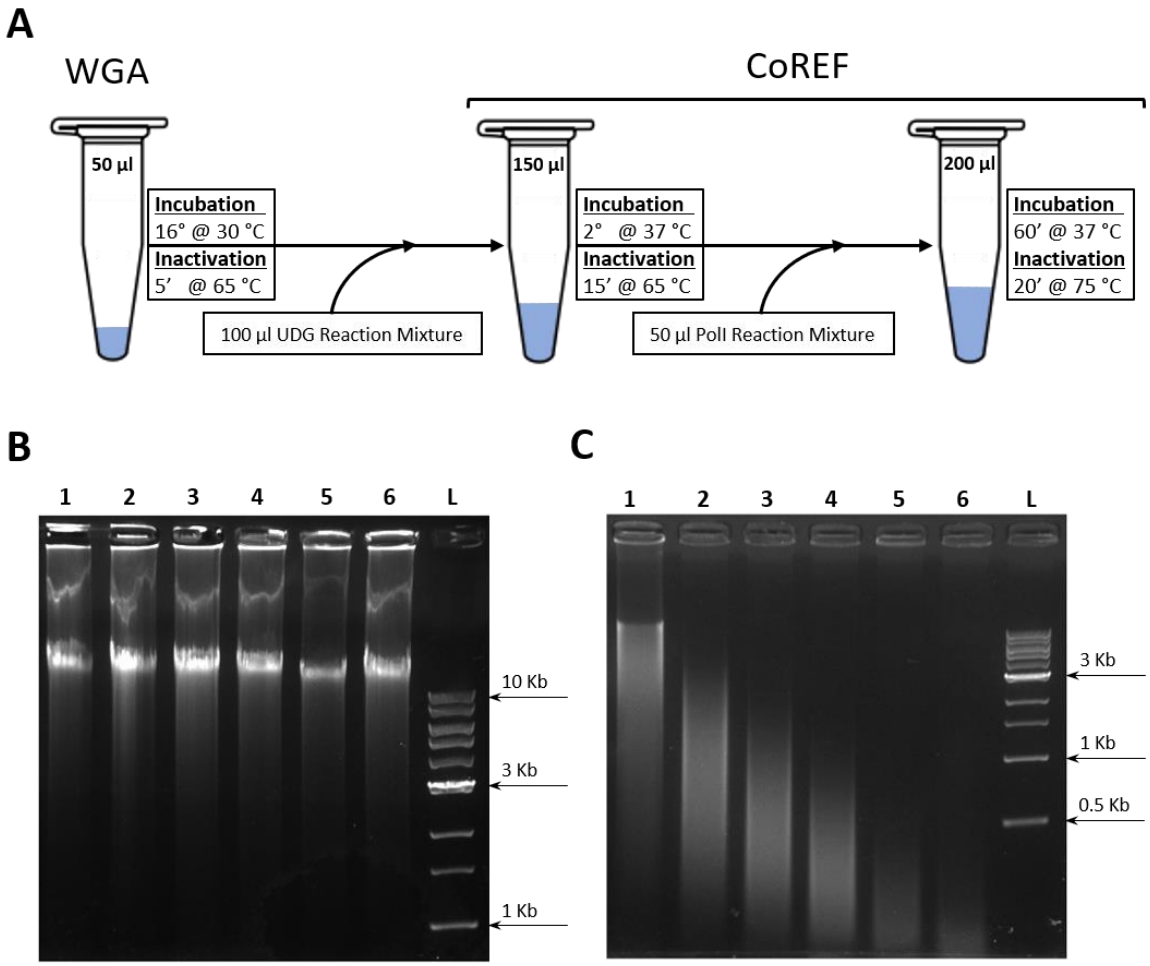


Figure 2.1. Streamlined CoREF process for DigiPico. (A) WGA is performed in the presence of 5 different dUTP concentrations (0, 2.5, 5, 10, 25, 50 µM) using REPLI-g mini kit following manufacturer's instruction on 50 ng of template DNA in 50 µl for 16 hours. Next, 100 µl of UDG reaction mixture (0.05 U/µl rSAP, 0.125 U/µl UDG, 0.25 U/µl EndoIV in 1.5x NEBuffer 3) was added to the heat inactivated WGA reaction and incubated at 37 °C for 2 hours followed by heat inactivation at 65 °C. Lastly, 50 µl of PolI reaction mixture (1 U/µl PolI, 0.3 mM dNTP, 10 mM $MgCl_2$, 1 mM DTT) was added to the previous reaction to complete the CoREF process. **(B)** Agarose gel electrophoresis of WGA products in the presence of various dUTP concentrations indicates that Phi29 DNA polymerase is not affected by the presence of dUTP in the reaction. **Lanes: [1] 0 µM dUTP, [2] 2.5 µM dUTP, [3] 5 µM dUTP, [4] 10 µM dUTP, [5] 25 µM dUTP, [6] 50 µM dUTP, [L] 10 Kb DNA Ladder**. **(C)** Agarose gel electrophoresis of CoREF products show a strong inverse correlation between the concentration of dUTP in the WGA reaction and the size of the DNA fragments following CoREF. 10 µM dUTP (Lane 4) generates the desired range of fragments for Illumina NGS platforms. **Lanes: [1] 0 µM dUTP, [2] 2.5 µM dUTP, [3] 5 µM dUTP, [4] 10 µM dUTP, [5] 25 µM dUTP, [6] 50 µM dUTP, [L] 10 Kb DNA Ladder.**

Resolving the resulting products on an agarose gel shows that CoREF performs well under these reaction conditions (Figure 2.1B-C). Therefore, it is possible to streamline the CoREF reactions without the need for any purification steps. Moreover, these results clearly confirm the strong inverse correlation between the final fragment size and the concentration of dUTP in the MDA reaction. Based on these results, addition of 10 µM dUTP to the WGA reaction seems to result in 200 – 400 bp sized fragments which is the desirable range of fragments for WGS libraries that are intended to be sequenced on Illumina sequencing platforms.

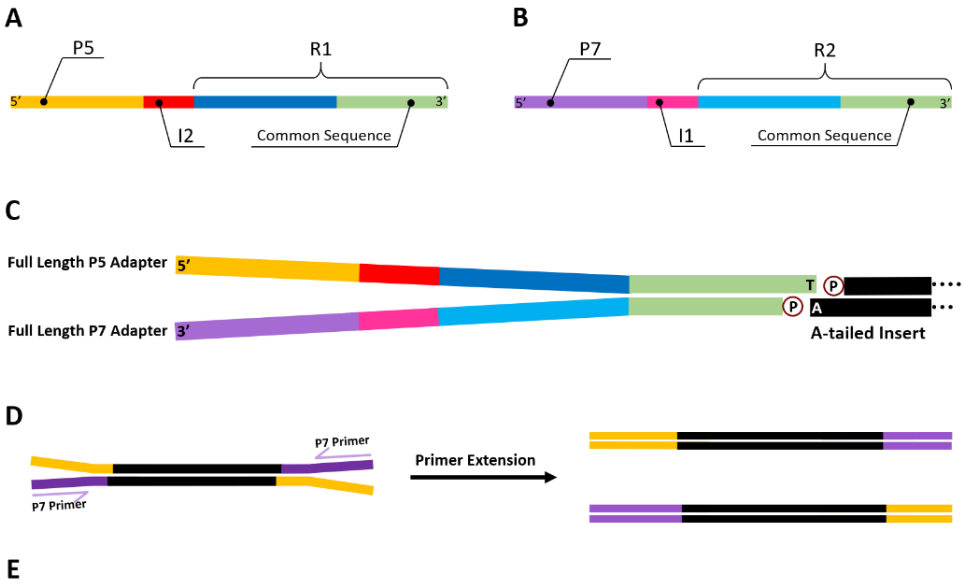
Table 2.1. Relative activity of CoREF and adapter ligation enzymes in various commercially available buffers.

Enzyme	NEBuffer 1	NEBuffer 2	NEBuffer 3	NEBuffer 4	CutSmart Buffer
rSAP	Active	Active	Active	Active	**Active***
UDG	Active	Active	Active	Active	Active
EndoIV	Inactive	Active	**Active***	Low Activity	N/A
E. coli PolI	Active	**Active***	Active	Active	Active
Klenow exo-	Active	**Active***	Active	Active	Active
T4 DNA Ligase	Active	Active	Active**	Active	Active

* Recommended buffer
** High salt concentration may interfere with the activity of enzyme. (New England Biolabs n.d.)

2.1.2 Adapter ligation optimization

Having optimized a series of compatible enzymatic reactions to obtain 200 – 400 bp DNA fragments from WGA DNA, the next step is to prepare the fragmented products for ligation to Illumina adapters. Illumina adapters serve several important functions for library preparation, sequencing and data analysis. Figure 2.2 summarises the important features of Illumina adapters, all of which, except for the indices, must be present on both ends of any DNA fragment that is needed to be sequenced on a standard Illumina sequencing platform. There are two main strategies to ligate Illumina adapters to fragmented DNA while ensuring that all fragments receive both the P5 and the P7 sequences. First strategy is to perform asymmetric ligation of full-length adapters (Figure 2.2C).



P5 Adapter: AATGATACGGCGACCACCGAGATCTACAC [i5 Index] ACACTCTTTCCCTACACGAC**GCTCTTCCGATC***T

P7 Adapter: [P]-**GATCGGAAGAGC**ACACGTCTGAACTCCAGTCAC [i7 Index] ATCTCGTATGCCGTCTTCTGCTTG

Figure 2.2. Anatomy of Illumina adapters used in DigiPico protocol. Any DNA fragment that needs to be sequenced on Illumina sequencing platforms must contain a P5 adapter sequence on one end and a P7 adapter sequence on the other end. These adapter sequences provide essential sequences that are necessary for the cluster formation, sequencing, and analysis of the NGS data. All the components except for the index sequences are essential for a standard sequencing procedure on Illumina sequencing platforms. Indices are short stretches of sequence (usually 6 – 8 bp) that can be used to identify a group of sequencing reads and therefore allow for pooling of multiple samples in the same run. **(A-B)** Full length P5/P7 adapters contain a P5/P7 sequence at the 5' end, an optional index (I2/I1) sequence in the middle, and R1/R2 sequences at the 3' end. P5/P7 sequences provide primer binding sites for P5/P7 primers to allow for enrichment of adapter ligated fragments during the library preparation using limited-cycle PCR. They also provide binding site for grafted oligos on Illumina flow cells to allow for bridge amplification and cluster formation. On the NovaSeq 6000, MiSeq, HiSeq 2500, and HiSeq 2000 platforms P5 also provides primer binding site to read the index 2 (I2) sequence. R1 and R2 provide binding sites for Read 1 and Read 2 sequencing primers, respectively. The first nucleotide following R1 and R2 sequences are the first nucleotides that will be sequenced and will be reported for reads 1 and 2, respectively. R2, in reverse, also provides primer binding site for sequencing of index 1 (I1). On the iSeq 100, MiniSeq, NextSeq, HiSeq X, HiSeq 4000, and HiSeq 3000 platforms, R1, in reverse, provides the primer binding site for sequencing of index 2. The 12 bp long *Common Sequence* at the 3' end of the P5 and P7 adapters provides a complementary region for the annealing of P5 and P7 adapter oligos which is required for their ligation to the target DNA fragments. **(C)** Asymmetric annealing of full-length T-tailed P5 and 5' phosphorylated P7 adapter oligos allows for TA-ligation of annealed adapters to A-tailed target fragments. **(D)** The asymmetry of ligation product can be resolved following at least one cycle of primer extension using P7 primer. Although in reality up to 10 – 12 cycles of PCR using P5 and P7 primers might be performed to enrich for adapter ligated products. **(E)** Sequence of the adapter oligos used in DigiPico protocol. In P5 adapter oligo the 'T' nucleotide extension is connected to the rest of the oligo using a phosphorothioate bond in place of a phosphodiester bond (Marked by *). Phosphorothioate bonds are resistant to exonuclease cleavage and therefore show a higher stability. This will ensure minimum loss of the 'T' extension. If lost, this can result in self-ligation of blunted adapters. To allow annealing to P5 adapter oligo, P7 adapter oligos are synthesised in reverse-complement and are phosphorylated at their 5' end for TA-ligation. P5 and P7 sequences are underlined. The *Common Sequence* is highlighted using bold typeface. The list of all i5 and i7 index sequences can be found in Appendix I.

In this method full length P5 adapter oligo with a 3' T nucleotide extension is annealed to the reverse complement of a full length and 5' phosphorylated P7 adapter oligo. The 12 nucleotide long *Common Sequence* at the 3' end of the P5 adapter oligo and 5' end of the P7 adapter oligo allows for partial annealing. The T nucleotide extension at the 3' end of the P5 adapter oligo then allows for TA-Ligation with A-tailed DNA fragments. Performing at least one cycle of PCR on the ligation products using P5 and P7 primers resolves the asymmetry and produces fragments that contain P5 on one end and P7 on the other end (Figure 2.2D). Second strategy starts with the ligation of a short, T tailed, and often looped common adapter sequence followed by at least 2 – 3 cycles of PCR to add the indices and P5/P7 sequences to the end of the ligation products. The second strategy is often more efficient but it requires at least 1 or 2 extra enzymatic reaction steps and is more likely to require a purification step before the indexing PCR. Since DigiPico is performed in a 384-well plate format any purification steps or additional enzymatic reactions can increase the complexity of the protocol. Therefore, I decided to employ the asymmetric ligation strategy for DigiPico.

2.1.2.1 Adapter design and indexing strategy

Full length P5 and P7 adapter sequences were derived from Illumina TruSeq HT adapter sequences. Figure 2.2E demonstrates the sequence and structure of both adapters. To achieve 384 different indices, one for each well of the plate, a dual indexing strategy was employed. For this purpose, 16 i5 index sequences and 24 i7 index sequences from Illumina's Nextera indexing system were selected and a full length P5 or P7 adapter oligo was designed and synthesised for each set of indices, respectively (Appendix I). As recommended by Illumina, in P5 adapter oligos the last phosphodiester bond (3' end) was replaced with a phosphorothioate bond and the 5' end of the P7 adapter oligos were phosphorylated (Figure 2.2E). All oligos were requested to be PAGE purified to ensure the highest purity of the full length oligos for each index. Combining each of the 16 i5 oligos with each of the 24 i7 oligos will result in 384 unique index combinations which is required for DigiPico sequencing.

2.1.2.2 A-Tailing and adapter ligation

Having designed a set of adapter sequences required for DigiPico, the possibility to ligate these adapters to the ends of DNA fragments from CoREF was evaluated. For this purpose, CoREF products were directly used in a series of enzymatic reaction as shown in Figure 2.3. In this test both commercially available adapters and home-brewed adapters were used to check for their ligation efficiencies. The resulting ligation products were then used in a PCR reaction using primers complementary to P5 and P7 sequences and the products were resolved using agarose gel electrophoresis. Although the results clearly show the successful ligation of both home-brewed and commercial adapters to the fragmented DNA, the large amount of self-ligated adapters in the products is likely to interfere with the downstream sequencing process (Figure 2.3B). This is because shorter products are much more efficient at forming clusters during bridge amplification and therefore, presence of minute amount of self-ligated adapters in the final library could result in a complete run failure.

2.1.2.3 Troubleshooting self-ligated adapters

To troubleshoot the issue of self-ligated adapters, I first tried to avoid their formation by reducing the concentration of adapter oligos in the ligation reaction. To check this, several dilution ratios of adapters were used in a set of ligation reactions and the products were used for PCR using P5 and P7 primers. The results (Figure 2.3C) show that although limiting the concentration of adapter oligos could potentially reduce the formation of self-ligated adapter but it will also reduce the ligation efficiency to target fragments, which, is not desirable.

Another approach to solve the issue of self-ligated adapter would be to remove any short-sized products from the reaction before performing the P5/P7 PCR. This is a viable strategy for DigiPico because after adapter ligation each well will already be uniquely indexed. Therefore, at this stage it is possible to pool all of the wells together and perform a purification step on the pooled products using SPRI (Solid Phase Reversible Immobilization) magnetic beads.

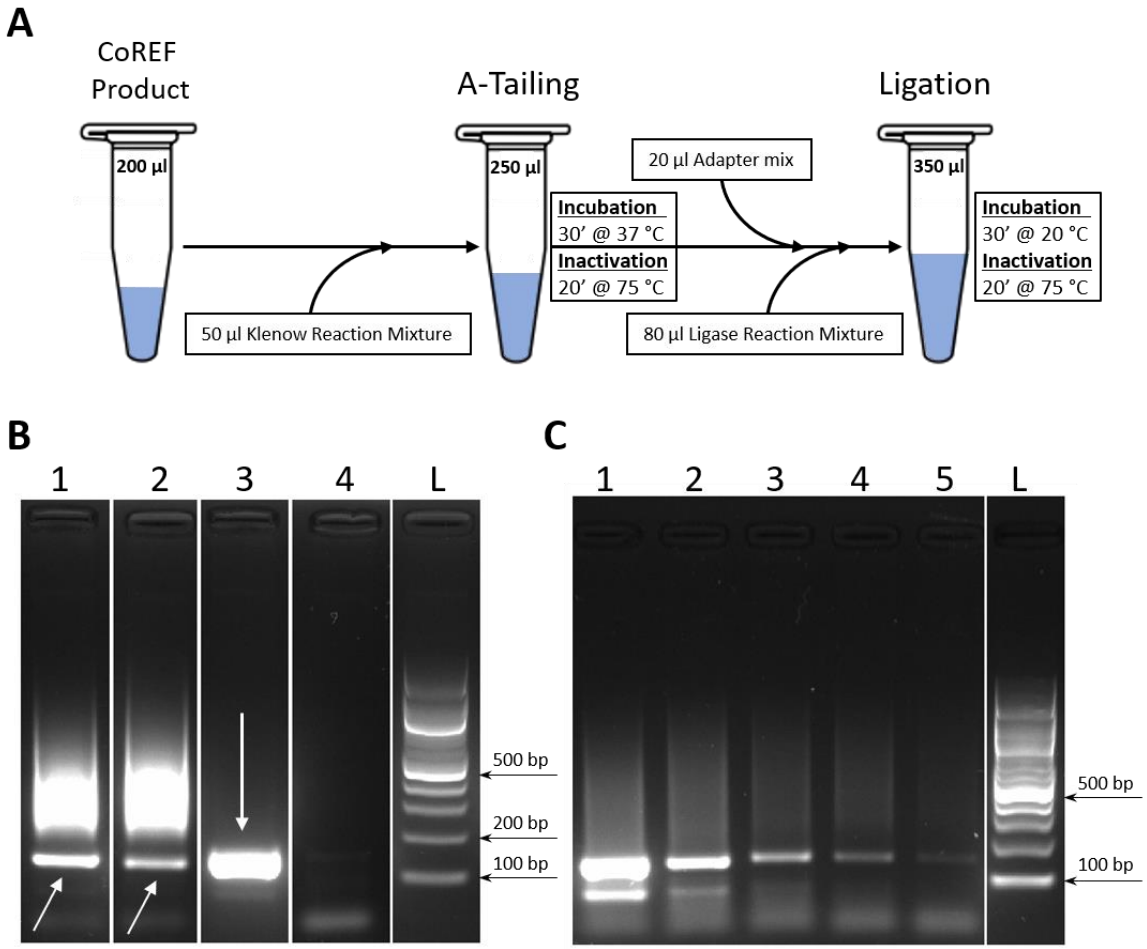


Figure 2.3. Streamlined A-tailing and adapter ligation for DigiPico. (A) To perform the A-tailing reaction, 50 µl of Klenow reaction mixture (0.5 U/µl Klenow exo-, 0.5 mM dATP, 10 mM $MgCl_2$, 1 mM DTT) was added to the heat inactivated CoREF reaction products. Following 30 minutes incubation at 37 °C and heat inactivation of Klenow fragment, 20 µl of Agilent Adapters or home-brewed Adapter mix (10 µM P5 adapter oligo and 10 µM P7 adapter oligo in water) was first added to the reaction and then 80 µl of Ligation Reaction Mixture (25 U/µl T4 DNA Ligase, 3.75 mM ATP, 6.25 mM $MgCl_2$) was added to perform the TA-ligation at 20 °C for 30 minutes. **(B)** 1 µl of heat inactivated ligation products were then directly used for 25 cycles of PCR using P5/P7 primer pairs. Visualizing the PCR products on agarose gel, it is clear that the process has worked. However, the large amount of self-ligated adapter (indicated by white arrows at around 150 bp) are likely to interfere with the sequencing process. **Lanes: [1] Home-brewed adapters, [2] Agilent Adapters, [3] No Klenow Control, [4] No Ligase control, [L] 100 bp DNA Ladder**. **(C)** To check if reduction of adapter oligo concentration in the ligation reaction can help reducing the self-ligated adapters, a 1:2 serial dilution of home-brewed adapters down to 1:80 was prepared and all of the dilutions were used for ligation. Visualising the P5/P7 PCR products of various adapter dilution ratios indicated that reduction of adapter oligos adversely affected the ligation efficiency and therefore cannot be used as a remedy to solve the problem of self-ligated adapter. **Lanes: [1] No dilution, [2] $\frac{1}{3}$ dilution, [3] $\frac{1}{9}$ dilution, [4] $\frac{1}{27}$ dilution, [5] $\frac{1}{81}$ dilution, [L] 100 bp DNA Ladder**.

SPRI beads are carboxyl coated, paramagnetic particles that can reversibly bind to DNA molecules in the presence of a crowding agent such as Polyethylene glycol (PEG) (DeAngelis, Wang and Hawkins, 1995). One interesting property of these particles is that they lose their affinity to shorter fragments of DNA at lower concentrations of PEG. Therefore, by lowering the concentration of PEG it is possible to selectively eliminate short fragments from the solution. As SPRI beads are commercially supplied in a solution of PEG 8000, rather than adjusting the PEG concentration, the volumetric ratio of beads to DNA solution is often used as the means to adjust the size selection. Figure 2.4A, illustrates the relationship between the bead ratio and the size of the fragments that can efficiently bind to the beads. Employing this property of the SPRI beads it is possible to select a specific range of DNA fragments from a crude solution through a process referred to as Dual-Size Selection (DSS). During DSS, larger DNA fragments are first bound to the beads by adding a low ratio of beads to the solution. The beads are then discarded and more beads are added to the supernatant to bind to the desired range of fragments. Lastly the supernatant containing undesired short fragments is discarded and the target fragments are eluted from the beads using water (Verrow *et al.*, 2016). For the purpose of DigiPico, DNA fragments in the range of 250 – 500 bp are required after ligation, therefore, in the first step a bead ratio of 0.45x can be used to remove all of the larger fragments from the solution. Then a further 0.35x beads (total of 0.8x) can be added to the supernatant to capture the fragments larger than 250 bp from the solution. The supernatant containing the short self-ligated adapters can then be discarded and the target DNA sequences can be eluted from the beads. To put this idea into test, ligation products of both home-brewed adapters and commercial adapters were used for DSS and the solutions from various stages were used for P5/P7 PCR (Figure 2.4B). The results indicate that DSS is indeed capable of removing all of the self-ligated adapters from the ligation products and therefore, can be used in the DigiPico library preparation protocol.

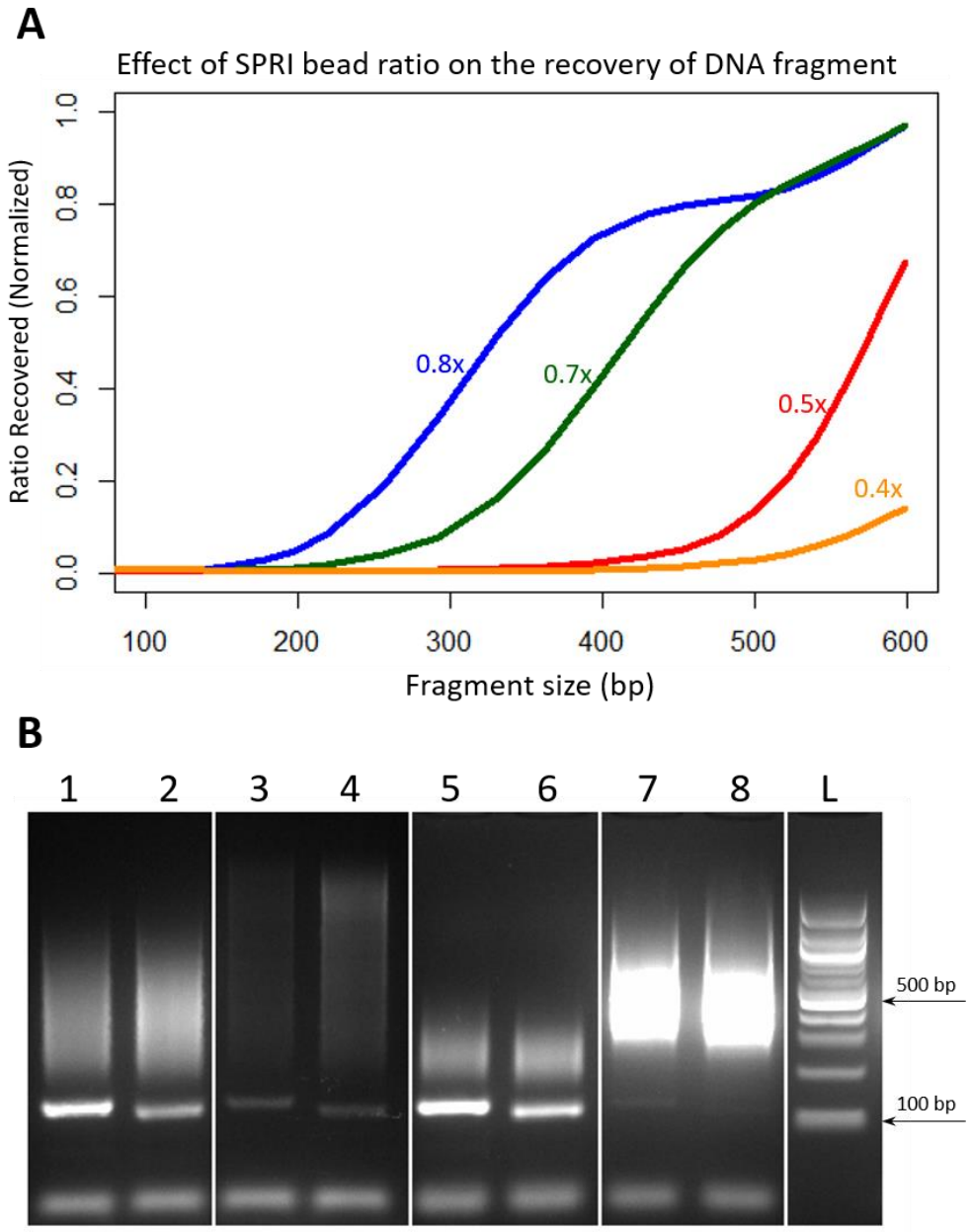


Figure 2.4. SPRI bead purification of ligation product to remove self-ligated adapters. SPRI beads show varying affinities to different sizes of DNA fragments depending on the **Bead:Solution** ratio. At higher bead ratios the affinity to smaller fragments increases. **(A)** Plot shows the relative affinity of SPRI beads to different fragment sizes at various **Bead:Solution** ratios. Data obtained from Beckman Coulter Inc. (Verrow *et al.*, 2016) **(B)** Self-ligated adapters can be efficiently removed by performing Dual-Size Selection using SPRI beads. Ligation products from home-brewed (HB) and Agilent (AG) adapters were used for Dual-Size Selection. 45 µl of beads were first added to 100 µl of ligation reaction (0.45x). After incubation and magnetic separation, supernatant was kept and the beads containing the large fragments were discarded. A further 32 µl of beads were then added to 140 µl of supernatant (0.8x). After incubation, supernatant containing small fragments was discarded and beads containing the desirable size range were washed with 80% ethanol and the DNA was eluted from them using water. Small aliquots from various stages were used for PCR using P5/P7 primers and the products were visualized on agarose gel to monitor the progress of the size selection. **Lanes: [1] HB-Ligation, [2] AG-Ligation, [3] HB-0.45x Beads, [4] AG-0.45x Beads, [5] HB-0.8x Supernatant, [6] AG-0.8x Supernatant, [7] HB-0.8x Beads, [8] AG-0.8x Beads, [L] 100 bp DNA Ladder**.

2.1.3 WGA optimization and automated library preparation

At this stage a streamlined process to produce adapter ligated DNA fragments from whole genome amplified DNA in a single tube without the need for any purification steps have been implemented. However, to this point, all of the reactions had been performed in standard reaction volumes and on recommended quantity of template DNA. Therefore, the next step would be to see if these reactions would still perform well in reduced reaction volumes and in the presence of lower amounts of template. Since DigiPico is planned to be performed in a single 384-well plate, the volume capacity of the wells determines the upper limit of the reaction volumes. The currently optimised protocol requires addition of 2 volumes of UDG reaction mixture, followed by 1 volume of PolI reaction mixture, 1 volume of Klenow reaction mixture, and finally 2 volumes of Ligation mix. This means that the volume of the final product would be 6 times more than the starting WGA reaction volume. As 384-well reaction plates have a maximum capacity of 35 µl, the WGA reaction must be performed in volumes of 5 µl or lower to allow for accommodation of all of the reaction steps. Moreover, as the entire process of DigiPico requires addition of low volume reaction mixtures to a 384-well plate, manual pipetting of every reaction step would extremely user intensive and prone to user error and cross contamination. Therefore, the use of an automated liquid handler seems necessary. At the time of implementation of DigiPico, access to two automated liquid handlers was possible. EpMotion 5075 and Mosquito HTS. EpMotion allows for dispensing 1 µl – 1ml of solution from individual tubes or plates to 96-well or 384-well plates. Mosquito on the other hand allows for aliquoting 50 nl – 1.2 µl of liquid from a reservoir 384-well plate to another 384-well plate. Both of these systems offer important features that can be useful for automation of DigiPico process.

2.1.3.1 Low volume WGA of low amounts of DNA

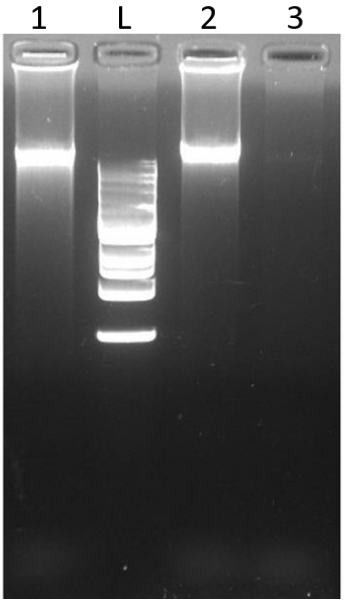
To begin with, a WGA experiment in 5 µl reaction volume showed that MDA using REPLI-g mini kit does not seem to result in any products in 10% of the recommended reaction volume (Figure 2.5A).

To further dissect the effect of volume on WGA using REPLI-g mini kit, a set of WGA reactions in various reaction volumes were performed. Results confirmed that the reaction volume seems to have a strong effect on the WGA outcome (Figure 2.5B). This is unexpected because all of these reactions have received the same amount of template DNA and all of the other conditions except for the reaction volume had been identical between them. One explanation for this observation might be that in MDA reactions, it is the total amount of the enzyme, rather than its concentration, that determines the outcome. Therefore, to check this hypothesis, a set of WGA reactions in the volume of 5 µl with various amounts of extra enzymes were performed (Figure 2.5C). These results clearly support the hypothesis that the total amount of Phi29 DNA polymerase is important for the completion of the reaction and that to reduce the reaction volume to 5 µl additional enzymes must be supplied to the reaction. This provides a reaction condition in which WGA on pg quantities of DNA can successfully be performed in 5 µl reaction volumes which is required for DigiPico protocol.

2.1.3.2 Small-scale, manual DigiPico library preparation

Having filled the last gap in the series of reactions that are required for DigiPico, I decided to put the whole process of DigiPico into test in a small-scale library preparation experiment using only 16 wells. For this experiment 120 pg of DNA was first denatured and diluted in 400 µl of water to achieve a template concentration that is comparable to what will be used in a full-scale DigiPico experiment. 1 µl of this solution was then used as template for a 5 µl WGA reaction in each well of the 16-well experiment. The WGA products were then directly used for CoREF, adapter ligation, bead purification and limited-cycle PCR amplification. The resulting library showed the correct size distribution when analysed on TapeStation and therefore, was used for sequencing on an Illumina MiSeq platform. Despite a good run performance, analysing the distribution of each of the indices in the sample revealed that only one of the index pairs have a strong presence in the library and all of the other indices seem to be present at very low frequencies (Figure 2.6A). This most likely indicates that ligation has worked efficiently in well 15 (N718-S511 sample) but all of the other wells seem to have had a very low adapter ligation efficiency.

A



B

Reaction Volume	Template amount	Template Concentration	WGA Product Concentration
50 µl	0.15 pg	0.003 pg/µl	8.56 ng/µl
20 µl	0.60 pg	0.030 pg/µl	0.42 ng/µl
10 µl	0.30 pg	0.030 pg/µl	0.13 ng/µl
5 µl	0.15 pg	0.030 pg/µl	0.08 ng/µl

C

Extra Enzyme in 5 µl Reaction	Phi29 Polymerase Provider	Average Product Concentration
0	-	0.05 ng/µl
0.9 µl	Qiagen	41.2 ng/µl
0.15 µl	NEB	43.2 ng/µl
0.4 µl	NEB	0.09 ng/µl
0.9 µl	NEB	0.09 ng/µl

Figure 2.5. Optimising WGA using REPLI-mini in 5 µl. (A) REPLI-g mini is able to WGA 50 ng (**Lane 1**) or 3 pg (**Lane 2**) of genomic DNA in a volume of 50 µl. However, an isolated 5 µl aliquot of the reaction in Lane 2 does not result in any products (**Lane 3**), indicating that MDA using REPLI-g mini kit might be sensitive to down-scaling. **(B)** To further understand the importance of reaction volume for efficiency of WGA using REPLI-g mini kit, a series of WGA reactions in various volumes were performed. After 16 hours incubation the concentration of DNA was measured using Qubit. Results clearly confirm that reducing the volume significantly affects the efficiency of WGA using REPLI-g min kit. **(C)** It is possible that the number of Phi29 DNA polymerase enzymes can be the limiting factor in the progression of WGA in smaller volumes. Therefore, to check this hypothesis, various amounts of excess enzymes from different sources (More enzyme from REPLI-g mini kit or Phi29 DNA polymerase from NEB) were added to a 5 µl WGA reactions in duplicates. The concentration of product DNA in the reactions was measured using Qubit after 16 hours incubation at 30 °C. Increasing the amount of polymerase to 10-folds the recommended quantity using kit's own polymerase (Qiagen) or addition of an extra 1.5 units of Phi29 DNA Polymerase from another source (0.15 µl NEB) seems to completely restore the WGA reaction in 5 µl (Highlighted rows). However, addition of extra amounts of Phi29 DNA Polymerase from NEB (0.4 and 0.9 µl) seems to inhibit the reaction, probably because of interference of enzyme's storage buffer with the reaction conditions.

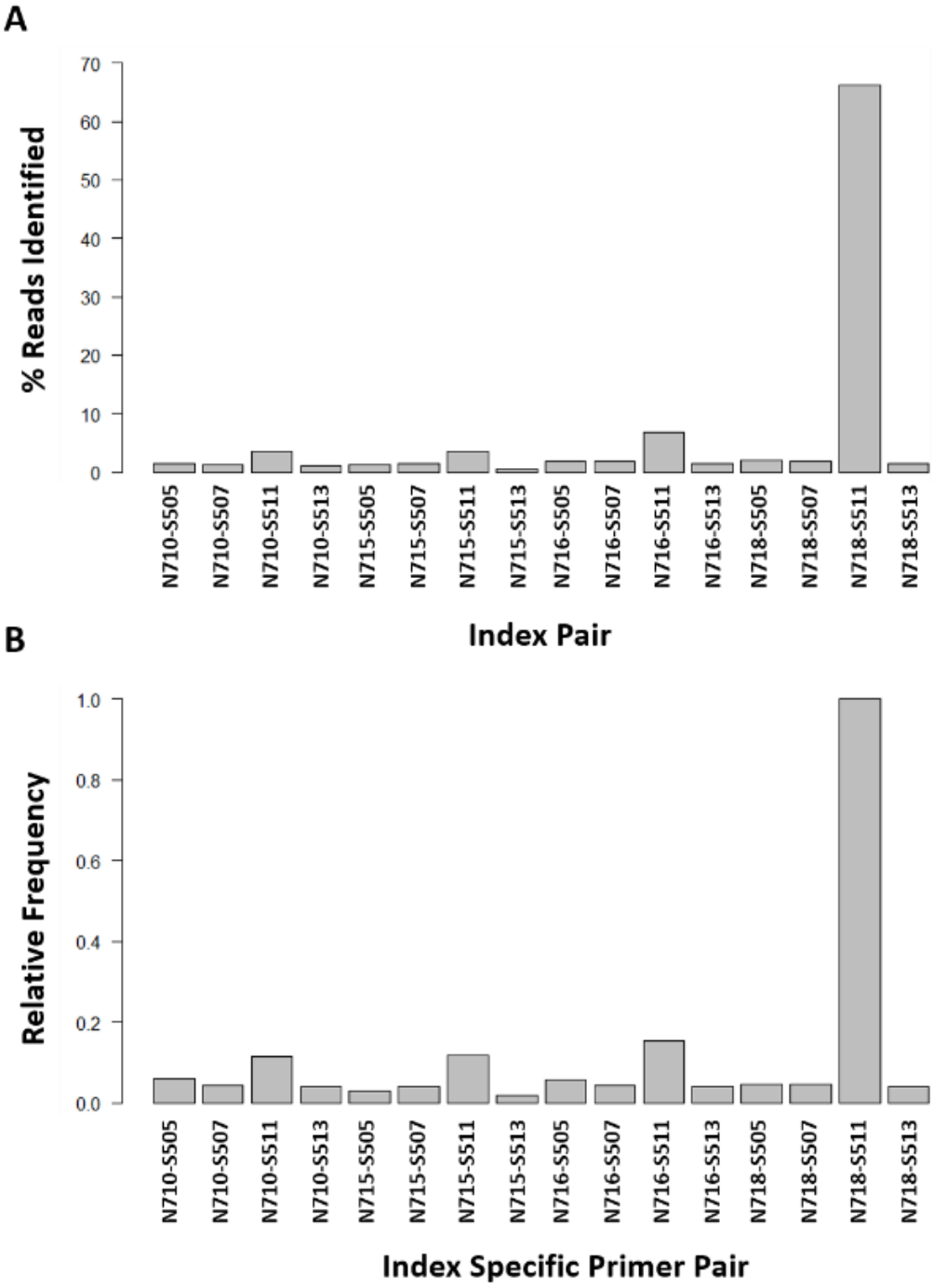


Figure 2.6. Inconsistent adapter ligation in small scale DigiPico library preparation. (A) Analysing the data from an NGS sequencing of a small scale DigiPico library preparation in 16 wells shows an inconsistent presence of reads from various wells with over 65% of the reads coming from only one of the wells. **(B)** To ensure that this bias is not caused at the level of sequencing or analysis, index specific qPCR primers were designed. A qPCR with all of the 16 combinations of adapter specific primers was then performed on the library. The qPCR results showed a complete agreement with the NGS results, indicating that inefficient and inconsistent ligation is likely to be the reason for this issue. qPCR results were normalized to N718-S511 for calculation of the relative frequencies.

To further ensure that the unequal distribution of indices had been present in the library and is not a result of sequencing or analysis bias, a set of adapter-specific primers were designed such that the 5' end of each primer would anneal to the P5 or P7 sequences and the 3' end would anneal to the 8 bp long i5 or i7 index sequences, respectively. This would ensure that each primer pair can only amplify fragments with perfect matching i5 and i7 indices. qPCR was then performed on the library as template using each combination of adapter specific primer pairs. The outcome was completely in agreement with the observed frequency of each index in the MiSeq sequencing data (Figure 2.6B) confirming that the bias is most likely due to issues at the ligation step. Therefore, to resolve this issue it is once more necessary to optimise the adapter ligation step to achieve the highest possible efficiency and consistency.

2.1.3.3 TA-ligation optimisation

TA-Ligation, inherently, is not an efficient process. Moreover, in this case, the small complementary region of 12 bp in the long adapter oligos could further reduce this efficiency. Hexamine Cobalt Chloride (HCC) and PEG 8000 are two compounds that are known to enhance the efficiency of blunt end ligation (Rusche and Howard-Flanders, 1985). To check the effect of these compounds on the adapter ligation efficiency in DigiPico protocol, in an experiment, 2% PEG, 1 mM HCC, or both were added to the ligation reaction mixture. Comparing the efficiency of adapter ligation in the presence of PEG and/or HCC using qPCR clearly indicated that while HCC doesn't seem to have any effects, PEG 8000 can significantly enhance the ligation efficiency (Figure 2.7A). Therefore, to check whether or not addition of PEG in the ligation reaction would also help to achieve a more consistent ligation for different adapter combinations another small scale DigiPico library preparation with 9 indices was performed. Comparing the frequency of each of the indices in the pool of the sample before and after limited-cycle PCR using adapter-specific qPCR, clearly indicates that presence of PEG 8000 in the ligation reaction significantly improves the outcome (Figure 2.7B).

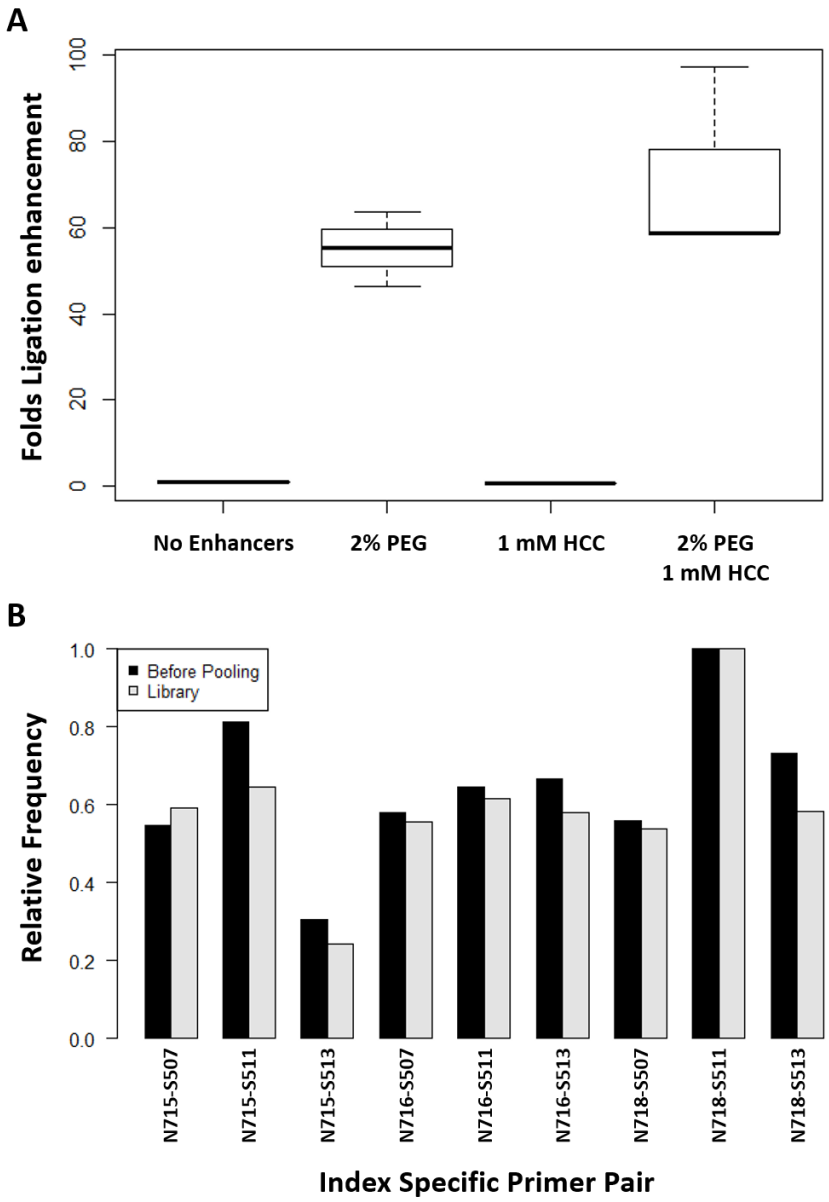


Figure 2.7. Addition of PEG 8000 can significantly enhance the adapter ligation efficiency. Polyethylene Glycol (PEG) and Hexamine Cobalt Chloride (HCC) are often used as enhancers to improve the efficiency of blunt end ligation. **(A)** Comparing the concentration of adapter ligated products using P5/P7 qPCR indicates that addition of 2% PEG 8000 to the ligation reaction can increase the ligation efficiency up to 50 folds. HCC, however, does not seem to have any effect on the adapter ligation efficiency in DigiPico protocol (n=3). **(B)** To check if the addition of PEG 8000 can improve the adapter ligation consistency, a small scale DigiPico library preparation using 9 indices was performed and the adapter ligation reaction was supplied with 2% PEG 8000. Ligation products from individual wells, before pooling, were used for qPCR using index specific primers of respective wells (dark bars). The final DigiPico library was also used as template for qPCR using all 9 adapter specific primer pairs (light bars). The results show a consistent presence of all of the indices in the library both at the pre-pool stage and at the library stage. qPCR results were normalized to respective N718-S511 data points for calculation of relative frequencies.

2.1.3.4 Failure of a full-scale, automated DigiPico library preparation

After confirming that enhancing the ligation efficiency using PEG would result in consistent adapter ligation across wells, next step would be to check whether or not a full-scale DigiPico library preparation in a 384-well plate using and automated pipetting system would result in a successful library. As the reaction volumes for this experiment are in the range of 5 – 10 µl, epMotion 5075 seems to be the better choice for setting up the reactions. EpMotion allows for two modes of liquid dispensing. In the 'free jet' mode liquid droplets are dispensed from the top of the plate into the wells without any contact with the contents of the well. In the 'mix' mode however, tips will enter the wells for each dispense. This mode also allows for a mixing step if required. Although the 'free jet' mode might not be ideal for DigiPico because in this mode it is not possible to mix the reaction contents with the newly added reaction mixture, I decided to use this setting for all of the reaction steps. This is because, while in the 'free jet' mode the robot takes about 10 minutes to dispense a reaction mixture in a 384-well, in the 'mix' mode aliquoting a solution into an entire 384-well plate takes over 1 hour to be completed. This makes it nearly impossible to use the 'mix' mode for DigiPico as with this setting the first few wells will be in the presence of the reaction mixture for over 1 hour at room temperature which could further increase the inconsistencies across the plate.

Therefore, using the 'free jet' mode of epMotion, a 384-well plate was set-up for the reactions. After the completion of the ligation step, the final products from each well were used as template for qPCR using P5/P7 primers. The results, however, showed an inconsistent amplification of adapter-ligated product across the wells (Figure 2.8A). While this inconsistency might be a result of poor reaction efficiencies because of lack of mixing of the reactions, there also is a chance that the WGA might have failed in some of the wells. Therefore, to ensure that whole genome amplified products are present in all of the wells, the DNA concentration in each well was quantified using the PicoGreen dye. The results show that although the DNA concentrations seems to be more consistent than the qPCR results, the WGA has failed in over 50% of the wells (Figure 2.8B).

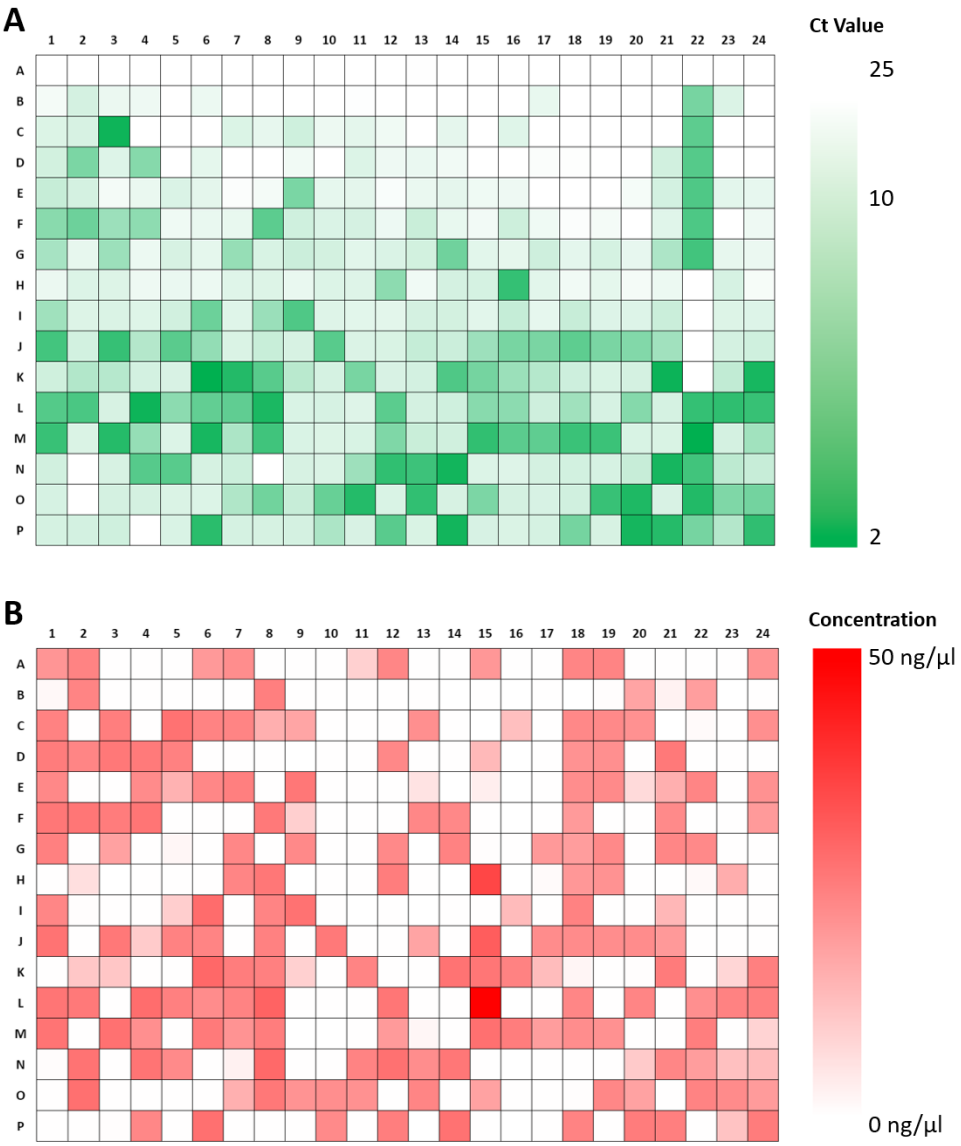


Figure 2.8. Heatmap presentation of a failed DigiPico library preparation. A full-scale DigiPico library was prepared using the epMotion 5075 liquid handler following the optimised DigiPico protocol starting from 5 µl WGA reactions in a 384-well plate. **(A)** After completion of the ligation step, 100 nl of the ligation products from each well were individually used for qPCR using P5/P7 primer pairs. Heatmap shows the Ct values obtained from each well in the original plate format. Majority of the wells seem to have failed to produce adapter ligated products. **(B)** Measuring the total amount of DNA in each well using PicoGreen assay shows that WGA had also been completed with varying outcomes across the plate. With some wells having as high as 50 ng/µl of whole genome amplified DNA while over 50% of the wells completely failing at producing any WGA products. Heatmap presents the concentration of DNA in each well across the plate in the original plate format.

Therefore, in a full-scale experiment in 384-plates using automated pipetting, not only the reaction efficiencies drop down significantly but also the WGA at low volume becomes inconsistent. This failure calls for another round of optimisation of reactions, starting from the WGA step.

2.1.3.5 Identifying the most suitable MDA kit for DigiPico

While the small-scale experiments with REPLI-g mini kit showed that this kit allows for amplification of pg amounts of DNA in 5 µl volume if supplied with extra enzymes, a large-scale test using automated pipetting in 384-well plate resulted in inconsistent amplification outcomes. Therefore, to find a more consistent MDA kit, I decided to compare the efficiency of 3 more commercially available MDA kits in a 384-well format. Moreover, to be able to monitor the progress of the WGA over time, EvaGreen was added to all of the reactions and the fluorescent intensity was monitored every 5 minutes in a qPCR instrument. EvaGreen is an intercalating dye that binds to the minor groove of the DNA and therefore, does not interfere with the isothermal WGA reactions (Hosokawa *et al.*, 2017). The results of this experiment showed that while neither of the REPLI-g mini and REPLI-g midi kits are able to whole genome amplify pg amounts of DNA in 5 µl (Figure 2.9A-B), both the REPLI-g UltraFast mini and REPLI-g Single Cell kits result in consistent amplification of DNA under similar conditions (Figure 2.9C-D). Although REPLI-g Single Cell kit took slightly longer than the REPLI-g UltraFast kit to WGA of the DNA, the REPLI-g Single Cell kit did not show any signs of unspecific amplification in the no-template control reaction (Figure 2.9D). These properties make REPLI-g Single Cell kit an ideal choice for DigiPico, because while it can consistently whole genome amplify pg quantities of DNA in small reaction volumes it also does not seem to produce any nonspecific products. Therefore, based on these results, I decided to use the REPLI-g Single Cell kit for WGA step of DigiPico library preparation protocol.

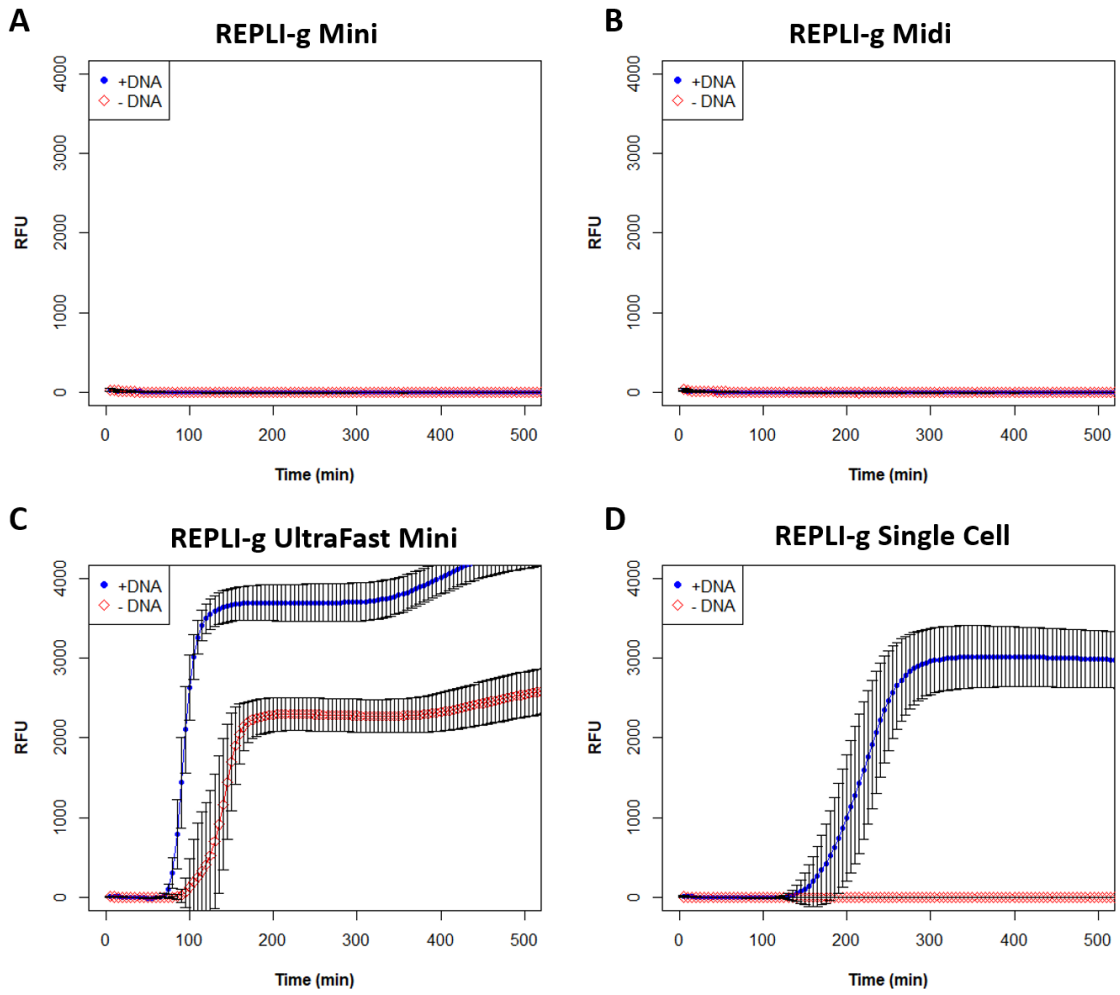


Figure 2.9. Comparison of WGA in 5 µl using four different MDA kits. 5 µl WGA reactions were set up using each of the four different MDA kits from Qiagen in a 384-well plate in 9 replicates with (blue) and without (red) template DNA. Reactions were supplied with 1x EvaGreen and the isothermal amplification was performed in a real-time PCR instrument. The fluorescent intensity was measured every 5 minutes. **(A-B)** REPLI-g Mini and REPLI-g Midi kit did not produce any products over the assay period. **(C)** REPLI-g UltraFast mini kit consistently produced a product within the first 2 hours of the reaction in the presence of template. However, with a small delay, unspecific products were also amplified in the control wells. Amplification of unspecific products is a result of self-annealing of random primers and the early onset of their appearance in this case is a likely indication for the presence of a high concentration of random primers in the reaction mixture. **(D)** REPLI-g Single Cell kit also produced WGA product in all of the replicates within 4 – 5 hours. The Single Cell kit, however, did not produce any unspecific products in the assay period which, makes it a more suitable choice for DigiPico. Error bars represent SD.

2.2 Reaction miniaturization

The results from the previous experiment showed that REPLI-g Single Cell kit can robustly amplify minute amount of template DNA in small reaction volumes. This suggest that this kit might not be as sensitive to the reaction volume, therefore, offering an opportunity to miniaturize the DigiPico reactions even further. Miniaturization of reaction volumes to below 1.2 µl for each step, not only would make DigiPico cost effective but also would allow the use of Mosquito liquid handler in place of epMotion. As Mosquito can dispense into 16 wells simultaneously, it can significantly reduce the pipetting time and therefore, would allow for mixing of the reaction mixtures in each step which could further improve the efficiency and consistency of all of the reaction steps.

2.2.1 WGA miniaturization

To use a liquid handler for DigiPico it is required to first add the WGA reaction mixture into each well and then add the template to the reaction. Mosquito liquid handler is able to consistently dispense volumes as low as 300 nl into an empty well. Lower volumes are not possible for the first round of pipetting into an empty well mainly because the small droplets may stick back to the tips because of surface tension causing inconsistency. However, when using a viscous solution such as WGA reaction mix, the minimum volume raises to about 800 nl. Since, in the WGA reaction, the reaction mixture constitutes 80% of the volume with the remaining 20% being used for the addition of the denatured template, with the use of Mosquito, it should be possible to start the DigiPico process with volumes as low as 1 µl.

2.2.1.1 Reduction of WGA reaction to 1 µl

To identify the lowest volume at which REPLI-g Single Cell kit can consistently WGA pg quantities of DNA in a 384-well plate, EvaGreen WGA assay was performed in several volumes (1, 2.5, 5 µl) using the REPLI-g Single Cell kit. The results show that the WGA reaction can efficiently progress to completion in volumes as low as 1 µl using the REPLI-g Single Cell kit (Figure 2.10A).

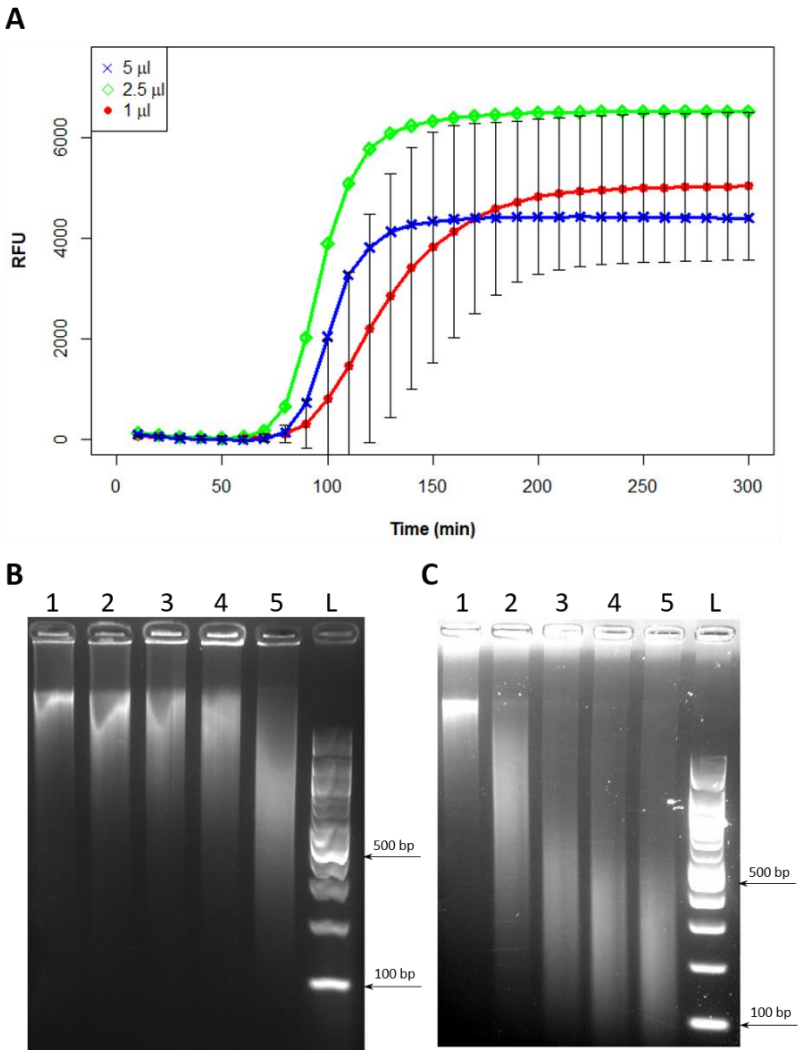


Figure 2.10. Reaction miniaturization and CoREF optimisation for REPLI-g Single Cell kit. The success of WGA in 5 µl using REPLI-g Single Cell kit hinted at the opportunity that this kit might be able to perform WGA in lower volumes as well. **(A)** To test this, WGA was performed at 3 different volumes (5 µl, 2.5 µl, and 1 µl), using REPLI-g Single Cell kit in 8 replicates. Progress of the WGA was monitored using EvaGreen assay with measurements every 10 minutes. The results show that REPLI-g Single Cell kit can easily tolerate the 50-fold down-scaling from the recommended reaction volume of 50 µl and consistently produces WGA products in 1 µl reaction volumes. The SD error bars are only presented for the 1 µl reaction volume. End point mean and SD values for 5 µl, 2.5 µl, and 1 µl reaction volumes are 4396±337, 6521±1855, and 5032±1471, respectively. **(B)** Products of CoREF were visualised on agarose gel from WGA products in the presence of **[1] 0 µM, [2] 2.5 µM, [3] 5 µM, [4] 10 µM, and [5] 25 µM dUTP**. REPLI-g Single Cell kit, compared to the REPLI-g Mini kit, seems to require a higher concentration of dUTP to produce 200 – 400 bp DNA fragments following CoREF. This is likely because the reaction mixture in the Single Cell kit contains a higher concentration of dNTPs and therefore to achieve a certain dUTP:dTTP ratio a higher concentration of dUTP is required. **(C)** Increasing the concentration of dUTP in the WGA reaction results in smaller DNA fragments following CoREF. **Lanes: [1] 0 µM, [2] 25 µM, [3] 50 µM, [4] 75 µM, and [5] 100 µM dUTP. [L] 100 bp DNA Ladder**. 75 µM dUTP seems to result in 200 – 400 bp product range which is required for DigiPico library preparation.

2.2.1.2 dUTP concentration optimisation for CoREF

As the WGA reaction conditions in REPLI-g Single Cell kit are likely to be different compared to the REPLI-g mini kit, it is important to check if CoREF can still efficiently take place with the previously optimised reaction mixtures. Therefore, to test this, WGA was performed in the presence of various concentrations of dUTP (0, 2.5, 5, 10, and 25 µM) in 1 µl. Then UDG and PolI reaction mixtures were added to the reaction and the resulting fragments were visualised on agarose gel. The results only showed fragmentation in the presence of the highest dUTP concentration with large products ranging from 1500 – 6000 bp (Figure 2.10B). This is most likely due to the presence of a higher concentration of natural nucleotides in the REPLI-g Single Cell kit compared to the REPLI-g mini kit. To overcome this issue, another round of dUTP concentration optimisation was performed using the REPLI-g Single Cell kit with dUTP concentrations in the range of 25 – 100 µM (Figure 2.10C). The results show that 75 µM dUTP results in production of 200 – 400 bp fragments, which is ideal for libraries intended for Illumina sequencing platforms.

2.2.2 Reaction volume adjustment for Mosquito

Since the maximum dispense volume for Mosquito is at 1.2 µl, with the current protocol two rounds of dispensing are required for the addition of UDG and Ligase reaction mixtures, as each require an addition of 2 volumes (2 µl) of reaction mixture to the previous step. Although this inherently is not a problem, but two rounds of dispensing in these steps will prolong the pipetting time which could result in inconsistent reaction performance in these steps. Therefore, to address this issue, a new workflow was designed (Figure 2.11A) in such a way that no step would require a dispensing of more than 1.2 µl of solution. Evaluation of the new workflow showed that CoREF can efficiently be performed with the new reaction conditions without any compromises (Figure 2.11B).

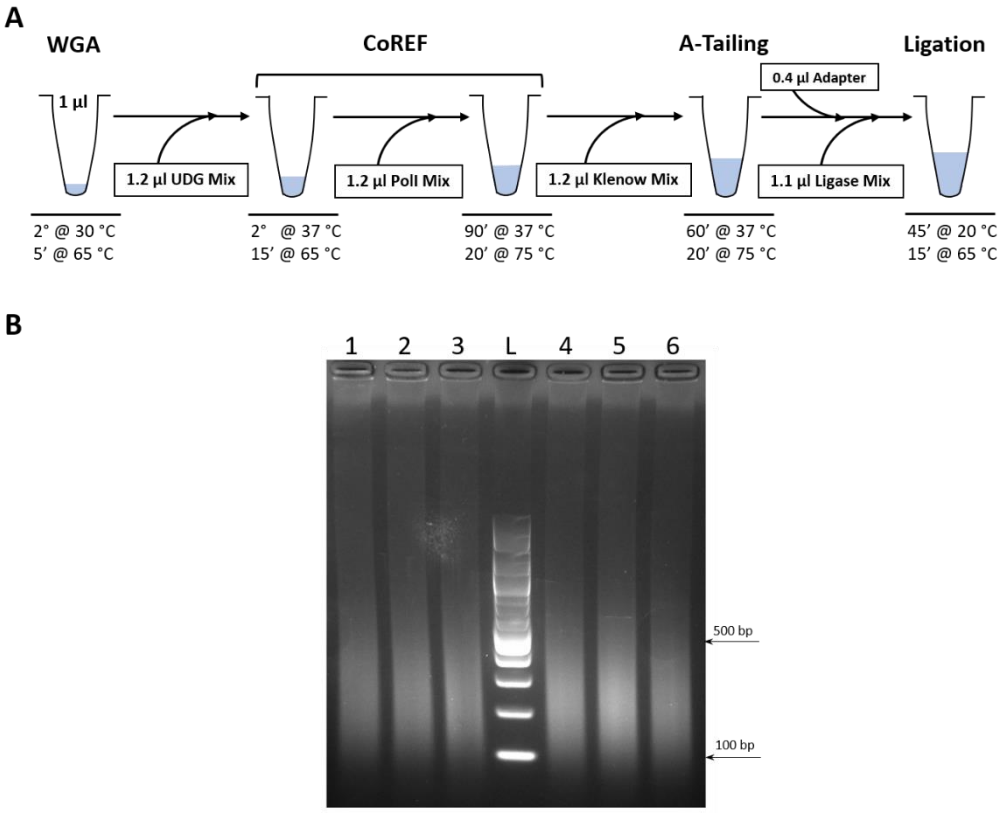


Figure 2.11. Miniaturized DigiPico workflow. To enable efficient use of Mosquito liquid handler for the preparation of DigiPico libraries a new DigiPico workflow was designed such that no step would require pipetting of more than 1.2 µl of liquid. This is because 1.2 µl is the upper limit of pipetting volume for Mosquito and higher volumes would require multiple rounds of pipetting which would be inefficient. **(A)** The workflow starts with addition of 800 nl of WGA reaction mix (0.58 µl Sc Reaction Buffer, 0.075 µl 1 mM dUTP, 0.04 µl EvaGreen 20x, 0.04 µl Sc Polymerase, and 0.065 µl water) into all the wells, followed by the addition of 200 nl of denatured template to each well. Because of the presence of EvaGreen dye in the reaction mixture, the progress of WGA can be monitored in a real-time PCR machine, measuring the fluorescent intensity every 5 – 10 minutes. After 2 hours of isothermal amplification at 30 °C the Phi29 DNA polymerase is heat inactivated at 65 °C for 5 minutes. Next, 1.2 µl of UDG Mix (0.08 U/µl rSAP, 0.2 U/µl UDG, 0.4 U/µl EndoIV in 1.8x NEBuffer 3) is added to each well, with three rounds of mixing, using Mosquito. After 2 hours incubation at 37 °C and heat inactivation, 1.2 µl of PolI Mix (0.4 U/µl PolI, 0.25 mM dNTP, 8 mM $MgCl_2$, and 0.8 mM DTT) is added with three rounds of mixing, followed by 90 minutes incubation at 37 °C, and 20 minutes heat inactivation at 75 °C. The product is then used for A-tailing by adding 1.2 µl of Klenow Mix (0.5 U/µl Klenow, 0.5 mM dATP, 8 mM $MgCl_2$, and 0.8 mM DTT), followed by 3 cycles of mixing using Mosquito and incubation at 37 °C for 1 hour. After heat inactivation of Klenow fragment, 400 nl of 20 µM unique adapter oligos is added to each well from an indexing plate and then 1.1 µl of Ligase Mix (40 U/µl T4 DNA Ligase, 5 mM ATP, 11.5% PEG 8000, and 6.8 mM $MgCl_2$) is added to all of the wells with 5 cycles of mixing. The ligation reaction is then completed by 45 minutes incubation at 20 °C and T4 DNA Ligase is heat inactivated at 65 °C for 15 minutes. At this point each well have received a unique barcode and therefore the contents can be pooled to finalise the library preparation. **(B)** To ensure that the new workflow does not affect the CoREF process negatively, in three replicates, old and new versions of the CoREF were performed on 1 µl of WGA DNA using REPLI-g Single Cell kit. The products were visualised using agarose gel electrophoresis and no significant difference in the size distribution of the fragments was detected. **Lanes: [1-3] CoREF using old protocol, [L] 100 bp DNA Ladder, [4-6] CoREF using new protocol**.

2.2.3 DigiPico library preparation and sequencing: Library 1110

2.2.3.1 Preparation of the first successful DigiPico library

With the success of WGA in 1 µl and the new 1.2 µl CoREF workflow, it is now possible to perform the entirety of the DigiPico protocol using the Mosquito liquid handler at a very low cost. Therefore, to test if the integrity of the protocol holds up in a full-scale experiment, a DigiPico library was prepared on DNA from a recurrence tumour sample (Patient 11152, site PT2R) following the latest optimised DigiPico workflow as illustrated in Figure 2.11A. Quality control of the WGA reaction and Ligation efficiencies confirmed a consistent completion of all the reactions (Figure 2.12A-C). Moreover, visualising the final library on TapeStation showed the expected size distribution without any traces of self-ligated adapters (Figure 2.12D).

2.2.3.2 Sequencing and preliminary analysis of library 1110

As the prepared library passed all the quality control measures, it was used for sequencing with a 300 Cycles High Output Illumina NextSeq kit which should result in a 30 – 40x coverage across the human genome (Library ID: 1110). The overall run performance seemed consistent with what is expected from standard WGS libraries. Raw data quality control of the reads using FastQC tool (Andrews, 2016) revealed that read duplication levels were slightly higher than standard sequencing data at about 15%, with no overrepresented sequences. Moreover, the per base sequence content analysis showed that the first 10 – 12 bases of each read have a slight deviation from the expected frequencies (Figure 2.13A). Because of this issue, DigiPico reads were trimmed at their 5' end before mapping. Preliminary analysis of the results also showed a consistent distribution of reads between the wells (Figure 2.13B-C). Next, the resulting trimmed reads from each well were individually mapped to the human hg19 reference genome using Bowtie2 (Langmead and Salzberg, 2012) which resulted in an average mapping rate of 91.35% across the wells (Figure 2.13D). The coverage analysis of the resulting 'bam' files using Samtools and Bedtools (Li *et al.*, 2009; Quinlan and Hall, 2010) showed that on average each well covers about 4.4% of the

genome with an average depth of 1.7x over the sequenced regions (**Figure 2.13**E-F). Lastly the analysis of the mapped reads indicated a bias against the GC-rich regions (**Figure 2.13**G) similar to what is observed in LFR technology (Peters *et al.*, 2012). These results overall confirm that DigiPico protocol allows for preparation of NGS libraries from minute amount of input DNA which can effectively be sequenced on Illumina sequencing platforms.

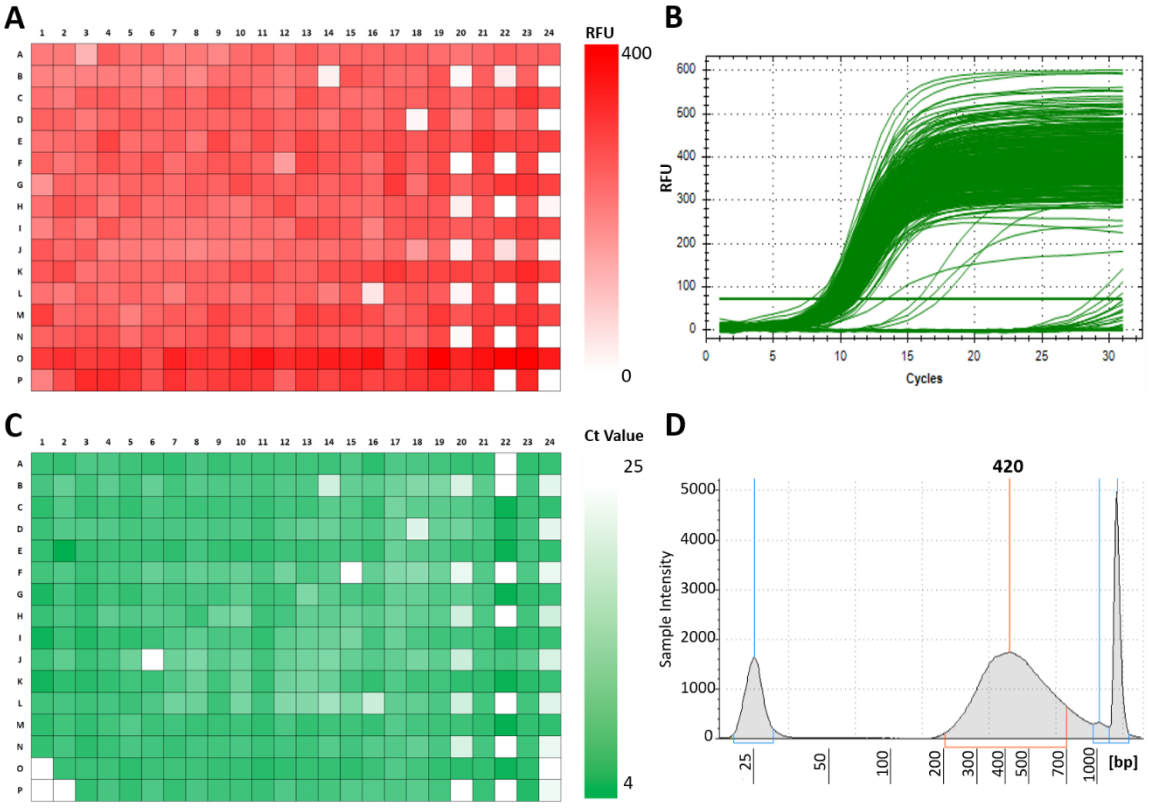


Figure 2.12. Library 1110 – First successful DigiPico library. Using REPLI-g Single Cell kit and with the help of Mosquito liquid handler, 300 pg of DNA from a recurrent tumour sample (patient 11152), was whole genome amplified after sub-haploid dilution in a 384-well plate. **(A)** Heatmap of the end point fluorescent values from the amplification indicates a consistent WGA across the plate. Because of an unexpected mechanical complication with Mosquito liquid handler, few of the wells at the end of the plate did not receive any template, which stand out by lack of amplification. **(B)** Monitoring the progress of WGA using EvaGreen assay using a real-time thermocycler shows a homogenous amplification across the wells consistent with the end point fluorescent data. Each cycle represents 10 minutes of incubation at 30 °C. **(C)** Following application of the latest DigiPico workflow on the whole genome amplified material, 100 nl of the adapter ligated end products were used as template for qPCR using P5/P7 primers. Heatmap of Ct values show a consistent and efficient ligation of adapters across the plate. **(D)** After pooling, isopropanol precipitation, Dual-Size Selection, limited-cycle PCR, and final purification, the resulting library was analysed on a TapeStation instrument. The strong homogenous peak at 420 bp indicates a high-quality library at the expected range. The sample was used for NGS on a NextSeq platform.

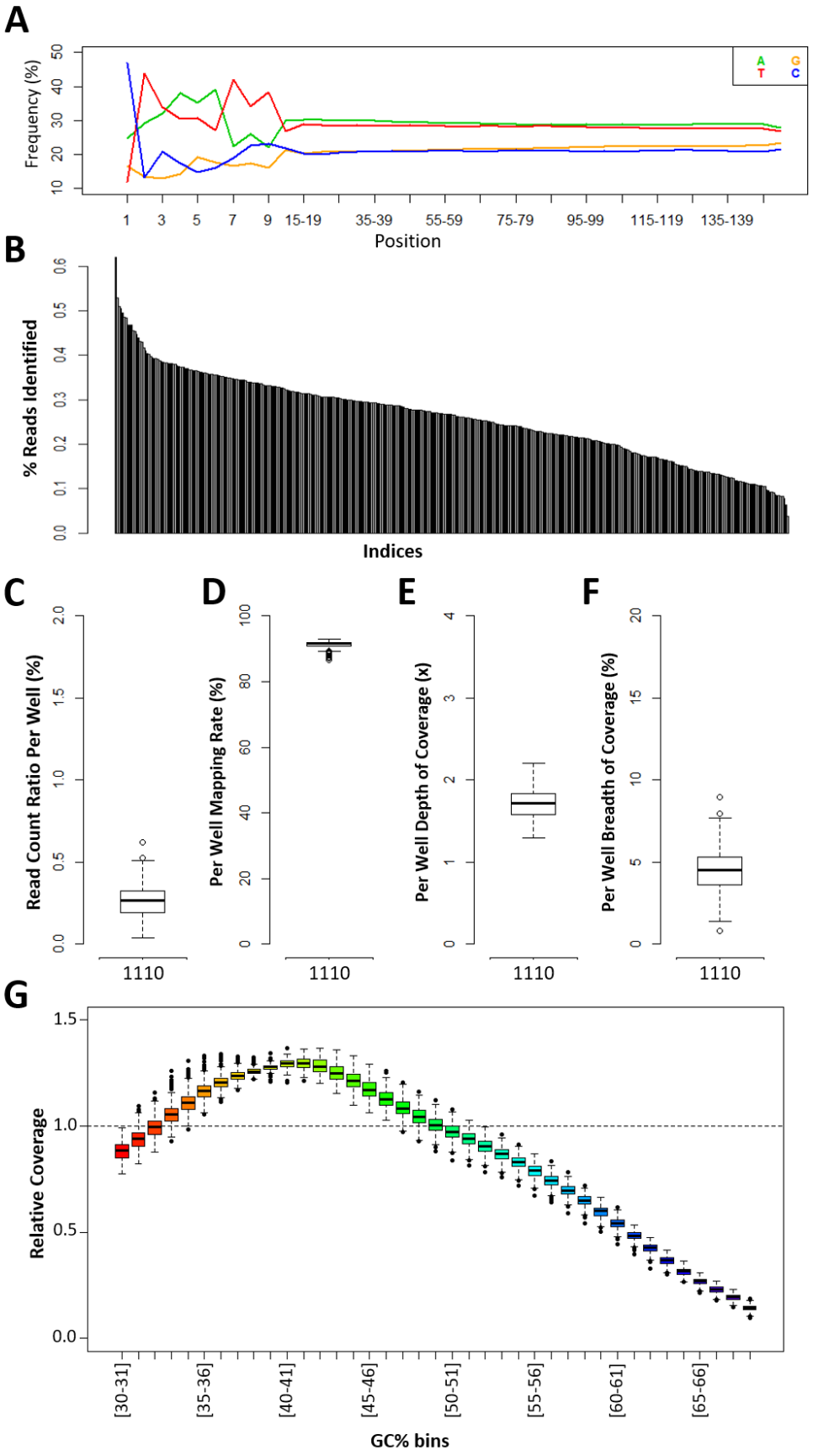


Figure 2.13 (Legend on next page)

Figure 2.13. Quality measures and sequencing statistics of library 1110. The sequencing data obtained from the first successful DigiPico library were thoroughly analysed with various tools to ensure the quality of the raw sequencing data before proceeding to higher level analysis. **(A)** Analysis of read qualities using FastQC tool revealed that the base composition of the first 10 – 12 sequenced nucleotides deviates from the expected 30:30:20:20 (A:T:G:C) ratios in a WGS library. This is likely due to the method of fragmentation (CoREF) which requires the collision of two opposing PolI enzymes on a nicked stretch of DNA. To avoid any errors that might result from this bias, the first 12 bp of all DigiPico reads will be trimmed for all of the downstream analysis applications. **(B-C)** Plotting the demultiplexing results show that there is a relatively homogenous distribution of reads among the different index combinations, which results in a fair distribution of reads across the wells. Bar plot represents the fraction of reads in each well (index combination) sorted from the well with the highest frequency to the lowest. **(D-F)** Reads from individual wells were mapped to human hg19 genome using Bowtie2. Samtools and Bedtools were used to calculate the coverage information from 'bam' files. Boxplots show the various basic mapping parameters obtained from each well. The fairly homogenous distribution of the values across the plate confirms the high quality of the data. Per well breadth of coverage indicates the fraction of the genome that is covered by at least one read in each well. Per well depth of coverage shows the average number of reads that cover the fraction of the genome that is covered by at least one read in each well. **(G)** Plot shows the relative coverage over various GC% contents across the genome. The bias against the GC-rich regions had previously been reported in similar sequencing approaches (Peters *et al.*, 2012) and is likely due to a bias at WGA step. The coverage values were obtained using 'CollectGcBiasMetrics' of 'Picard Tools' ('Picard Tools', 2018). All boxplots are generated from all of the 384 data points (all wells) in the library.

2.3 Direct application of DigiPico on LCM samples

Since it is nearly impossible to perform DNA extraction on rare populations of cells that can be obtained from Fluorescence-Activated Cell Sorting (FACS) or LCM of frozen tissue sections, to perform DigiPico on such samples direct template denaturation on the crude lysate seems to be the only solution. This is likely to be achievable because DigiPico, as part of its process, requires extreme dilution of the sample which could potentially also dilute out all of the amplification inhibitors that might be present in the tissue lysate. Being able to perform DigiPico library preparation directly on sorted cells and frozen sections can significantly expand the versatility of the DigiPico applications. Therefore, to check the feasibility of this idea, DigiPico library preparation was directly performed on a frozen tumour section.

2.3.1 DigiPico on a frozen tumour section: Library 1111

The first step at attempting to perform DigiPico directly on frozen sections would be to estimate the amount of the input DNA from these sections. This is to ensure that adequate quantities of template DNA is supplied to the WGA reactions. Lower quantities of DNA would reduce the complexity of the data and may even fail at the WGA level. Excess amounts of DNA, on the other hand, would result in co-occurrence of multiple copies of the same locus in the same well which can interfere with the fundamental idea of sub-haploid dilution in LFR sequencing. If the template amount is estimated correctly, the rest of the process can be followed similar to section 2.2.3.

2.3.1.1 Direct template preparation from frozen tumour sections

A 15 µm section from an OCT embedded recurrent tumour tissue (Patient 11152, site PT2R) was directly placed in a microtube containing 100 µl of DNase-, RNase-, and DNA-free PBS. Sample was vortexed to dissociate the nuclei into the solution. 3.5 µl of the suspension was then directly mixed with 3.5 µl of REPLI-g Single Cell denaturation solution (Buffer D2). After 10 minutes incubation at 65 °C, 3.5 µl of *Stop Solution* from the kit was added to the mix. Next, 5 µl of the crude, denatured

DNA solution was used for quantification using Qubit. Based on the results, 5 µl of the crude DNA solution, estimated to contain 300 pg of denatured DNA, was diluted by 155 µl of water and was used as template for WGA in a 384-well plate using Mosquito liquid handler.

2.3.1.2 DigiPico library preparation and sequencing from a frozen tumour section

WGA, CoREF and adapter ligation of the crude template was performed as previously explained in section 2.2.3. Analysis of the end point fluorescent values following WGA, and qPCR results after adapter ligation indicated a fairly consistent WGA and adapter ligation reactions across the plate, comparable to what had been observed in library 1110 which was prepared from purified DNA (Figure 2.14A-B). Analysis of the final library (Library ID: 1111) on TapeStation resulted in a strong, homogenous peak at 465 bp which further confirms the good quality of the library. The library was therefore, sequenced at 30 – 40x on a NextSeq instrument.

2.3.1.3 Preliminary analysis of library 1111

The quality of the raw reads was similar to that of run 1110 with a duplication rate of around 15% and an imbalance of base compositions at the beginning of the reads. The read distributions across the wells was found to be homogenous and consistent (Figure 2.14C). Mapping the demultiplexed reads to human hg19 genome resulted in an average alignment rate of 94.27%, with an average per-well breadth of coverage of 4.6% at an average per well depth of 2.1x over the sequenced regions (Figure 2.14D-F). The distribution of mapped reads across the genome showed an expected bias against the GC-rich regions of genome (Figure 2.14G). These results indicate that high quality DigiPico libraries can also be obtained from crude DNA samples obtained from direct denaturation of frozen sections. The quality control measures of such libraries seem comparable to what can be obtained from purified DNA templates (Figure **2.13**).

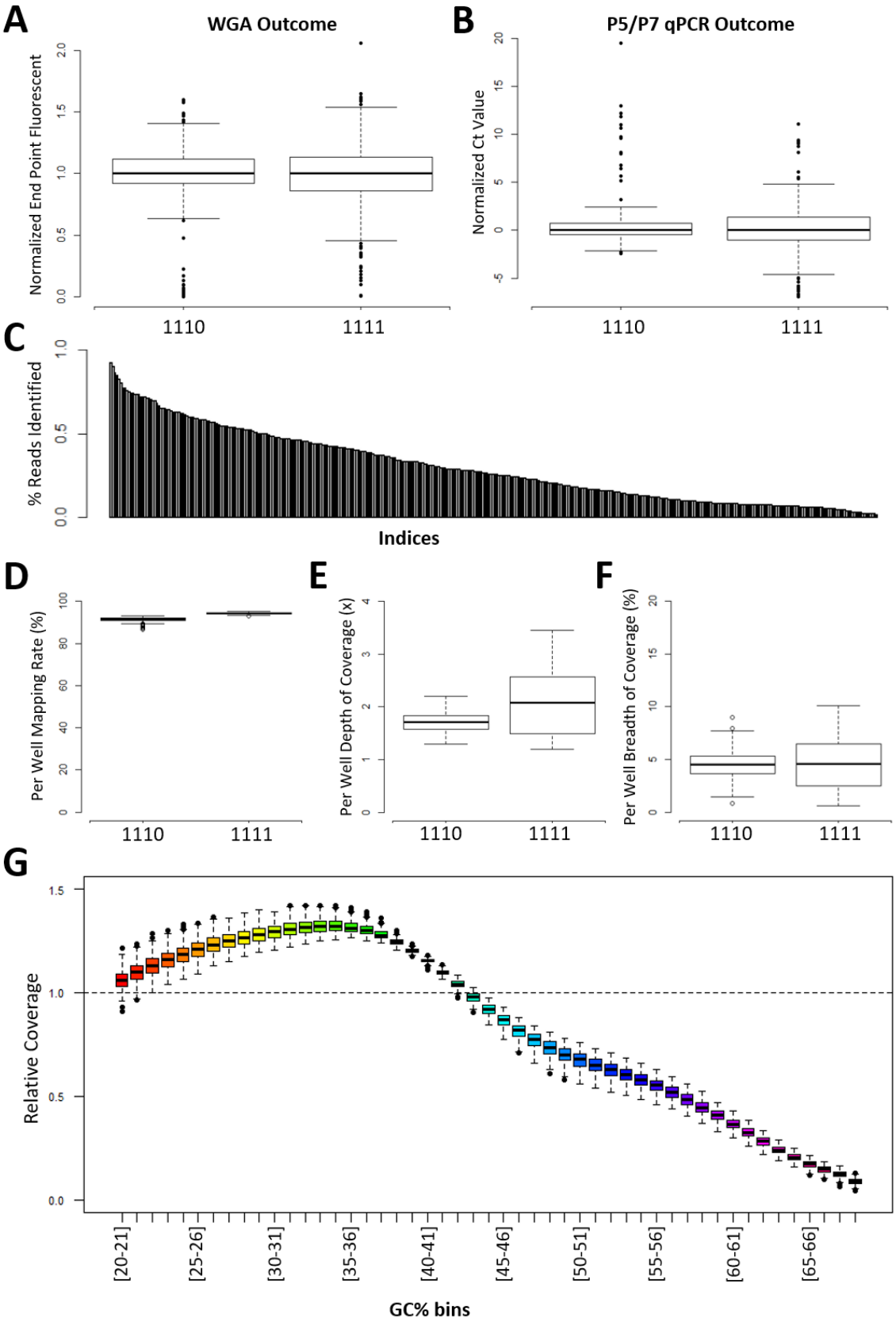


Figure 2.14 (Legend on next page)

Figure 2.14. Quality measures and sequencing statistics of library 1111. (A) Comparison of WGA outcomes between a library prepared from purified DNA (1110) and a library prepared from a frozen section (1111) of the same tumour sample (Patient 11152, Site PT2R). End point EvaGreen fluorescent measurements from each well were normalised by dividing the values by mean in each sample. Results indicate a homogenous amplification across the plates as had previously been shown for library 1111 (Figure 2.12A-B). **(B)** Quantification of adapter ligated fragments in individual wells using qPCR with P5/P7 primers can be used as a reliable quality control measure to ensure consistency and homogeneity across the plate. Ct values were normalised by subtracting the sample's mean Ct value from the Ct value of each well. Tight distribution of normalised Ct values around 0 is an indication of a successful adapter ligation across the plate. **(C)** Consistent WGA and adapter ligation results in the fair distribution of reads among the index combinations (wells). Plot represents the fraction of reads that correspond to each index, ranked from index with the highest frequency to the lowest **(D-F)** Comparison of basic mapping statistics between 1110 and 1111 libraries. Values for library 1110 are those of Figure **2.13**D-F. Similarity of the values and distribution of these parameters between these two libraries indicates that direct application of DigiPico library preparation protocol on unpurified DNA samples from crude lysates does not adversely affect the quality of the libraries. **(G)** The coverage bias against GC-rich regions follows the expected pattern as previously explained.

2.3.2 Quantification of crude DNA from LCM samples

While performing DigiPico directly on frozen tissue sections seem to be feasible, there still exists an obstacle to expand this application to LCM samples. Quantifying the DNA content of crude lysates from frozen sections was possible because of the large availability of material. This, however, will most likely not be the case when dealing with small LCM samples. Considering the worst-case scenario, an LCM sample may contain about 200 pg of DNA. To be able to perform DigiPico on this sample, no more than 10% of the sample must be consumed for DNA quantification. Therefore, to correctly quantify this sample and be left with enough material to perform DigiPico sequencing a quantification method with sensitivity in the range of 10 – 20 pg of input material is required.

Qubit ssDNA assay if adopted to plate readers might be able to cover this range. To check if the quantification assay behaves linearly in this range, a series of serially diluted denatured human genomic DNA samples were used for quantification using this kit. For this test the assay volume was reduced to 20 µl and the spectroscopy was performed in a 384-well assay plate. The results did not show any correlation between the data points indicating that Qubit ssDNA assay is not a suitable option (Figure 2.15A). Among the other commercially available quantification strategies, SpectraMax® Quant™ AccuBlue™ Pico dsDNA assay kit, seemed to be the most suitable option with an advertised detection range of 5 – 3000 pg of dsDNA. To check the linear range for this kit, a similar serial dilution experiment but with re-natured human genomic DNA was performed. The results clearly confirm the potential of this kit for accurate quantification of crude DNA from LCM samples with a minimal material consumption (Figure 2.15B). Considering the sequencing outcome of direct application of DigiPico on frozen tissue sections and the success in quantify 10 – 20 pg of re-natured DNA using AccuBlue™ Pico kit, it is now safe to assume that DigiPico can be extended to be used on LCM samples obtained from frozen tumour sections.

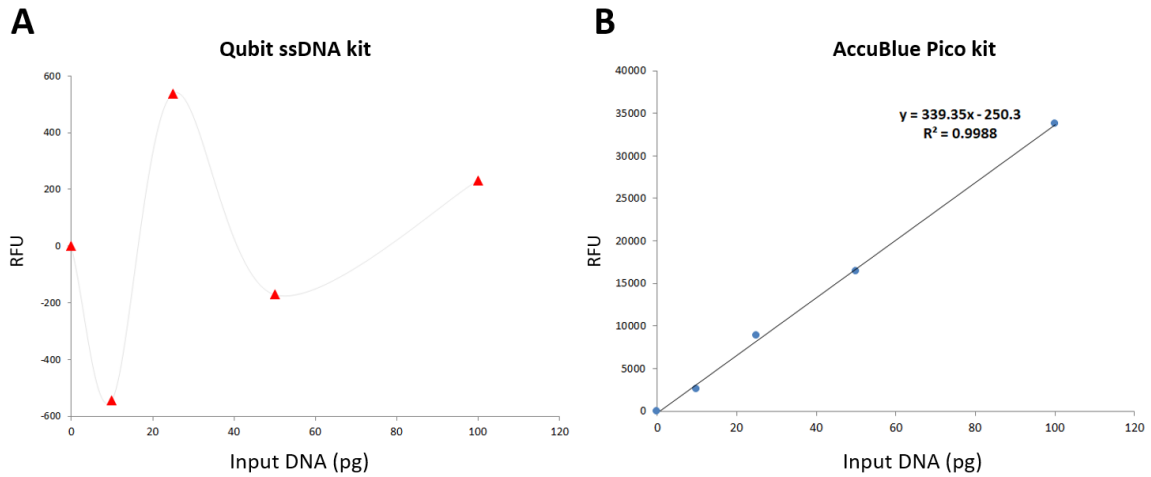


Figure 2.15. Evaluation of quantification assays for detection of pg quantities of DNA from crude lysates. **(A)** To check whether Qubit ssDNA assay kit can be used as a means to quantify 10 – 20 pg of alkaline denatured DNA, first 1 ng of human genomic DNA from HEK293T cells was added to 9 µl of lysis buffer (D2 buffer) from REPLI-g Single Cell WGA kit (The buffer consists mainly of KOH which results in alkaline denaturation of DNA). The 100 pg/µl sample was then serially diluted in the D2 buffer to achieve 50, 25, and 10 pg/µl denatured DNA solutions. 1 µl from each sample was then added to 20 µl of assay solution in a 384-well black assay plate (1 µl D2 buffer was used for baseline reading) and fluorescent spectroscopy was performed (Ex: 490, Em: 527). Results did not show a linear correlation between the fluorescent intensity and the concentration of input DNA in the desired range. Test was performed twice, results are shown from one replicate. **(B)** Results from AccuBlue Pico kit, however, showed a linear correlation between the DNA amount and the fluorescent intensity in the 5 – 100 pg DNA concentration range. For this test 100, 50, 25, and 10 pg/µl denatured DNA dilutions were prepared similar to the experiment with the Qubit assay. However, as AccuBlue dye only binds to dsDNA, 1 µl *Stop Solution* from REPLI-g Single Cell kit was first added to 1 µl of DNA sample and then the resulting 2 µl re-natured DNA solution was added to 50 µl of AccuBlue Pico assay solution in a black 384-well assay plate. Fluorescent spectroscopy (Ex: 468, Em: 507) was performed after 5 minutes incubation at room temperature in dark. Test was performed twice, results are shown from one replicate.

2.4 Increasing the multiplexabilty of DigiPico libraries

With the current dual indexing strategy, 16 i5 indices and 24 i7 indices are combined to uniquely barcode all of the 384 wells of a DigiPico reaction plate. This strategy, although effective, provides no opportunity for combining multiple DigiPico libraries on the same flow cell and also limits the use of DigiPico libraries only to those Illumina platforms that allow for dual-indexing. Therefore, to make DigiPico libraries more versatile a new indexing strategy was developed for DigiPico which expands the pool of potential index combinations to 147,456 unique barcodes. This was achieved by generating a list of 384 unique 8 bp sequences that meet the following criteria:

1) Balanced distribution of bases in each position.
2) Minimum edit distance of 3 between any two pair of indices.
3) Balanced G/C content.
4) No sequences with homopolymers.

To achieve this, first, a list of all possible 8 bp sequences was generated. Then, any sequences with a GC% of less than 50% or above 65% were eliminated. All sequences containing at least one pair of similar neighbouring bases were removed. Lastly, the edit distance matrix of all the remaining sequences was generated and 384 indices that meet all of the aforementioned criteria were selected. Using these 384 index sequences (Appendix II) it theoretically would be possible to uniquely barcode 147,456 individual samples in a dual indexing system. For the purpose of DigiPico, using these indices, it is possible to uniquely barcode every well of a 384-well plate only using index 1, and Index 2 can be used as a plate barcode, allowing for multiplexing of various DigiPico libraries in a single run. Therefore, the list of 384 indices was used to generate 384 unique P7 adapter sequences and the respective 5' phosphorylated oligos were synthesised in a 384-well plate format for use in DigiPico library preparation (Appendix II).

2.5 Further advancements with DigiPico sequencing

2.5.1 Expanding the DigiPico datasets to three patients

Having successfully established the DigiPico library preparation pipeline, to ensure the robustness and reproducibility of the protocol, four more DigiPico libraries from two more patients were prepared and sequenced. Figure 2.16A summarises the details of these libraries and the two previously sequenced libraries. Similar to the previous runs, all libraries generated high quality paired end reads when sequenced on HiSeq X Illumina sequencing platform. All runs passed basic quality control measures and the preliminary analysis of the data showed consistency with previous runs (Figure 2.16). These results further confirm the robustness of the DigiPico library preparation pipeline and proves its reproducibility across various sample types and sources.

2.5.2 Effect of sequencing depth on DigiPico data

To understand the effect of sequencing depth on the behaviour of DigiPico sequencing results, library 1111 was used for another round of sequencing at 30x on a HiSeq 4000 instrument (Run 1112). Combining the data from runs 1111 and 1112 a total depth of 60 – 70x can be achieved for this library. Moreover, to see the effect of lowering the depth, data from 2 lanes or 1 lane out of the 4 lanes of run 1111 were also analysed individually to simulate 20x and 10x sequencing depths, respectively (Figure 2.17A). Plotting the per well depth and breadth of coverage for various sequencing depths of the same DigiPico library indicated that 30 – 40x of coverage is most likely not enough to capture all the complexity of the data in DigiPico libraries (Figure 2.17B-C). Therefore, if an in-depth understanding of genomic data from a sample becomes necessary, increasing the sequencing depth of the library is likely to be beneficial.

A

Library ID	Patient ID	Sample	Template
1110	11152	Tumour (PT2R)	DNA
1111	11152	Tumour (PT2R)	Section
6011	OP1036	Blood	DNA
6311	OP1036	Tumour (RPCG)	Section
E011	11513	Blood	DNA
E111	11513	Tumour (Ascites)	DNA

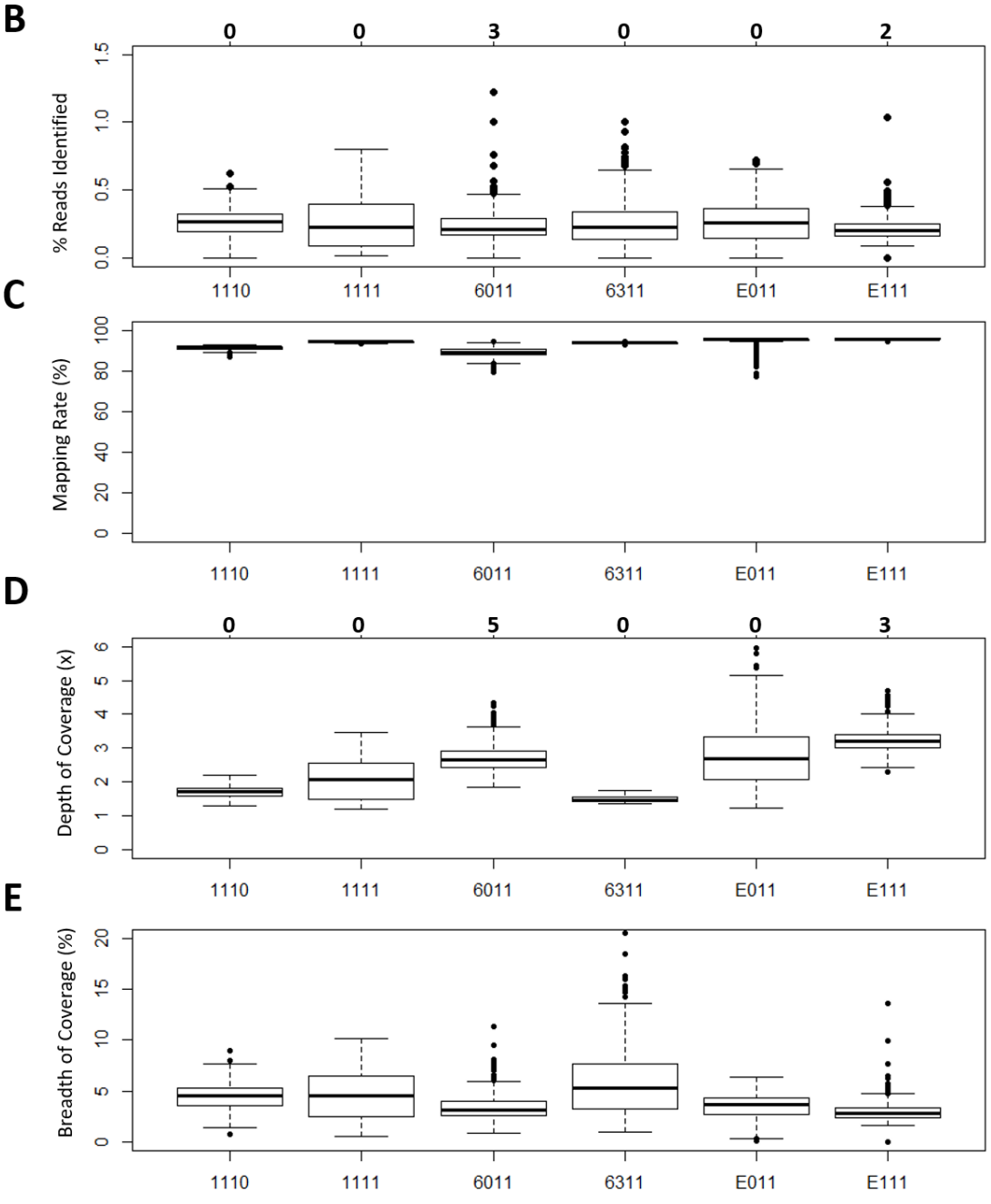


Figure 2.16 (Legend on next page)

Figure 2.16. Comparison of basic sequencing statistics of various DigiPico libraries. **(A)** Besides 1110 and 1111 libraries from patient 11152, four more DigiPico libraries from two other patients were prepared and sequenced to assess the versatility and robustness of the DigiPico protocol. Patients 11152 and 11513 were recruited under GoTarget study group and patient OP1036 was recruited under OXO-PCR-1 clinical trial study. PT2R is a recurrent tumour biopsy from pelvic node. RPCG is a pre-chemotherapy tumour biopsy from Right Paracolic Gutter. For blood and Ascites DigiPico libraries, DNA was first purified from the respective samples and then was diluted to 200 pg/µl to be used for DigiPico sequencing. **(B-E)** Boxplots show the consistency of read distribution, mapping rate, breadth and depth of coverage, within a library and among various libraries confirming the robustness and reproducibility of the DigiPico protocol. Numbers above the plots B and D indicate the number of out of range outlier data points. Each boxplot is generated from all of the 384 data points of the respective libraries.

A

Library ID	Sequencing Platform	Number of Lanes	Total number of Reads	Total Depth over hg19 Genome	Total Breadth of Coverage
1111	NextSeq	1	111,382,175	~10x	84.15%
1111	NextSeq	2	229,210,351	~20x	88.65%
1111	NextSeq	4	463,937,096	~40x	91.08%
1112	HiSeq 4000	-	337,140,877	~30x	90.44%
1111+1112	-	-	801,077,973	~70x	92.32%

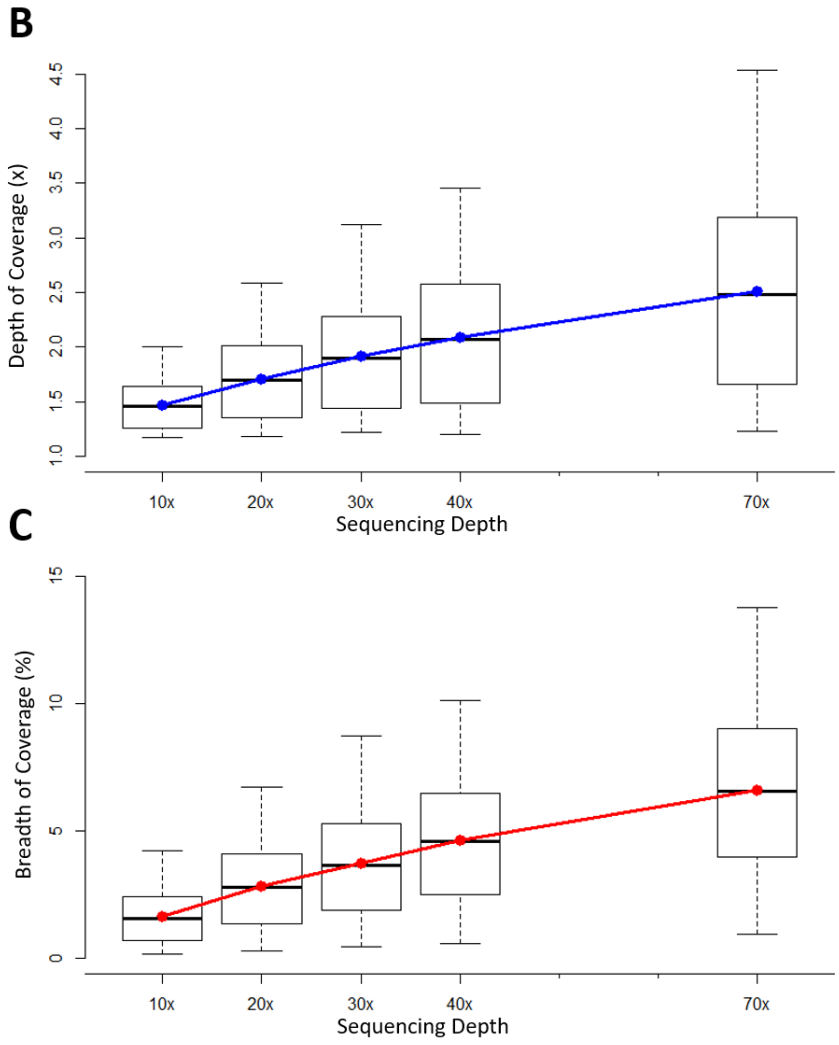


Figure 2.17. Effect of sequencing depth on the per well breadth and depth of coverage in DigiPico. (A) To obtain an estimate of the complexity of the DigiPico sequencing data, sequencing data at various depths were generated by either sub-sampling or combining the available data for DigiPico library 1111. Run 1112 is a second run with the same library as run 1111 to obtain a higher sequencing depth. **(B)** As expected, increasing the sequencing depth results in an increase in the depth of coverage over the sequenced regions of the genome. Boxplots show the per well depth of coverage over the sequenced region of all of the 384 wells for various sequencing depths. **(C)** If the sequencing depth is close to saturation, unlike with the depth of coverage, the breadth of coverage is not expected to increase by an increase in the sequencing depth. However, the per well breadth of coverage of the DigiPico data also seems to increase with the increase in the sequencing depth. This indicates that the sequencing depth is not yet high enough to saturate the library and therefore to capture the full complexity of the library higher sequencing depths might be required.

2.5.3 DigiPicoExome

Further to the previous observation, it seems beneficial to combine the DigiPico protocol with targeted sequencing approaches such as exome sequencing. Targeted sequencing could allow for a substantial increase in the sequencing depth in regions of interest without incurring extra sequencing costs. It can therefore allow to capture the full complexity of the data if required. To evaluate the feasibility and potential of target-enrichment of DigiPico libraries, an exome enrichment experiment was performed on library 1111 (Library 11E1). Analysing the sequencing data shows an on-target coverage of over 90% with over 60% of the target regions being covered by at least 50 reads (Figure 2.18A). However, with a nearly 10-fold enrichment over the target region only about 35% of the reads were found to be mapped to the target region or its soundings (Figure 2.18B). This high rate of off-target contamination is most likely due to the fact that target enrichment strategies heavily rely on perfect blocking of adapter sequences using 'Blocking Oligos' during the capture process. In case of DigiPico however, the use of custom adapter and index sequences could potentially interfere with the blocking efficiency, which would therefore, increase the off-target rate. These values although not ideal when compared to standard whole exome libraries, still are within an acceptable range and could potentially be useful for better understanding of the target regions when required.

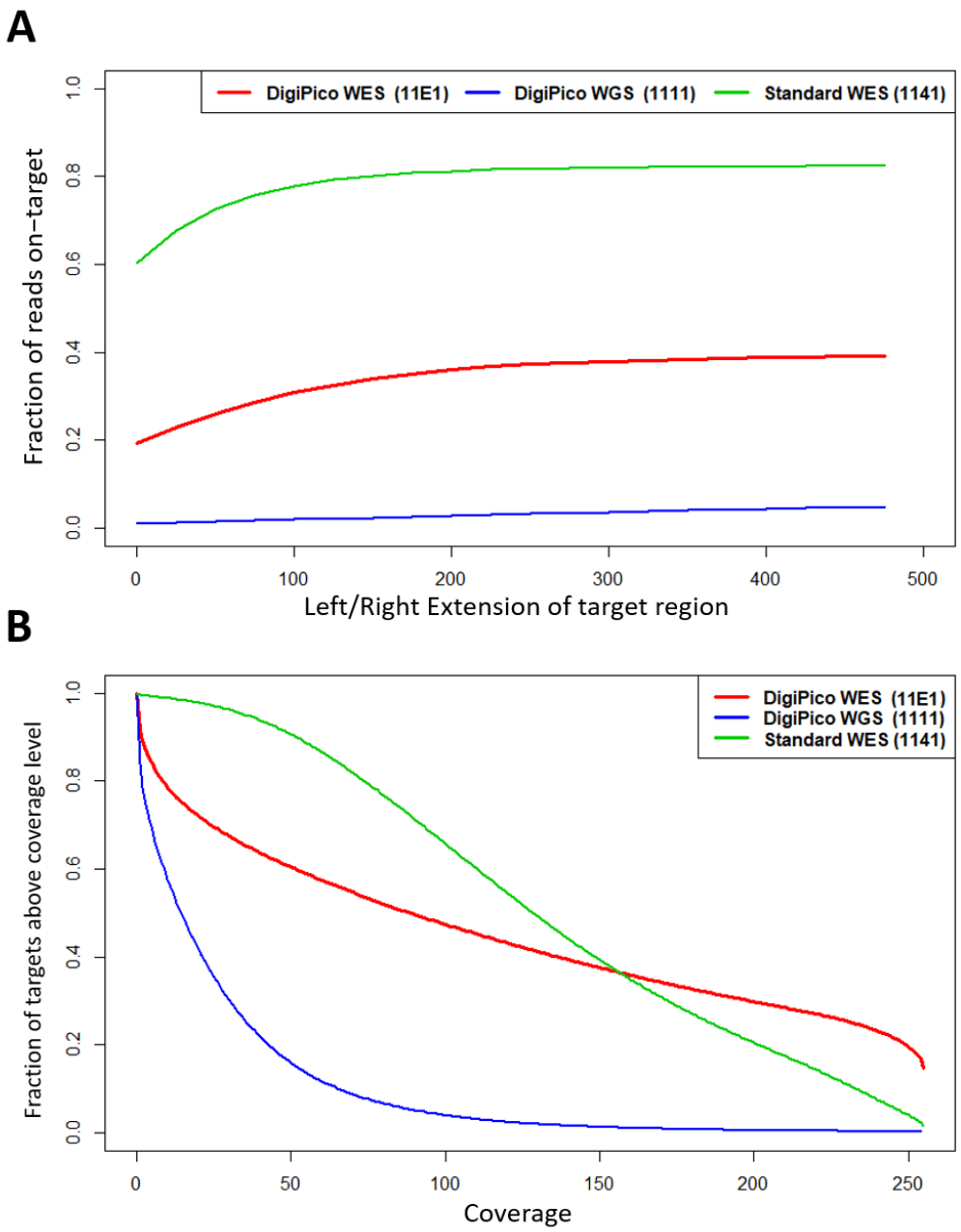


Figure 2.18. Target enrichment efficiency in DigiPico Exome. DigiPico library 1111 was used for target enrichment using SureSelect Human All Exon V7. The resulting DigiPicoExome library (11E1) was then sequenced on a Mid Output 300 cycles NextSeq kit. In case of a perfect capture this sequencing depth should result in an average depth of coverage of 500x over the nearly 50 Mbp target region. **(A)** Comparing the on-target rate of the DigiPicoExome library (11E1) with the standard DigiPico library (1111) shows a near 10-fold enrichment of the target regions. However, comparing the results with a standard WES library from the same tumour sample (Library 1141) prepared with the same target enrichment kit, it is clear that the enrichment efficiency in DigiPicoExome is suboptimal due to off-target contamination. **(B)** Despite the contamination from off-target reads, comparison of depth of coverage over the target regions clearly shows a significant increase in the depth of coverage at regions of interest. In DigiPicoExome (11E1) nearly 45% of the target regions are covered with at least 100 reads. In the standard DigiPico (1111), however, only about 4% of the target regions have a depth of coverage that exceeds 100x.

Chapter 3 : DigiPico Analysis Pipeline

Preliminary analysis of DigiPico data confirms that DigiPico library preparation protocol can be used to robustly and reproducibly obtain high quality sequencing data from picogram quantities of tumour DNA. Next step is to see whether these data can be used for reliable genomics analysis of such samples. As DigiPico datasets have a unique structure compared to standard WGS data it is crucial to first find a set of tools that can optimally process DigiPico data. Standard Illumina WGS is often performed in a 2x150 cycles (paired-end) format with 400,000,000 reads. When mapped to human genome, the result would be a single 'bam' file, containing all the reads with respective coordinates of their potential genomic location, with an average depth of coverage of 30 – 40x over 90 – 95% of the genome. DigiPico data, on the other hand, with the same sequencing data output, results in 384 individual sets of data, each of which can be used to generate an independent 'bam' file with an average coverage of 1 – 3x over 2 – 10% of the human genome (**Figure 2.16**D-E). Because of the initial sub-haploid dilution of the template DNA into individual wells, each of the resulting 'bam' files are expected to carry the information from a single allele of a particular locus. Therefore, by counting the number of the wells that contain a particular allele, it is possible to infer the frequency of that allele in the original DNA sample. This digitised well count information is expected to provide a more accurate representation of original allelic frequencies compared to the standard read count values. This is mainly because compartmentalization of template prior to WGA can make the well count information immune to WGA-associated biases such as unequal amplification and allelic drop-outs. Therefore, to preserve this valuable information in DigiPico datasets it is important to analyse the data from each well independently. Consequently, as there currently are no sets of bioinformatics tools that specifically deal with such a dataset, it is important to first optimise an analysis pipeline for DigiPico data to ensure a robust and reliable analysis.

In general, DigiPico data can potentially be used for two main categories of genomic studies to reveal the evolutionary history of ovarian cancer:

1. **Copy Number Analysis.** As explained previously CNAs are one of the key features of HGSOC tumours and play an essential role in tumour initiation, progression, and evolution (Etemadmoghadam *et al.*, 2009; TCGA, 2011). Therefore, CNA studies are an inseparable part of ovarian cancer genomics studies. Identification of CNAs from limited amount of material is often challenging mainly because nearly all WGA approaches result in unequal amplification and allelic drop-outs. DigiPico, however, is expected to reduce the effect of unequal amplification and allelic drop-outs through compartmentalisation and by digitising the information.

2. **De novo Variant Identification.** The study of tumour progression and evolution often requires the identification of the driver mutations as well as the clonal and sub-clonal mutations. Moreover, identifying the order at which certain mutations have occurred can also allow for better understanding of the tumour initiation and evolution. Lastly, *de novo* identification of clone-specific variants can further reveal the underlying evolutionary processes within a tumour mass. Therefore, accurate and sensitive identification of variants from limited amount of clinical materials can be extremely beneficial for better understanding of the tumour origin, progression, and evolution. However, standard WGA of templated DNA from picogram amounts of DNA, often results in a large number of artefact variants (Dong *et al.*, 2017) which could interfere with the aforementioned evolutionary analysis. DigiPico however, by digitising the variant allele information could allow for selective removal of FP calls, resulting in a more accurate variant calling outcome.

3.1 CNA prediction from DigiPico data

There are several strategies that can be employed for the detection of CNAs from NGS data. These mainly include (1) Paired-End Analysis, (2) Split Read Analysis, (3) Read Depth Analysis, and (4) De Novo Genome Assembly (Zare *et al.*, 2017). Paired-End and Split Read methods rely on paired reads that span the break points. These reads show an irregular pattern when mapped to the genome, and this information can be used to predict structural variants as well as small to medium size deletions and duplications. Read Depth analysis, on the other hand, uses the information from the number of reads that cover a certain area of the genome to infer the CNA status and is suitable for both medium and large deletions and duplication. Lastly, Genome Assembly methods detect the CNAs from mapping de novo assembled contigs to the genome and can be used for detection of structural variants as well as CNAs. While each of these methods have their own specific cons and pros, Read Depth analysis seems to be the most suitable method for the analysis of DigiPico data from ovarian cancer tumours. This is mainly because chimeric DNA molecules that are a by-product of extensive WGA will interfere with all the other methods but will have little to no impact on the read depth information.

Originally Read Depth analysis was employed for the analysis of data from SNP arrays and Comparative Genome Hybridisation (CGH) experiments. Therefore, classically bioinformatic algorithms were designed to perform copy number profiling from such sources of data based on LRR (Log R Ratio) and sometimes also BAF (B Allele Frequency) values. LRR, the normalized intensity of markers for each SNP or location, correlates with the number of DNA copies at the respective genomic position. While higher LRRs indicate potential duplications, lower LRRs often mark potential regions of LOH. BAF, on the other hand, is the relative intensity of the variant allele to reference allele at a certain SNP position. For heterozygous SNPs, in the regions of copy number imbalance, this value will deviate from the expected 1:1 to other ratios such as 1:2 or 1:0 in the areas of duplication or LOH, respectively. With the advancement of NGS technologies, several of

the classic Read Depth analysis algorithms were adopted and new algorithms were developed for the identification of copy number alterations from various sources of NGS input data. These inputs can range from ultra-low coverage WGS data (Scheinin *et al.*, 2014; Kader *et al.*, 2016) to high coverage targeted sequencing data such as WES (Magi *et al.*, 2013). Irrespective of the exact algorithm however, nearly all of these tools follow a similar approach to identify the CNAs. Initially several statistical values, such as BAF, and LRR are inferred from the sequencing data. Next, taking various types of biases (mainly mappability and GC% bias) into account, these parameters will be normalized to remove the effect of any such biases in the data. Finally, the normalized data will be segmented using various models or model combinations and a copy number profile will be generated. While the basic principle of analysis is shared among majority of these algorithms, their input data requirements, segmentation method and the way they handle heterogeneity may vary significantly. As DigiPico data follows a particular structure, it is important to see which of the currently available algorithms is better suited for CNA analysis of DigiPico data. DigiPico data is composed of 384 independent 'bam' files which if needed could be merged together to generate a single file. Although this merged file is the desirable input format for all of the available CNA analysis tools, but it does not carry the digitised information that is the main feature of DigiPico sequencing method. This is mainly because simple merging of the 'bam' files will likely result in an increase in the noise levels because of unequal amplification and uneven distribution of reads between the wells. Hence, where possible, it would be beneficial to see if the digitised well information can also be supplied to the analysis algorithms to potentially improve the CNA prediction quality. Therefore, to find a suitable algorithm for the analysis of DigiPico data, the performance of several commonly used Read Depth based CNA analysis algorithms were evaluated on DigiPico data through comparison of outcome with that of WGS datasets.

3.1.1 QDNASeq

QDNASeq (Scheinin *et al.*, 2014) had been developed for the detection of DNA copy number changes from ultra-low pass WGS data obtained from fresh or formalin-fixed samples. It had recently been used in an extensive analysis of copy number signatures in ovarian cancer from fresh and archival samples (Macintyre *et al.*, 2018). The method solely relies on read count information and does not require a normal sample for the analysis. The analysis pipeline starts by dividing the WGS data into non-overlapping bins followed by normalisation of read count values of each bin for mappability and GC content. The pipeline will then remove the problematic bins from the dataset and the final data will be used for segmentation and copy number profile inference. To check the applicability of this method to DigiPico data, WGS data from patients 11152 and 11513 as well as DigiPico runs 1110, 1111, 111X, and E111 were analysed using the QDNASeq R package. As the algorithm only accepts one 'bam' file for the analysis, it only was possible to use the merged 'bam' file as input for DigiPico data. Comparing the genome-wide QDNASeq copy number profile plots between DigiPico and WGS data clearly shows that QDNASeq is not a suitable algorithm for copy number analysis of DigiPico data (Figure 3.1).

3.1.2 Control-FREEC

Control-FREEC (C-FC) (Boeva *et al.*, 2012) is one of the most commonly used bioinformatic tools for the detection of CNAs from NGS data. It can perform CNA analysis on both WGS and WES data and does not necessarily require a paired normal sample in WGS mode. It automatically performs all the analysis steps directly on the alignment files and returns the predicted copy number and BAF profiles as output. This offers a user-friendly platform for the analysis of NGS data but the streamlined process will limit the applicability of the tool to DigiPico data as it will not be possible to input the digitised well count information to this algorithm. However, because of the versatility that this method offers, its performance on DigiPico data was evaluated.

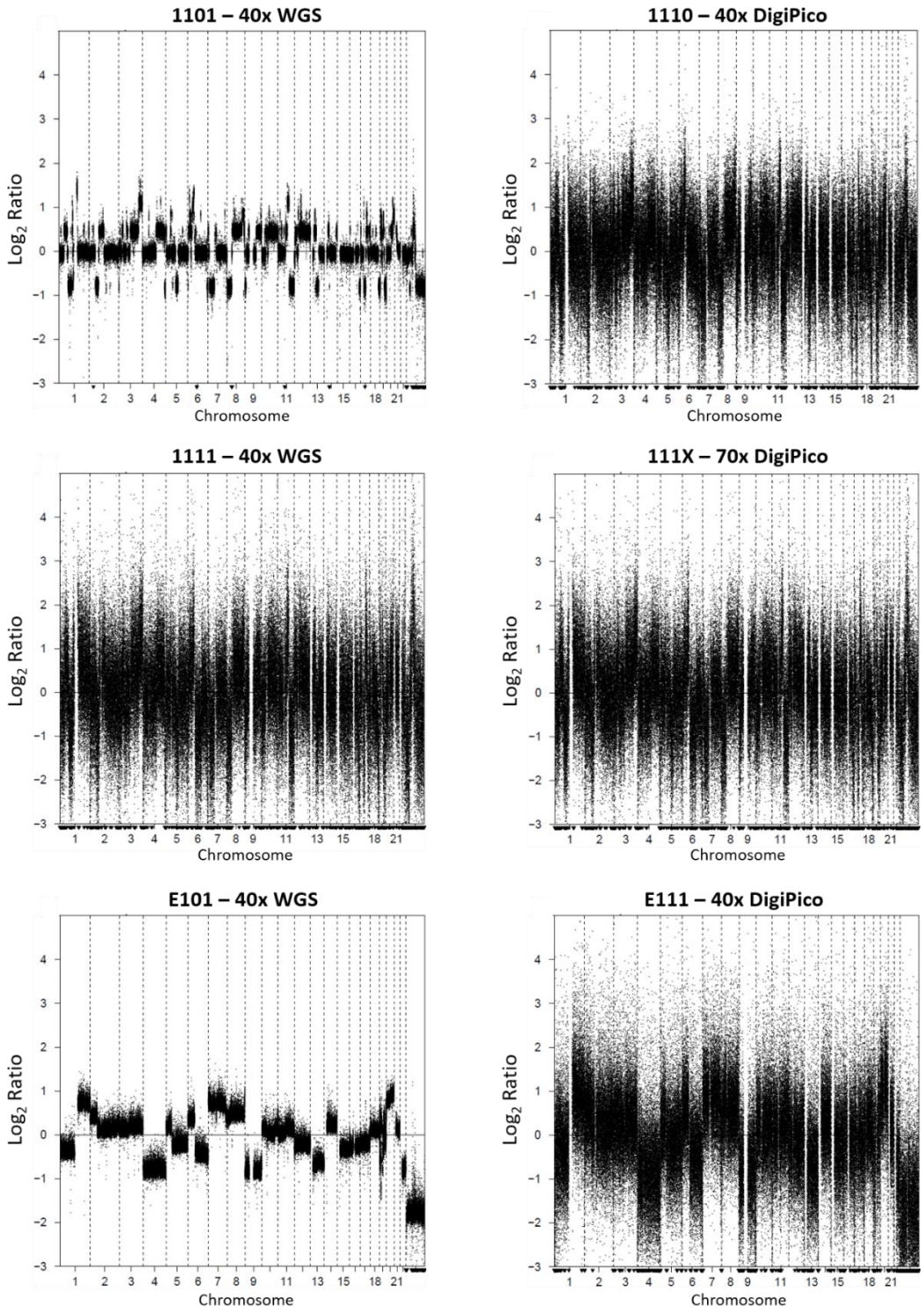


Figure 3.1. Results of QDNASeq analysis of various DigiPico runs. QDNASeq R package was applied to 40x WGS results from patients 11152 and 11513 (runs 1101 and E101 respectively). Results clearly show a high-quality estimation of copy numbers based on read count data. Application of this algorithm to respective DigiPico data of various depths did not result in any conclusive outcomes indicating that QDNASeq algorithm is likely not suitable for the analysis of DigiPico data.

3.1.2.1 Application of Control-FREEC to WGS data

Firstly, to obtain a prior knowledge of the copy number profiles of the samples under study, C-FC was applied to standard WGS data from tumour/normal pairs 1001/1101 and E001/E101 from patients 11152 and 11513, respectively. The genome-wide copy number profiles show an extensive copy number alteration across the genome typical to HGSOC tumours (Figure 3.2Figure 3.4). While the Ascites from patient 11513 seems to carry a large number of deletions across the genome (blue), the recurrent tumour from patient 11152 shows a high level of duplications (red). Moreover, overlaying the BAF plots (Figure 3.3Figure 3.5) with copy number profiles, clearly shows the presence of a high number of Copy-Neutral LOH (CN-LOH) events. These regions can be identified as areas with a predicted ploidy of greater than one but with BAF values clustered around 0 and 1. However, although the result of C-FC analysis of WGS data from both patients seem conclusive, closer inspection of the result shows some inconsistencies with the segmentation algorithm. For instance, despite a clear accumulation of BAFs around the middle line of chromosome 6q in patient 11513 (Figure 3.5 - chr6), the algorithm has predicted an LOH in that region which does not match with the observed BAF values. These inconsistencies are most likely a result of large number of CNA events in HGSOC genomes which alter the average ploidy of the cell and therefore, will affect the normalisation factors used in CNA analysis which, consequently, leads to wrong assumptions and predictions by the model.

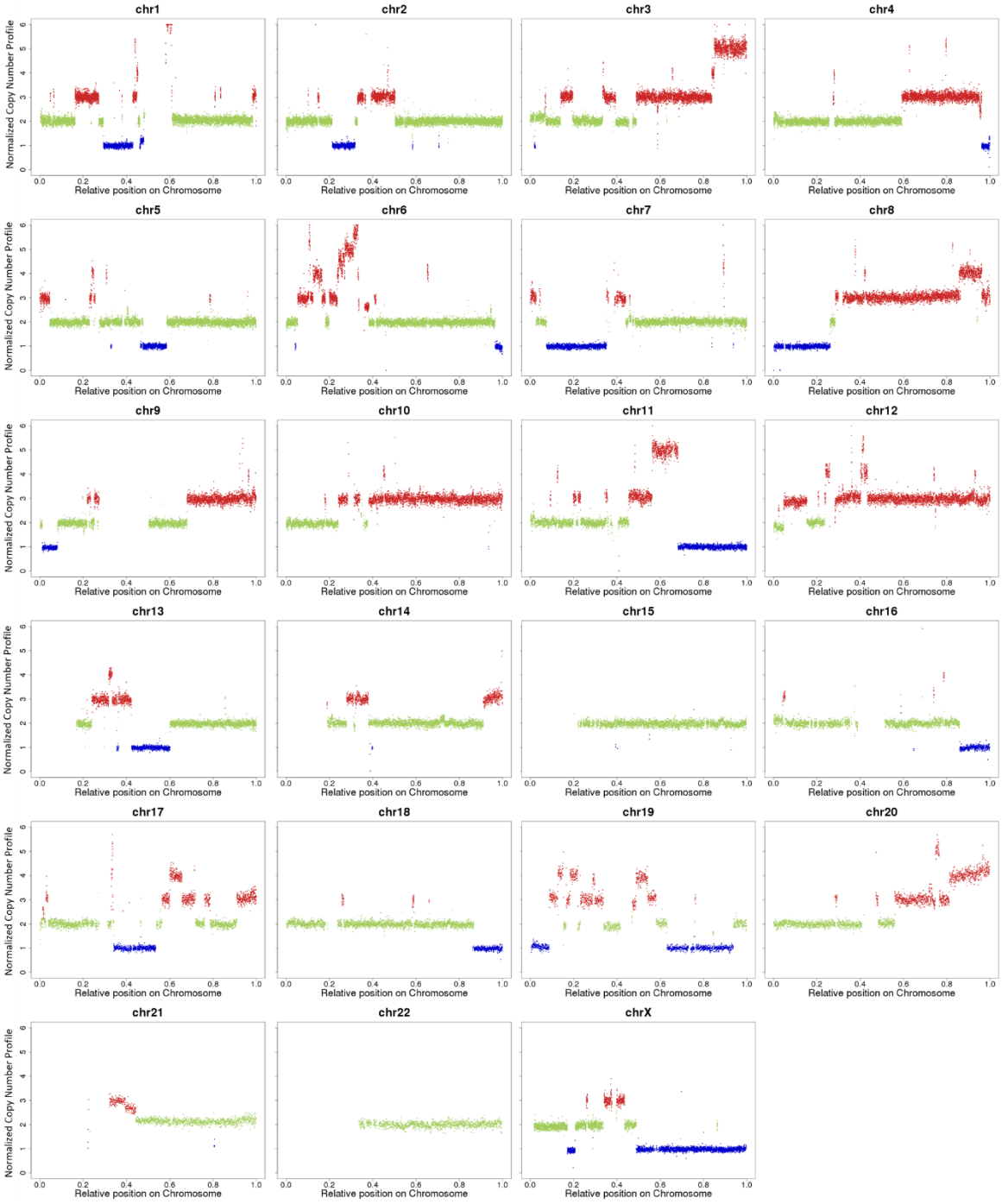


Figure 3.2. Copy number profile estimate of standard WGS run 1101 using Control-FREEC. Control-FREEC algorithm was applied to 40x WGS data from run 1101. Results show a significant amount of small and large duplications across the genome. Each plotting area shows the copy number estimation for regions across the respective chromosomes. Regions labelled as red indicate areas of duplication, blue marks areas of deletion, and green shows areas with neutral copy number.

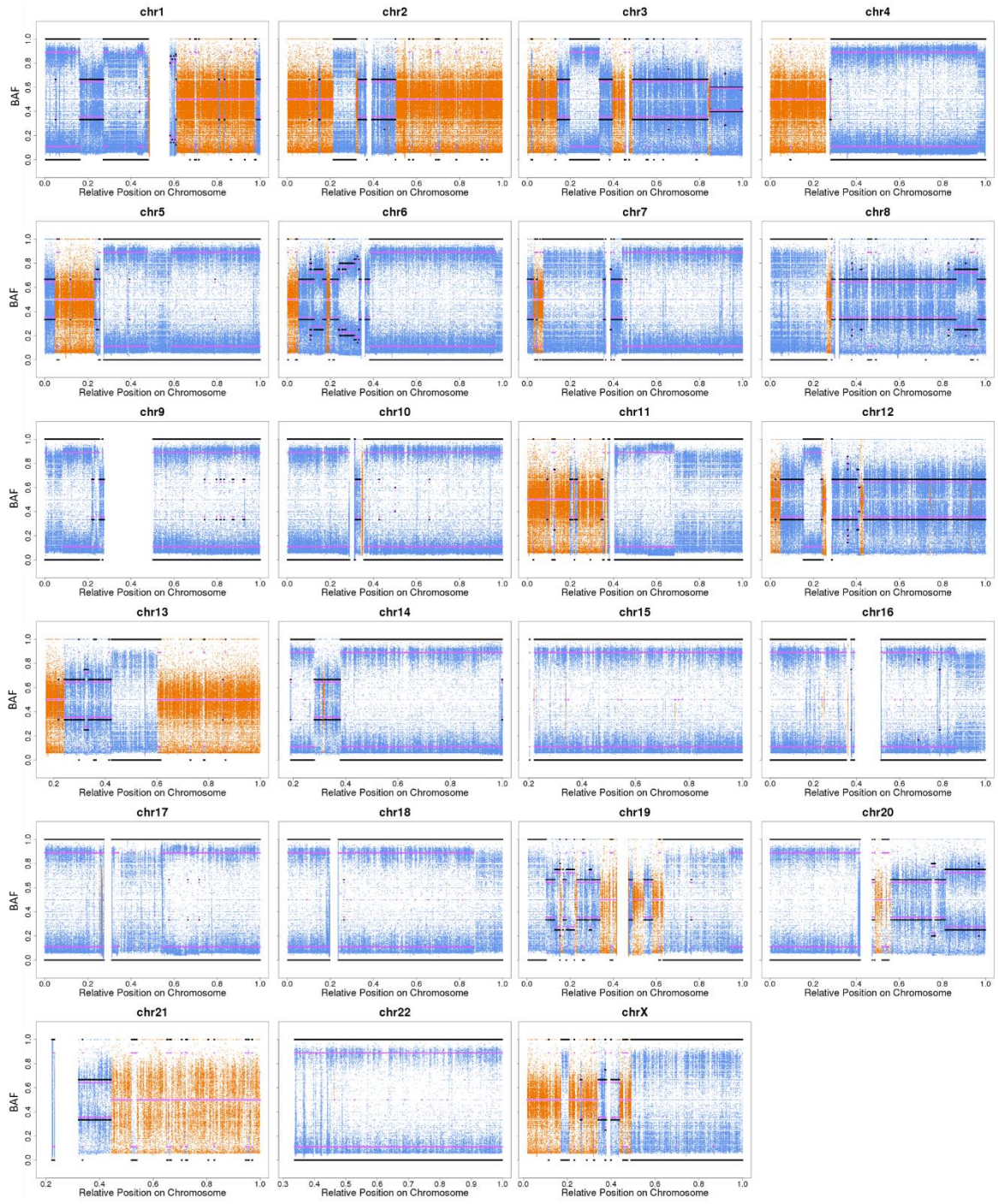


Figure 3.3. BAF plots of standard WGS run 1101 from Control-FREEC. Each panel shows the allele frequencies at heterozygous SNP positions for respective copy number estimation panels in **Figure 3.2**. Areas where allele frequencies deviate from the middle line represent potential areas of duplication or deletion. The exact nature of the change can be inferred by combining the BAF information with the copy number values. For instance, areas with a copy number estimation of 2 and BAF values of 0 and 1 represent regions with copy number neutral LOH events (e.g. chr15 and chr16).

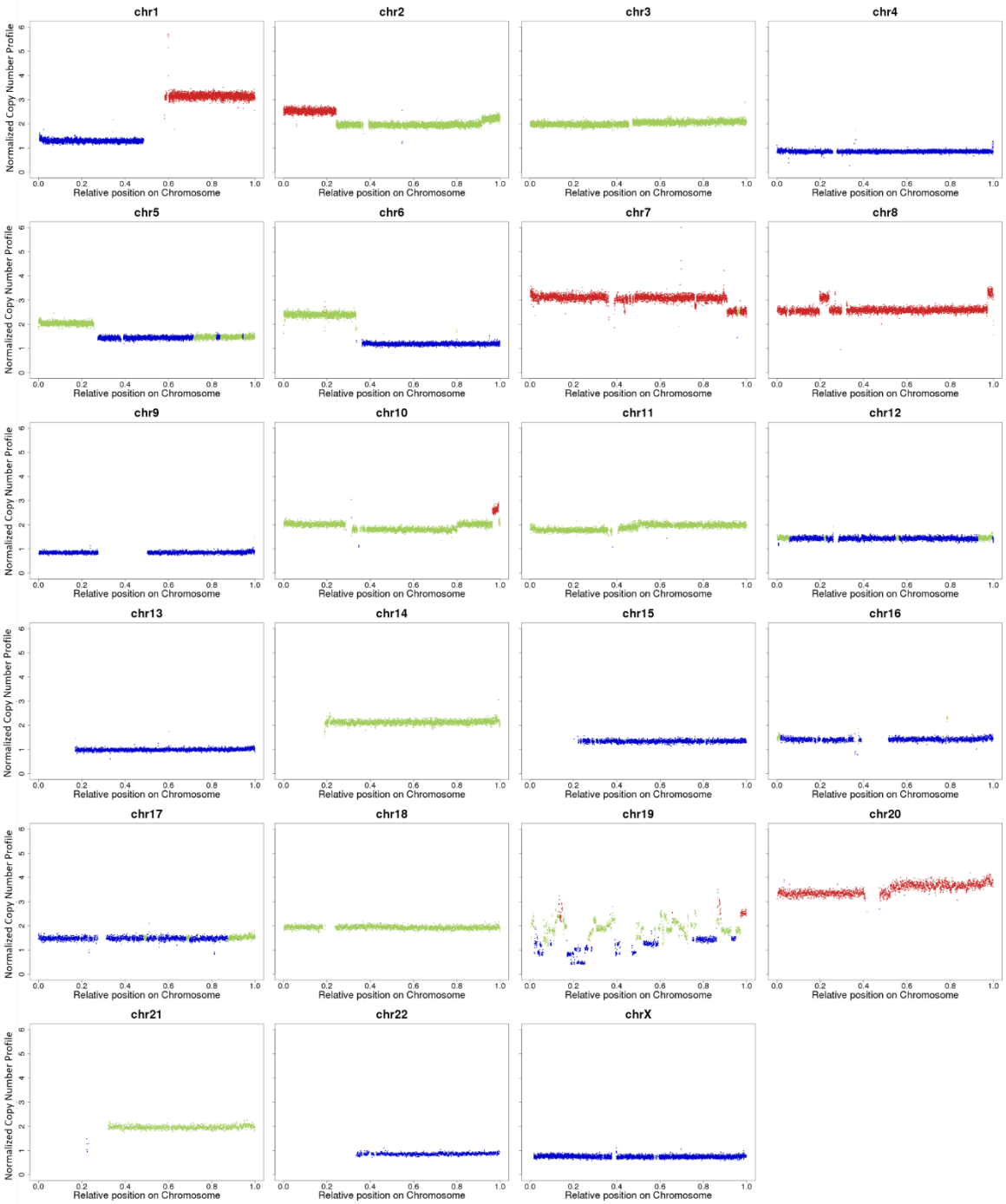


Figure 3.4. Copy number profile estimate of standard WGS run E101 using Control-FREEC. Similar plots to that of **Figure 3.2** but from run E101 of patient 11513. Except for chr19 all the other regions seem to only contain large CNA events as opposed to what had been observed in run 1101 from patient 11152 (Figure 3.2). This is likely because run 1101 is from a recurrence tumour while run E101 is from pre-chemotherapy ascites samples.

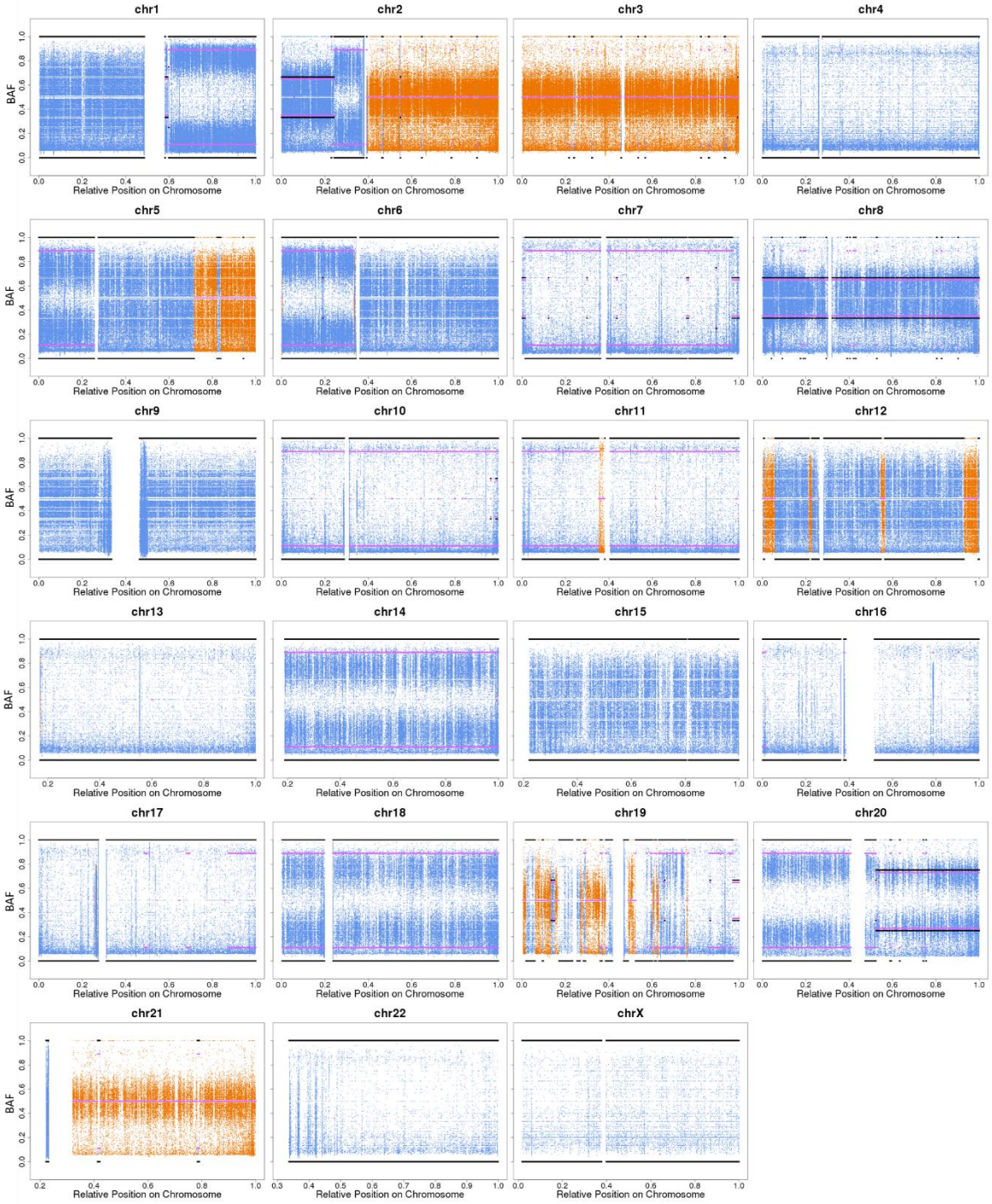


Figure 3.5. BAF plots of standard WGS run E101 from Control-FREEC. Similar plots to that of **Figure 3.3** but from run E101 of patient 11513. Although the results seem to have a good quality, there seem to be some inconsistencies between the copy number prediction and BAF values in this sample. For instance, in the q arm of chr6 there seem to be a balanced allelic distribution, indicating a copy number neutral area. However, Control-FREEC predicts a haploid state for this region. Such discrepancies could be a result of wrong estimation of sample's overall copy number because of the extensive copy number alteration events.

3.1.2.2 Application of Control-FREEC to DigiPico data

The outcome of previous analysis shows that C-FC may not be a suitable algorithm for the CNA analysis of NGS data from HGSOC tumours. However, to better understand how DigiPico data may affect the outcome of CNA analysis, C-FC was used for the copy number analysis of various DigiPico datasets from both patients. Comparing the copy number profiles of DigiPico runs 1110 and 1111 with that of 1101 show that while in some regions there are some discrepancies between the analysis outcome of DigiPico data and standard WGS data, C-FC can still achieve a close estimation of the actual copy number profile from the DigiPico data (Figure 3.6). Moreover, comparison of copy number profiles between runs E102 and E111 shows that at least in this case, DigiPico data slightly outperform the data that are obtained from WGS of standard WGA DNA (Figure 3.7). Interestingly, it appears that the miss-prediction at chromosome 6q of patient 11513 gets resolved in DigiPico data (Figure 3.7), further confirming that C-FC is likely to behave inconsistent when analysing data with large number of CNAs such as data from HGSOC. These observations cumulatively suggest that C-FC is not a suitable algorithm for the analysis of DigiPico data obtained from HGSOC.

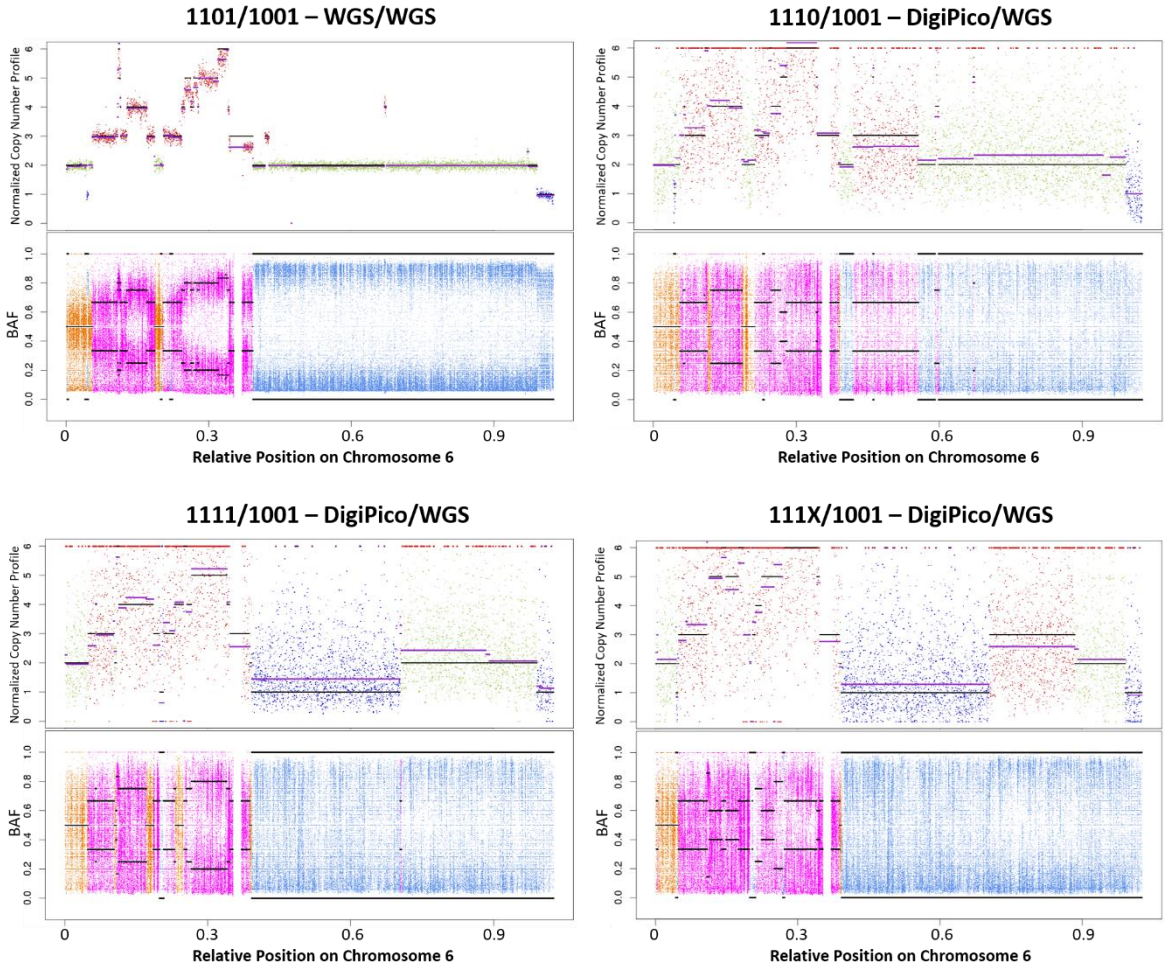


Figure 3.6. Copy number Profile estimation from merged DigiPico runs of patient 11152 using Control-FREEC. Control-FREEC algorithm was applied to merged DigiPico data of runs 1110, 1111 and 111X. For all of the analyses standard WGS data from the blood of the same patient was used as the 'normal sample' control. For simplicity only results from chromosome 6 are presented. Although the results from DigiPico data follow a similar trend to that of standard WGS data (top left panel) there is a significant amount of noise in the outcome that could interfere with any further comparative analysis.

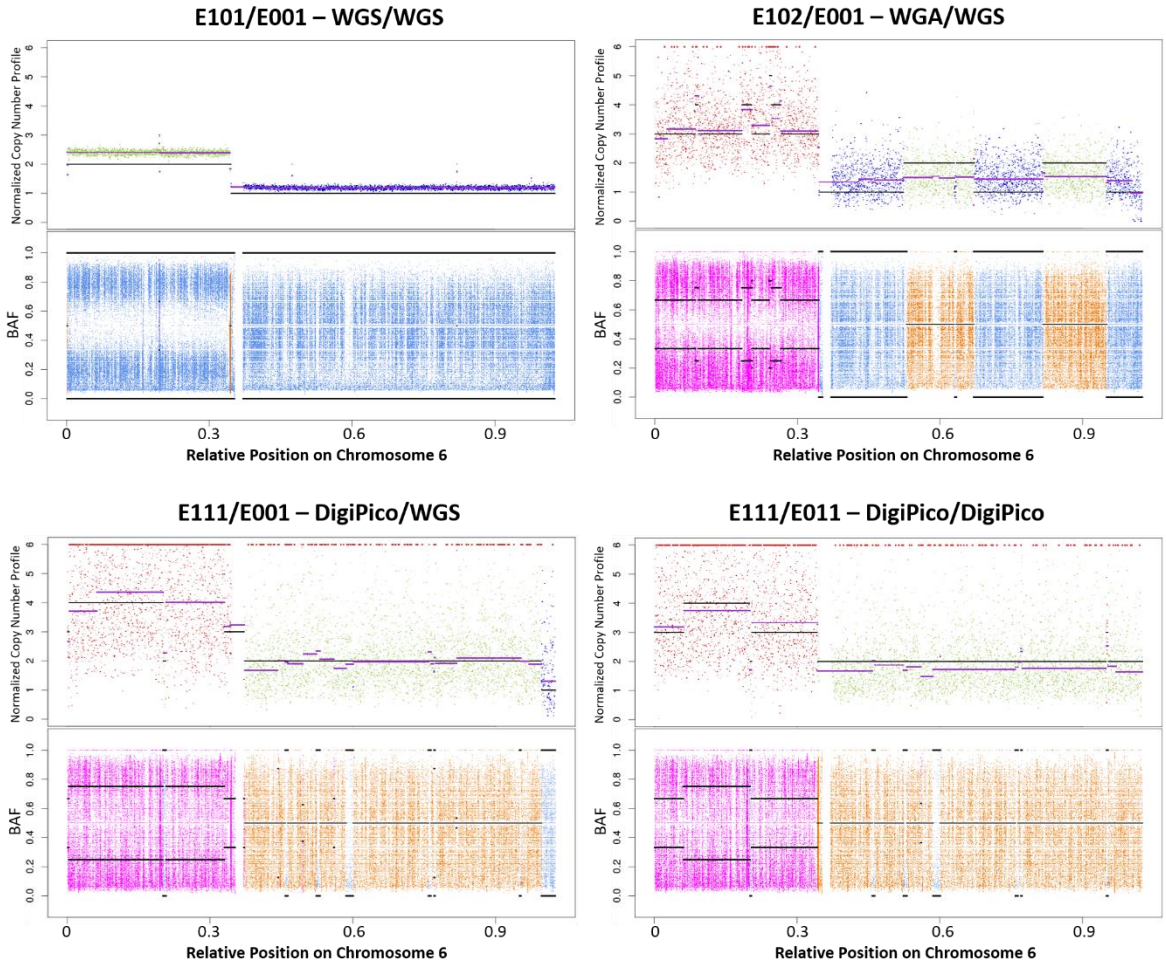


Figure 3.7. Copy number Profile estimation from WGA DNA and merged DigiPico runs of patient 11513 using Control-FREEC. Similar analysis as **Figure 3.6** conducted on standard WGA WGS data (top right panel) as well as DigiPico run E111 (bottom panels) from patient 11513. For the analysis of DigiPico run E111, both standard WGS data from blood (bottom left panel) and merged DigiPico data from blood (bottom right panel) were used. Although DigiPico data seem to have performed slightly better than standard WGA WGS in the copy number estimation part of the analysis, however, it has resulted in a significantly noisier BAF plot. This hampers any further downstream analysis as no accurate estimation of copy number changes can be made from these results.

3.1.3 TitanCNA

TitanCNA is a powerful analysis pipeline designed for prediction of copy number changes in clonal tumour cell population from NGS data. This method has previously been used for the analysis of ovarian cancer tumours and therefore, it is more likely to be compatible with genomes with complex copy number profile (Boeva *et al.*, 2012; Ha, 2014). As part of its pipeline, TitanCNA relies both on the *Titan* algorithm and various other algorithms for read depth analysis, genomic segmentation, and estimation of tumour cell fraction. The analysis pipeline starts from a pair of matched normal-tumour 'bam' files, and (1) identifies heterozygous SNP positions from the normal sample, (2) calculates the BAF for all of the identified heterozygous SNP positions, (3) Obtains the read depth and normalizes the values using *ichorCNA* (Adalsteinsson *et al.*, 2017), (4) Uses *Titan* algorithm to construct copy number profiles, to perform segmentation, and for calculation of the normal cell contamination, and lastly, (6) estimates model parameters based on the read counts and normalized read depth values for inference of sub-clonal CNAs and LOHs. This modular structure of TitanCNA workflow offers the possibility to manipulate the input structure to some extent which could make it compatible with DigiPico data. Therefore, to evaluate the performance of TitanCNA on DigiPico data, it was first applied to standard WGS data and the outputs were compared with that of DigiPico data.

3.1.3.1 Application of TitanCNA to merged DigiPico data

Figure 3.8 and Figure 3.9 show the comparison of results obtained from application of standard TitanCNA workflow on WGS and merged DigiPico data from patients 11152 and 11513, respectively. The results from the analysis of WGS data show that, as expected, unlike C-FC, TitanCNA pipeline results in a more consistent segmentation outcome when applied to genomes with complex copy number profile. However, from the plots it is clear that both *Titan* and *ichorCNA* algorithms are unable to clearly identify any patterns from the analysis of merged DigiPico data (Figure 3.8, Figure 3.9).

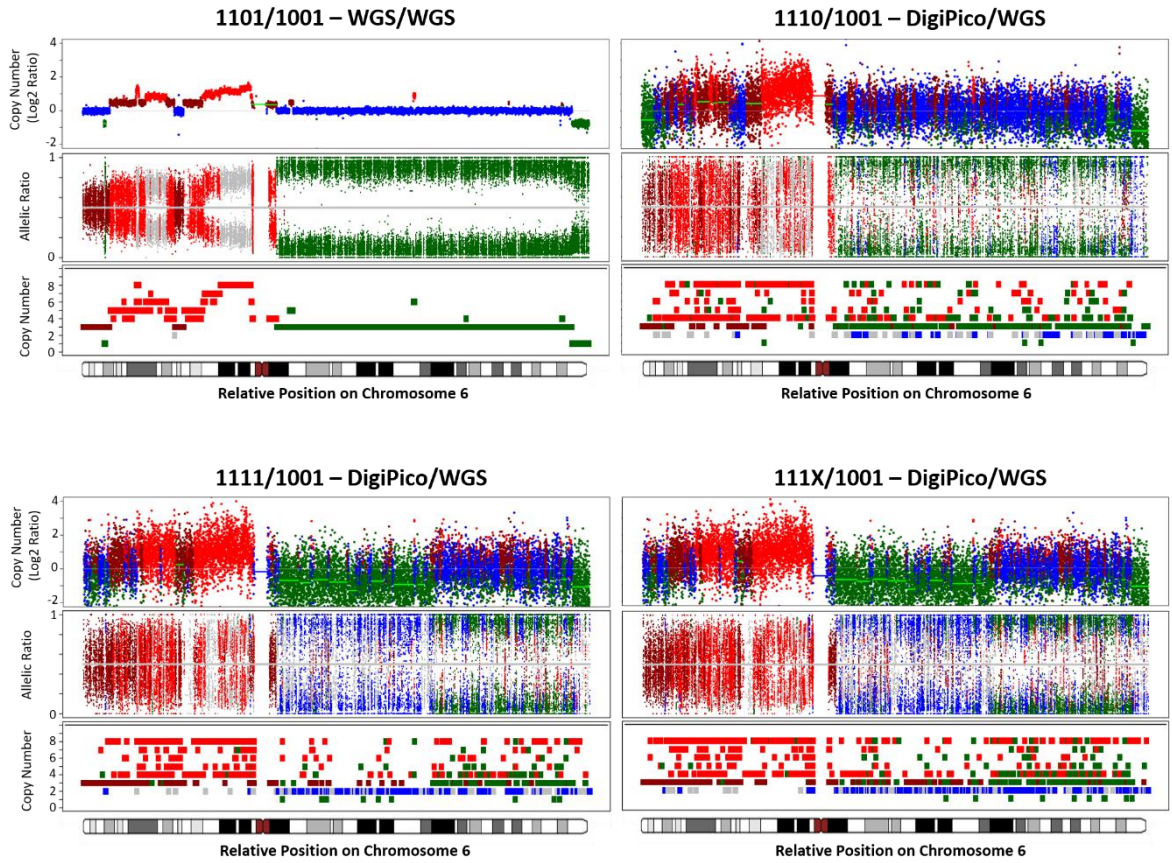


Figure 3.8. Copy number profile estimation from WGS and merged DigiPico data of patient 11152 using TitanCNA. TitanCNA algorithm was applied to 40x standard WGS data (run 1101) as well as merged DigiPico data (runs 1110, 1111, and 111X) from recurrence tumour sample of patient 11152. For all of the analyses, standard WGS data from blood of the same patient was used as 'normal sample'. It is clear that although very accurate estimations of copy number changes can be made from standard WGS data, analysis of merged DigiPico data result in an inconclusive, and noisy prediction outcome.

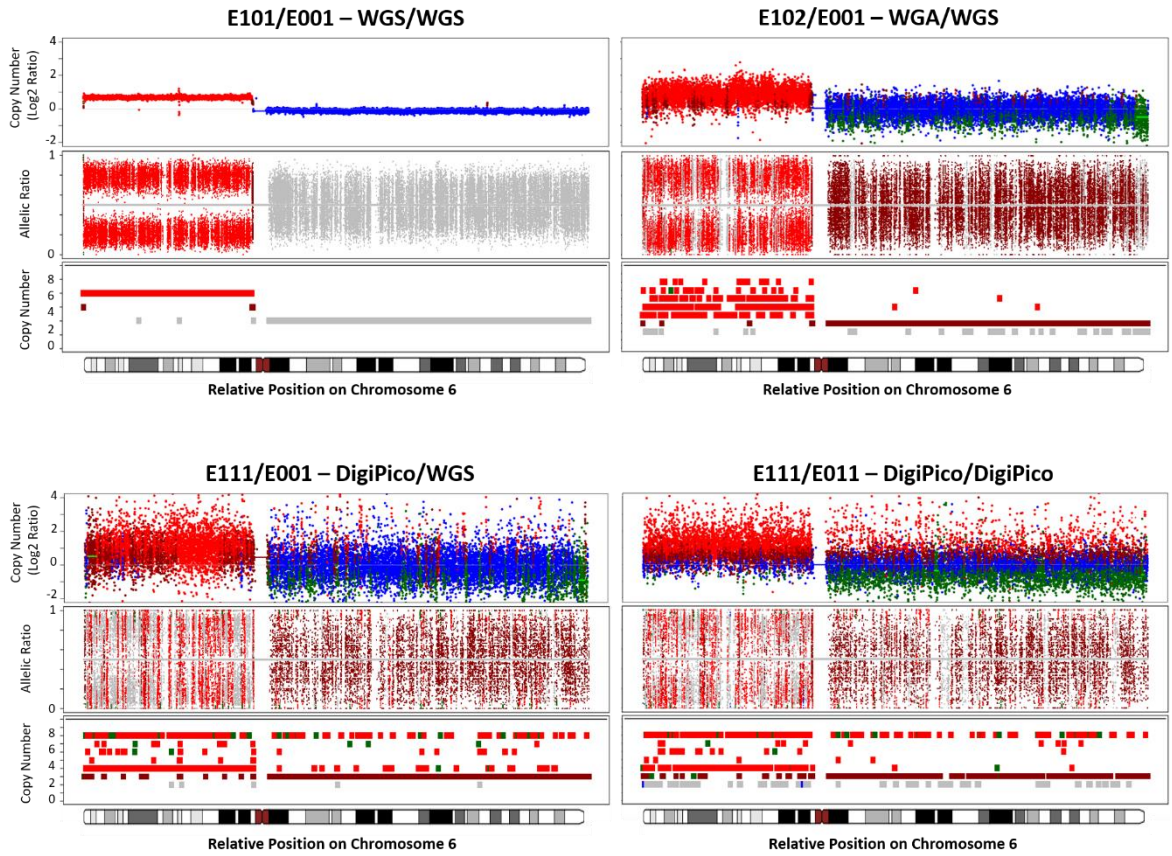


Figure 3.9. Copy number profile estimation from WGS, standard WGA WGS, and merged DigiPico data of patient 11513 using TitanCNA. Similar analysis as in **Figure 3.8** but on sequencing data from ascites sample of patient 11513. Similar to Figure 3.8 TitanCNA algorithm seems to have resulted in reliable estimation of copy number changes from standard WGS data. It also appears that data from standard WGS of WGA template (top right) also results in a reasonable prediction outcome. Merged DigiPico data (bottom panels) however do not seem to result into a conclusive prediction of copy number changes. However, analysing merged DigiPico data from run E111 with merged blood DigiPico data as normal sample, seems to result in a slightly less noisy copy number profile in comparison with when standard WGS data is used as normal sample. This slight improvement, however, does not appear to be sufficient to resolve the inconclusive copy number change prediction results.

Moreover, it seems that analysis of data from WGS of standard WGA DNA from the tumour (run E102) with *ichorCNA* results in a relatively more consistent copy number profile than that of DigiPico data (Figure 3.9). These results clearly show that although TitanCNA seems to be a suitable algorithm for the CNA analysis of NGS data from HGSOC tumour samples, in its standard form, it is not useful for the analysis of DigiPico data.

3.1.3.2 Customisation of TitanCNA pipeline for DigiPico data

As standard TitanCNA workflow only accepts one 'bam' file for each of the tumour and normal samples, it is not directly possible to input the digitised data from DigiPico to TitanCNA pipeline. However, step-wise monitoring of the workflow revealed two points at which digitised DigiPico data can be inputted to the algorithm. Firstly, it seems that after the identification of heterozygous SNP positions from the normal sample, TitanCNA generates a read count file for all of these SNP positions from the tumour 'bam' file. This read count file will later on be used for the calculation of the BAF values for the tumour sample. Therefore, intercepting the pipeline's workflow to replace these read count values with respective well count values from the original DigiPico dataset would partially provide the algorithm with the better-quality digitised data. Secondly, the pipeline calculates the number of reads in non-overlapping 10 Kb windows for both normal and tumour samples. These depth values will then be used by *ichorCNA* algorithm for mappability and GC% normalisation. To supply the pipeline with digitised depth values, the DigiPico data from individual wells were first binned into non-overlapping windows. Each bin from each well was then valued 1 or 0 based on the absence or presence of reads in that specific bin. Next, the values of any specific bin from all of the 384 wells were summed together to obtain the digitised depth values. These digitised well count values were then supplied to *ichorCNA* for the calculation of normalised depth values which will later be used by *Titan* to estimate the copy number profiles. Combining these strategies together, a new copy number analysis pipeline based on *ichorCNA* and *Titan* algorithms, termed DigiTitan, was developed.

Comparing the results of standard TitanCNA workflow (Figure 3.10B) with the results of a preliminary DigiTitan workflow (Figure 3.10C), clearly shows a significant reduction in the noise level of the copy number profile. It however, does not seem to enhance the BAF profile in a noticeable level. This is likely due to the fact that on average only about 10-15 wells may contain information about a specific SNP position. In such a low depth BAF values are expected to appear noisier as any small change in the count values can have a larger effect on the fraction values. One noise reduction strategy in such situations is to bin the BAF values in non-overlapping windows and average the BAF values in each bin. The average bin values are expected to be significantly less noisy as random fluctuations in the data will cancel each other out throughout the averaging process. One flaws of this strategy however, is that in areas where multiple sub-groups of values are present averaging across all of the values will artificially merge the sub-groups, eliminating valuable information from the data. In the case of BAF values, in regions of LOH for instance, averaging the allele frequencies will mostly result in the mid value of 0.5 for majority of the bins. This issue however, can be overcome if the prior knowledge for the sub-groups is available so that for each bin, averaging only takes place amongst the values within the same sub-group. In the case of BAF values, the sub-group information can be obtained from the phase information of the genome. This is ideal for DigiPico data as sub-haploid compartmentalisation of input prior to WGA and library preparation results in preservation of phase information in DigiPico data. This information can be extracted by calculating the number of times that two particular nearby SNPs co-occur in the same compartment (Peters *et al.*, 2012). The more the number of the co-occurrence events, the more likely it is that the two SNPs are within the same group. Using this logic, a simple short-phase estimation algorithm was developed for DigiPico data and the phase group information of every SNP within a range of 5-50Kb was obtained. Using this information, a binning-and-averaging noise reduction algorithm was then applied to the digitised well count information and the denoised data was inputted to TitanCNA for copy number profiling (Figure 3.10D).

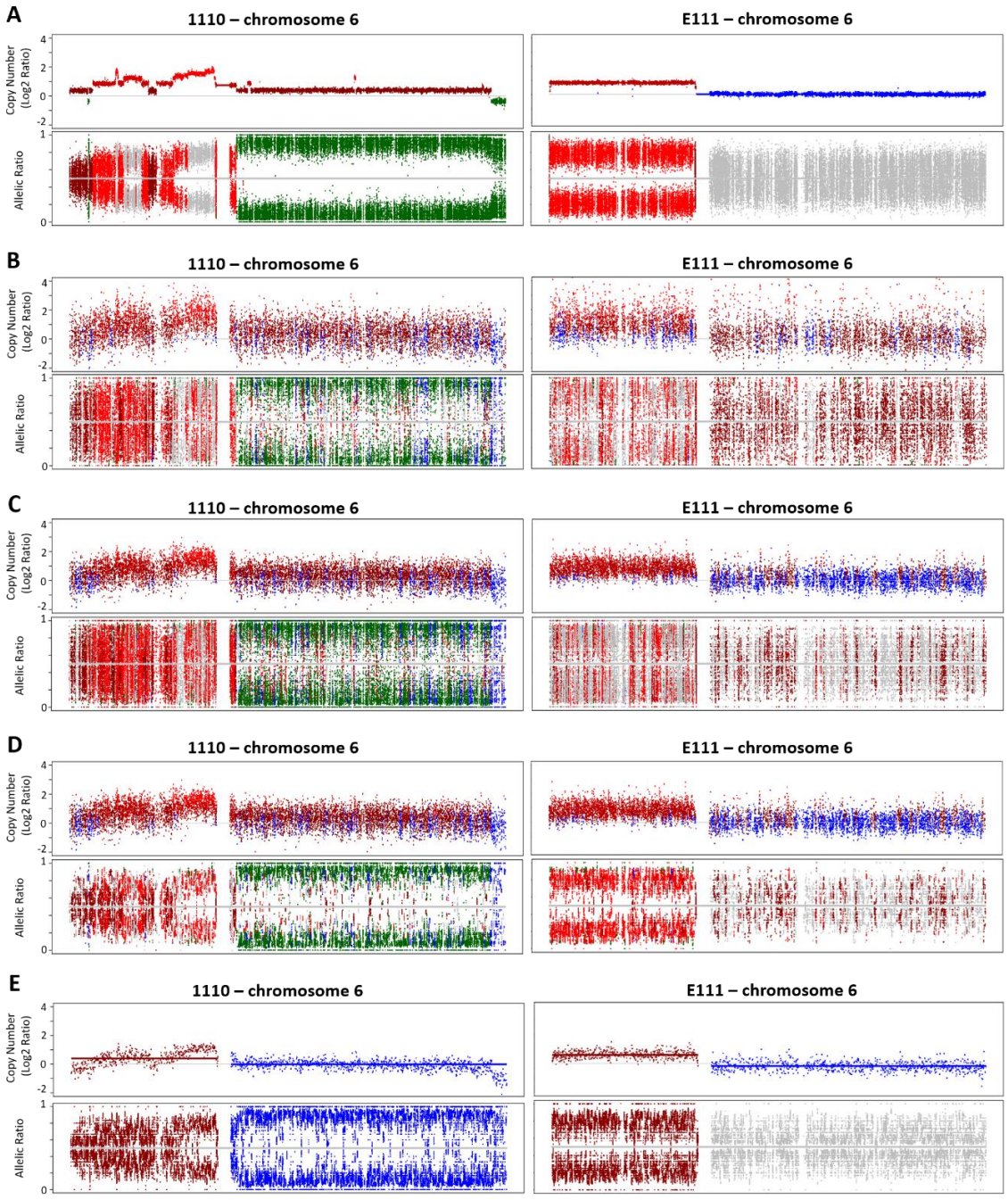


Figure 3.10. Effect of supplying digitised and phased DigiPico data on the accuracy of copy number estimation using Titan algorithm. (A) Copy number and BAF profiles of standard WGS run 1101 and E101 estimated by TitanCNA workflow. **(B)** Copy number and BAF profiles of merged DigiPico data from runs 1110 and E111 estimated by standard TitanCNA workflow. The significant amount of noise in both copy number profile and the BAF profile will not allow for any further analysis on these data. **(C)** Digitising the input allele frequency and read count values for DigiPico data significantly improves the quality of copy number profile estimation by *ichorCNA*. **(D)** Using the phase information extracted directly from DigiPico data to average the phased allele values in non-overlapping bins remarkably improves the BAF profiles that are obtained from digitised DigiPico data. The improvement is so significant that even subtle structures in the data have now become visible. **(E)** Lastly, the final DigiTitan workflow, uses a larger window size for the binning of read count data in *ichorCNA* (100 Kb instead of 10 Kb) which minimises the noise level in the copy number profile data allowing for reliable analysis and copy number change estimation from DigiPico data.

The results show a significant improvement in the BAF profiles of DigiPico data with fine structural changes becoming visible in complex regions (Figure 3.10D). Lastly, to further enhance the quality of the copy number profiles, the default read count binning window size of 10 kb was increased to 100 kb in *ichorCNA*. This increase in the bin size would still allow for reliable identification of CNVs as small as 1 Mb (supported by at least 10 non-overlapping windows) and also results in a significant improvement in the copy number profile of DigiPico data (Figure 3.10E). Comparing the copy number and BAF profiles of standard WGS data (Figure 3.10A) with the final outcome of DigiTitan workflow on the respective DigiPico data (Figure 3.10E) shows that DigiTitan workflow is indeed an effective strategy for reliable estimation of copy number changes from DigiPico data.

3.2 De novo variant calling from DigiPico data

Despite the advancements in variant calling algorithms, de novo identification of somatic variants from WGA DNA still seems to be extremely challenging. This is mainly because of uneven amplification, allelic drop-outs, and introduction of new mutations, notably C>T transitions, during the WGA process (Dong *et al.*, 2017). In recent years several algorithms have been developed to resolve these issues computationally (Zafar *et al.*, 2016; Dong *et al.*, 2017). DigiPico sequencing however, because of the digitised nature of its data output, offers a new solution for tackling the issues with both uneven amplification and allelic drop-outs. Moreover, as the likelihood for the occurrence of the same artificial mutations in several independent compartments is extremely low, DigiPico also offers a novel solution for identification of artefact mutations that are introduced during the WGA process. While theoretically these characteristics of DigiPico could allow for accurate *de novo* identification of variants from picogram quantities of DNA, in practice it still requires the implementation of a pipeline that could process DigiPico data. This would then allow for a comparative evaluation of DigiPico method to confirm its potentials.

To optimise a pipeline that can accurately identify somatic variants from DigiPico data, it is important to firstly find a variant calling algorithm that can best process the DigiPico data. For this purpose, various variant calling methods were tested for their performance on DigiPico data:

1. **SCcaller**: This algorithm is specifically for accurate variant calling from NGS data obtained from whole genome amplified DNA from a single cell (Dong *et al.*, 2017). It calculates the level of amplification bias at every locus based on the read depth at known heterozygous SNP positions. Using this information, it can then evaluate whether a particular SNV is likely to be an artefact of WGA or to be a true SNV. As SCcaller had been designed for the analysis of individual single cell data it can only process a single 'bam' file as input and therefore, for the purpose of DigiPico only merged data from all of the wells can be used for this algorithm.

2. **MonoVar**: This method employs a statistical SNV calling algorithm that combines the data from multiple single cell data for the discovery of high confidence SNVs (Zafar *et al.*, 2016). Considering the effect of allelic drop-outs as well WGA specific false positive errors, the algorithm can then assign a posterior probability to each identified SNV. As this approach takes multiple datasets as input, it is compatible with DigiPico data, whereby the data from every well can individually be inputted for the analysis.

3. **Platypus**: Platypus is a standard haplotype-based variant caller for high throughput analysis of NGS data (Rimmer *et al.*, 2014). Compared to other similar variant callers, such as GATK and Strelka, it is more suited for the purpose of DigiPico, because it allows for join variant calling of a large number of independent samples. Moreover, Platypus allows the user to adjust various variant calling parameters and provides a comprehensive list of statistical value which allow for a high degree of filter customization.

3.2.1 Evaluation of variant detection sensitivity of different algorithms

To identify the best variant caller for the analysis of DigiPico data, initially the variant detection sensitivity of each of the three candidate variant callers was evaluated. The SCcaller algorithm was applied to a merged 'bam' file from all of the wells, while individual 'bam' files from all wells were passed to MonoVar and Platypus as inputs. The identified variants from each method were then compared to a list of high-quality variants that were obtained from WGS data by independent algorithms (GATK and Strelka). The comparison was performed on multiple DigiPico datasets and in various categories to ensure a robust evaluation of each of the three methods. The results clearly indicate a very high false negative (FN) rate for all three of the algorithms even for homozygous germline SNPs (Figure 3.11). But analysing the Platypus data without the application of PASS filter to the output data improved the sensitivity to over 85%. This indicates that the PASS flag in Platypus is not necessarily a reliable flag for variant filtering. However, using all of the variants from Platypus without application of any filters is likely to drastically increase the rate of false positive (FP) calls. Therefore, to use Platypus for *de novo* detection of variants from DigiPico data, an optimised variant filtering algorithm to effectively remove the FP calls while preserving a high sensitivity will be required. On the other hand, comparing the results from different variant categories in SCcaller confirms the presence of a strong correlation between detection sensitivity and variant frequency (Figure 3.11). Moreover, it seems that analysing 111X data, which is from the same library as run 1111 but at a higher sequencing depth, results in a better detection rate in SCcaller. These observations indicate that the performance of SCcaller is likely to improve with a higher depth of sequencing making it a potential candidate for the analysis of DigiPicoExome data.

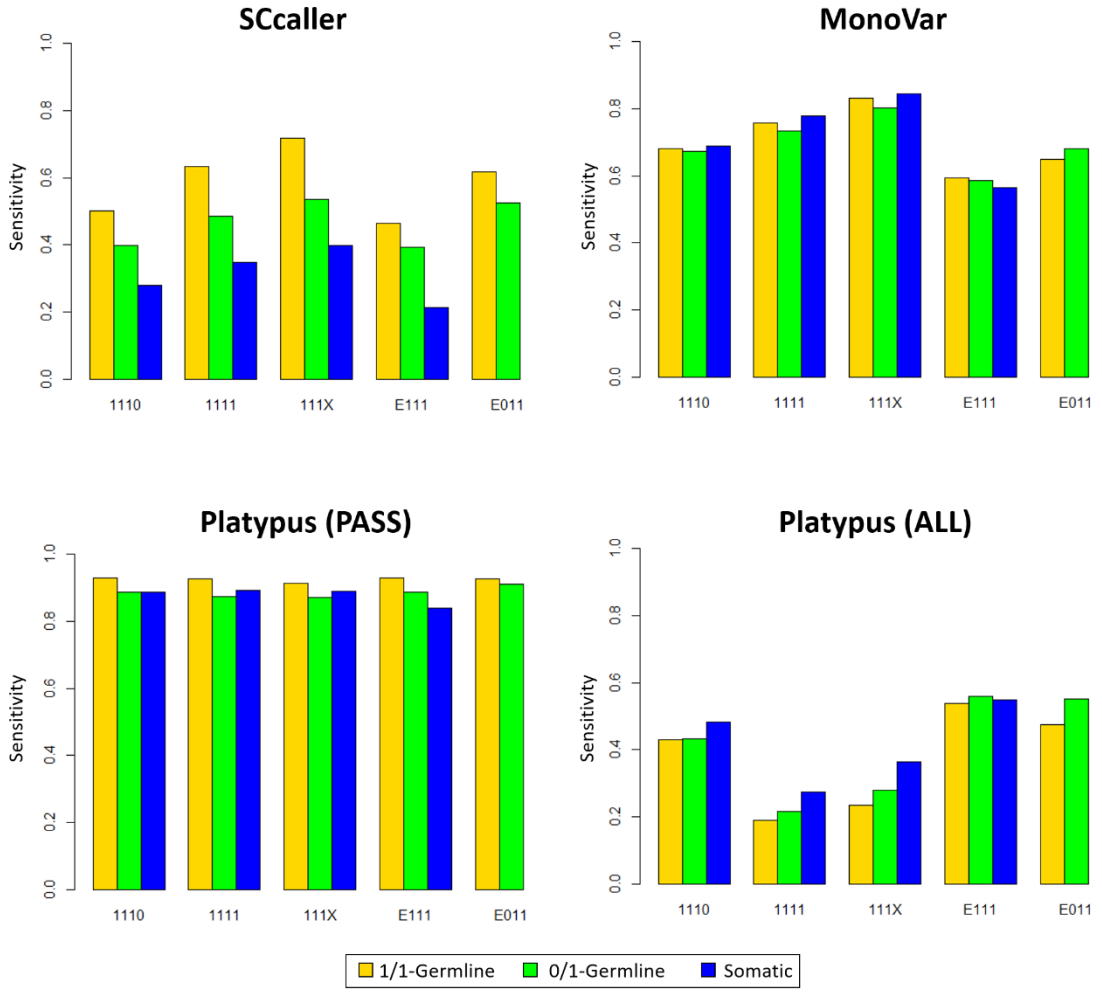


Figure 3.11. De novo variant calling sensitivity from DigiPico data. Various variant calling algorithms (SCcaller, MonoVar, and Platypus) were used for de novo variant calling from DigiPico data. For Platypus, variant calls were either filtered based on the tool's PASS filter (PASS) or all of the identified variants were used for the analysis (ALL). The resulting variants from each method were then compared with a list of known variants from various categories - homozygous germline (1/1), heterozygous germline (0/1), and somatic. The list of known germline variants was obtained from variant calling using GTAK best practices from the normal sample of respective patients. Known somatic variants were obtained by performing somatic variant calling on respective tumour samples using Strelka2 (Kim *et al.*, 2018). Sensitivity was calculated by dividing the number of identified variants to the number of *callable* variants from each category in each DigiPico run. A variant was considered callable if the respective position was covered by at least 5 wells in the DigiPico data and that at least 1 of the wells supported the variant. As run E011 is a DigiPico run on normal genomic DNA, the somatic variant category is not applicable for this run.

3.2.2 Evaluation of false positive rate of different algorithms

As DigiPico library preparation requires extensive WGA of picogram quantities of DNA, a high number of FP calls are expected to arise from the analysis of its data. Such FP calls had previously been described in data from WGA DNA and are likely to be a result of deamination of C nucleotides (C>T transitions) and other artefacts of WGA process (Mendez *et al.*, 2017). To evaluate the FP rate of each of the variant detection algorithms, the list of variants from each variant caller was first used for genotyping of standard WGS data from blood and tumour samples. Any variants with no supporting reads in either of the standard WGS data was then classed as 'Unique To DigiPico' (UTD). It is important to note that UTD variants are not necessarily FP calls, as there could truly be variants that are only identifiable in DigiPico data. These variants are in particular the point of interest when clonal heterogeneity of the tumour is under investigation. Since the variants that are private to small tumour islets will be present at low frequencies in the bulk of the sample, they are less likely to be identified in the standard WGS data, mainly because of coverage limitations. However, such mutations are expected to be highly prevalent within clonal areas, and therefore, true UTD variants can only arise from DigiPico data that are obtained from small regions of the tumour (Runs 1111, 111X, and 11E1). This is unlike the DigiPico data from DNA that is extracted from blood (Runs E011 and E0E1) or bulk of tumour cells (Runs 1110, E111, 11E0, and E1E1), mainly because in these samples, variant frequencies in the DigiPico data represent the variant frequencies of the bulk data and therefore, by definition, true UTDs in such data are most likely FP calls. Comparison of the number of uniquely identified SNVs using different algorithms on several DigiPico data (Table 3.1) shows that SCcaller identifies the lowest number of UTDs in the data and therefore outperforms both of the other algorithms in this regard. Although, Platypus results in a higher number of potential FPs, given the high detection sensitivity that Platypus offers, optimising an algorithm to filter out the FP calls could be beneficial for the analysis of DigiPico data. MonoVar on the other hand, shows an extremely high number of UTD variants. Also, considering the low detection sensitivity, it is safe to assume that MonoVar is not suitable for the analysis of DigiPico data.

Moreover, calculating the relative frequency of different mutation types of UTD variants from DigiPico runs 1110 and E111 reveals that both SCcaller and Platypus UTDs are enriched with C>T transitions and C>A transversions (Figure 3.12). The consistent bias across various samples and experiments indicates that majority of UTDs are likely WGA artefacts.

Table 3.1. Number of 'Unique To DigiPico' variants identified in various DigiPico runs using different variant calling methods.

	1110	**1111**	**111X**	**E111**	**E011**
SCcaller	754	945	1,078	1,716	810
MonoVar	1,398,239	4,809,717	13,146,344	1,299,532	1,008,342
Platypus	563,219	606,021	962,879	411,603	470,952

3.2.3 Filter optimisation for Platypus variant caller

To optimise a variant filtering pipeline for the removal of FP calls from Platypus output, it is first important to better understand the characteristics of FP and TP variants in DigiPico datasets. Platypus outputs various variant quality measures for each individual variant which could be analysed for this purpose. As it is expected that the UTD list will be enriched with FP calls and that the list of known germline and somatic variants is enriched with TP calls, any quality value that significantly differs between these two lists could potentially be used as a filtering parameter. For this analysis, several of these quality measures were extracted from Platypus output files of three DigiPico runs that are least expected to have 'true' UTD variants (1110, E111, E011). Values from each quality parameter were individually plotted for UTD variants as well as the list of known germline and somatic variants for each DigiPico run (Figure 3.13). From the plots it is clear that Posterior Probability value (PP), Failed Read fraction (FR), Homopolymer runs around the variant (HP), the ratio of variant quality to depth (QD), and the binomial P-value of strand bias (SbP), are most likely to be able to separate the reliable variant calls from artefacts.

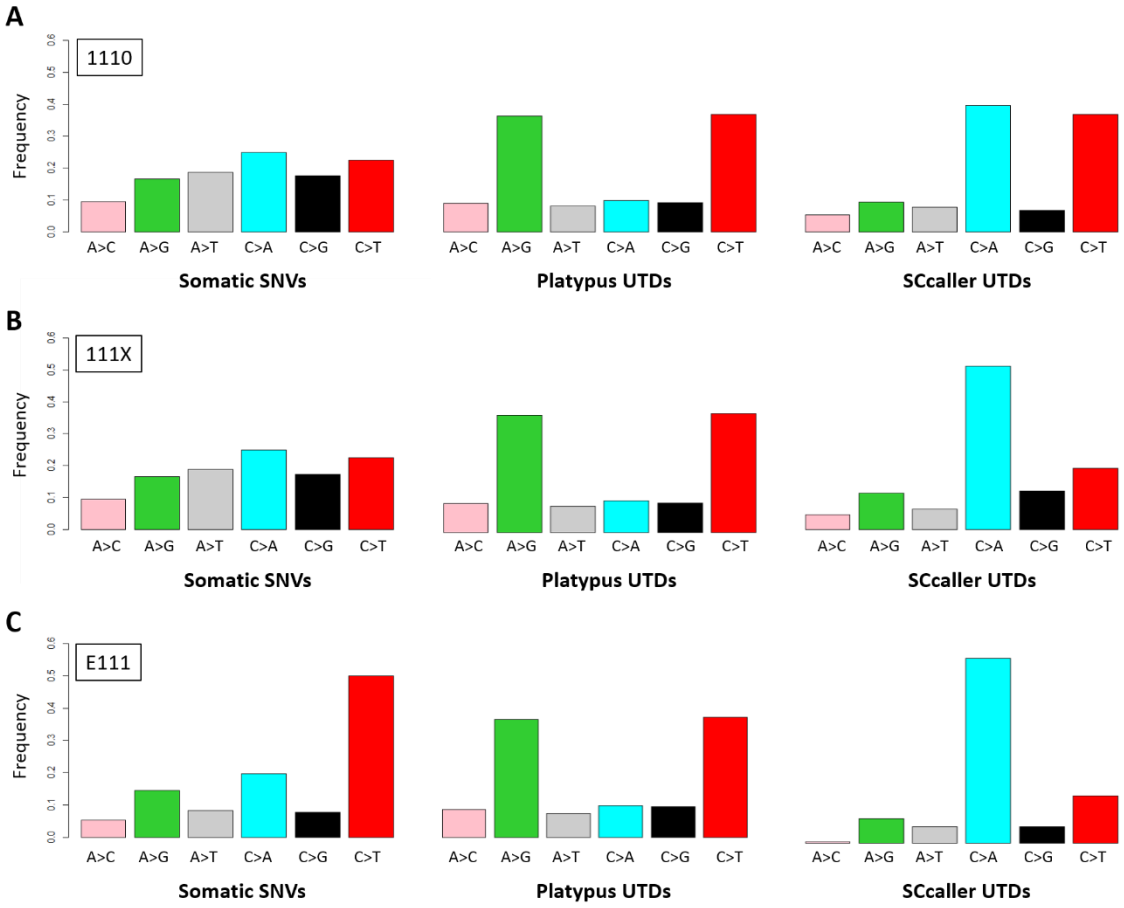


Figure 3.12. Analysis of mutation type distribution of 'Unique to DigiPico' variants from SCcaller and Platypus. Platypus and SCcaller UTD variants from three different DigiPico runs of two different patients, run 1110 **(A)**, 111X **(B)** and run E111 **(C)**, were analysed for the frequency of different types of mutations. Comparing the distribution of mutation types of known somatic variants (left panels) with UTD variants from Platypus (middle panel) or SCcaller (right panel) clearly shows that DigiPico data is specifically enriched with A>G, C>A and C>T variants. These variants are most likely FP calls and an artefact of variant calling, WGA or DigiPico library preparation process.

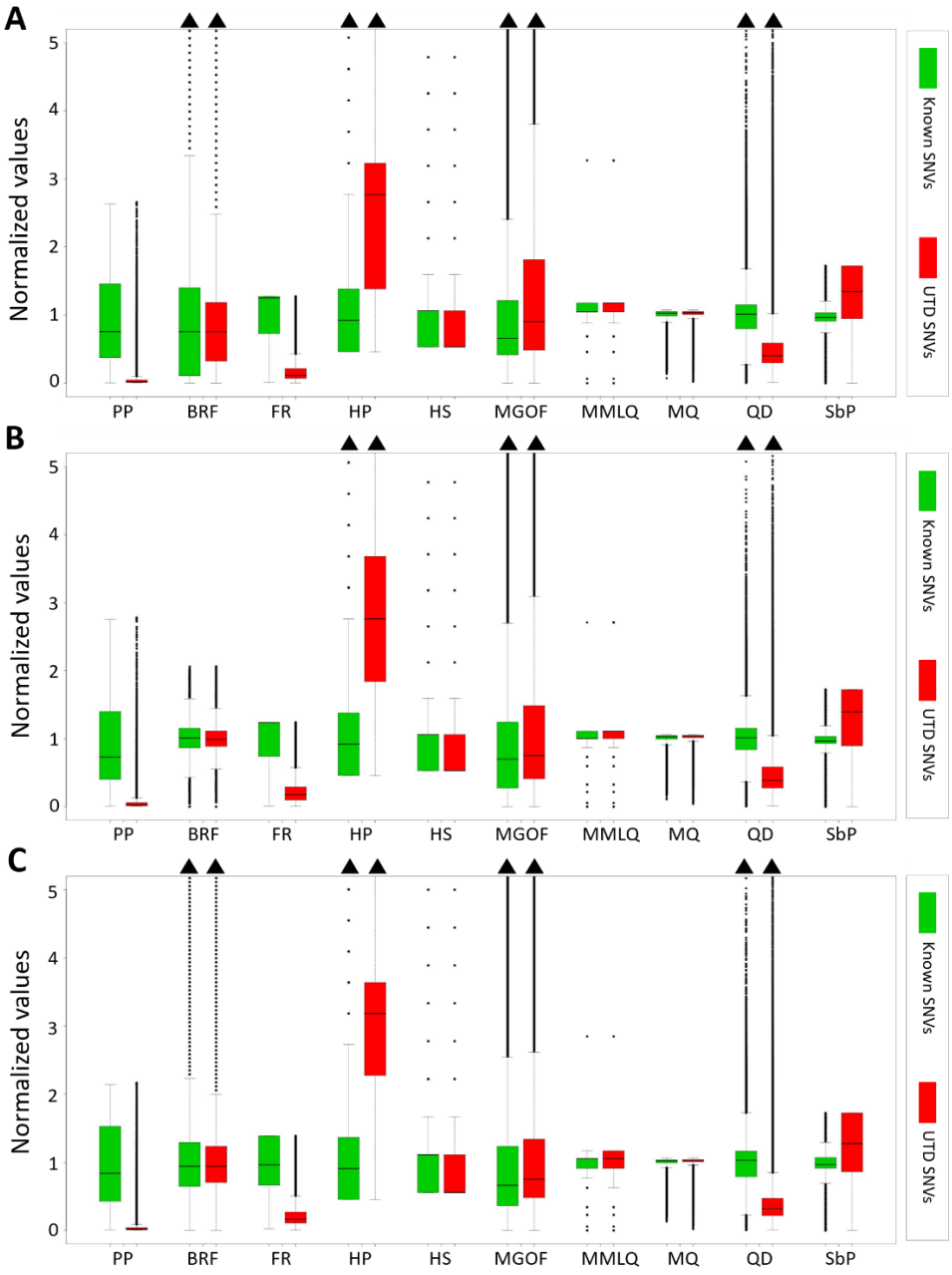


Figure 3.13. Comparing the distribution of values from various Platypus quality measure parameters between known SNVs and UTD SNVs. To identify a set of parameters that can be used to differentiate between low- and high-quality Platypus variant calls, values for various quality measure parameters were extracted from Platypus output. Next, the distribution of values for each parameter was compared between UTD variants (which are expected to be enriched with FP calls) and known SNVs (which are expected to enriched with TP calls) from 3 different DigiPico runs, 1110 **(A)**, E111 **(B)**, and E011 **(C)**. To generate the boxplots, for each parameter the values were normalised to the mean of values from the 'known SNV' category. It is immediately clear that the bulk of UTD variants are a distinct category of variants when compared with known SNVs. This confirms that majority of the UTD variants are FP calls. Among the parameters, PP, FR, HP, QD, and SbP seem to have the highest potential to be used as filtering criteria to specifically remove FP calls from UTDs. Abbreviations: **PP**, Posterior-probability. **BRF**, Fraction of failed reads around the variant. **FR**, Haplotype population frequency. **HP**, Homopolymer run. **HS**, Haplotype score. **MGOF**, Goodness-of-fit. **MMLQ**, Median minimum base quality around the variant. **MQ**, Root mean square of mapping qualities. **QD**, Variant-quality/read-depth. **SbP**, Binomial P-value for strand bias.

Next, to find potential filtering criteria, various threshold values for each of these parameters were examined for their effect on the number of UTD variants and the detection rate of known variants. The results clearly show that, for each of these parameters, a threshold can be selected that specifically removes the UTD variants while having little to no effect on the detection sensitivity (Figure 3.14). Based on this analysis, a Quality Filter (QF) with threshold values for each of the investigated parameters was developed. The QF was then applied to various DigiPico data and its effect on the number of UTDs was evaluated. The results show that using QF up to 99% of the UTD calls were removed from the data while retaining a high average detection rate of over 75% for the known variants (Table 3.2). Although this outcome is promising, the number of UTDs that are identified by Platypus still remains an order of magnitude higher than that of SCcaller (Table 3.2). Therefore, to further investigate the Platypus UTD SNVs, reads that are supporting some of such variants from the run E111 were viewed using IGV (Robinson *et al.*, 2011; Thorvaldsdóttir, Robinson and Mesirov, 2013). Visualising hundreds of randomly selected UTD variants from their source 'bam' file revealed that while a few were seemingly legitimate variant calls (Figure 3.15A), others were clearly a result of misalignment or inaccurate variant calling (Figure 3.15B-C). Therefore, to further improve the accuracy of variant calling from DigiPico data using Platypus it is essential to correctly identify and filter out these low-quality FP variant calls.

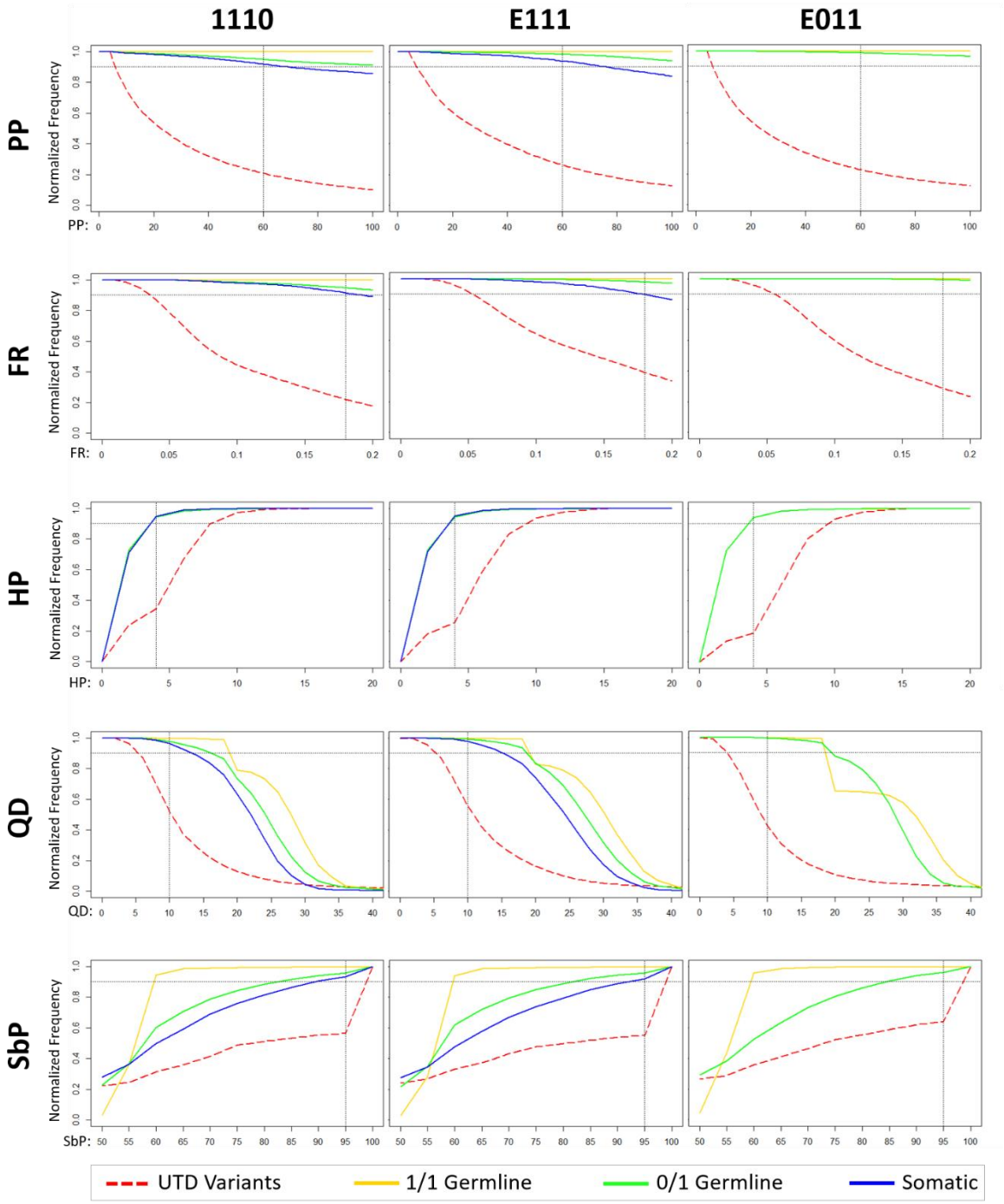


Figure 3.14. Threshold optimisation for filtering Platypus variant calls from DigiPico data. Plots show the effect of various cut-off values for each quality parameter (PP, FR, HP, QD, and SbP) on the relative frequency of each variant category (UTD, homozygous germline, heterozygous germline, and somatic) in 3 different DigiPico runs. Quality Filter (QF) thresholds (vertical dotted lines) were selected to maximise the reduction of UTDs with minimal impact on the detection sensitivity of known variants. As run E011 is a DigiPico run on normal genomic DNA, the somatic variant category is not applicable for this run.

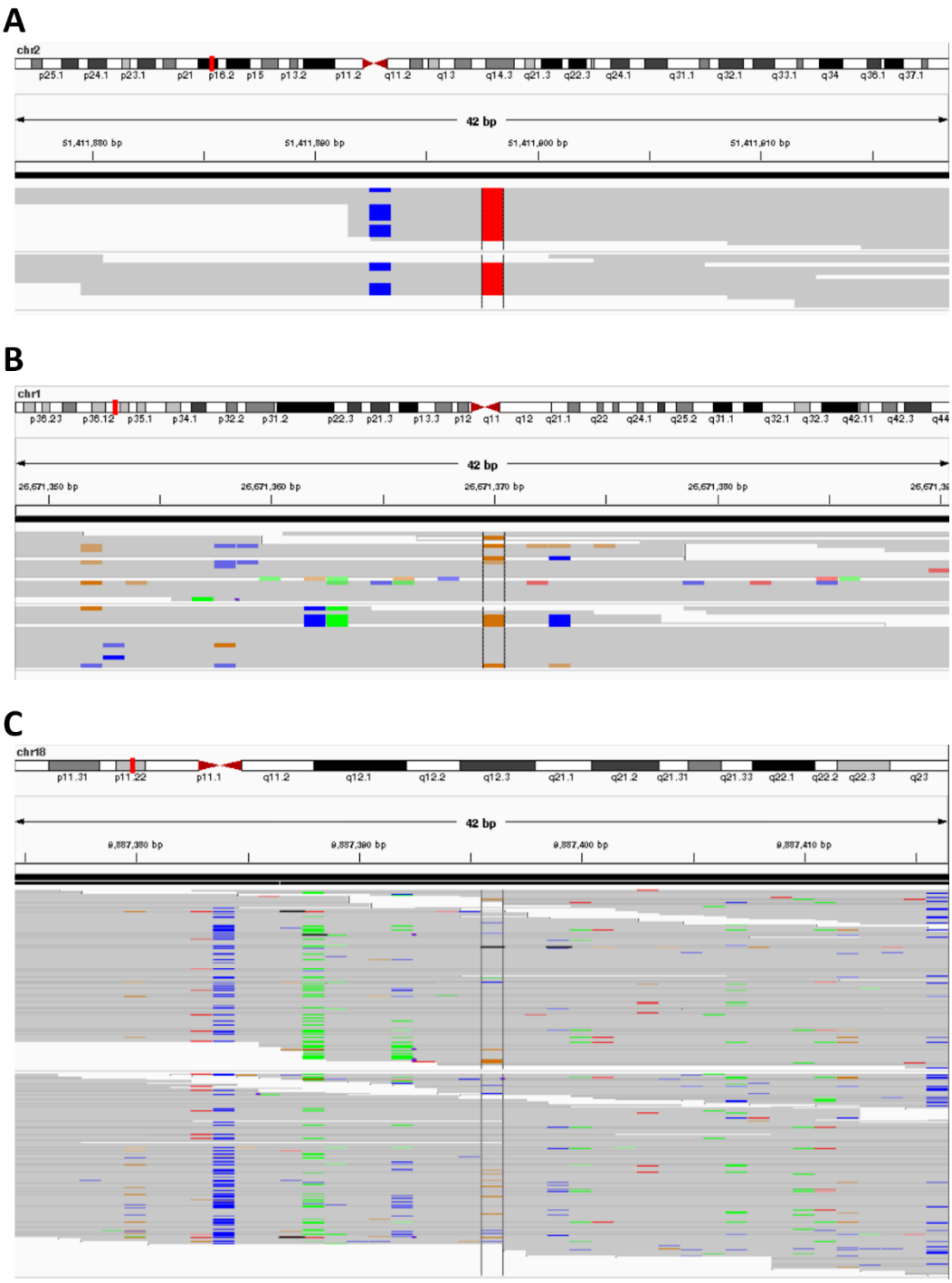


Figure 3.15. Visualisation of Platypus UTD variants from run E111. (A) Some of the UTD variants seem to be high quality variants when visualised using IGV tools. **(B,C)** However, majority of these variants are a result of misalignment to complex regions which can clearly be identified upon visualisation.

3.2.4 MutSwipe and visual processing of UTD variant calls

Having observed the potential of using the visual representation of variants for variant quality confirmation, I decided to use this information for the identification of low-quality variant calls. To do so firstly an automated pipeline was prepared to snap and store IGV images of a set of SNVs from a target merged DigiPico 'bam' file. Next, to facilitate the visual processing and classification of the images, a mobile application, MutSwipe, was developed that would allow the user to assort images to 'Good' and 'Bad' categories by swiping the images to right or left, respectively (Figure 3.16A). Using MutSwipe, out of 6,200 UTD variants from run E111, a definite decision could be made for 70% of the variants, 1,540 of which were classed as high-quality variants, while 2,845 were identified to be low-quality variants that are most likely a result of misalignment. This observation suggests that a large number of the UTD variants are in fact a result of misalignment and false variant calling by Platypus rather than an artefact of DigiPico sequencing. Identification and removal of such variants will not only improve the accuracy of variant calling but would also allow for a better evaluation of DigiPico method. While using MutSwipe for classification of variants could be useful to extract a list of high- or low-quality SNVs from a pool of called variants, it does not seem to be an appropriate method for variant filtering. This is mainly because images do not necessarily provide a certain outcome for every mutation (about 30% were left undecided in the case of E111) and that the use of MutSwipe is prone to user bias and inconsistency. Therefore, an alternative method for accurate identification of these low-quality variants is required.

3.2.5 Development of MutLearn algorithm for filtering low-quality DigiPico variants

Although MutSwipe cannot directly be used for identification of low-quality variant calls, but the large labelled datasets obtained by it could be used to train a predictive model using a supervised machine learning method such as artificial neural networks (ANN). Using the resulting predictive model, it will then be possible to automatically identify low-quality variants in other datasets.

Supervised machine learning approaches are composed of three simple steps: Firstly, a labelled dataset is generated. Each labelled dataset is composed of several data points with a label that defines the category to which each data point belongs. Moreover, each data point has several features that describe various qualities of it. In the case of MutSwipe, data points are each of the variant calls. Labels are the results of MutSwipe that define whether a variant is a high-quality variant (right swipe) or a low-quality variant (left swipe). Features in this case can be various quality and count parameters of each variant call such as allele fractions, read counts, Platypus quality parameters, etc. A subset of this labelled dataset will then be used by a machine learning algorithm for training. The training is expected to find a model that uses a linear or non-linear combination of the given features to predict the labels, with the highest possible sensitivity and specificity. After the training, the model will be tested against the rest of the labelled dataset, consisting of previously unseen data points. In a binary classification scenario, the model will assign a probability to each data point, defining the likelihood that this data point is positive (high-quality variants in this case). Lastly, the result of this test is usually presented in a Receiver Operating Characteristic (ROC) plot (Figure 3.16B). ROC curves present the relationship between the False Positive Rate (FPR; ratio of FP calls to the total number of negative labels) and True Positive Rate (TPR; ratio of TP calls to the total number of positive labels). A second Y-axis on the ROC plot, shows the threshold value that results in each of the plotted FPR-TPR values. The shape of this plot can very well show the diagnostic ability of the trained model. A good model will have a curve that significantly deviates from the X=Y line with an under the curve area of close to 1. If the design of the algorithm and the training are successful the model can then be applied to an unlabelled dataset with the same features to calculate the probability values for each data point. A threshold value can then be selected based on the ROC plot to categorise the data points into their respective categories.

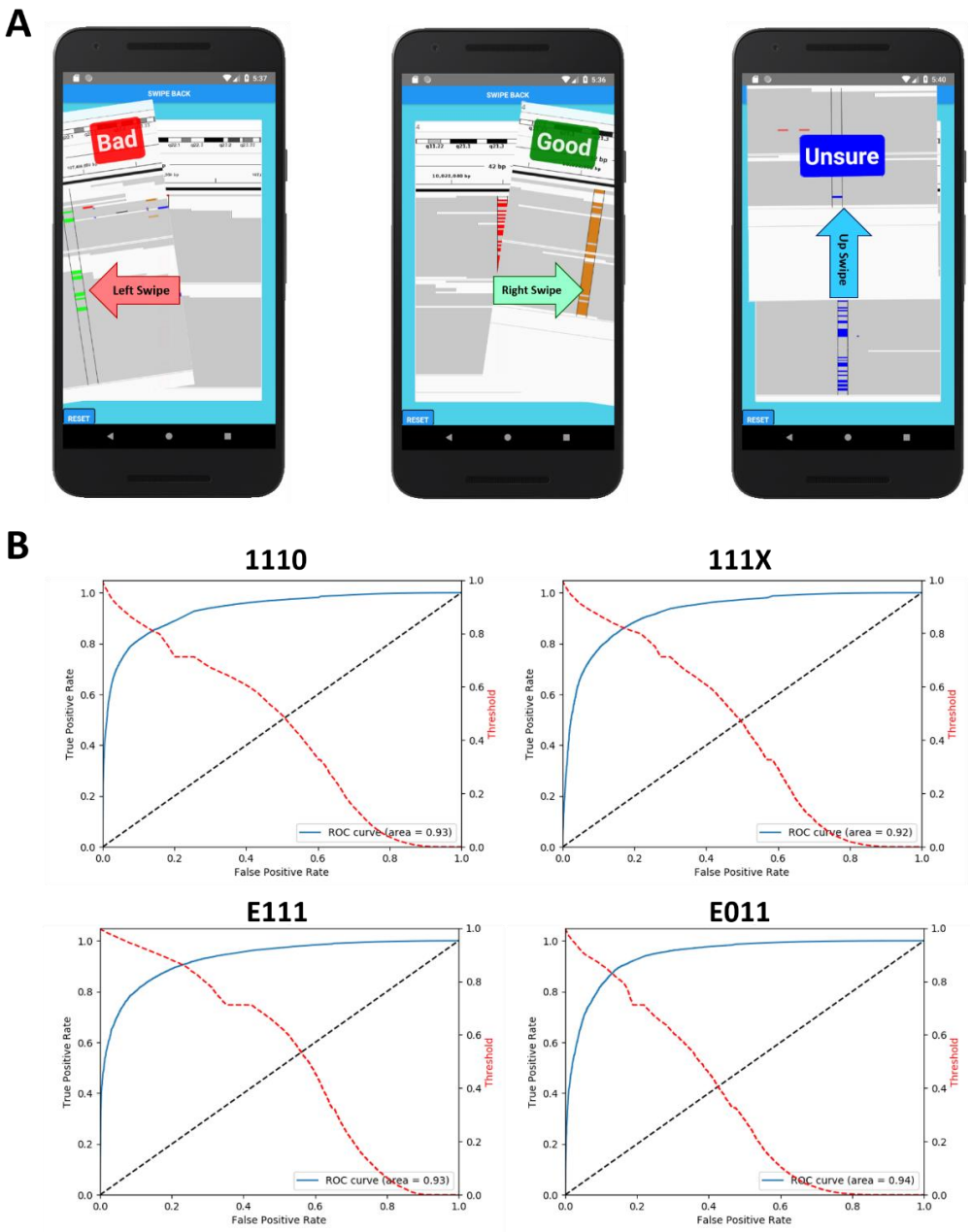


Figure 3.16. Using visual information and machine learning to filter FP calls from DigiPico data. (A) To facilitate the visual classification of UTD variants, a mobile application, MutSwipe, was developed that would allow the user to categorise the images of mutations to 'Good' or 'bad' by swiping the image to right or left respectively. If definite decision based on the images could not be made swiping the mutation to up or down would classify the variant as 'unsure'. **(B)** MutLearn model was trained using a labelled dataset from MutSwipe using an Artificial Neural-Network (ANN). Application of the model to various DigiPico data clearly shows that although the model had been trained only based on data from run E111, it performs equally well on other DigiPico runs. Receiver Operating Characteristic (ROC) plots for each run show the relationship between the False Positive Rate (FPR) and the True Positive Rate (TPR) of the test at each threshold value. To select a threshold value, a horizontal line from the desired threshold value must be drawn to intersect with the threshold curve (red). A vertical line from the intersection point, will indicate the FPR and TPR values of the chosen threshold at the intersection point with the ROC curve (blue). Based on these ROC curves a lenient threshold of 0.75 was used to filter out the low-quality variant calls from the DigiPico data. The structure of the MutLearn ANN was designed and implemented by Mr. Sahand Sharifzadeh.

ANNs are one of the most effective machine learning approaches in modelling high dimensional datasets. Therefore, for the purpose of developing a model that can automatically distinguish low-quality variant calls from high-quality variant calls of Platypus, in collaboration with Mr. Sahand Sharifzadeh, an ANN model, termed MutLearn (MutL), was developed. Training was performed on the labelled MutSwipe database from run E111 using a multilayer perceptron (MLP). Features that were used in the training, were extracted from quality control parameters of Platypus, as well as from the well count and read count data of each variant. The model was then applied to four different DigiPico datasets (Figure 3.16B). The results show that the MutL model is successfully able to identify the low-quality variants in Platypus variant calling data. Applying a threshold value of 0.75 to the outcome of MutL on these DigiPico runs resulted in an average 65% reduction in the number of UTDs while preserving a high sensitivity of 75% on average (Table 3.2). Visualising the remaining UTDs after application of MutL filter revealed that all such variants are high-quality variants that visually and based on quality values are indistinguishable from the known germline and somatic variants. Therefore, to be able to reliably confirm whether or not these variants are 'true' variants, an independent experimental approach would be required.

Table 3.2. Effect of variant filtering on the average detection rate of known variants (top rows, highlighted) and number of UTDs (bottom rows).

Data Type	1110	1111	111X	E111	E011
Platypus (No Filter)	90.15%	89.76%	89.16%	88.49%	91.88%
	563,219	606,021	962,879	411,603	470,952
Platypus (QF)	77.00%	77.65%	77.23%	76.25%	84.02%
	12,829	7,791	7,751	8,400	5,505
Platypus (QF + MutL)	63.87%	64.41%	67.24%	68.64%	77.47%
	2,357	1,383	1,999	2,745	984
SCcaller	35.76%	44.63%	50.88%	33.58%	57.17%
	754	945	1,078	1,716	810

3.2.6 Experimental validation of high-quality UTD variant calls

Having established a *de novo* variant calling pipeline that efficiently removes all of the low-quality variants from Platypus variant calls of DigiPico data, it is now important to check whether the high-quality UTD variants are indeed present in the original DNA sample or if they are an artefact of WGA. For this purpose, 40 high-quality UTD variants from run E111 were selected for a targeted sequencing validation experiment. Each target region was amplified using one pair of primers from the same DNA template that was used for the preparation of E111 library. Amplification was also performed on the E111 DigiPico library as well as genomic DNA obtained from the blood of the same patient as control. All amplicons were then used for library preparation and the libraries were sequenced on an Illumina MiSeq platform. The resulting data were mapped to human hg19 genome and for each target variant, the number of reads that support the reference or the variant allele were computed using AlleleCounter (Cancerit, 2018). Results of this validation experiment clearly show that none of the studied UTD SNVs had been present in the original template (Table 3.3), and therefore, confirming that all of these UTDs are likely to be artefacts of DigiPico library preparation.

Table 3.3. Results of targeted sequencing validation of high-quality UTD variants from run E111.

Position	Mutation	Total Well Coverage	VAWF (%)	% Reads supporting the variant		
				Ascites gDNA	Blood gDNA	E111 Library
chr1:198871416	C>G	11	54.55	0.00	0.00	2.62
chr1:205392201	G>A	26	38.46	0.02	0.00	0.04
chr1:212152502	C>A	28	21.43	0.04	0.03	1.49
chr2:165954947	C>A	14	42.86	0.01	0.02	4.37
chr2:217085031	G>T	34	11.76	0.04	0.01	0.80
chr3:55936181	C>A	14	78.57	0.02	0.05	5.66
chr3:80372726	C>A	9	66.67	0.01	0.02	0.24
chr3:148085551	C>T	10	40.00	0.00	0.01	0.02
chr3:166385800	G>A	15	46.67	0.00	0.01	1.37
chr4:21723721	C>T	8	75.00	0.02	0.00	5.05
chr4:114452967	A>C	16	68.75	0.00	0.00	0.00
chr4:187590704	G>T	9	88.89	0.02	0.00	34.26
chr5:24745295	C>A	11	36.36	0.01	0.01	1.05
chr6:29409126	C>T	21	33.33	0.01	0.02	0.01
chr6:89320707	C>A	7	57.14	0.01	0.02	0.09
chr6:120042047	C>A	12	33.33	0.00	0.00	0.01
chr7:11409022	C>T	9	55.56	0.01	0.03	0.02
chr7:16023731	C>A	21	71.43	0.07	0.02	19.36
chr7:140434437	G>T	28	39.29	0.01	0.00	1.46
chr8:28991687	G>T	31	25.81	0.02	0.03	3.24
chr8:47633895	G>T	25	16.00	0.02	0.02	1.13
chr8:56078651	G>T	24	20.83	0.01	0.00	3.67
chr8:133086518	A>T	16	31.25	0.03	0.01	0.20
chr8:134979562	G>T	11	36.36	0.04	0.07	2.29
chr10:68909833	C>A	4	100.00	0.04	0.01	5.56
chr11:5653934	C>A	26	19.23	0.01	0.03	1.00
chr11:119730031	C>T	43	13.95	0.01	0.00	2.34
chr12:128816427	G>T	9	55.56	0.01	0.02	21.32
chr13:80467921	G>T	8	62.50	0.03	0.01	1.22
chr15:34394338	G>A	10	60.00	0.01	0.01	15.84
chr15:81086223	C>T	28	78.57	0.02	0.01	0.00
chr16:80389783	G>T	11	72.73	0.04	0.02	18.68
chr17:10440734	C>A	15	26.67	0.03	0.01	0.53
chr17:57290575	T>C	28	35.71	0.00	0.00	7.40
chr18:24891922	C>A	21	76.19	0.06	0.01	0.00
chr18:34756888	G>T	23	78.26	0.04	0.05	17.09
chr18:52260039	G>T	20	75.00	0.02	0.01	0.11
chr19:3977859	G>T	4	100.00	0.02	0.00	0.01
chr21:40855807	T>A	10	70.00	0.01	0.01	0.64
chr22:25928538	A>T	15	73.33	0.00	0.02	0.05

Despite the failed validation, few important observations can be made from this experiment: (1) While no evidence were found to support the presence of any of the high-quality UTD variants in the original template, over 60% of the UTDs were found to have been present in the E111 library. This further confirms that these UTD variants have been generated during the DigiPico library preparation, most likely at the WGA step. (2) Surprisingly, the frequency of targeted sequencing reads that support the UTD variants do not seem to match with the well fraction values obtained from the analysis of DigiPico data from E111 library. This discrepancy could be a result of various factors such as unequal WGA in different wells, allelic drop-outs, or biased targeted amplification. (3) Interestingly, majority of the studied UTD variants in this experiment had been present in more than 5 to up to 22 independent compartments. As the likelihood for co-occurrence of the same artefact mutations in multiple individual compartments is extremely low (in the range of 10^{-6N} for Phi29 DNA polymerase, where N is the number of compartments containing the artefact variant) this observation contradicts the consensus understanding that WGA artefacts occur at random genomic locations. (4) Since 26 out of 40 UTD variants in this experiment (Highlighted in Table 3.3) were also called by SCcaller, these results show that SCcaller, despite its promise, is also unable to reliably identify and remove the FP calls that are formed as an artefact of WGA from DigiPico data.

3.2.7 Dual-Allelic Compartment filter for further filtering of Platypus UTD calls

While the results of the validation experiment indicate that majority of UTD SNVs are most likely an artefact of DigiPico library preparation, it also offers and opportunity for identification and removal of such FP variant calls from the DigiPico data. Assuming that these artefacts form during the WGA process, it can be imagined that both the variant allele and the reference allele must be present in the same compartment that contains the artefact mutation. However, because of sub-haploid dilution of template during the DigiPico library preparation process, for majority of the 'true' variants, each individual compartment may only contain the reference or the variant allele, and not both of them at the same time. Therefore, the number of dual-allelic compartments for each variant could potentially be used as a marker for identification of FP artefacts.

To check this hypothesis, the number of dual-allelic compartments for each of the 40 UTDs from the validation experiment were calculated. This analysis revealed that 29 out of the 40 studied UTD variants have at least 1 dual-allelic compartment. Therefore, using this information and also considering that variants with higher frequencies are more likely to have a dual-allelic well, a Dual-Allelic Compartment (DAC) filter was designed. Application of this filter on other DigiPico runs showed that, using this filter, it would be possible to remove up to 40% of the high-quality FP calls from the DigiPico data at the expense of a 5-15% reduction in sensitivity (Table 3.4).

Table 3.4. Effect of the DAC filter on the variant detection sensitivity (top row, highlighted) and the number of Platypus UTD calls (bottom row).

Data Type	1110	1111	111X	E111	E011
Platypus (QF + MutL)	63.87%	64.41%	67.24%	68.64%	77.47%
	2,357	1,383	1,999	2,745	984
Platypus (QF + MutL + DAC)	58.67%	56.55%	57.04%	64.13%	70.73%
	1,584	908	1,200	2,073	651
SCcaller	35.76%	44.63%	50.88%	33.58%	57.17%
	754	945	1,078	1,716	810

3.2.8 Experimental validation of DAC filtered UTD calls

Having eliminated a large number of high-quality UTD calls using the DAC filter, a second round of experimental validation of UTD calls was performed on run 1111. Unlike run E111, run 1111 is expected to have some number of 'true' UTD variants as its input template had been obtained from a cross section of a recurrent tumour sample (site PT2R from patient 11152). Therefore, it is plausible to assume that the few numbers of cells that were used for the preparation of this library, might have had some sub-clonal mutations that are only present at very low frequencies in the bulk DNA of the PT2R sample. Consequently, at standard WGS sequencing depths of 30-40x these variants will not be detectable from the bulk sequencing data. However, by performing targeted sequencing on the bulk DNA extracted from the PT2R site, there is a chance to detect such variants.

To test whether these speculations are correct, 15 DAC filtered UTD variants from run 1111 were selected and amplicons containing these variants were amplified from the blood and PT2R DNA extracts. Moreover, as a control, DNA from another recurrence tumour site in the same patient (PLNAR) was also used for this experiment. While 3 out of the 15 target sites seemed to have a high noise level in the data (Table 3.5, highlighted in yellow), the results of this validation analysis show that 9 out of the remaining 12 analysed UTD variants are present in at least 0.2% of the reads from the PT2R gDNA sample (Table 3.5). This indicates that these variants are very likely to be 'true' clone-specific variants in the region from the tumour that the section had been collected. Further detailed analysis of this validation analysis led to three interesting observations:

1) One of the confirmed variants (**chr1:207222565.G>C**) with 43% variant allele well frequency (VAWF) in the DigiPico data (Table 3.5, highlighted in green), is also present in 10% of the reads from the PT2R site as well as in 18% of the reads from the PLNAR site. Considering that this mutation resides in a diploid region of chromosome 1q (Figure 3.2) it is likely that this mutation has a 100% prevalence in the clonal area where the section for run 1111 had been obtained.

2) Presence of **chr1:207222565.G>C** mutation in both of the analysed recurrence tumour samples in relatively high frequencies suggests that this mutation is very likely a sub-clonal variant that had been present in the population of the cells that has initiated the recurrence tumour. Analysing the WES data from the PT2R site (run 1141) confirms that this variant is indeed present in the PT2R site but at a low frequency (Figure 3.17). Moreover, analysing the standard WGS data from a pre-chemotherapy tumour sample of the same patient (run 1201) confirms the pre-chemotherapy presence of this variant in the tumour at about 20% frequency (Figure 3.17). Further indicating that this mutation is likely to be a sub-clonal mutation from a clone of cells that had contributed to the emergence of the recurrence.

3) Calculating the distribution of mutation types of the 9 validated UTD variants reveals that more than half of these variants are of the C>T type. As these variants are very unlikely to be FP calls after being experimentally validate, and also considering that 10 out of the 30 described *signatures of mutational processes* (Alexandrov *et al.*, 2013) are dominantly presented by C>T mutations, this observation suggests that there likely is a tumour specific molecular process behind this bias towards the C>T variants in this tumour sample.

4) On the other hand, all three of the unvalidated UTD variants are also of the C>T type (Table 3.5, highlighted in red). Moreover, analysing the mutation type distributions after each of the filtering steps shows that there remains a significant bias in the variant distributions, particularly towards the C>A variants (Figure 3.18). These observations suggest that there still are a large number of FP calls among the UTD variants that would require to be eliminated using further filtering strategies.

Table 3.5. Results of targeted sequencing validation of high-quality, DAC filtered UTD variants from run 1111.

Position	Mutation	Total Well Coverage	VAWF (%)	% Reads supporting the variant		
				Blood gDNA	PLNAR gDNA	PT2R gDNA
1:207222565	G>C	7	42.86%	0.35	18.00	9.56
1:84044398	T>C	14	78.57%	0.02	0.02	0.85
2:33010682	A>G	36	58.33%	0.00	0.01	0.77
2:123789390	C>T	28	46.43%	0.00	0.01	0.27
3:129524269	G>A	20	45.00%	0.00	0.00	0.31
8:110503277	G>A	9	22.22%	0.02	0.05	0.48
13:50163982	T>C	5	100.00%	0.03	0.02	0.20
17:66520311	C>T	18	38.89%	0.06	0.05	0.46
17:66523728	C>T	33	39.39%	0.02	0.03	0.35
4:76733361	C>T	11	27.27%	0.03	0.02	0.03
17:66521158	C>T	22	31.82%	0.01	0.01	0.09
X:84510666	G>A	19	21.05%	0.00	0.01	0.01
1:168274409	C>G	14	35.71%	0.33	0.26	0.49
4:171617451	G>A	31	41.94%	0.17	0.07	0.29
11:113723225	G>T	13	15.38%	0.21	0.16	0.26

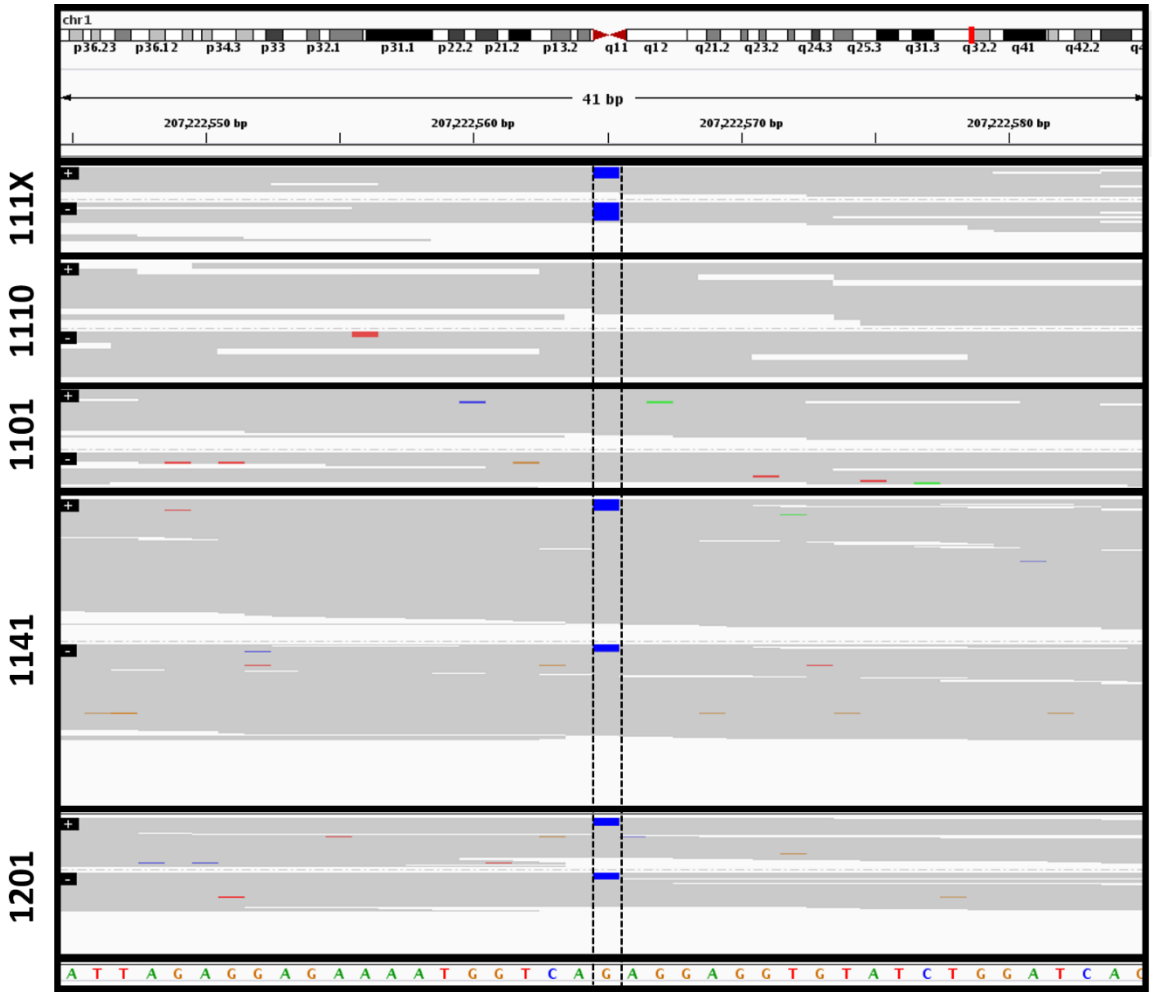


Figure 3.17. IGV visualisation of the variant position chr1:207222565 in sequencing data from patient 11152. The validated **chr1:207222565.G>C** variant seem to be a sub-clonal mutation from a clone of cells that had partially contributed to the emergence of recurrence tumour sample. The variant seems to be present at a very high prevalence in DigiPico data obtained from a tumour section of the PT2R recurrence tumour (111X). It however, does not seem to appear in the 30-40x WGS data (1101) or DigiPico data that had been obtained from extracted DNA of the same site (1110). The variant however, can be detected in 200x WES data of the PT2R site (1141) as well as 30x WGS data of an omental pre-chemotherapy tumour sample of the same patient (1201).

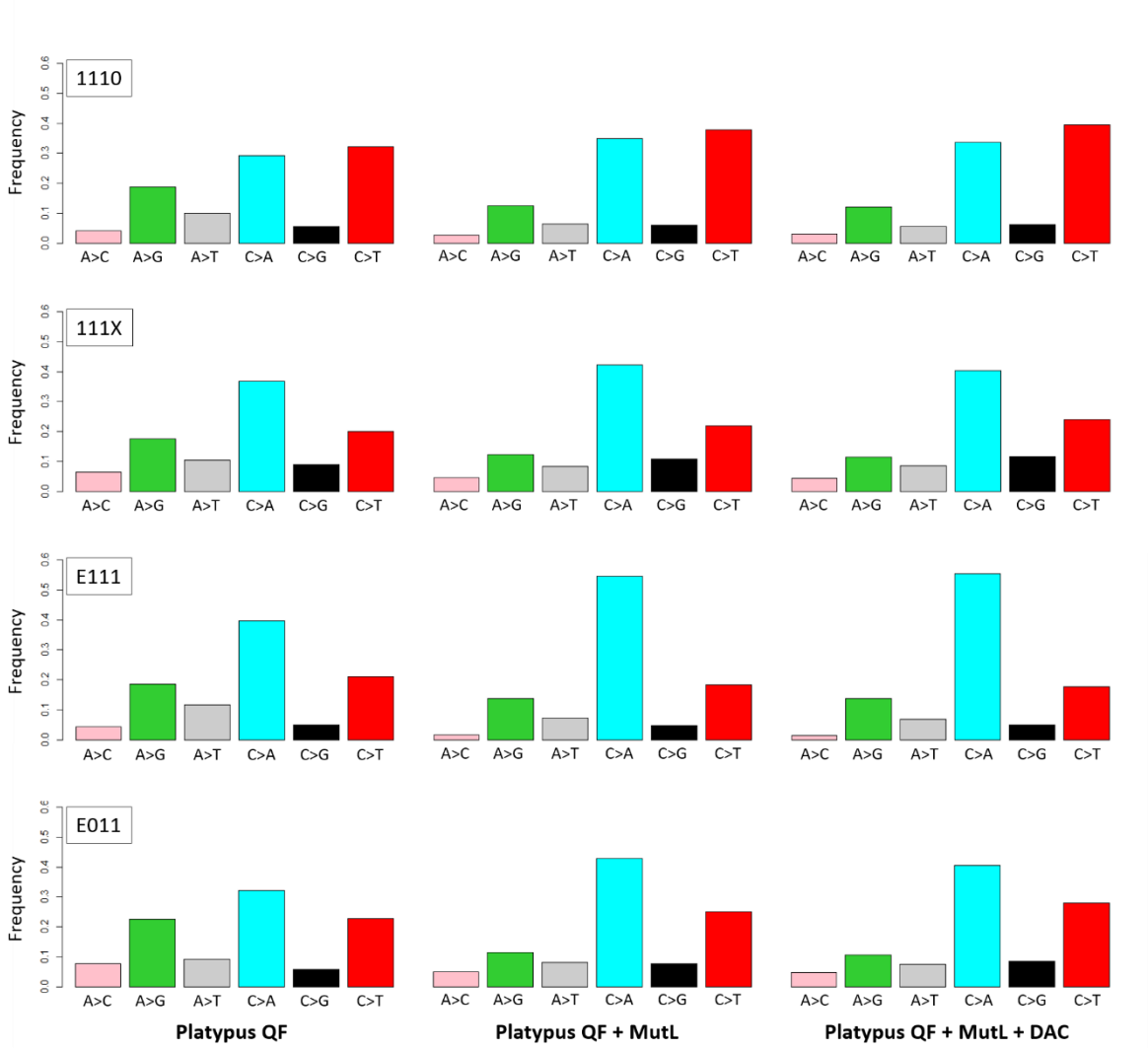


Figure 3.18. Effect of various filtering criteria on the mutation type distribution of Platypus UTD variants. Bar plots representing the frequency of mutation types after each filtering level of Platypus UTD variants of various DigiPico runs. It is clear that, although majority of FP A>G variants had been removed from the Platypus data using the QF filter, but even after application of QF, MutL and DAC filters there still is a significant bias towards the C>T and C>A mutations. This indicates that majority or even all of the UTD variants are very likely to be FP calls generated during WGA process.

3.2.9 Development of MutLearn2 algorithm for filtering high-quality FP UTDs

Recapping on the hypothesis that UTDs are mainly a product of polymerase errors during WGA, it is also expected that the distribution of reads in wells and among wells differ from that of 'true' variants. As the exact patterns that may arise from WGA mistakes is not immediately inferable from the data, once more I decided to employ the power of machine learning to see if it would be possible to further remove the noise from the DigiPico data. For this purpose, a new ANN based machine learning algorithm, MutLearn2 (MutL.2), with features that were all related to read distributions among the wells was designed (Section 6.4.7.3). Unlike with MutL, an independent and accurately labelled database of high-quality WGA artefacts is not available for MutL. However, comparing the frequency of mutation types in UTD variants after each level of filtering, clearly shows that there still is a strong bias towards the C>A variants in the UTDs from all of the runs, which together with the results of the validation tests further confirms that a large proportion of these UTDs are likely to be FP calls (Figure 3.18). Moreover, as read distributions in DigiPico data can vary based on various library preparation factors, such as the quality and quantity of input DNA or WGA efficiency, a model that is trained on one dataset may not necessarily be applicable to other datasets. Therefore, assuming that majority of the UTD variants are FP calls and that all of the variant calls of known variant positions are TP calls, it would be possible to use the read distribution data from a subset of UTDs and known sites from each run to train an ANN model for MutL.2. This model can then be applied to the remaining variants of the same dataset to determine whether or not they are FP calls. To evaluate this approach, all the UTDs from the unfiltered Platypus variant calls of run 1110 were combined with a similar number of variant calls from the known SNVs of the same run. Then 10% of this dataset was used to train the MutL.2 ANN and the remaining 90% were used for model validation. The validation results showed that the trained model is extremely effective at differentiating FP calls from TP calls with the area under the ROC curve reaching to 0.99 (Figure 3.19A). This is not surprising because even with the QF filter, which arose from simple comparison of quality measures, it was possible to filter up to 99% of the FP calls

with a minimal loss on the TP calls. However, the fact that MutL.2 has achieved this accuracy only by comparing the read distributions, without any input from any of the quality control measures, is a great indication that the selected read distribution features are fit for the purpose of this filtering approach. However, to ensure that MutL.2 is not over-fitting a model into the training dataset, a new training was performed on the QF filtered UTD variants. If overfitting occurs, it is expected that the ROC curve resulting from the QF filtered data matches with the ROC curve resulting from the training on the unfiltered data. The results, however, indicate that MutL.2 is unlikely to be over-fitting a model into the training dataset and hence, it is effective solution for further filtering of Platypus UTD variants (Figure 3.19B). Therefore, to evaluate the applicability of MutL.2 to DigiPico data, for each DigiPico dataset, a new run-specific model was generated which was then applied to the DAC filtered data of the respective run. Having removed up to 90% of the remaining UTD calls from the data with a sensitivity loss of only about 15%, these results clearly confirm the effectiveness of MutL.2 in identifying FP calls (Table 3.6).

Table 3.6. Effect of application of MutLearn2 variant filtering model on the number of UTDs (bottom rows) and average detection rate of known variants from DigiPico data (top rows, highlighted).

Data Type	1110	1111	111X	E111	E011
Platypus (No Filter)	90.15%	89.76%	89.16%	88.49%	91.88%
	563,219	606,021	962,879	411,603	470,952
Platypus (QF)	77.00%	77.65%	77.23%	76.25%	84.02%
	12,829	7,791	7,751	8,400	5,505
Platypus (QF + MutL)	63.87%	64.41%	67.24%	68.64%	77.47%
	2,357	1,383	1,999	2,745	984
Platypus (QF + MutL + DAC)	58.67%	56.55%	57.04%	64.13%	70.73%
	1,584	908	1,200	2,073	651
Platypus (QF + MutL + DAC + MutL.2)	56.84%	55.80%	55.76%	55.81%	67.62%
	516	423	432	182	108
SCcaller	35.76%	44.63%	50.88%	33.58%	57.17%
	754	945	1,078	1,716	810

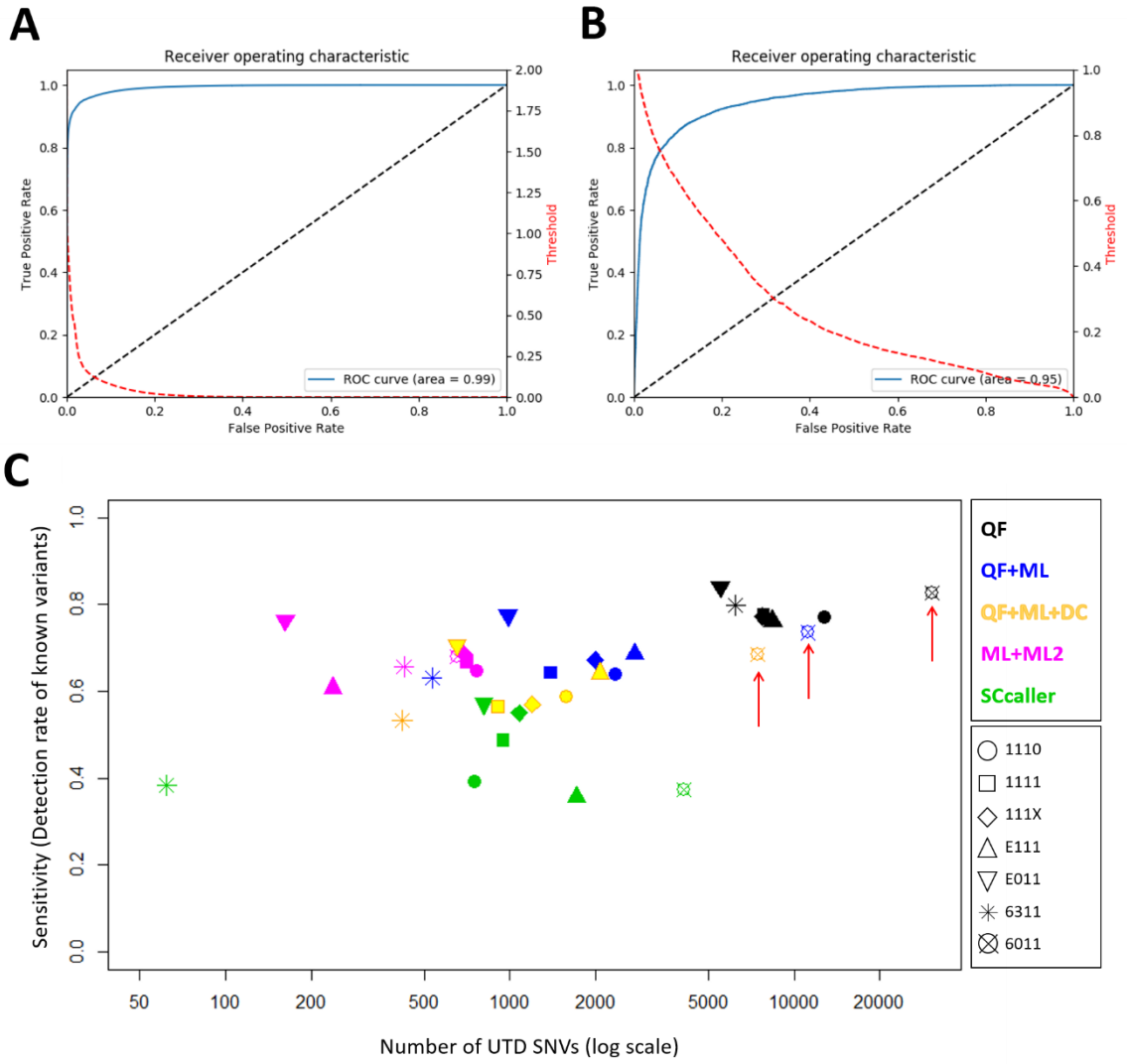


Figure 3.19. Development and application of MutLearn2 algorithm for improved variant filtering of DigiPico data. (A) Performing training on 10% of the unfiltered DigiPico data and testing on the remaining 90% using MutL.2 Artificial Neural-Network (ANN) algorithm results in an extremely effective model. Although such results might be worrying in terms of over-fitting, it does not raise any concerns in this case. This is mainly because over 99% of the UTD variants are very low-quality variants that can immediately be identified and removed from the dataset, therefore significantly improving the training outcome. **(B)** Application MutL.2 on the Quality Filtered (QF) DigiPico data, however, results in the training of a high-quality model that is expected to be able to efficiently remove FP calls from the DigiPico data. **(C)** Scatter plot presenting the detection sensitivity of known variants against the number of UTD variants from various DigiPico runs with different filtering criteria. It is clear that application of filters more specifically removes the UTD variants (note that X-axis is presented in log scale), this further highlight that majority of UTD variants are likely to be FP calls. Although application of 3 levels of filtering to Platypus data (QF, MutL and DAC filters) has resulted in a significant reduction in the number of UTDs, these filters seem to be behaving inconsistently on at least one of the DigiPico datasets. As indicated with red arrows on the plot, number of UTD variants for run 6011 seem to be significantly higher than other runs under similar QF, MutL and DAC filtering criteria. The MutL.2 ANN, however, seems to have been able to fully address this issue. Moreover, it is clear that combination of MutL.2 and MutL results in a remarkable improvement in the filtering outcome of Platypus DigiPico data. Abbreviations: **QF**, Quality Filter. **ML**, MutLearn filter. **DC**, DAC filter. **ML2**, MutLearn2 filter.

Considering the very high efficiency of MutL.2, it was also evaluated to see if application of MutL.2 to data from earlier stages of filtering could improve the sensitivity while maintaining a low FP rate. Results indicated that as long as both MutL and MutL.2 filters are applied to the data the overall outcome is similar and significantly better than when either of the machine learning algorithms are absent (Figure 3.19C, Table **3.7**). Lastly, to ensure the generalisability of these observations, all of these filtering conditions were applied to two new sets of data (Runs 6011 and 6311 from patient OP1036) which has not been used in any of the previous optimisation stages. Surprisingly, the results show that QF, MutL and DAC filters are unable to effectively filter the UTDs from run 6011 (Figure 3.19C, red arrows). This is most likely because the manual threshold optimisation for QF and DAC filters as well as using the data from only one DigiPico run for the training of MutL has resulted in over-fitting of the filtering parameters to certain DigiPico experiments. MutL.2 however, seems to be robust to these run-to-run variabilities. This is likely due to the fact that in MutL.2, a new model is generated for each run which allows MutL.2 to compensate for such of variabilities.

Table 3.7. Effect of combining MutL.2 with various other variant filtering criteria on the average detection rate of known variants (top rows, highlighted) and number of UTDs (bottom rows).

Filters on Platypus data	1110	1111	111X	E111	E011
QF, MutL, DAC, MutL.2	56.84%	55.80%	55.76%	55.81%	67.62%
	516	423	432	182	108
QF, MutL, MutL.2	61.42%	63.35%	64.95%	57.65%	72.12%
	666	603	626	200	149
MutL, MutL.2	64.70%	66.81%	68.57%	60.86%	76.25%
	770	705	689	238	162
QF, MutL.2	68.39%	74.20%	71.62%	59.02%	74.03%
	1,039	1,278	900	234	187
Only MutL.2	73.54%	81.42%	77.88%	63.15%	78.94%
	2,338	7,203	3,044	629	236

3.2.10 MutLearnX: Development of an unbiased, effective and customisable variant filtering algorithm for DigiPico data

The observation in Figure 3.19C raises the concern that dependency of MutL to a pre-built model based on MutSwipe outcomes might result in poor performance on a subset of DigiPico runs, which, are dissimilar to the DigiPico runs that were used for the training of the MutL model. However, robustness of MutL.2 to these run-to-run variabilities suggests that training run-specific models could potentially be used as a solution to compensate for variabilities in quality measures of different DigiPico datasets. Therefore, combining the MutL and MutL.2 strategies, a new ANN-based pipeline termed MutLearnX (MutLX) was developed.

MutLX is composed of two main branches. A quality-measure based branch (Q), composed solely of quality measure features, and a read-distribution based branch (R), composed solely of read distribution values. For each DigiPico dataset, MutLX trains a Q-model and an R-model from a small subset of UTD and known variant calls. When applied to the remaining majority of the data in the same dataset, each of these models generate a probability value for each variant. Variant calls can then be filtered using a combination of R- and Q-thresholds to achieve the required level of noise reduction that best fits the scopes of the analysis. Before applying this strategy to DigiPico data, it is essential to ensure that MutLX does not over-fit a model into the UTD variants which would hamper the identification of 'true' UTD variants. To test this and to evaluate the effectiveness of MutLX, multiple synthetic datasets were generated by artificially labelling various ratios of the known SNVs in the dataset as UTDs (UTD*) from the DigiPico data of run 1111. For each synthetic dataset, 20% of the data were used by MutLX for training and the resulting model was then applied to the remaining 80% of the variants in the respective synthetic dataset. If over-fitting occurs, it is expected that UTD*s, despite being 'true' good quality variants, will be removed at a much higher rate compared to the rest of the know variants. Comparing the ROC plots of various UTD* ratios shows that, UTD* at low frequencies do not exert any significant effect on the training of the model, but may reduce the sensitivity of the model at higher ratios (Figure **3.20**A).

This however, is not an issue when dealing with unaltered DigiPico data, as the frequency of 'true' UTD variants is not expected to exceed the tolerated levels that had been examined in this analysis. Moreover, 3D representation of the results of the analysis for 2% UTD* ratio, clearly indicates that, as expected from a correctly trained model, increasing the threshold values, increases the PPV of UTD* variant, despite the fact that during the training they had artificially been labelled as 'negative' data points (Figure **3.20**B). Further analysis of the results also confirms that although at higher UTD* frequencies the efficiency of the model diminishes, but at all of the examined UTD* ratios the True Positive Rate (TPR) of UTD* variants equates with the average TPR of heterozygous and somatic variants, as expected when over-fitting doesn't occur (Figure 3.21A).

Extending this analysis to other DigiPico samples also confirms that the TPR of UTD*s consistently matches with the TPR of unaltered known variants further proving the robustness of the MutLX algorithm (Figure 3.21B). Combination of these analysis outcomes nicely confirms that MutLX does not overfit a model into the UTD variants and therefore, is able to correctly identify about 50-60% of the rare 'true' UTD (clone-specific) variants from among thousands of FP UTD variant calls, despite having them labelled as 'negative' during the training stage. Based on the success of these evaluations, MutLX was then applied to QF filtered data from all of the available DigiPico datasets and the outcomes were compared with previous filtering attempts (Figure 3.22A). The result illustrates that, MutLX algorithm, can achieve a comparable outcome to that of the previous variant filtering attempts in this project. However, the added benefit of MutLX is that it allows for custom selection of threshold values which could be adjusted to achieve any level of filtering from DigiPico data as required.

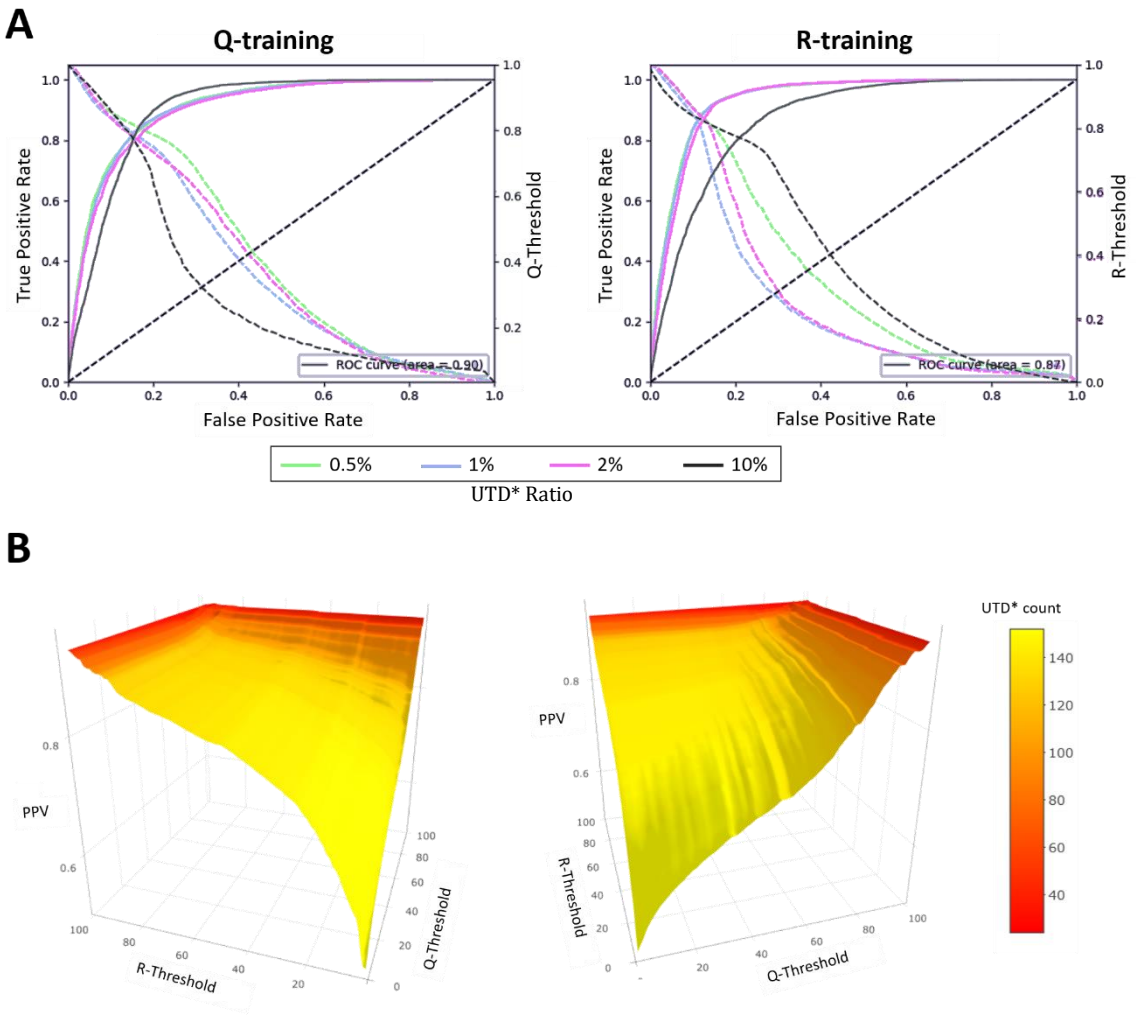


Figure 3.20. Validation of MutLearnX Artificial Neural-Network pipeline using UTD* variants. To ensure that MutLX, despite being trained on the UTD variants, is still able to identify 'true' UTD variants, various ratios of known heterozygous and somatic variants were artificially labelled as UTD variants (UTD*) and the resulting altered databases were individually used to train the MutLX ANN. **(A)** The ROC plots of Q- and R-training attempts on these artificial databases clearly show that at UTD* ratios below 10% the effect on the training efficiency is negligible. **(B)** 3D surface representation of the results, coloured by the number of correctly identified UTD*, indicates that increasing values of Q- and R-thresholds while reducing the number of correctly identified UTD*s, increases the Positive Predictive Value (PPV) of UTD*. This indicates that MutLX ANN does not seem to over-fit a model into the training data and is able to correctly identify 'true' UTD variants from among a large number of FP UTD variants.

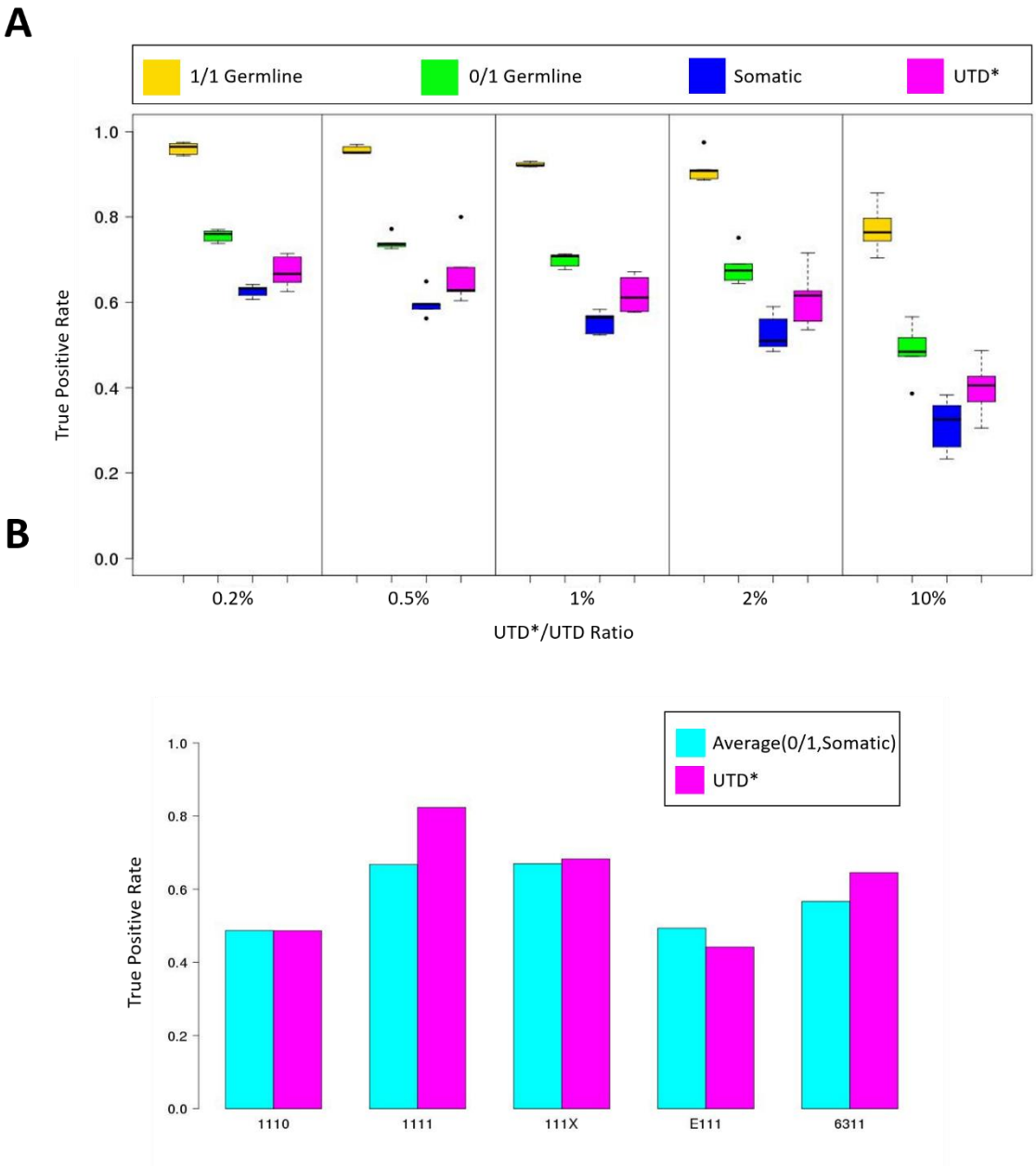


Figure 3.21. Comparison of TPR of UTD* and unaltered known variants. (A) Comparison of the True Positive Rate of true variants that had been labeled as negative during the training (UTD*) with 'true' variants that had been labelled 'positive' during the training (1/1 Germline, 0/1 Germline, and Somatic) shows that the UTD* variants are identified at similar rates to what had been expected from them had they been labelled correctly in the training dataset. As the UTD*s are a 1:1 mix of heterozygous germline and somatic variants, their TPR is expected to be the average TPR of the two known variant types. The results for each ratio are from 5 different randomly selected sets of UTD*s. Boxplots represent the mean and standard deviation of the 5 replicate analysis. **(B)** Comparison of the the TPRs at 0.5% UTD*/UTD ratio of four more DigiPico samples also confirms that the TPR of the UTD* variants is unaffected by their presence among the 'negative' categry of the training dataset.

Having been validated as a reliable variant filtering algorithm, MutLX was then applied to DigiPico library 1111 for further evaluation and optimisation. MutLX automatically selects an 'optimal' value for each of the Q- and R-thresholds based on their respective ROC curves, but this only occurs at a 2D level for each threshold without taking the effect of the other threshold into account. Although, this would be sufficient for general analysis purposes, to achieve an optimal filtering, selecting the best threshold values would be essential. To facilitate this, contour plots of both threshold values (as X and Y axis) against the TPR (as level plot) and the number of remaining UTDs (as contour plot) and vice versa were generated (Figure 3.22B). Using this plot, it is now easily possible to select Q- and R-threshold values that would best serve the purpose of the analysis.

3.2.11 Application of MutLX algorithm to DigiPicoExome data

Next, to also demonstrate the versatility of the MutLX algorithm, it was applied to four DigiPicoExome datasets from patients 11152 and 11513. DigiPicoExome libraries, 11E0, 11E1, E1E1, and E0E1, had been directly prepared from 1110, 1111, E111, and E011 libraries and therefore, their genetic contents within the capture region must be identical. As DigiPicoExome data have a significantly smaller number of total variants, application of MutLX to already Quality Filtered DigiPico data is not possible as the number of datapoints in the training set would be so small that an effective training will not be achieved. Therefore, for all of the DigiPicoExome data, the Q- and R-trainings were performed on the 20% of the unfiltered data. The results of the analysis were then used to generate Q- and T-threshold contour plots (Figure 3.23). The outcomes clearly indicate the MutLX is indeed able to effectively train a mode to filter the DigiPicoExome data.

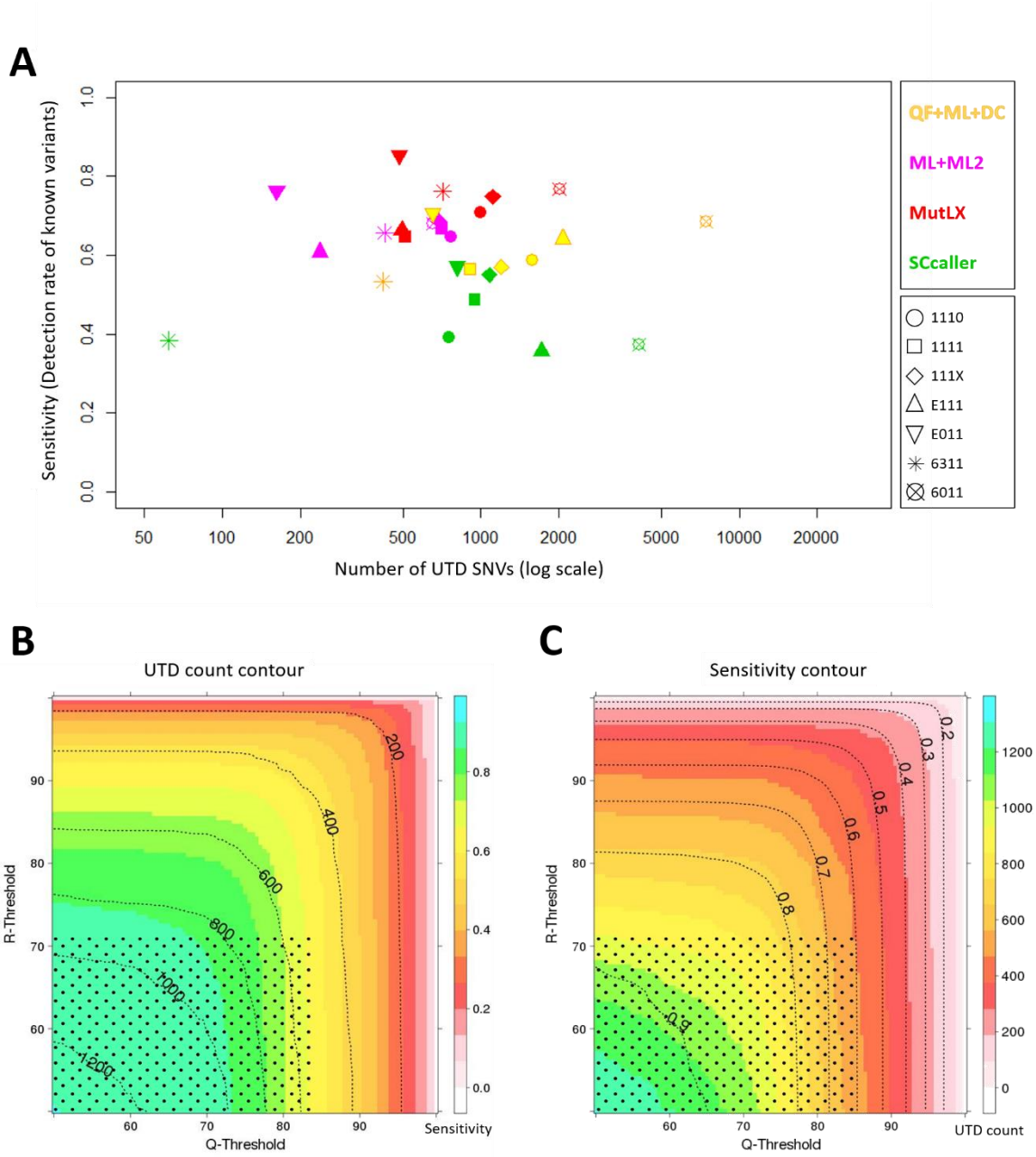


Figure 3.22. Application of MutLX to DigiPico data. (A) MutLX algorithm was applied to all of the DigiPico datasets that had previously been analysed. Q- and R-threshold values for each run were selected based on the 2D ROC curve of each model. For each model, the cross section of the ROC curve and the threshold curve was used as the threshold value for that model. Results indicate that MutLX performs in a similar level as the other successful variant filtering algorithms used in this study. MutLX, however, offers flexibility in the selection of two threshold values, combination of which can be used to optimise the analysis outcome to better match with the requirements of the experiment and project. This can be achieved by evaluating the Q- and R-threshold contour plots. **(B,C)** Contour and level plots from the result of analysis of run 1111 using MutLX algorithm. The two plots are alternative representation of the same data. Dotted area indicates the region within which the previously described 'true' UTD variant from run 1111 (**chr1:207222565.G>C**) (Figure **3.17**) is correctly identified as a true variant using MutLX.

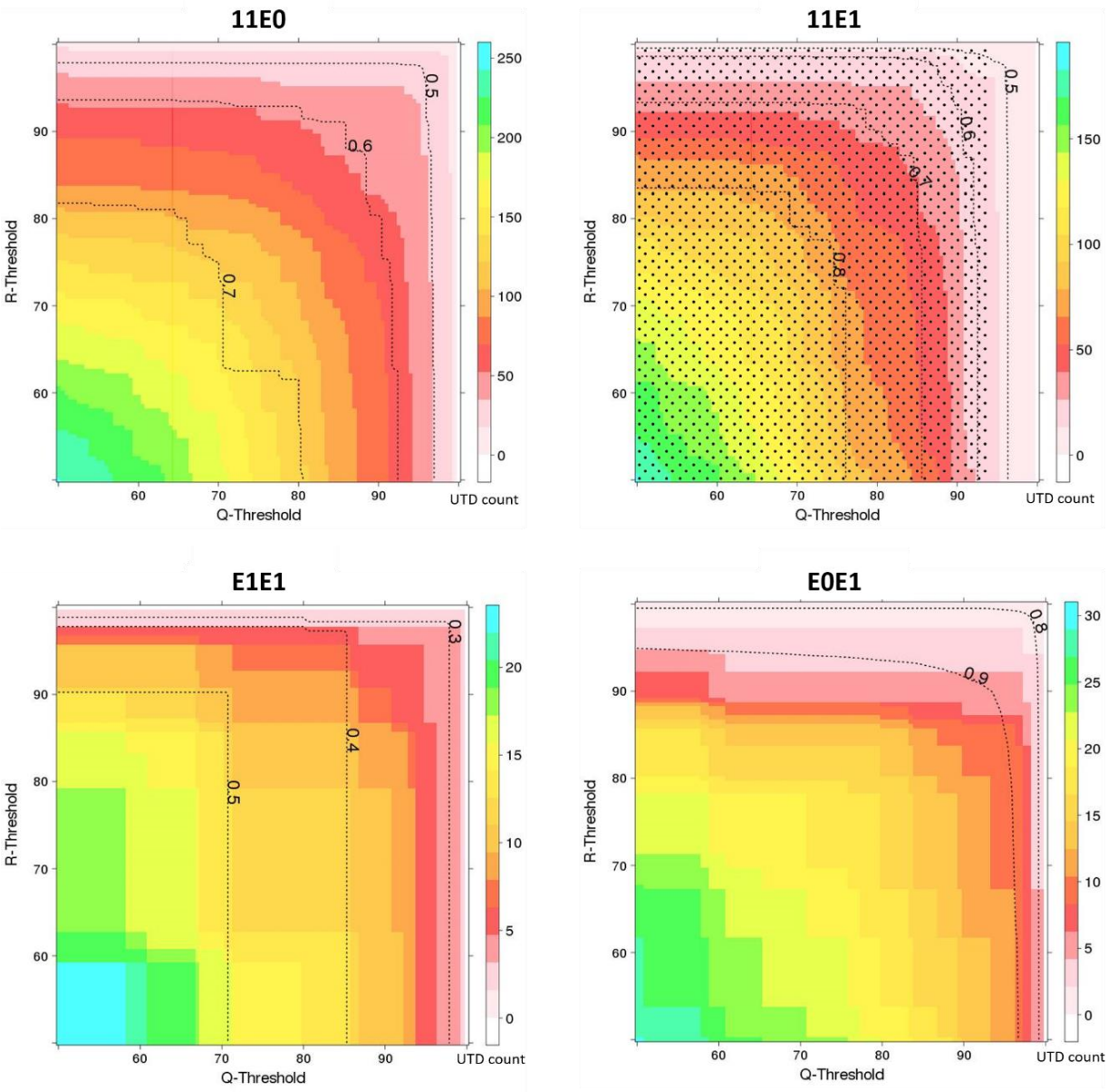


Figure 3.23. Application of MutLX algorithm to DigiPicoExome data. To evaluate the effectiveness of MutLX algorithm in filtering the DigiPicoExome data, it was applied to DigiPicoExome runs 11E0, 11E1, E1E1, and E0E1. Training was **performed** on 20% of the UTD variants from the unfiltered data. The resulting contour plots clearly indicate that increasing the threshold values to up to 95%, would result in elimination of majority of the FP UTD calls, with a small cost on the detection sensitivity. Interestingly, highlighting the area within which the chr1:207222565.G>C can be identified from the 11E1 data, indicates that DigiPico exome data are probably more likely to retain the 'true' UTD variants even if high threshold values are applied.

3.2.12 Accurate prediction of the status of clone-specific SNVs using MutLX

Lastly, visualising the frequency of the different mutation types from the MutLX filtered list of variants indicates that, some of them still seem to present a strong bias in the distribution of variant types which indicates majority of these calls are likely to be FP calls (Figure 3.24). In an attempt to identify a set of Q- and R- threshold values that can specifically be used to address the bias in the mutation frequencies, a series of level plots showing the relationship between the relative frequencies of each of the mutation types with Q- and R-thresholds were generated (Figure 3.25). Interestingly, the results show a clear distinction between the pattern of C>A mutations in run 1111 (where true UTD variants are expected) and run E111 (where true UTD variants are not expected). To check if similar patterns can also be observed in other DigiPico data, adjusting a uniform scale for all of the samples, the level plots of the C>A mutations were generated for all of the available DigiPico datasets (Figure 3.26). The results clearly indicate the appearance of orange-yellow (C>A frequencies of 15-20%) levels with the increase in the Q- and R- threshold values in DigiPico samples obtained from tumour cross section, where true UTDs are expected (1111 and 6311). On the other hand, however, such patterns do not seem to appear in DigiPico libraries that had been prepared from blood or DNA extracts, where true UTDs are not expected (1110, E011, E111, and 6011). This observation suggests that although MutLX is not able to completely remove all of the FP calls from a sample, however, based on the behaviour of C>A mutations in response to increasing Q- and R-threshold values, it would be possible to predict whether a particular sample is expected to have true clone-specific variants or not. This is extremely important when small tumour samples obtained via LCM are being investigated. Whether or not a specific sample has true clone-specific variants can determine the level of SNV heterogeneity in the tumour under study.

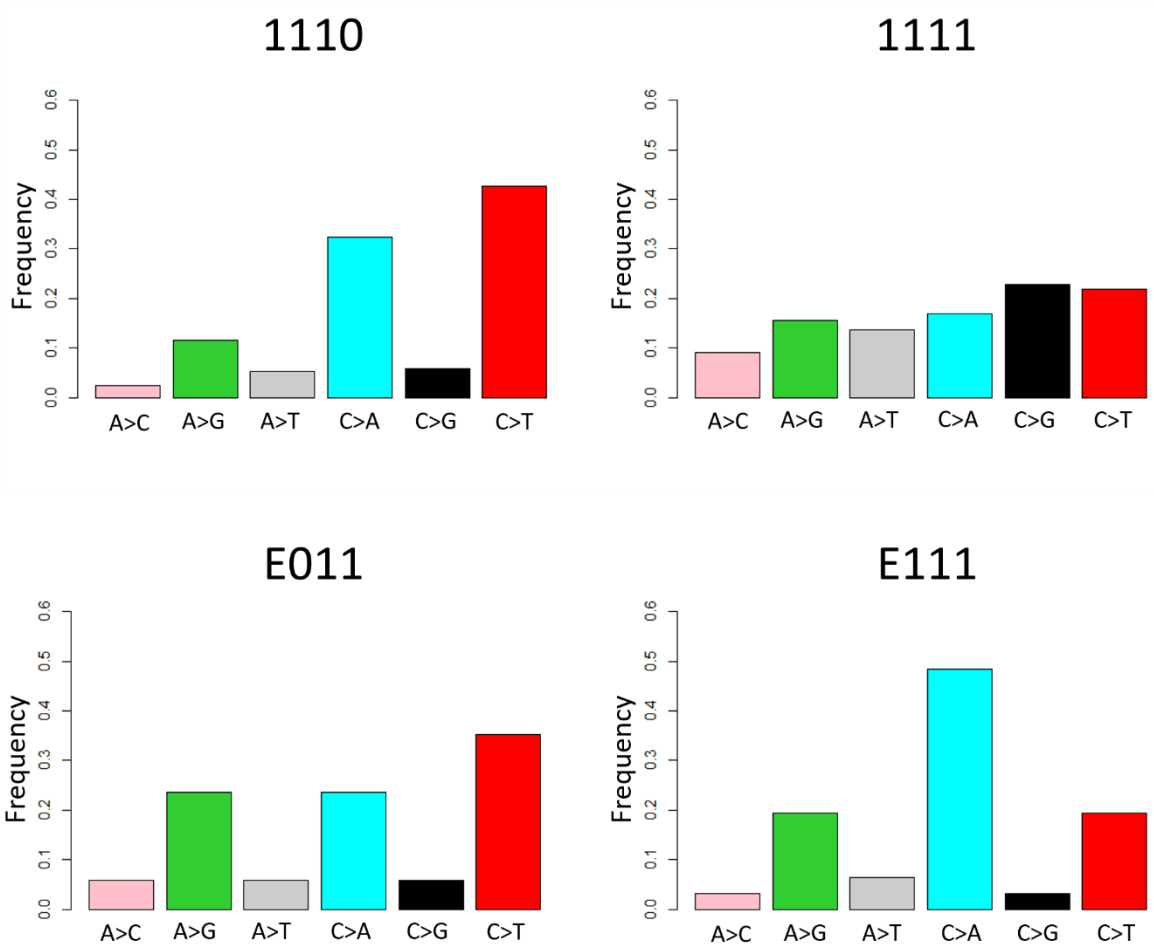


Figure 3.24. Mutation type distribution of MutLX filtered UTD variants. Analysing the distribution of mutation types across various DigiPico samples indicates that while in one case (1111) the bias toward certain types of variants has completely vanished, other runs still seem to suffer from this issue.

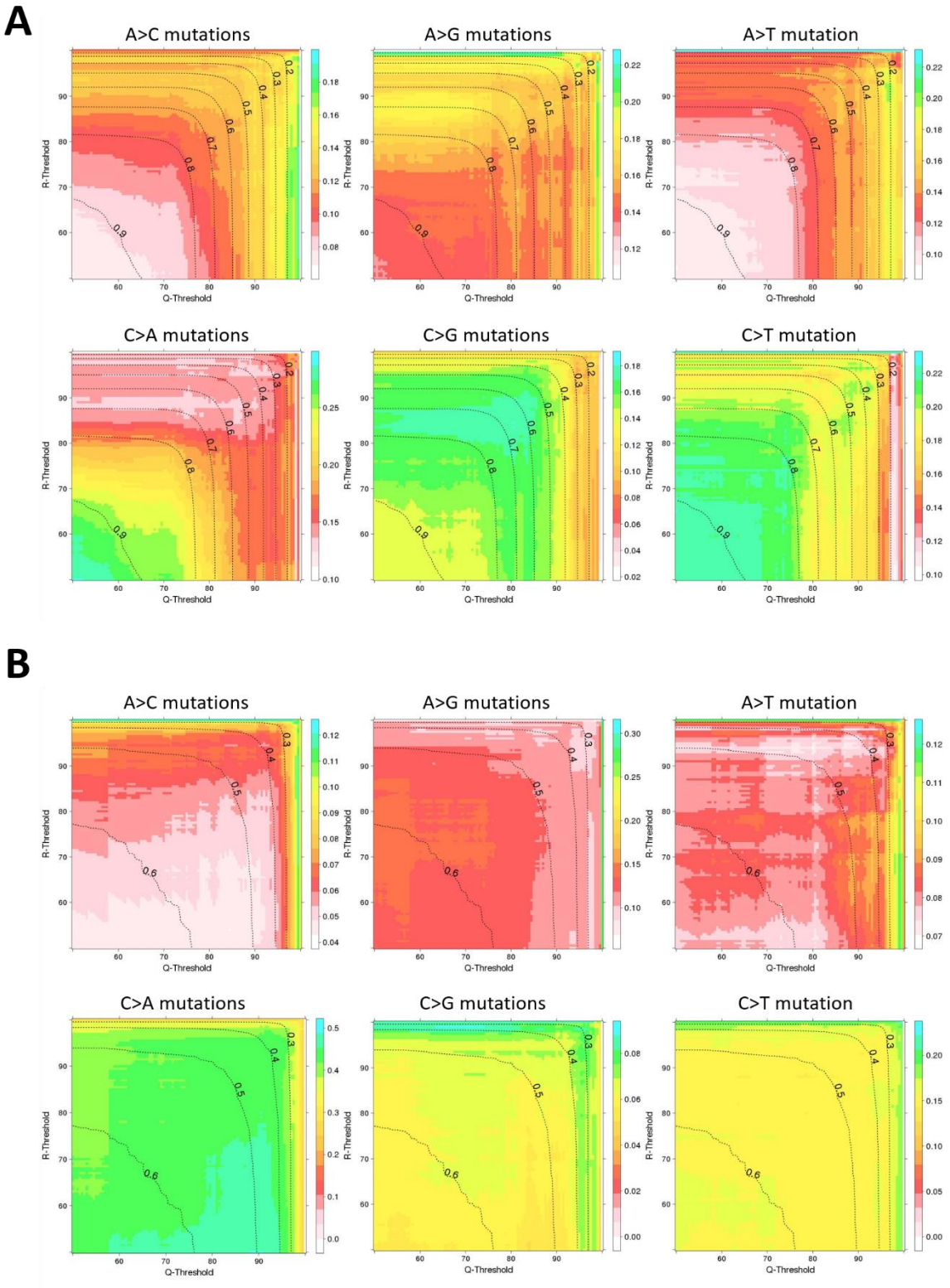


Figure 3.25. Effect of MutLX threshold values on the frequency of different mutation types. Level plots were generated for the relative frequency of each of the 6 mutation types in respect to increasing Q- and R-threshold values for samples 1111 **(A)** and E111 **(B)**. Comparing the patterns of change of frequencies indicates that there is a significant and consistent difference between how C>A mutations are behaving in sample 1111 (where true UTDs are expected) and E111 (where true UTDs are not expected). Contour plots show the sensitivity of detecting known variants at the respective Q- and R-thresholds.

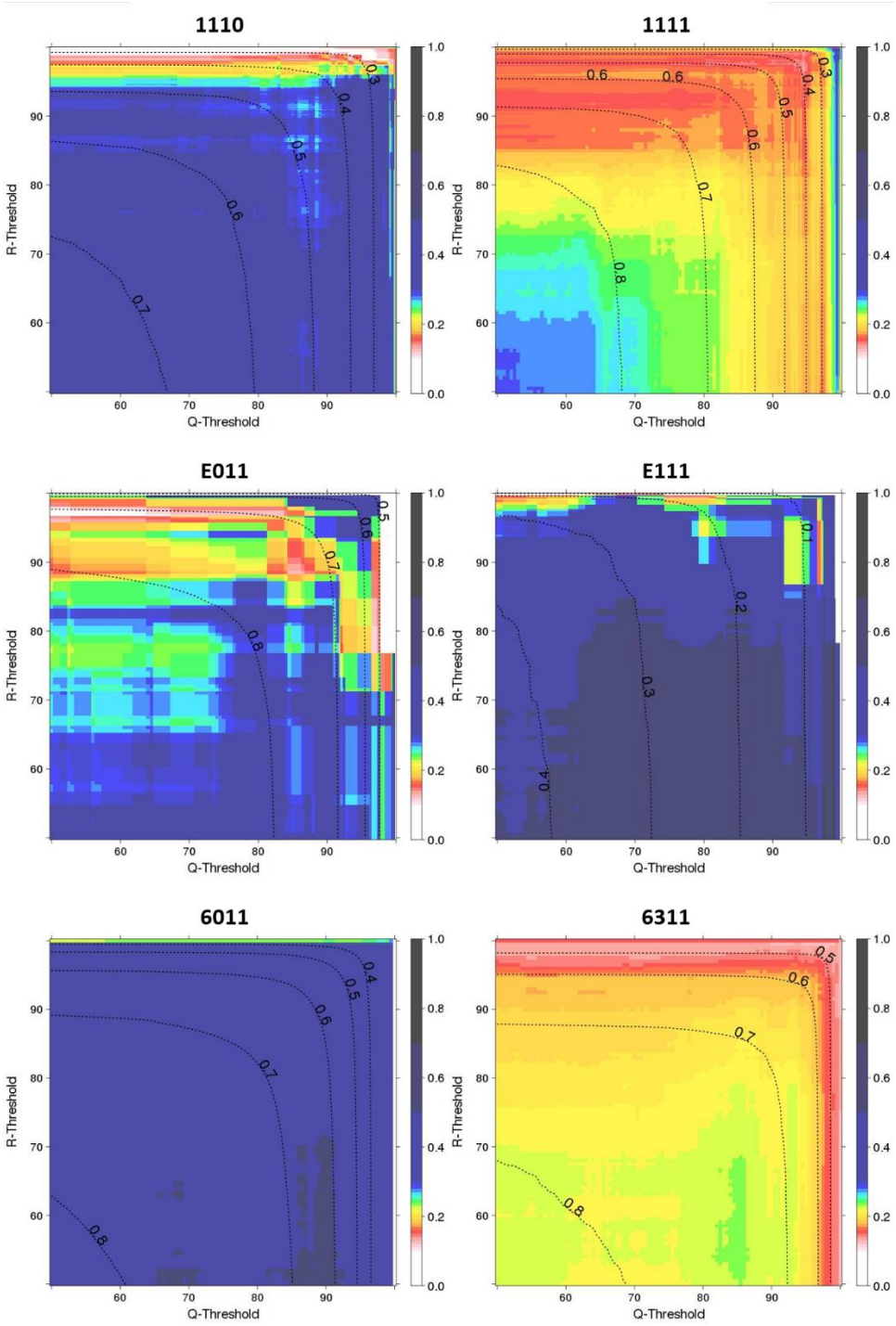


Figure 3.26. Comparison of C>A mutation level plots in various DigiPico libraries. Based on the observation in **Figure 3.25** a unified colour scale was defined for the C>A mutations to better highlight the differences in the range of 10-20% frequency values. Level plots were generated for all of the available DigiPico data. Among these only runs 1111 and 6311 that were prepared directly from cross sections of tumour are expected to have true clone-specific variants. The results clearly show that an increase in the levels of yellow and orange colours in response to an increase in the MutLX threshold values can be used as an indicator to decide whether or not a specific DigiPico sample is expected to have clone-specific variants.

Chapter 4 : Inferring the Evolutionary History of a HGSOC using DigiPico: A case study

In the previous two chapters I aimed at establishing a high throughput, robust and reproducible library preparation workflow for DigiPico sequencing as well as an accurate and sensitive analysis pipeline for DigiPico data. Having achieved the aforementioned goals, I decided to apply the DigiPico method to several samples across a diverse temporal and spatial spectrum from one patient in an attempt to infer the evolutionary history of a high-grade serous ovarian tumour in a comprehensive case study. Screening the HGSOC patients in GO-Target study, patient 11152 was selected for this study. Following the initial diagnosis, the patient has undergone laparoscopic surgery and a small omental biopsy containing visible tumour masses had been collected. This biopsy will be referred to as pre-chemotherapy Omental Mass (OM) in this analysis. At the time of sample collection, a surgical clip had also been placed close to the site of biopsy to mark the area. After 3 cycles of neoadjuvant chemotherapy (NACT) the patient showed a complete response with no visible masses remaining. Following the debulking surgery, the marked area in the omentum, where the initial biopsy had been taken from, was used for microscopic inspection. This site will be referred to as post-chemotherapy Microscopic Residual Disease (MRD) in this study. The results of microscopy indicated the presence of minimal MRD in this site and therefore, 3 more cycles of chemotherapy were administered to the patient. Two years following the completion of the chemotherapy, patient was presented with extensive recurrence tumour in the pelvic and abdominal area. At this point two separate recurrence tumour sites, one from a pelvic mass (PT2R) and one from a pelvic lymph node (PLNAR) were biopsied from this patient. This collection of samples makes a complete series of spatially and temporally related tumour samples which perfectly matches with the scopes of this case study. Therefore, to decipher the evolutionary history of the tumour from patient 11152 and to evaluate the importance of MRD in the emergence of recurrence, two sets of experiments were performed on the frozen tumour samples from this

patient: Firstly, various LCM samples from the pre-chemotherapy OM site were used for DigiPico library preparation and sequencing (Figure 4.1). To study the evolutionary history of this tumour, data from these samples were then compared with the DigiPico data from the recurrence tumours for both CNA and SNV mutations. Next, all of the pre-chemotherapy DigiPico libraries from the OM were used for exome capture to prepare their respective DigiPicoExome libraries. Moreover, several MRD islets were obtained from the matched post-chemotherapy OM using LCM which were also used for preparation of DigiPicoExome libraries. The sequencing results from these samples were then used to study and infer the relationship between MRD and tumour recurrence in this patient.

4.1 Study of copy number alterations

To decipher the evolutionary history of a tumour, it is first important to understand the make-up of genetic changes within the tumour under study. HGSOC tumours are known to harbour a large number of structural variants (SVs), as well as single nucleotide variants (SNVs) (TCGA, 2011). Therefore, to obtain a complete perspective of evolutionary changes of a HGSOC tumour it is crucial to simultaneously study both SV and SNV mutations of the tumour. Although such studies can easily be performed on the WGS and WES data obtained from the bulk of tumour, these methods often require a large amount of template DNA which may not be obtainable from certain types of samples, therefore, limiting the applicability of these methods. Moreover, analysing the data from bulk sequencing data suffers from a main disadvantage: In tumours with a complex clonal composition, studying the bulk of the tumour is likely to mask sub-clonal genetic changes especially in the case of SVs. DigiPico sequencing on the other hand, promises to provide high-quality SV and SNV information from picogram quantities of input material.

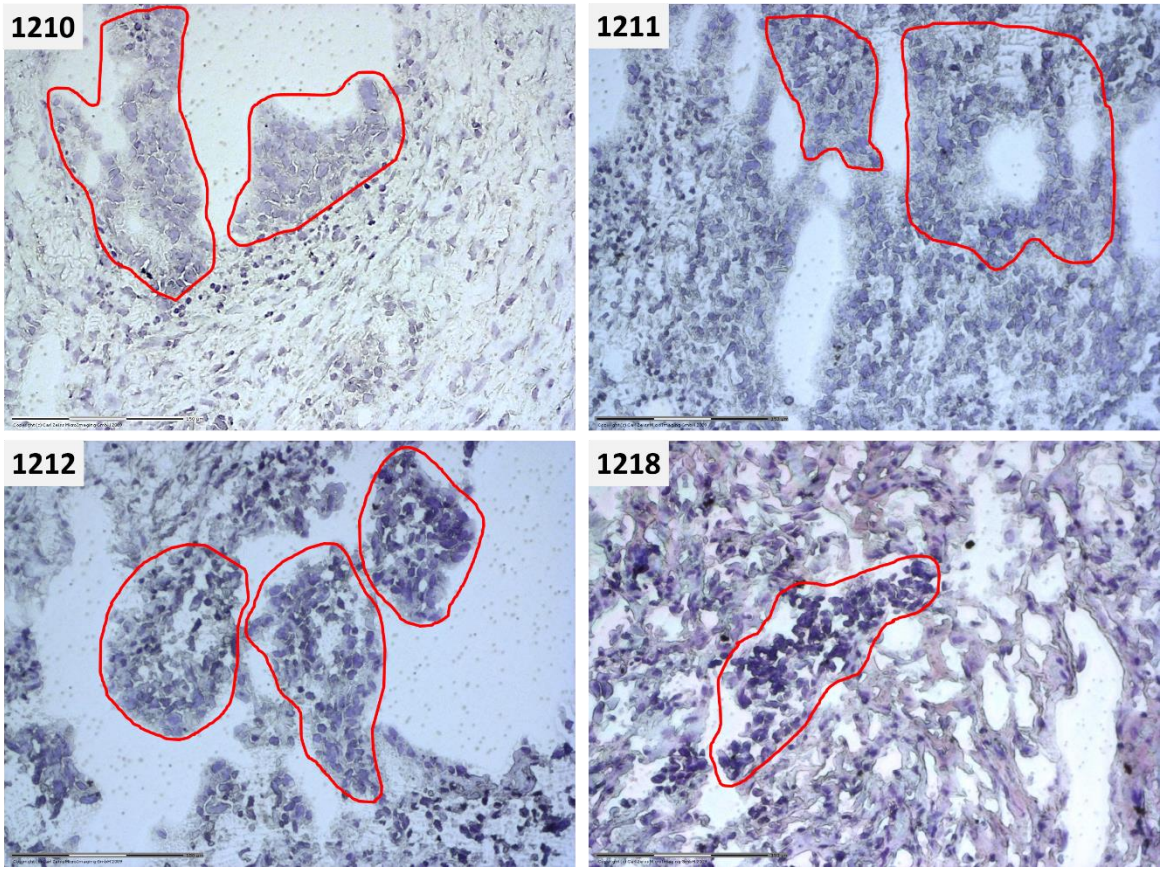


Figure 4.1. Laser-Capture Microdissection of tumour islets from a pre-chemotherapy mass of tumour. Frozen OCT-embedded tumour sections at a thickness of 10-15 µm were prepared using a cryostat microtome. Sections were stained using an alcohol-based dye (cresyl violet) to avoid damage to nucleic acids. A Zeiss laser-capture microscope was used to isolate the marked tumour islets in each image. Images are representative of samples used in this study. All of the marked areas in each image were combined to make one library. Numbers in the top left corner indicate the sample and library identity number. The scale indicator at the bottom left of each image represents 150 µm. From the images it is clear that even in a pre-chemotherapy tumour mass a large number of stromal normal cells can surround the tumour cells. This explains the high rate of normal cell contamination that is often observed in bulk WGS of tumour masses. In DigiPico however, the specific isolation of tumour islets using LCM addresses this issue.

Therefore, using DigiPico, rather than sequencing the bulk of the tumour, it would be possible to obtain sequencing data from small clonal *tumour islets* that reside within the mass of a tumour. As, in solid tumours, nearby cells are very likely to be clonally related, the study of individual tumour islets can reveal intra-tumour heterogeneity that may exist in a tumour mass. Hence, to better understand the SV and SNV mutations of the pre-chemotherapy tumour from patient 11152, DigiPico sequencing was performed on nine tumour islets (runs 1210-1218) isolated from the frozen sections of the pre-chemotherapy OM via LCM (Figure 4.1). After pre-processing of the data, eight out of the nine DigiPico runs were found to have a high enough quality to be used for further analysis (run 1217 was eliminated). Moreover, as a control, sections from the same pre-chemotherapy mass of tumour were used for bulk DNA extraction and standard WGS (run 1201).

4.1.1 Micro-heterogeneity of CNAs in the pre-chemotherapy HGSOC

Initially to obtain a general understanding of the CNAs and allelic imbalances in the pre-chemotherapy and recurrence tumour samples of patient 11152, TitanCNA algorithm (Ha, 2014) was applied to the standard WGS data of runs 1201 (pre-chemotherapy Omental Mass; OM) and 1101 (recurrence Pelvic Tumour; PT2R). The results show that, while some of the copy number changes are shared between the OM and PT2R sites, others seem to be strictly present in the PT2R recurrence (Figure 4.2). This observation suggests that OM and PT2R masses might have emerged from the same clones of cells and further progression of the recurrence tumour has resulted in accumulation of more extensive copy number changes in the PT2R site. Moreover, from these observations, it might be inferred that the tumour at the pre-chemotherapy stage has had a more stable genomic structure and the complexity of the copy number changes has increased over time with the evolution of the recurrence tumour. However, analysing the DigiPico data of the eight tumour islets from OM site indicates that these conclusions may not necessarily be true.

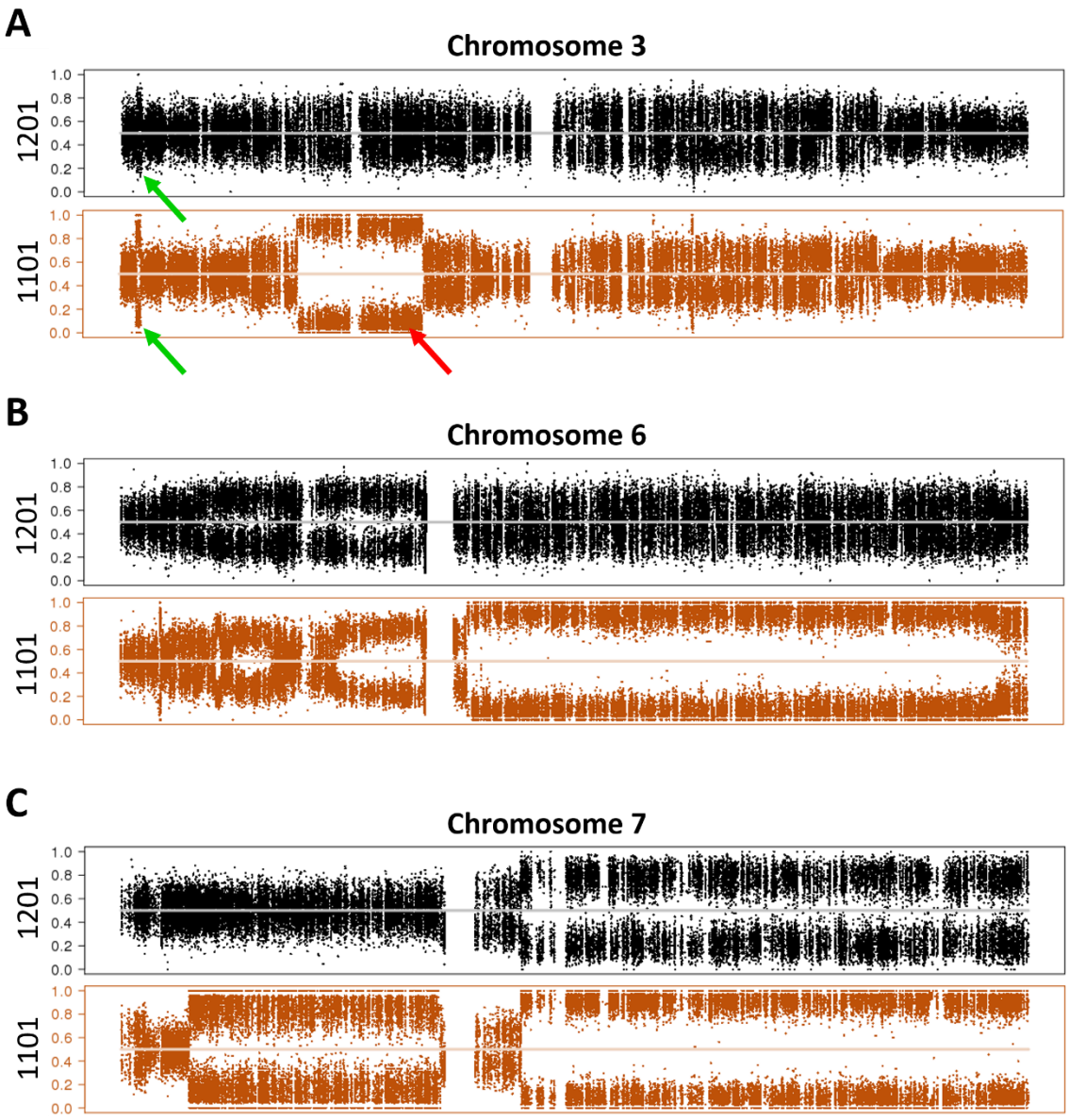


Figure 4.2. Copy number analysis of bulk sequencing data from pre-chemotherapy and recurrence tumour samples. TitanCNA algorithm was applied to bulk WGS data from OM (1201) and PT2R (1101) sites. Only BAF profiles from chromosomes 3, 6, and 7 are presented as examples. In the BAF plots, X axis represents the relative positions across the respective chromosomes and the Y axis shows the frequency of the heterozygous variant alleles. In normal diploid state, BAF values accumulate around the middle line at 50% frequency. At the areas of deletion, loss of heterozygosity (LOH) or duplications, the BAF values deviate from the middle line towards the extremes depending on the extent and the type of the CNA. In these examples, some of the CNAs appear to exist in both samples e.g. small deletion at the distal end of chromosome 3p (**A**, green arrows) or the copy-neutral LOH (CN-LOH) of chromosome 7q (**C**). However, a large number of CNAs only seem to be present in the recurrence sample, e.g. large deletion in 3p (**A**, red arrow) and CN-LOH of chromosome 6q (**B**). This can indicate that the recurrence tumour has experienced an extensive series of copy number alterations through its evolution after the chemotherapy and that the pre-chemotherapy tumour has had a relatively more stable genomic structure.

Comparing the BAF profiles of various DigiPico data obtained via DigiTitan workflow (Section 3.1.3.2) reveals that unlike what had been observed from the analysis of the standard WGS data of the OM site, there seem to be a great degree of diversity in the copy number profiles of various pre-chemotherapy tumour islets (Figure 4.3). This diversity of the copy number changes appears to be in the form of an intra-tumour micro-heterogeneity, whereby, different tumour islets appear to have different sets of copy number changes at varying levels (Figure 4.3A-B). Interestingly, comparing the BAF profiles of tumour islets at regions of heterogeneity with the BAF profiles of the recurrence tumour reveals that some of the copy number changes that were seemingly absent in the bulk sequencing data of the OM are indeed present in some of the tumour islets (Figure 4.3A-B, highlighted in magenta). Moreover, further comparisons also reveal the presence of islet-specific CNAs that are only present in one or few of the islets but are absent in the recurrence tumour or the bulk data of the OM (Figure 4.3A-B, highlighted in green). Lastly, there also appears to exist copy number changes that are present in the recurrence, bulk of the OM, and in all but one of the pre-chemotherapy tumour islets (Figure 4.3B, highlighted in orange). The combination of these observations indicates that, in this tumour, the pre-chemotherapy omental mass is composed of a heterogeneous mix of various clones with diverse sets of CNAs. As some of these heterogeneous CNAs are also present in the recurrence tumour, it is plausible to assume that clones with such CNAs are likely to be more closely related to the recurrence tumour than what would have been inferred from the analysis of the bulk sequencing data of the OM. Nevertheless, the possibility of convergent evolution in the case of some of these CNAs must also be considered. For instance, the LOH event in the region of chr3:p14.1-p22.2 is present in both, the recurrence sample (1101) and in the pre-chemotherapy islet 1218, however, only a smaller version of this LOH in the region of chr3:21.2-p22.2 appears to be present in the 1211 islet (Figure 4.3A). This can suggest that the most recent common ancestor (MRCA) of 1101/1218 is closer to the recurrence than the MRCA of 1101/1211.

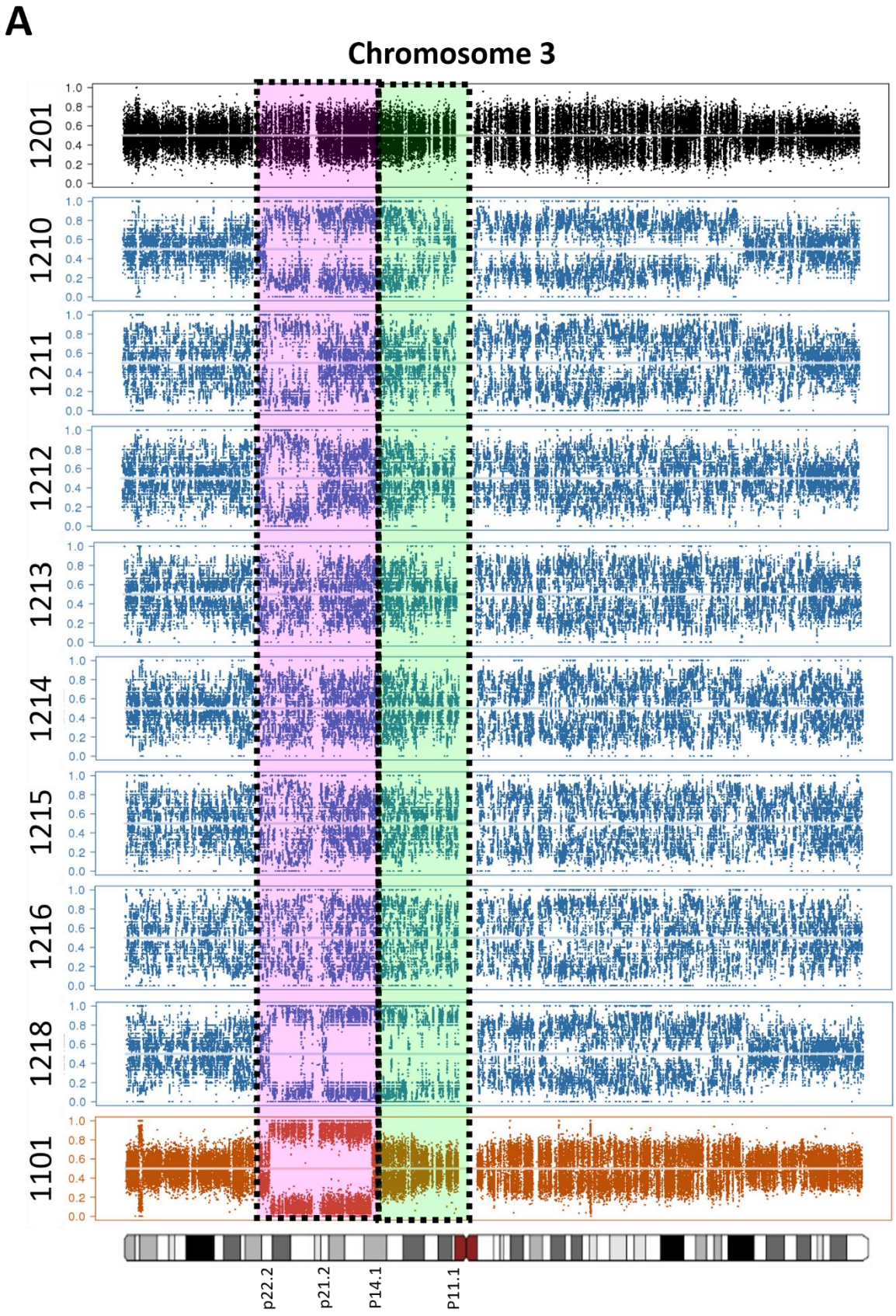


Figure 4.3. Analysis of CNAs across various pre-chemotherapy tumour islets using DigiTitan.
(Panel B follows in the next page)

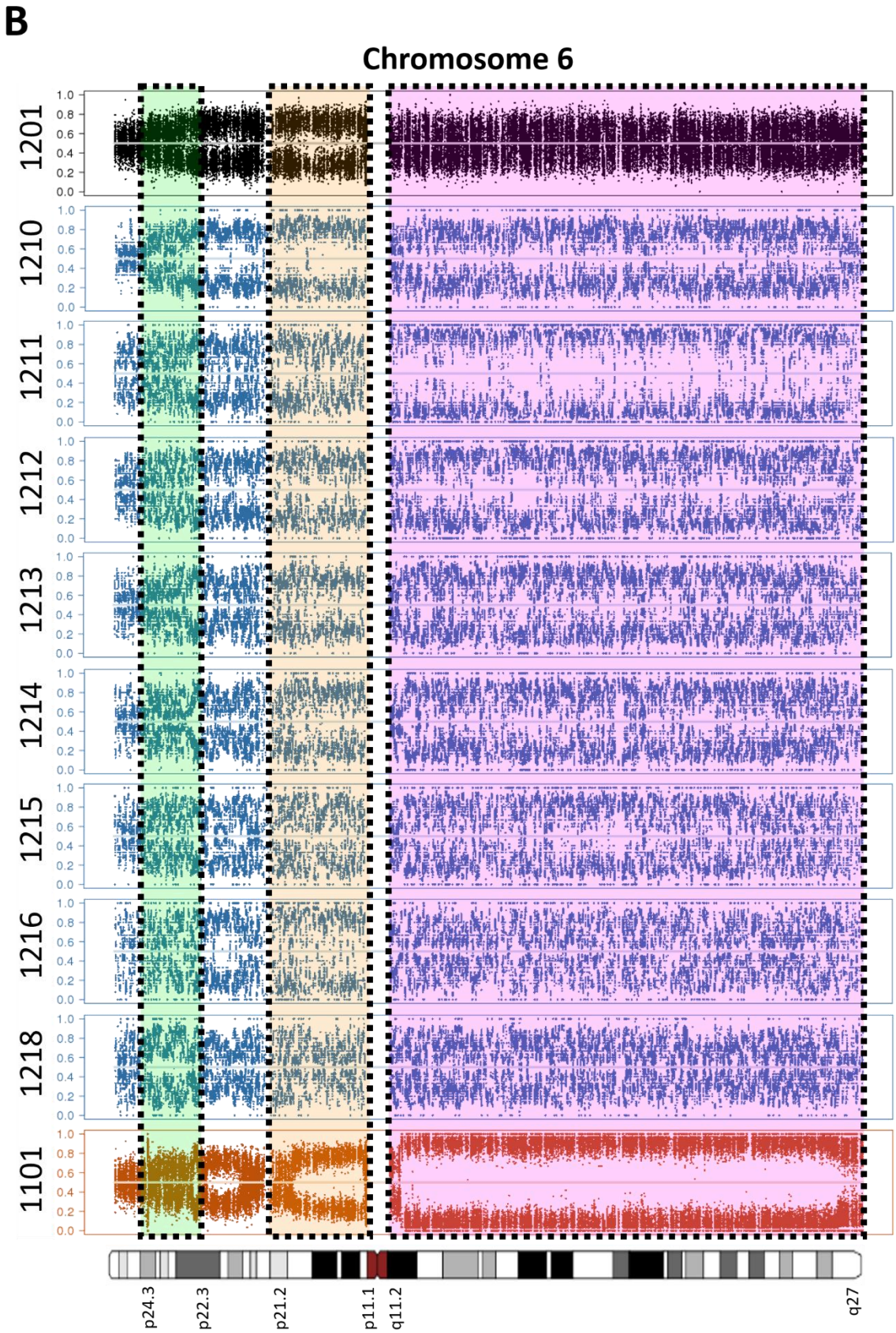


Figure 4.3. **Analysis of CNAs across various pre-chemotherapy tumour islets using DigiTitan. (Cont'd.)**
(Panel C follows in the next page)

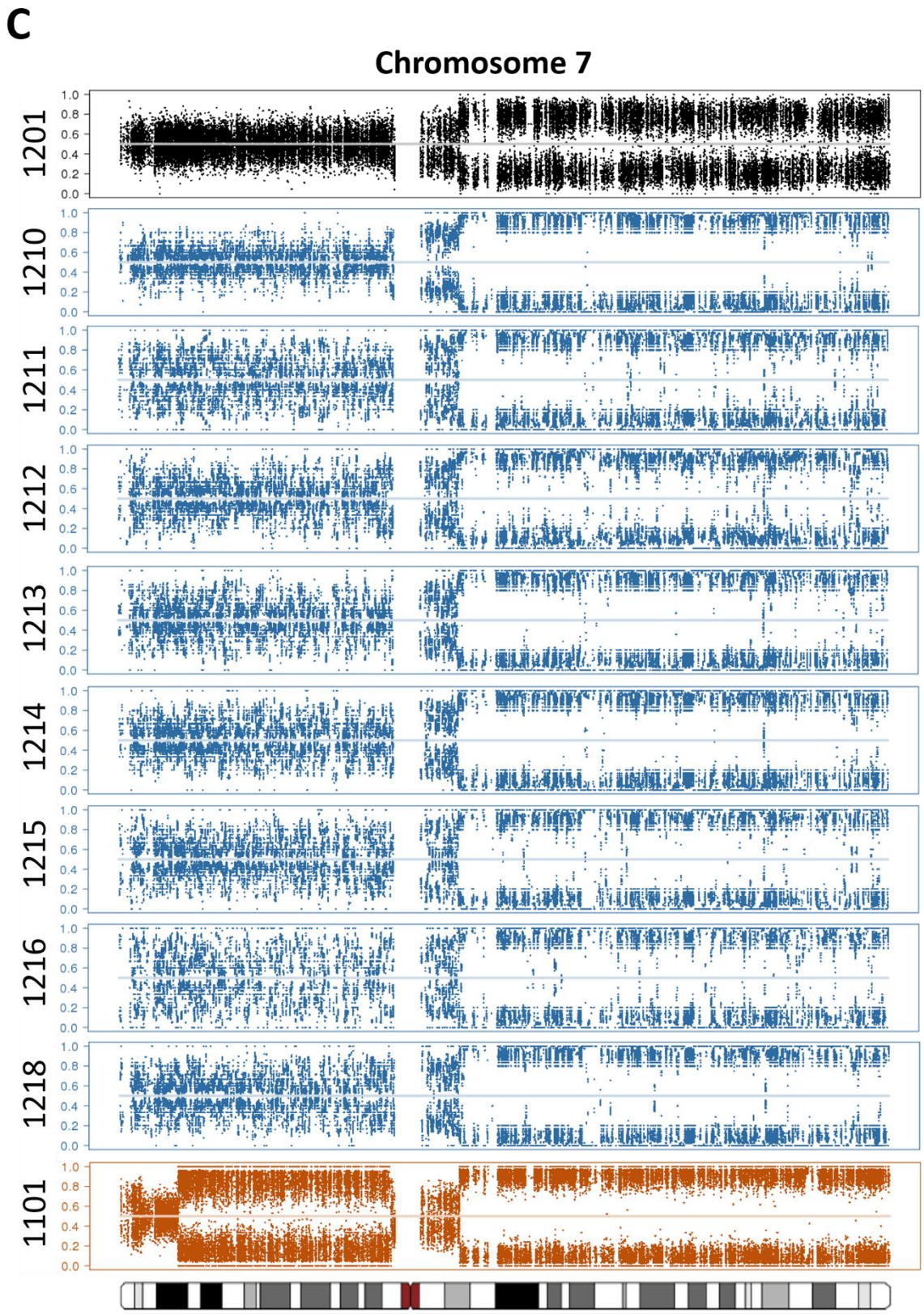


Figure 4.3. Analysis of CNAs across various pre-chemotherapy tumour islets using DigiTitan. (Cont'd.)
(Legend follows in the next page)

Figure 4.3. Analysis of CNAs across various pre-chemotherapy tumour islets using DigiTitan. Eight pre-chemotherapy tumour islets were isolated from the OM using LCM and were used for DigiPico sequencing followed by DigiTitan analysis. Comparing the resulting BAF profiles from the tumour islets (blue), with the BAF plots obtained from bulk WGS data of OM (black) and PT2R (orange) clearly shows that unlike what had been observed in the bulk sequencing data of OM, some of the apparently recurrence specific CNAs are indeed also present in some of the pre-chemotherapy tumour islets. **(A)** The large deletion event on the short arm of chromosome 3 (p14.1-p22.2) which shows a complete absence in the bulk sequencing results of OM site, appears to be present in majority of tumour islets from the OM site (highlighted in magenta). However, although there is an apparent consistency in the position of the distal breakpoint of this deletion, the proximal breakpoint seems to be extremely diverse in different samples (highlighted in green). This either indicate a series of consecutive deletions or shows that these events are a result of convergence evolution. In the latter case, the area close to the distal breakpoint is likely to be target of this CNA event. **(B)** Only about half of the tumour islets seem to present the 6q CN-LOH event that appears in the recurrence (highlighted in magenta). This suggests a likely closer evolutionary relationship between some of the tumour islets and the recurrence tumour. On the other hand, a potential duplication event at the distal end of the chromosome 6p appears to only be present in tumour islet 1218 (highlighted in green) while a proximal event on the same chromosome arm seems to be absent in this islet (highlighted in orange). This potentially indicates that islet 1218 is evolutionary more distant to the recurrence tumour and further confirms the presence of an extensive CNA heterogeneity in the pre-chemotherapy OM. **(C)** Chromosome 7 appears to have a large CN-LOH event nearly at its entirety in the recurrence sample, but the pre-chemotherapy tumour islets seem to consistently have only lost on copy of the chromosome 7q. This indicates that while the loss of 7q may be a clonal event that has occurred early in the evolution of the tumour, the loss of 7p has likely occurred during the progression of the recurrence sample. Moreover, the noise-free profiles of 7q CN-LOH indicates that all of the tumour islets in this study appear to be devoid of normal cell contamination.

On the other hand, based on the events on chromosome 6 (Fig. 4.3B), the opposite of this hypothesis can be inferred as the large LOH of chromosome 6q is only present in the recurrence sample and islet 1211 and seems to be absent in 1218 islet. Such contradictions can however, be addressed if it is assumed that some of the similarities between the CNAs of different sample can be a result of convergent evolution. If true, this further indicates that there is a high level of CNA heterogeneity in the pre-chemotherapy tumour of this patient.

4.1.2 CNA diversity in the recurrence HGSOC tumour

Following to the previous analysis, to better understand the status of the CNAs in the recurrence tumour, the BAF profiles from various DigiPico data of the PT2R site (runs 1110 and 111X) were compared with the results from the bulk WGS data of the same site (run 1101). The results clearly show a high degree of consistency between the BAF profiles obtained from the DigiPico and bulk WGS data with a near perfect match between the BAF profiles (Figure **4.4**). Considering the noise-free BAF profiles from the bulk WGS data as well as the consistency of DigiPico results with the bulk results, it can be concluded that the PT2R tumour mass is most likely homogenous for CNAs. Therefore, suggesting that this recurrence site might have emerged from a subset of clonally related cells that have already had a high level of CNAs with little to no new CNAs occurring after the establishment of the tumour mass. Moreover, to also evaluate the diversity of CNAs between different recurrence sites, the BAF profiles from DigiPico data of a second recurrence tumour from a different site (Pelvic Lymph Node; PLNAR) were compared with respective profiles from PT2R and OM sites (run 1511). The results of this comparison showed a strong divergence between the two recurrence sites revealing a large number of CNAs that are exclusively present in the PT2R site (Figure **4.4**B, red arrows). Although PLNAR site also seems to have some distinguishing CNA events, comparison with 1210 pre-chemotherapy islet shows that majority of these CNAs had also been present in the pre-chemotherapy tumour (Figure **4.4**B, green arrows).



Figure 4.4**. Genome-wide comparison of copy number alterations in recurrence tumours.**
(Panel B follows in the next page)



Figure 4.4. Genome-wide comparison of copy number alterations in recurrence tumours. (Cont'd.)
(Legend follows in the next page)

Figure 4.4. Genome-wide comparison of copy number alterations in recurrence tumours. (A) Comparing the BAF profiles of two different DigiPico runs with the BAF profiles obtained from a respective bulk WGS data across the genome shows a near perfect match between the results. This not only shows that DigiPico can be used for accurate prediction of CNAs from picogram quantities of DNA but also indicates that, unlike the OM site, the PT2R recurrence tumour has a homogenous structure in terms of CNAs. 111X plot had been obtained by combining data from runs 1111 and 1112. **(B)** Genome wide comparison of copy number alterations between PT2R recurrence, PLNAR recurrence, and islet 1210 of pre-chemotherapy shows that the PLNAR sites is more closely related to the OM site than the PT2R site. Therefore, the distinctive patterns of CNAs between the two chemotherapy sites suggests that their post-chemotherapy evolution might have been from two independent events. Red arrows mark the CNA events that are present in PT2R but are absent in PLNAR. Green arrows mark the CNA events that are present in PLNAR but are absent in PT2R.

This observation suggests that the two recurrence sites are most likely a result of two independent recurrence emergence events, with the PT2R site having diverged from the OM at an earlier time point in comparison to the PLNAR site. Moreover, in agreement with the previous observations, the presence of both, a PT2R specific CNA and a PLNAR specific CNA on chromosome 3 of the 1210 islet strongly indicates that at least one of these events must have been a result of convergent evolution (Figure 4.4B). Although, based on the assumption that the PLNAR site is evolutionarily closer to the OM site, and the more extensive 3p LOH in islet 1210 compared to PT2R, it is more likely that the 3p LOH event in the pre-chemotherapy islet 1210 is analogous to the 3p LOH event in the recurrence. This conclusion, if true, indicates that 3p LOH must confer a great evolutionary advantage for the growth of this tumour (Gao *et al.*, 2017). Such an advantage can for instance arise from the elimination of the functional copy of a tumour suppressor with an already existing heterozygous mutation or might be secondary to an essential genomic rearrangement.

4.2 Study of single nucleotide variations

4.2.1 Inference of evolutionary relationship of tumour islets based on SNVs

To obtain a general perspective on the status of SNVs in this tumour, somatic variant calling was performed on the WGS data of both the pre-chemotherapy OM and the recurrence PT2R samples. Screening of the variant calling results for somatic mutations in genes previously reported to be important in ovarian cancer (TCGA, 2011) resulted in the identification of a pathogenic mutation in the TP53 gene (C275F) as well as a point mutation in the BB5 SOX2-regulatory-region (chr3: 182218102.C>G) (Hellner *et al.*, 2016). Moreover, a pair of somatic variants, upstream to the FAT3 gene (chr11:92084283.T>C, chr11:92084337.T>A), were also identified in both of the samples. Although, deleterious DNA breakpoints were identified in both the NF1 and RB1 genes, none of the other investigated ovarian cancer related genes (BRCA1/2, CSMD3, CDK12, BRAF, PIK3CA, N/KRAS, CDKN2A, PTEN, FANCA, CTNNB1, ERBB2, AKT1) were found to have somatic SNVs in this patient.

Next, the four aforementioned SNVs as well as six more somatic variations that had a relatively good coverage in all of the sequencing data were used to generate a heatmap of variant allele frequencies for each sample. The results of this analysis show that some of the studied somatic variants, such as mutations in TP53 or BB5, are consistently present in all of the studied tumour samples (Figure 4.5A). These SNVs are likely to be clonal SNVs that have occurred early on during the evolution of the tumour. On the other hand, other variants, such as those in the upstream region of FAT3, are likely to have occurred at a later stage during the evolution of the tumour as they are only detectable in some of the pre-chemotherapy islets as well as in the recurrence. Similar to the observations from the study of CNAs, these results suggest that some of the tumour islets in the pre-chemotherapy OM site might have a more direct evolutionary relationship to the recurrence tumours compared to what would have been estimated from the analysis of the bulk WGS data. The extent of these relationships could potentially be inferred from an evolutionary tree that is drawn based on the absence or presence of somatic SNVs from across all of the samples. To obtain this evolutionary tree, initially, *de novo* somatic variant calling was performed on all of the available sequencing data from this patient. After application of MutLX algorithm (Section 3.2.10) to the DigiPico data, the resulting variants from across all of the samples were combined to generate a combined list of somatic variants. This list was then used as a reference for variant re-calling from all of the samples to determine the genotypic status of every somatic mutation in each sample. Next, after removing the variant positions that are not covered by at least 10 wells in all of the samples, the results were used to generate an evolutionary tree using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm (Figure 4.5B). The resulting tree was found to be in agreement with the pre-knowledge about these samples as well as some of the observations from the analysis of the CNAs. For instance, as expected, all of the PT2R sequencing data (1101, 1110, 1111, and 1112) clustered together into a distinct branch. Moreover, the organisation of the tree suggests that the PLNAR (1511) sample is likely to be an independent recurrence event and appears to be more closely related to some of the tumour islets than to the PT2R recurrence.

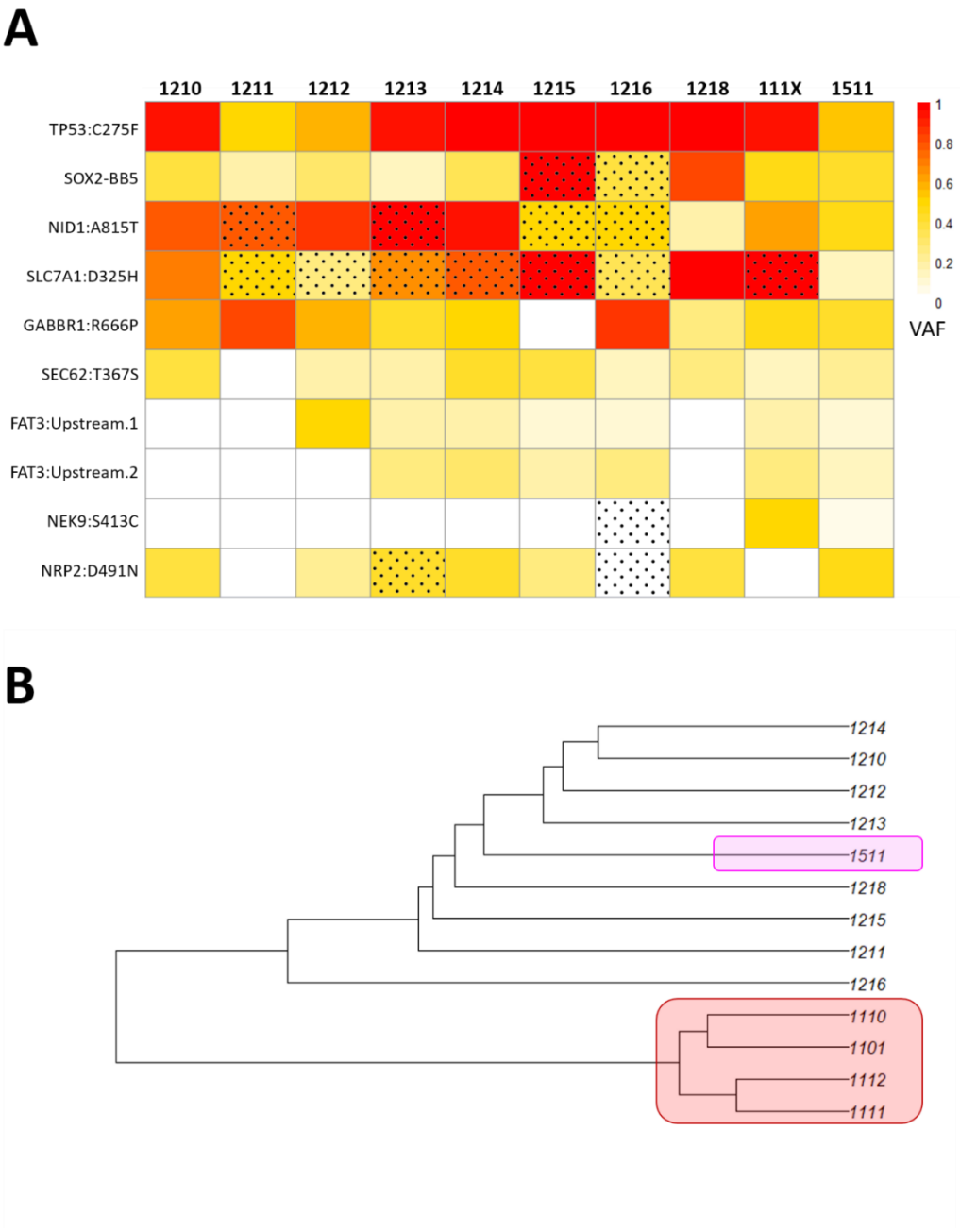


Figure 4.5. Inference of evolutionary relationship of tumour islets with recurrence based on SNVs. Various somatic single nucleotide variants were identified from the WGS data of PT2R and were used for genotyping of samples with DigiPico data. **(A)** The VAF of each of these variants is presented in a heatmap. It is clear that while some of these SNVs such as the TP53:C275F mutation seem to be clonal and present in all of the samples, others appear to be specific to only some of the samples. Dotted areas indicate samples with low coverage (less than 5 wells) at the respective locus. **(B)** Analysing the status of various SNVs across all of the samples followed by hierarchical-clustering of the samples using UPGMA algorithm results in an evolutionary tree that can reflect the evolutionary history of the tumour. As expected, all of the samples from PT2R site cluster together (samples 1111 and 1112 were analysed individually rather than in the form of 111X). Moreover, consistent with the observation from the CNA analysis (**Figure** 4.4**B**) the PLNAR (1511) recurrence seems to be closer to some of the pre-chemotherapy islets, indicating that the PT2R and PLNAR might have been a result of two independent recurrence emergence events.

Nevertheless, despite the apparent consistency that is observed in the outcome of these analyses, this tree does not necessarily reflect the exact series of events that have occurred during the evolution of this tumour. This is mainly due to the fact that some of the pre-chemotherapy islets in this study might have been formed from a mix of multiple clonal lineages. Such islets are likely to appear on the wrong branches when using hierarchical clustering methods which, to some extent, interferes with the evolutionary interpretations that could be drawn from such analyses.

4.2.2 17p LOH might precede the TP53 mutation in patient 11152

While the results of the SNV analysis of this tumour seem to agree with standard assumptions in regards to tumour evolution, the visualisation of VAFs of somatic mutations in all of the studied samples (Figure 4.5A) has revealed an interesting peculiarity in regards to the TP53 mutation. Because of a CN-LOH (Copy Neutral LOH) event at chromosome 17p in this patient, it is expected that all of the germline variants, as well as all of the clonal somatic mutations (such the TP53:C275F) in this region to have VAF values at frequencies close to 1. While majority of the studied tumour islets appear to have VAF values close to 1 for TP53, three of the DigiPico samples, 1211, 1212, and 1511, seem to deviate from this expected frequency. Classically, such observations are usually attributed to normal cell contamination of the tumour sample (Labidi-Galy *et al.*, 2017). In this case however, based on the observations from the CNA analysis (Figure 4.3C), all of the tumour islets are expected to be composed of 90-100% tumour cells and therefore, normal-cell contamination is unlikely to be the reason for the observed frequencies in runs 1211 and 1212. To further confirm this, the BAF plots of chromosome 17 from the DigiTitan analyses of these islets were compared with the respective BAF plot from run 111X as control (Figure 4.6A). The results show that while the low VAF of the TP53 mutation in run 1511 can be because of normal cell contamination, runs 1211 and 1212 do not appear to suffer from this issue.

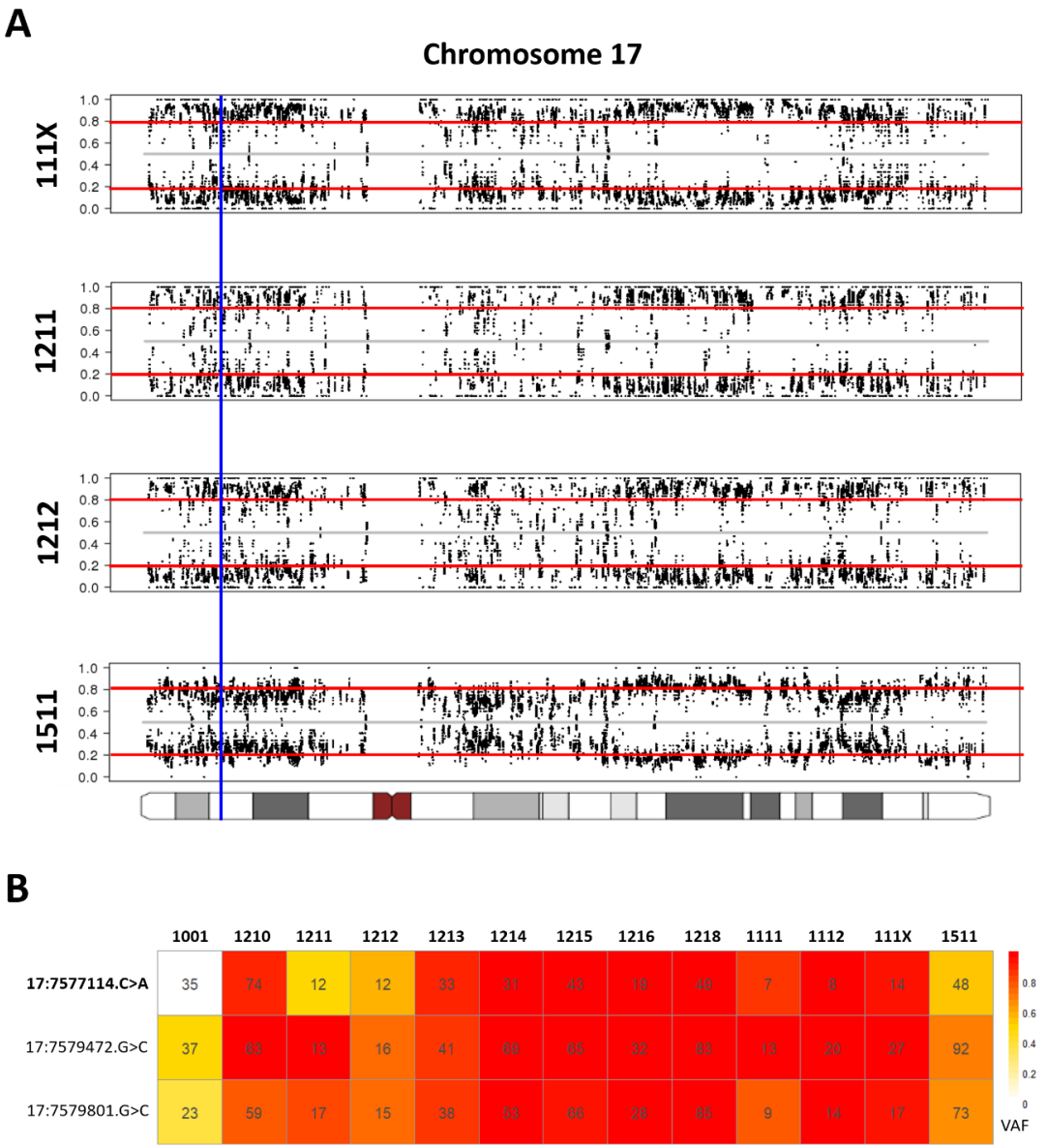


Figure 4.6. Investigation of the TP53:C275F status in 1211, 1212, and 1511 samples. Since the VAF of the TP53:C275F mutation in samples 1211, 1212, and 1511 does not seem to match with the expected frequencies (**Figure 4.5A**), these samples were studied further to investigate the cause for this discrepancy. **(A)** Since BAF plots represent the frequency of heterozygous mutations across the whole chromosome, when visualised at regions of LOH, they can be used to estimate the extent of normal cell contamination in a given sample. In the regions of LOH, the more VAF values deviate from 50%, the less likely it is that the respective sample had been contaminated with normal cells. Analysing the BAF plot of chromosome 17 from samples 1211, 1212, and 1511 as well as 111X as control, reveals that while sample 1511 might have some level of normal cell contamination, runs 1211 and 1212 are unlikely to suffer from this issue. Red horizontal lines indicate 0.8 and 0.2 allele frequencies. Blue vertical line shows the approximate position of TP53 gene on chromosome 17p. **(B)** Further to the previous observation, the VAF of two nearby phased mutations, both of which were also on the TP53 gene, were analysed across all of the pre-chemotherapy tumour islets. Results show that while run 1511 shows clear signs of normal cell contamination, islets 1211 is very likely to be composed of only tumour cells. These results indicate that in islet 1211, there likely exist malignant cells that have a 17p LOH without the presence of the TP53:C275F pathogenic somatic mutation. Numbers in the heatmap indicate the depth of coverage at the respective locus.

Assuming that these results are not an artefact of DigiPico sequencing, it can be hypothesized that, in this patient, the mutation in TP53 gene has occurred after the 17p LOH and that islets 1211 and 1212 are composed of multiple clonal lineages, some of which still have an intact version of the TP53 gene on one chromosome arm while having lost the other arm of the chromosome 17p. To test this hypothesis, two nearby heterozygous SNPs from this patient (both of which also resided within the TP53 gene) were used to generate a new VAF heatmap from various DigiPico libraries (Figure 4.6B). The results confirm that the lower TP53 VAF for runs 1511 and 1212 could be due to normal cell contamination. However, as it appears that in islet 1211 both of the heterozygous SNPs in the TP53 gene have high VAF of values, it can be concluded that the 17p LOH might have already occurred in the majority of the cells in this islet, while only some of these cells seem to have obtained the TP53:C275F mutation.

4.2.3 Presence of SNV micro-heterogeneity in the recurrence but not the pre-chemotherapy HGSOC tumour

Further to the analysis of SNVs in this tumour, the extent of intra-tumour heterogeneity of SNVs amongst the studied tumour islets was evaluated. For this purpose, the contour plots of C>A mutation frequencies against Q- and R-threshold values from MutLX were prepared for all of the pre-chemotherapy tumour islets as well as for all of the DigiPico runs from the recurrence tumours (Figure 4.7). As shown previously (Section 3.2.12) the structure of these plots can indicate whether or not 'true' clone specific SNVs can be identified in the respective DigiPico data. The resulting plots clearly show that, all of the recurrence tumour data are expected to have reliable clone specific SNVs that are not present in either of the WGS data from OM and PT2R. On the other hand, however, the results from tumour islets show that none of the pre-chemotherapy tumour islets in this study appear to have an identifiable set of reliable clone-specific SNVs. This indicates that nearly all of the 'true' somatic variants in all of the tumour islets have some extent of representation in the bulk WGS data from the OM and PT2R and that the islet-to-islet SNV differences appear to be negligible among the pre-chemotherapy tumour islets.

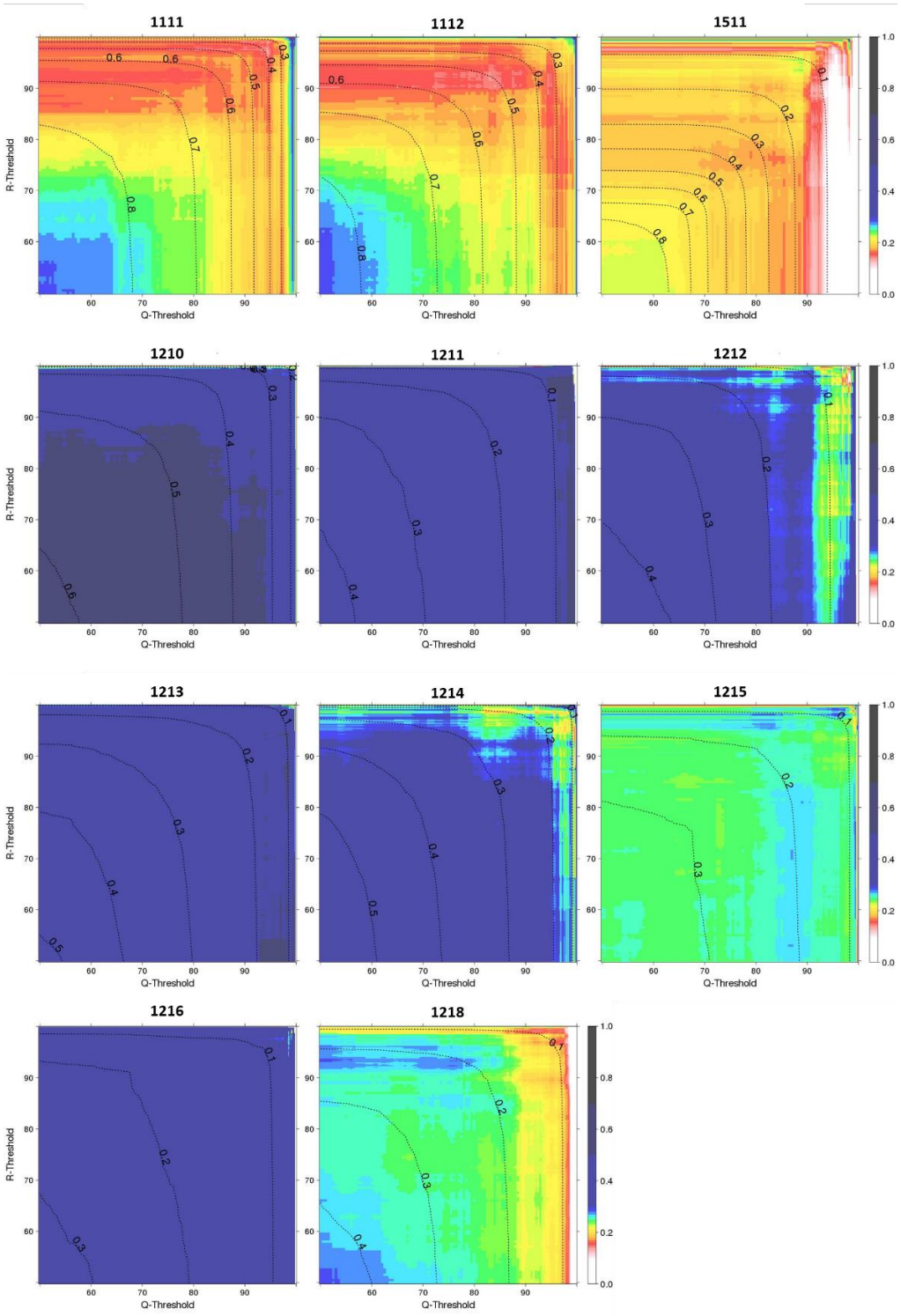


Figure 4.7. MutLX analysis of clone-specific somatic mutations of pre-chemotherapy and recurrence DigiPico data. As shown in section 3.2.12, the behaviour of C>A mutation frequencies in response to increasing Q- and R-threshold values from MutLX, can be used as an indicator for the clone-specific mutational load of the sample under study. From these results it can be interpreted that only the recurrence samples 1111, 1112, and 1511 are likely to have 'true' clone-specific somatic variants. For this analysis, any somatic variants that had been present in the WGS data of PT2R or OM were classed as 'shared' variant. The remaining of the somatic variant calls from each sample were used to generate these plots. Contour lines in each plot represent the ratio of known somatic variants that were identifiable at the respective threshold values (True Positive Rate).

These observations cumulatively suggest that while an extensive SNV micro-heterogeneity is present in the PT2R recurrence site, the pre-chemotherapy OM is relatively homogenous in term of SNVs across the tumour mass.

4.2.4 Investigation of active mutational processes in recurrence tumours

Furthermore, in an attempt to decipher the underlying mechanisms behind the SNV heterogeneity in the recurrence tumour the somatic mutations from WGS data of OM and PT2R as well as in the MutLX filtered variants of all of the recurrence DigiPico data were analysed for mutation signatures (Alexandrov *et al.*, 2013; Lee *et al.*, 2018). The results of this analysis show that signature 3 is most likely the dominant mutagenic process in this tumour (Figure 4.8A). This signature has frequently been observed in ovarian cancers and is believed to be associated with failure in double-strand break repair mechanisms that employ homologous recombination (Alexandrov *et al.*, 2013). However, Comparing the mutation signature plots of the pre-chemotherapy OM and recurrence PT2R sites from the bulk data reveals that the recurrence has a significantly higher number of C>T mutations, majority of which occur in a TCT context (Figure 4.8B). Although such a specific type of mutation does not fit into any of the described signatures of mutational processes, the extreme bias towards a specific context suggests that a potential molecular process might be responsible for the high rate of TCT>TTT mutations in the recurrence tumour. The involvement of Cytidine nucleoside in these mutations suggests that this process is very likely to be related to the deamination of Cytosine via a specific member of the APOBEC family of enzymes (Alexandrov *et al.*, 2013). Therefore, based on the specific presence of these mutations in the PT2R site, two possible explanation for the onset of activation of this process can be proposed:

1) **Acquired.** The mutational process might have become activated after the chemotherapy and during the progression of the PT2R site. In this situation it is expected that majority of the TCT>TTT mutations in the PT2R site to be clone-specific.

2) **Intrinsic.** The mutational process might have already been present in some of the cells in the pre-chemotherapy tumour and these cells have survived the chemotherapy resulting in relapse. In this situation, majority of the TCT>TTT mutations are expected to be present across all of the tumour.

Although to differentiate between these two hypothesis DigiPico sequencing of multiple tumour islets from the PT2R site would be required, however, analysing the available DigiPico data from the PT2R site (111X) revealed that a large number of clone-specific mutations in this sample are in the form of TCT>TTT and TCA>TTA transitions (Figure 4.8B). This suggests that at the time of sample collection, the mutagenic molecular process had still been active in the PT2R site. Moreover, analysing the clone-specific mutations from the DigiPico data of the PLNAR recurrence site reveals the presence of a bias towards a specific group of T>A mutations consistent with signature 22 (Figure 4.8). Signature 22 had been found to be associated with exposure to aristolochic acid in various cancers including ovarian carcinomas (Alexandrov *et al.*, 2013).

Additionally, the bulk WGS data of OM and PT2R were also analysed for the distribution of somatic variants across the genome in a *rainfall plot* to investigate the presence of potential localised hypermutation events (kataegis) in this tumour (Nik-Zainal *et al.*, 2012). Kataegis is represented by the presence of a high number of somatic mutations in a small region of the cancer genome that has originally been reported in various breast cancers. The process had been proposed to be a result of deamination of C nucleotides at the sites of double-strand break and is likely to be fuelled by the activity of APOBEC family of cytidine deaminases (Taylor *et al.*, 2013). The results of the analysis clearly confirm the presence of a localised hypermutation event on chromosome 19p of the pre-chemotherapy sample (Figure 4.9, blue arrow). Moreover, using the recurrence specific somatic variants for this analysis reveals the presence of another kataegis event on chromosome 6p, specific to the PT2R site (Figure 4.9, magenta arrow). This observation suggests that the 6p kataegis is likely to have occurred during the evolution of the recurrence tumour in the PT2R site.

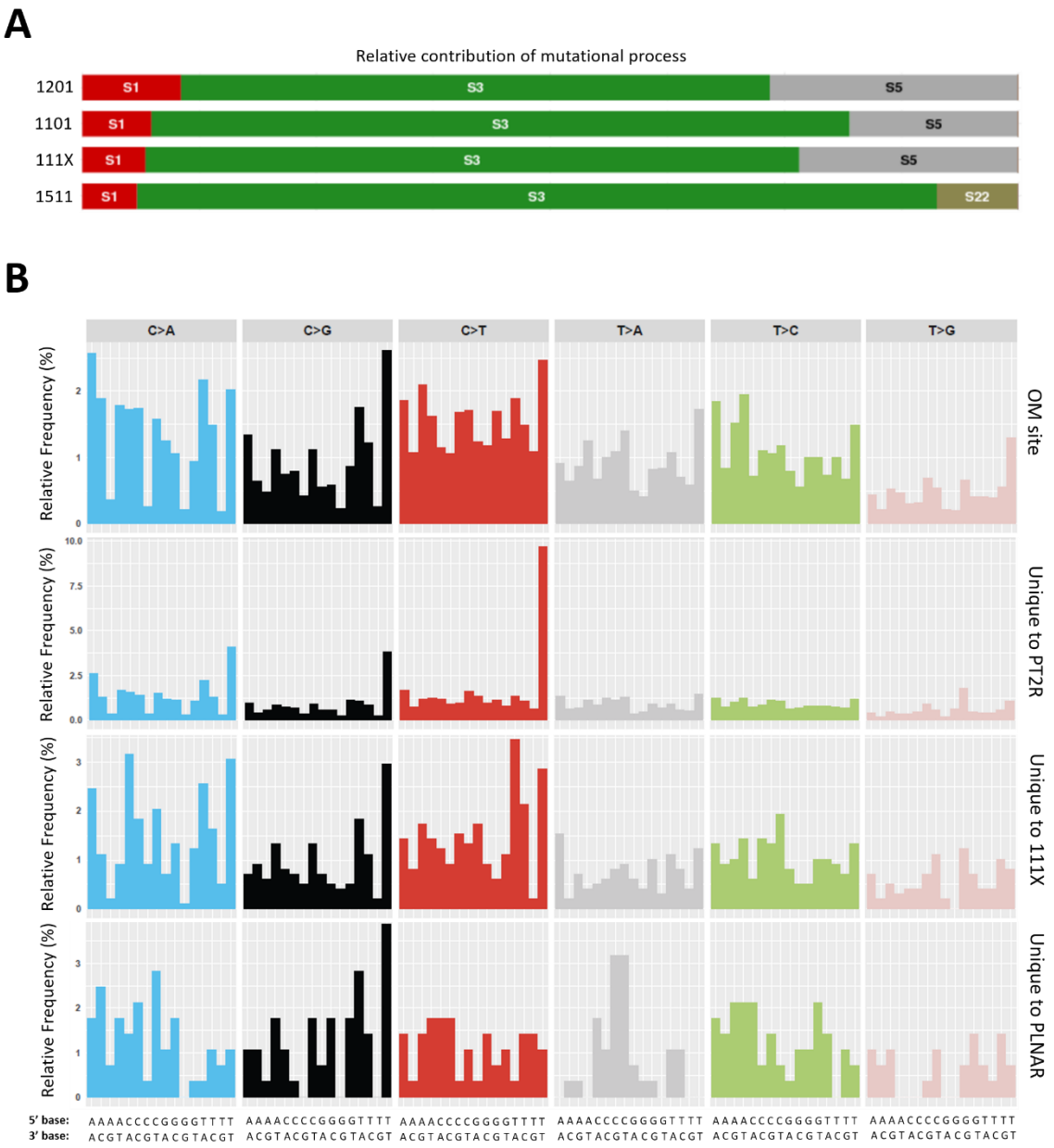


Figure 4.8. Analysis of the mutational process signatures of the pre-chemotherapy and recurrence tumours. (A) Analysis of the somatic mutations from the pre-chemotherapy OM (1201), as well as tumour/clone specific somatic variants of different samples for the signature of mutational processes using Mutalisk tool reveals that signature 3 is dominantly active in all of the studied samples. For samples 1101, 111X, and 1511, in this analysis, the clone specific variants were obtained by removing all of the variants that were present in sample 1201 or any of the other samples. **(B)** Detailed comparison of the frequencies of various mutation types (C>A, C>G, C>T, T>A, T>C, T>G) in all of the possible 3 nucleotide contexts reveals that PT2R and PLNAR sites each have their own specific patterns of unique mutations. This further highlights the possibility of their independent emergence and evolution. the 5' and 3' bases to each mutation in all of the 3 nucleotide contexts are represented at the bottom of every bar plot. All plots were directly obtained from the analysis of somatic variants with Mutalisk online tool (http://mutalisk.org).

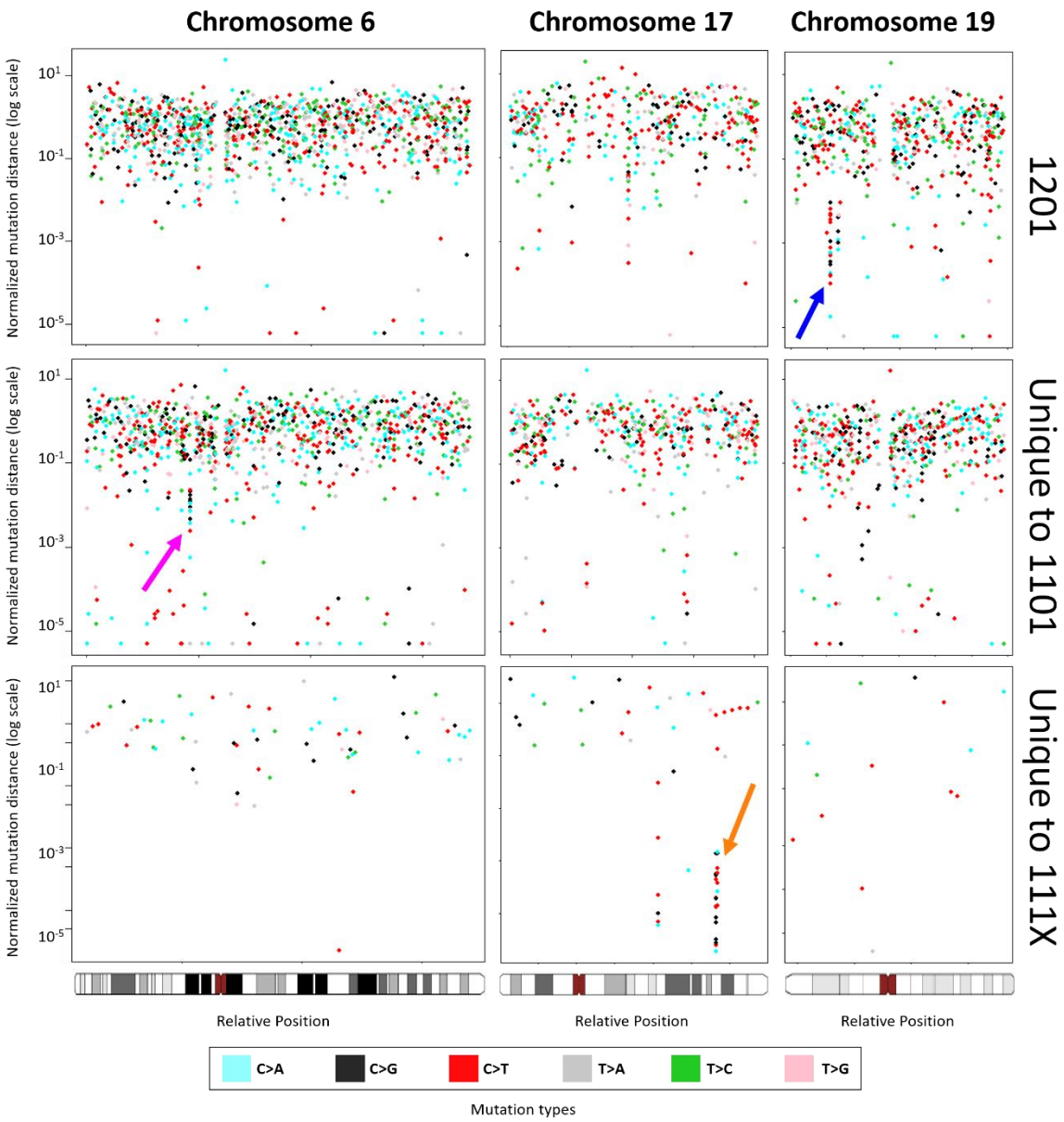


Figure 4.9. Rainfall plots of mutation distances reveals the presence of clone specific kataegis events. To generate the rainfall plots, the distances between pairs of consecutive somatic mutations across the genome were first normalised to the average distance of mutations in each dataset and the normalised values were plotted against their relative genomic position. The presence of clusters of localised mutations indicate kataegis events. These areas are often enriched with C>G and C>T mutations which highlights the potential involvement of APOBEC cytidine deaminases. In these plots, each dot is coloured based on the mutation type of the second mutation in the pair in respect to the hg19 human reference genome. The results of analysis in the pre-chemotherapy sample indicates the presence of a kataegis event at chromosome 19p. The analysis of PT2R specific somatic variants (all of the shared variants with 1201 were removed) results in the identification of a recurrence specific kataegis event at chromosome 6p. Moreover, the analysis of clone-specific mutations (all of the variants from 1101 and 1201 were removed) from the DigiPico data of a small cross section of the PT2R tumour reveals the presence of a strong sub-clonal kataegis event only in a fraction of the tumour cells in the PT2R sample. This can explain the observed SNV micro-heterogeneity of the PT2R site.

Interestingly analysing the clone-specific somatic variants from the DigiPico data of the PT2R site, also reveals the presence of another clone-specific kataegis event on chromosome 17q which does not seem to be detectable when bulk WGS data are used for the analysis (Figure 4.9, green arrow). Combining the results of this analysis with previous observations suggest that different mutational processes might be governing the evolution of different tumour sites in ovarian cancer. With the high level of C>T mutations in the PT2R site as well as clone-specific kataegis in this site, it is plausible to assume that a combination of break inducing phenomena, such as telomere crisis (Maciejowski *et al.*, 2015) combined with a faulty double-strand break mechanism and involvement of APOBEC family of enzyme might be the dominant mutagenic process in this site. On the other hand, the parallel recurrence tumour at the PLNAR site seems to show an independent pattern of mutations, consistent with the idea that these recurrence events might have evolved independently in this patient.

4.3 Study of MRD could reveal early mutations in HGSOC

Lastly, to evaluate the importance of MRD in the progression and evolution of ovarian cancer, various MRD islets were collected from the matched post-chemotherapy OM site of patient 11152 and were used for DigiPicoExome sequencing. Moreover, to allow for a more robust comparison, where possible, exome capture was also performed on previously studied DigiPico libraries to generate their respective DigiPicoExome libraries. As a quality control step, the resulting sequencing data were first used for variant re-calling of known germline and somatic variants that had previously been identified in patient 11152. Table 4.1 summarises the variant re-calling rate of each of the variant categories in all of the available DigiPicoExome libraries in this study.

Table 4.1. Variant re-calling rate of all of the available DigiPicoExome libraries from patient 11152**.**

DigiPico Run	1/1 Germline	0/1 Germline	Somatic
12E0	71.87%	48.11%	36.27%
12E1	77.79%	55.15%	33.56%
12E2	82.15%	62.36%	40.00%
12E3	76.78%	55.07%	38.98%
12E4	79.63%	57.82%	40.68%
12E5	79.27%	58.95%	38.64%
12E6	74.98%	52.98%	35.25%
12E8	61.82%	39.72%	19.66%
18E1	7.12%	4.07%	0.68%
18E2	76.38%	66.53%	0.34%
18E3	89.19%	84.36%	16.95%
18E4	9.66%	5.40%	0.68%
18E5	23.25%	13.32%	6.10%
18E6	56.51%	36.22%	23.73%
18E7	6.55%	3.43%	1.02%
18E8	35.88%	21.56%	11.86%
11E0	88.64%	74.80%	74.58%
11E1	85.79%	74.32%	70.17%

The results of this analysis indicate that 18E1, 18E4, and 18E7 libraries from the MRD islets suffer from low coverage and were therefore, removed from the analysis. Moreover, sample 18E2 seems to be composed mainly of normal cells and was also not considered as an MRD sample. Interestingly, from this data, it appears that MRD samples might have a lower somatic mutational load when compared with the pre-chemotherapy islets. This could suggest that chemotherapy has selectively eliminated tumour cells with higher mutation load or that it has a weaker effect on tumour cells that, in terms of mutational load are at earlier stages of the tumour evolution.

However, previous analyses showed that new mutations are constantly being generated in this tumour, and consistent with other studies (Patch *et al.*, 2015), it is expected that, over time, every cell division result in an increase in the overall mutation load of the tumour. Hence, the observation that post-chemotherapy MRD islets have a lower mutational load when compared with the pre-chemotherapy islets of a matched tumour site suggests that the MRD islets are most likely composed of dormant (less actively dividing) cells and that they might inherently be resistant to chemotherapy. Therefore, if this hypothesis is true, by studying the mutations in MRD, it should be possible to identify the earliest genomic changes that have initiated this tumour. Further to this finding, plotting the normalised mutational load of the MRD samples show that islet 18E3, despite having the highest germline re-call rate, has the lowest somatic mutational load amongst all of the studied samples (Figure 4.10A). Analysing the status of the TP53:C275F mutation in this islet reveals that only about 25% of the alleles in this sample carry this mutation. However, similar to what had previously been observed in islet 1211, the phased nearby SNPs showed an average VAF of about 90%. Assuming that no other unknown biasing factors had been involved, this observation suggests that either this MRD sample is composed of only about 25% cancer cells or it is likely that, in agreement with the previous observation in islet 1211, the CNA event at chromosome 17p have preceded the TP53 mutation which resides in this region. To further investigate these hypotheses, a heatmap of VAF values for two nearby phased and two nearby unphased SNPs was generated for the MRD samples as well as for the DigiPicoExome libraries of runs 1111, 1210, and 1211 (runs 11E1, 12E0, and 12E1 respectively) as control (Figure 4.10B). The results show that while islet 18E8 might mostly be composed of normal cells, the behaviour of VAF values in 18E3 and 18E5 MRD samples is clearly different from the control tumour samples. Therefore, to evaluate whether contamination with normal cells could explain the observed behaviour of VAFs, a Chi-squared test was performed to test the null hypothesis that: "*Normal cells can only have the S^PP^R/S^UP^R genotype while cancer cells only show the S^PP^M/S^PP^M genotype with no intermediate genotypes at this locus.*"

Where S is the SNP locus with P/U superscripts indicating the phased/unphased alleles in respect to the TP53:C275F mutation, respectively; and P is the TP53:C275 locus with R/M superscripts indicating the Reference/Mutated alleles. Performing the Chi-squared test analysis only on the samples with high TP53:C275F VAF values resulted in a p-value of 0.970 (X^2=1.33, DF=6) confirming the null hypothesis. However, the analysis of 12E1, 18E3, and 18E5 samples resulted in a p-value of less than 0.00001 (X^2=161.2, DF=6), indicating that it is unlikely for the null hypothesis to be true for these samples. This suggests that, if no other biasing factors had been involved, a third malignant cell type, likely with a $S^PP^R/$-- genotype (LOH at 17p), must have been present in these islets. This further supports the hypothesis that, the malignant cells from these MRD islets are likely to be from a more primitive lineage of the tumour and hence, the somatic mutations in these islets might represent the earliest genetic changes that have initiated this carcinoma. Therefore, in an attempt to identify these early events, variant calling was performed on islets 18E3 and 12E1 and all of the exonic mutations with a frequency higher than 5% in 18E3 were used to generate a VAF heatmap (Figure 4.10C). Amongst the 18 identified mutations, aside from TP53, 7 were found more likely to be relevant in the context of HGSOC (Figure 4.10C, highlighted in bold typeface). Most notably SKA1 (essential component of kinetochore) and PALB2 (scaffold for BRCA1/BRCA2) mutations, if with functional consequences, are extremely likely to be among the earliest driver mutations of this tumour. Confirmation of this however, would require functional studies of these mutation in the right context.

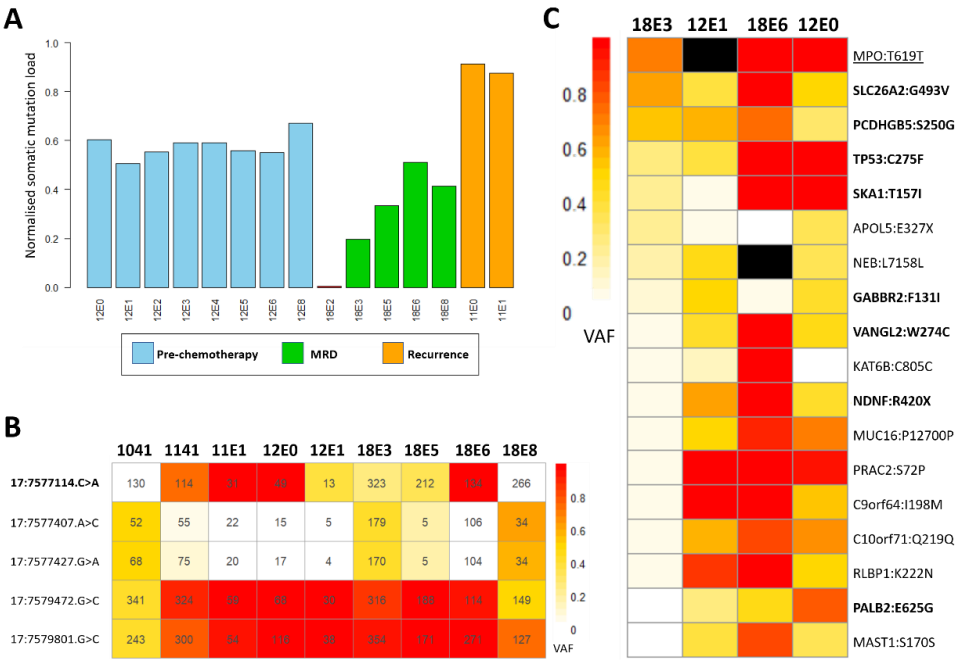


Figure 4.10. Study of somatic SNVs of MRD. (A) Comparing the normalised somatic mutational load of various DigiPico libraries from pre-chemotherapy islets, post-chemotherapy MRD, as well as recurrence shows that MRD samples appear to have a lower somatic mutational load compared to pre-chemotherapy and at recurrence. Normalisation of values was performed by dividing the somatic re-call rate of each sample by its germline re-call rate to compensate for lower breadth of coverage in some of the samples. Sample 18E2 appears to only contain normal cells and was therefore not considered in further analysis. **(B)** Analysis of VAFs of two phased (being present on the same chromatid as the TP53:C275F mutation) and two unphased (being present on the opposite chromatid as the TP53:C275F mutation) mutations in the vicinity of the TP53:C275F mutation of various samples reveals that, while 18E8 might have a high level of normal cell contamination, the behaviour of samples 12E1, 18E3, and 18E5 is unlikely to be a result of simple normal cell contamination. 1041 and 1141 samples are standard WES libraries from blood and PT2R of the same patient, respectively. **(C)** Investigation of somatic variants from MRD 18E3 identifies a set of somatic variants, majority of which seem to be present at high frequencies in other samples. Among these mutations, eight were found to be more likely to be relevant to the evolution of the tumour. **SLC26A2** (sulphate ion transport) had been show to downregulate the expression of receptors involved in extrinsic apoptosis pathway (Dimberg *et al.*, 2017). **PCDHGB5** (Protocadherin Gamma Subfamily B, 5) is a cell adhesion molecule that had been shown to be highly expressed in ovarian cancer (*Human Protein Atlas*, 2018). **TP53** is the main mutated tumour suppressor gene in HGSOC (Ahmed *et al.*, 2010). **SKA1** (Spindle and kinetochore-associated protein 1) is an essential component of kinetochore and its malfunction could result in mis-segregation of chromosomes during Anaphase. **GABBR2** (Gamma-Aminobutyric Acid Type B Receptor Subunit 2) is a GABA receptor, although no direct links had been reported for this gene, another GABA receptor, GABRA6, had previously been shown to be recurrently mutated in ovarian cancer (TCGA, 2011). **VANGL2** (planar cell polarity) is highly expressed in ovarian cancer (*Human Protein Atlas*, 2018) and is involved in the maintenance of cell polarity. It had previously been shown to play a role in the progression of breast cancer (Hatakeyama *et al.*, 2014). **NDNF** (neuron derived neurotrophic factor) is a neurotrophic factor, that similar to BDNF (Brain-derived neurotrophic factor) might be able to affect the tumour microenvironment (Radin and Patel, 2017). **PALB2** (Partner and localizer of BRCA2; FANCN) is an essential component of the BRCA1-PALB2-BRCA2 DNA damage response complex (Zhang *et al.*, 2009) and its defect can hamper the homologous recombination pathways. Moreover, a synonymous COSMIC mutation (COSM191540) in the MPO (Myeloperoxidase) gene had been found to be present in all of the tumour samples at very high frequency. Whether this mutation is a passenger mutation or not needs to be studied via functional studies.

Lastly, to further evaluate the importance of these potentially early mutations, their frequency in all of the available exome sequencing data from patient 11152 was evaluated (Figure 4.11). The results clearly indicate that, although at varying frequencies, all of these mutations are consistently present in all of the studied tumour sites from this patient. This further confirms the likelihood that at least some of these alterations might have occurred at the pre-malignancy stage and might have contributed to the initiation of this tumour.

In conclusion, this case study has shown how detailed analysis of homogenous islets of tumour can allow for identification of subtle intra-tumour heterogeneities, both at the level of CNAs and SNVs. Majority of these alterations would have remained undetected if only bulk sequencing data were to be used for the analysis, mainly because of the extreme complexity of such events in HGSOC. Moreover, employing DigiPico to study extremely small post-chemotherapy MRD islets, resulted in identification of genetic alteration events that are likely to have been involved in the initiation of the tumour. Future functional studies on these mutations could result in identification of novel early diagnostic markers for HGSOC. Overall, these analyses show the importance and potential of studying the small tumour islets alongside the bulk sequencing data in deciphering the evolutionary history of HGSOC and potentially other solid tumours, and validates DigiPico as a reliable means for such studies.

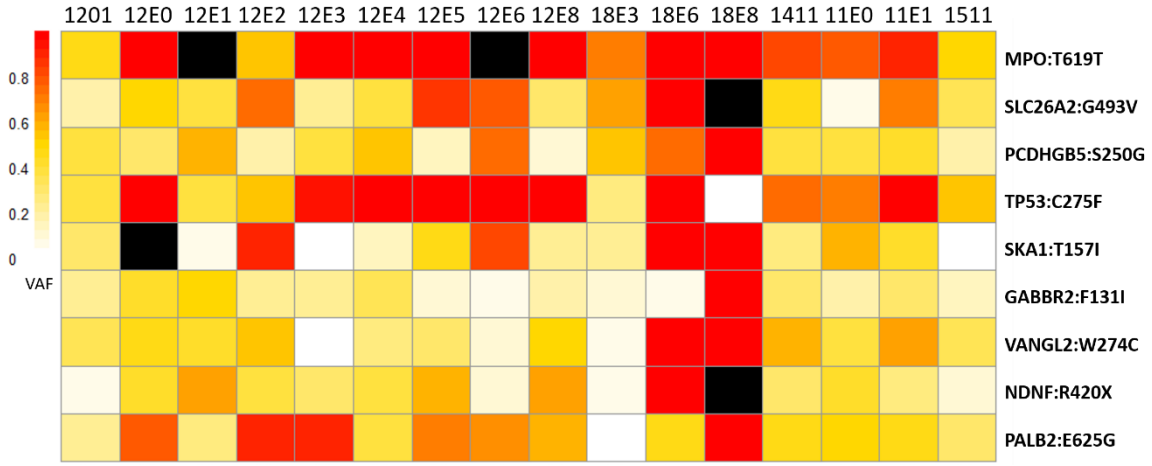


Figure 4.11. VAF of potential early mutations in all of the studied tumour samples. Analysing the VAF of potentially relevant early mutations in all of the available sequencing data from this tumour reveals that all of these variants are consistently present across the spatially and temporally diverse set of tumour samples in patient 11152. This further confirms that these variants are likely to be amongst the first mutations that had initiated this tumour. Black tiles indicate lack of coverage for the specific mutation in the respective sample. Samples 1201 and 1511 had been found to suffer from normal cell contamination and therefore the lower VAF values in these samples are expected.

Chapter 5 : Discussion and Conclusions

5.1 Discussion

5.1.1 Optimisation of DigiPico sequencing and analysis pipeline

In this thesis I initially aimed at optimising DigiPico, a library preparation method that would result in high quality next generation sequencing (NGS) data from picogram quantities of template DNA sourced from laser capture micro-dissected clinical samples. The implementation of this and similar methods is an extremely important step in the evolutionary study of high-grade serous ovarian cancer (HGSOC), and potentially other solid tumours. This is mainly because the most critical points of evolution of a tumour, i. e. tumour initiation and relapse, happen at the scale of tens to maybe hundreds of cells. Therefore, to study these processes, a sequencing approach that can provide reliable genomics data from such a low number of cells is essential. In chapter 2, I successfully implemented a robust protocol that reproducibly produces high quality data from tumour islets which are clusters of tens to hundreds of tumour cells that had been micro dissected from cross-sections of a solid tumour mass via LCM. The sequencing data that can be obtained from DigiPico sequencing has a unique structure, composed of 384 individual sequencing datasets, and in this structure, it carries a wealth of information that is not common in other sequencing approaches. While essential for the purpose of my thesis, this however, meant that none of the available NGS analysis pipelines and algorithms were able to effectively deal with DigiPico sequencing data. Therefore, in chapter 3, I designed and optimised an analysis pipeline for DigiPico sequencing and demonstrated its high accuracy and sensitivity in identifying copy number alterations (CNAs) and single nucleotide variants (SNVs) from DigiPico data. While the optimisation of CNA identification was simply achieved by modifying the input structure of a previously published algorithm, the accurate analysis of SNVs using existing methods remained to be a challenge. In overcoming these challenges, it is important to first realise that nearly all of the variant calling algorithms, and in general computational methods, at their core, are based on mathematical probabilistic models.

This means that there always is a trade-off between true positive rate (TPR; sensitivity; probability of detecting a true event) and false positive rate (FPR; probability of incorrectly identifying a negative event as positive). When analysing genomics data, variant calling is often performed to achieve two related but different goals: Firstly, variant calling may be performed to confirm the presence or absence of an already known variant in a given sample. This method often referred to as 'variant re-calling' or 'genotyping' requires to have an extremely high TPR. However, FPR, although important, but is not a significant issue. On the other hand, the second goal, *de novo* variant calling, requires an extremely low FPR, while TPR may not be that important. This is mainly because in *de novo* variant calling the main target is to identify novel variants without any pre-knowledge about them. When dealing with a genome as large as human genome, with over 3,000,000,000 base pairs, any significant FPR value would mean thousands to millions of false positive calls, which, would significantly interfere with any form of downstream analysis. As expected, variant re-calling from DigiPico data was found to be sensitive enough without any significant need for optimisation. However, similar to previous reports on variant calling from whole genome amplified DNA, an extremely large number of FP calls were detected in the DigiPico datasets. Therefore, to tackle this issue, I designed a machine learning approach, MutLX, to specifically address the issue of FP calls in DigiPico datasets. Using MutLX, I implemented an analysis pipeline that could accurately predict whether or not a specific tumour islet is expected to have clone-specific variants, and to identify them if they truly exist.

5.1.2 To better understand HGSOC using DigiPico

Having optimised and implemented the DigiPico sequencing method and analysis pipeline, I eventually found the opportunity to perform experiments that are required to try and address the two fundamental unanswered questions about HGSOC: "What exactly starts it and why does it come back so often?". As a step forward, I performed DigiPico sequencing on various tumour islets and MRD samples from one patient in a comprehensive case study. This study highlighted the importance of genomics studies on tumour islets, demonstrated the power of DigiPico for this

purpose, and resulted in interesting and controversial observations about HGSOC. While the results from this single case study using a newly developed method does not necessarily confirm or reject any hypothesis in regards to HGSOC, it however, allowed me to obtain a novel and unbiased insight towards the evolution of HGSOC. Since a part of the results from this case study seem to contradict some of the already accepted notions about the origin and progression of HGSOC, I hereby present them only as reliable observations from a single HGSOC case. However, I also provide logical explanation and a unifying hypothesis that can explain both the previously reported data as well as the observations made in this study.

5.1.2.1 Importance of tumour heterogeneity in HGSOC

Employing DigiPico sequencing method, firstly, I showed that at least in this case, the pre-chemotherapy tumour is enriched with heterogeneity at the level of structural variants. The reason behind the absence of similar observations in previous studies might be reflected in the fact that majority of the previous studies have been performed on the bulk WGS data of pre-chemotherapy tumours. In this study I showed that bulk WGS data fails to preserve the detailed information regarding these intra-heterogeneities. Although it might be argued that CNV heterogeneity could computationally be inferred from bulk sequencing data, but these methods can only identify a limited degree of diversity and heavily depend on assumptions that may not necessarily be true. Therefore, to identify such an extensive CNV heterogeneity in a tumour mass, the analysis of individual clonal populations is required. To my knowledge, the only available comprehensive study of pre-chemotherapy HGSOC tumours at single cell level addresses the clonal heterogeneity mainly based on SNVs and therefore its results cannot be compared with observations from the current study. Next, by performing somatic variant calling on a diverse set of tumour samples, in agreement with previous studies, it was shown that recurrent tumours might have evolved independently of each other and therefore the recurrence disease might be a result of multiple unrelated events. These results also indicated that at least one of the recurrent tumours (PLNAR) have obtained a new mode of mutagenesis which seemed to be absent in another recurrent tumour sample (PT2R).

Moreover, employing MutLX algorithm, I confirmed the presence of an intra-tumour heterogeneity in the recurrent tumour, but not in the pre-chemotherapy sample. Combining the results from the CNA and SNV analysis, it appears that while the main mode of intra-tumour heterogeneity in the pre-chemotherapy sample is in the form of CNAs, the recurrent tumour appears to be heterogeneous only at the SNV level. Consistent with previous studies, this observation indicates that the primary tumour has a defective DNA repair mechanism which has resulted in genomic instability. But the recurrent tumour, seems to have acquired resistance to platinum-based therapy through reactivation of the DNA repair pathways, restoring its genomic integrity. On the other hand, however, the recurrent tumour seems to have obtained new means to generate genetic diversity at the SNV level. Although the exact molecular mechanisms behind these processes remain to be determined, the unique mutational signatures that are observed in different recurrent sites suggest that maintenance of genetic diversity and therefore heterogeneity must be an important aspect of tumour evolution in HGSOC.

5.1.2.2 Early mutations in HGSOC

Furthermore, by critical analysis of the data from tumour islets and MRD, I observed that, the loss of heterozygosity (LOH) at chromosome 17p appears to have preceded the occurrence of TP53 mutation in this patient. This hypothesis seemed to be valid in at least three tumour islets, one from the pre-chemotherapy sample, and two from the MRD. Although this might contradict some of the presumptions about the origin of HGSOC but the observation can biologically be explained: While an LOH at the scale of a whole chromosome arm is often associated with several consequences, some of which might be advantageous for cell survival and growth, a single heterozygous missense mutation in the TP53 gene, is very unlikely to provide any growth advantages for the cell. Therefore, if it is only left to chance, the occurrence of a deleterious TP53 mutation, followed by the random loss of the other allele through LOH, would be so unlikely that under which the high incidence rate of HGSOC cannot be explained. Although the Dominant-Negative Effect (DNE) of TP53 missense mutants might be offered as a potential counter argument

in here (Willis *et al.*, 2004), there however are multiple flaws with this theory in the currently discussed context. Firstly, majority of the initial experiments that have shown the DNE of TP53 had been performed in a non-stoichiometric setting, where the mutant TP53 protein had been overexpressed in a mammalian cell. Further experiments in stoichiometric ratios however, have often failed to identify a relevant DNE for some of the missense mutations that had previously been thought to have a DNE (Stoczynska-Fidelus *et al.*, 2011). Secondly, it had been shown that in majority of the cases, DNE, even when strongly present, can only affect some of the TP53 functions (Aurelio *et al.*, 2000), and is only rarely able to cause transformation on its own (Wijnhoven *et al.*, 2007). This might explain why in majority of the cancers, including HGSOC, the inactivation of both of the TP53 alleles is essential for tumorigenesis. If TP53 DNE was a universal phenomenon, such as in the case with EGFR and IDH1 (Lynch *et al.*, 2004; Zhao and Guan, 2010), the loss of the other allele would not have been necessary (Stoczynska-Fidelus *et al.*, 2011). Lastly, the specific TP53 mutation in patient 11152, C275F, has previously been shown to have no DNE at all (Marutani *et al.*, 1999). Hence, at least in this case it is safe to assume that a heterozygous TP53 mutation would not have been transformative for the cells. Therefore, based on my observations and the above mentioned previous scientific works, I propose that, the earliest genetic alterations in the tumorigenesis of HGSOC target the integrity of the genome, likely by defecting the DNA repair mechanisms (Figure 5.1). Since aside from TP53 no other universally mutated genes have been identified in HGSOC, I also propose that the aforementioned early mutations must be in the regulatory regions supported by a previous observation regarding SOX2 overexpression in HGSOC (Hellner *et al.*, 2016). Next, the loss of genomic integrity accelerates the rate of large-scale mutations and genomic rearrangements. Such mutations on chromosome 17 could benefit the cell by activating any of the several oncogenes on this chromosome, such as HER2, or by rearranging any of the HOX genes in the HOXB cluster (Chen *et al.*, 1991; Nunes *et al.*, 2003; Han *et al.*, 2015; Bhatlekar, Fields and Boman, 2018; Miller *et al.*, 2018).

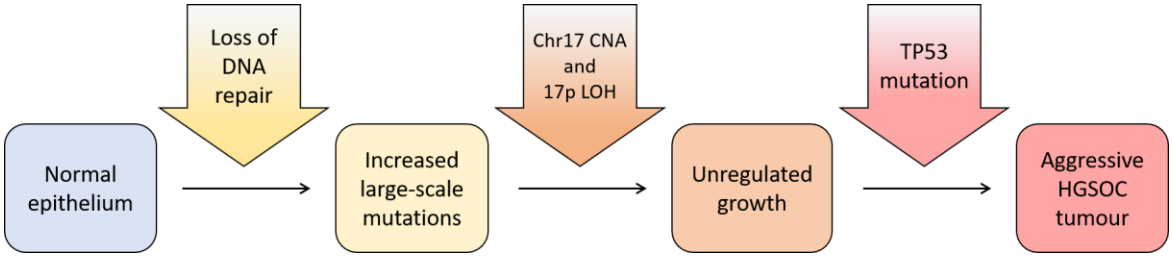


Figure 5.1. Proposed tumourigenesis pathway of HGSOC. Based on previous reports on somatic mutations in pre-malignant lesions of fallopian tubes (Hellner *et al.*, 2016) and following the identification of populations of cancer cells carrying an intact copy of the TP53 gene with 17p LOH in a HGSOC case, a new model of tumourigenesis for HGSOC is proposed. According to this model, initially mutations in regulatory regions of the genome may lead to an impairment in the DNA repair machinery of fallopian tube epithelial cells. This in turn leads to a drastic increase in genome-wide large-scale mutations, some of which may lead to growth inducing CNAs and rearrangements involving chromosome 17, resulting in 17p LOH. Lastly, mutations in the remaining copy of the TP53 gene results in the loss of TP53 function in these hyperplastic lesions and will lead to development of aggressive HGSOC.

During these processes, chromosome 17p may get lost, in which case, a second hit, in the form of a missense mutation in the TP53 gene could transform the cells. Therefore, the importance of a missense mutation in TP53, rather than its DNE on the intact TP53 molecules, might reside in the interaction of this protein with other proteins and is likely related to the invasiveness of the tumour, rather than initiation (Quartuccio *et al.*, 2015; Billant *et al.*, 2016). A missense mutation in TP53, in the absence of an intact allele, results in TP53 overexpression because of the loss of its negative feedback loop (Harris and Levine, 2005). The overexpression of an inactivated TP53, is likely to have negative effects on other tumour suppressors proteins such as P63 or P73 (Billant *et al.*, 2016) which can contribute to the tumour aggressiveness.

Although my hypothesis, claims the opposite of what is currently believed about the order of mutations in the tumorigenesis of HGSOC, it however, does not contradict any of the available data in publications that have offered the 'TP53 mutation comes first' hypothesis. To be more specific, below I summarise the results and interpretations from two of the most influential scientific publications in the field:

1) The study by Lee et al. in 2007 (Lee *et al.*, 2007), for the first time identifies the presence of focal lesions in the fallopian tubes of patients with germline BRCA1/2 mutations that overexpress TP53 and terms them 'P53 signature'. They then perform a systematic genetic analysis of the TP53 mutations between the P53 signatures, STICs and HGSOC of the same patient using sanger sequencing. In this study they show that identical TP53 mutations are present in STICs and in tumour. While they identify TP53 mutations in 'P53 signatures' of seven cases, none of these data is accompanied by TP53 sequencing data from tumour. In only one case, patient 7, they identify the same TP53 mutations in a 'P53 signature' and a STIC lesion but not in the tumour. In patient 7, however, the STIC and the 'P53 signature' were reported to be contiguous, which raises the concern of a possible cross contamination. Moreover, in two cases, they identified multiple TP53 mutations in different samples from the same patient.

Although the results of this study show a relationship between STICs and HGSOC, which can also be explained by metastasis to fallopian tube (Eckert *et al.*, 2016), no clear indication for a relationship between 'P53 signature' and HGSOC is present in their data. Moreover, because of the absence of any information on the frequency of identified mutations in each sample it is not immediately possible to conclude that the 17p LOH has occurred after the TP53 mutation. Interestingly, however, the presence of multiple TP53 mutations in the same patient can be explained by considering the 17p LOH as an earlier event. Therefore, although the results of this study do not contradict my hypothesis, the conclusions that had been made based on these results do.

2) A more recent study by Labidi-Galy et al. (Labidi-Galy *et al.*, 2017), has used a WES approach to compare the exonic somatic mutations obtained from various sites within the same patient. They have been able to identify similar sets of somatic mutations, including mutations in TP53 gene, between 'P53 signature' and tumour sample in two patients. However, as their data suffered from normal cell contamination, they have used the variant allele frequency of TP53 as their normalisation factor. As doing so, automatically requires the assumption that TP53 mutation precedes the 17p LOH event, naturally, any conclusions from this study would indicate that TP53 mutation have preceded the 17p LOH. Therefore, while the assumptions of this study do not agree with my hypothesis, the results neither validate nor reject the observations made in the case study of patient 11152.

5.1.2.3 Stem cell-like behaviour of MRD

Lastly, analysing the somatic mutation load of tumour islets, I realised that one of the MRD islets has a significantly lower somatic mutation load compared to other islets. Showing that the low variant allele fraction of TP53 mutation in this sample is unlikely to be due to normal cell contamination, I conclude that this MRD islet is likely to be more closely related to the *cell of origin* of this HGSOC tumour. The low somatic mutation load in this islet indicates that these cells are likely to have a slower cell division rate and a more protected genome.

Therefore, considering the slow division, stable genome, and the fact that these cells have survived the chemotherapy, it is plausible to assume that these cells might have stem cell-like characteristics that protects them from chemotherapy. This assumption had also been supported by previous observations that some of the MRD islets in this patient had been overexpressing stem cell markers such ALDH1 and members of the ABC transporters family. While it is appealing to assume that these stem cell-like survivors of chemotherapy are responsible for tumour relapse, it is extremely unlikely that these cells have directly contributed to the formation of the recurrent tumour. This is mainly because of the identity of the TP53 mutation between pre-chemotherapy and tumour recurrence and lack of a TP53 mutation in these potentially stem cell-like MRD. However, it can be hypothesised that these stem cell-like malignant cells, with an LOH on chromosome 17p but an intact TP53 gene, accompany the TP53 mutated aggressive cancer cells through their metastasis. Upon application of chemotherapy, nearly all of the cancer cells die, leading to an apparent complete response. However, since epithelial cells are inter-connected through their cytoplasm, cancer cells in the vicinity of these stem cell-like cells survive the chemotherapy under their protection. After chemotherapy, these cancer cells regain the opportunity to regrow and cause a more resistant recurrent tumour consistent with what had been observed. Although this hypothesis is consistent with all of the observations in this study and previous analysis on patient 11152, more case studies as well as extensive functional *in vitro* and in *vivo* experiments are required to reject or confirm it.

5.2 Conclusions

In conclusion, based on the experimental evidence in this study and previous observations I offer three theses in regards to the evolution of HGSOC:

5.2.1 On the origin

The earliest genetic changes in the *cell of origin* of HGSOC include regulatory mutations that target DNA repair mechanisms. As a result, catastrophic chromosomal rearrangements on chromosome 17 gain growth advantages for the cell. This may coincide with LOH of 17p. Consequently, a deleterious mutation in TP53 results in transformation and tumorigenesis.

5.2.2 On the progression

Cells that have undergone 17p LOH but have not yet attained the TP53 mutation, may exhibit some stem cell-like characteristics. These cells are able to accompany their TP53 mutated metastatic counterparts during metastasis. Application of chemotherapy, eliminates all cancer cells, but those that are in the vicinity of the stem cell-like cells. As a result, these cells get selected under the weak influence of the chemotherapy and contribute to relapse.

5.2.3 On the evolution

A maintained clonal heterogeneity and genetic diversity is a hallmark of HGSOC. In the absence of DNA repair pathways, it is achieved through large scale mutations and genomic instability. If under the pressure of chemotherapy genomic stability is regained, new means of mutagenesis evolve within the recurrent tumours resulting in heterogeneity at the single nucleotide level.

5.3 Future perspectives

5.3.1 DigiPico sequencing

Although DigiPico library preparation pipeline in its current form performs well, it however, can further be optimised to reduce the cost, hands-on time and to improve the efficiency and consistency even further. Potential points of improvement in the protocol are as follow:

- Combining the last step of CoREF with the A-tailing reaction
- Improving the ligation efficiency using commercially available ligation enhancers
- Increasing the number of compartments by using 1536-well plates
- Using custom blocking oligos for DigiPicoExome to improve capture efficiency

Furthermore, with new advances in WGA technologies use of new methods that may be able to incorporate the well specific indices during the WGA process could significantly improve the versatility and ease-of-use of the DigiPico protocol.

5.3.2 DigiPico analysis

Further improvements on Titan algorithm specific to DigiPico data can improve the accuracy of CNA calling. This however, would require a collaboration with developers of Titan algorithm. Moreover, MutLX, is currently based on an ANN that had originally been intended for MutL method. Since the main goal of MutLX, rather than generating a generalisable method, is to identify and remove the noise from the data, switching to ANN algorithms that have specifically been designed to deal with noisy data might improve the efficiency of MutLX even further.

5.3.3 Study of HGSOC

In this thesis I offered new hypotheses in regards to the origin, progression, and evolution of HGSOC based on the observations from a case study in one patient. Here I propose the following experiments for the examination of these hypotheses:

5.3.3.1 Further studies on the origin of HGSOC

Defining whether or not the observation that the 17p LOH may occur before the TP53 mutation in HGSOC has important consequences on how future studies on the origin of HGSOC must be performed. To clearly answer this question, I propose a systematic and unbiased study of fallopian tubes obtained from BRCA1/2 patients that have undergone prophylactic surgery. Two sets of experiments may be performed to examine the proposed hypothesis:

- Single-cell fallopian tube epithelial cells could be sorted in 384-well plate for single-cell WGA followed by targeted sequencing of TP53 and nearby heterozygous SNPs. The DNA from cells that show the presence of TP53 mutations and/or 17p LOH (based on the analysis of heterozygous SNPs) can then be used for WGS and further analysis.
- Cross sections from frozen fallopian tubes can be used to systematically LCM across the fallopian tube epithelium in bin sizes of 100 – 200 cells. After lysis, half of the lysate can be used to confirm the presence or absence of TP53 and/or 17p LOH using targeted sequencing. The remaining material from samples that show the desired genotypes can then be used for DigiPico sequencing for further validation.

5.3.3.2 Further studies on the progression of HGSOC

To examine the hypothesis in regards to the stem cell-like phenotype of MRD tumour islets, after validation of the same observation in other patients, following experiments can be performed to confirm or reject the proposed hypothesis:

- The early mutations that had been identified in these cells can be introduced to normal human fallopian tube epithelial cell cultures to investigate their effect on the phenotype of these cells.
- To investigate whether or not MRD are indeed capable of causing relapse, it is essential to attempt and derive xenografts from MRD. If successful, genomics study of the xenografts could reveal whether or not MRD contains stem cell-like cells with only the earliest mutations.

5.3.3.3 Further studies on the evolution of HGSOC

To confirm that maintenance of heterogeneity and genetic diversity is indeed one of the hallmarks of HGSOC, it is essential to first perform a similar study with tumour islets and DigiPico on multiple patients. If confirmed, further genetic studies to identify and validate the molecular mechanisms behind these *mutagenic events* could allow for development of new therapeutic strategies that will specifically target these *drives or the evolution* in HGSOC.

Chapter 6 : Methods

6.1 Ethical approval, patient consent, and sample collection

Patients 11152 and 11513 provided written consent for participation in the prospective biomarker validation study Gynaecological Oncology Targeted Therapy Study 01 (GO-Target-01) under research ethics approval number 11/SC/0014. Patient OP1036 participated in the prospective Oxford Ovarian Cancer Predict Chemotherapy Response Trial (OXO-PCR-01), under research ethics approval number 12/SC/0404. Necessary informed consents from study participants were obtained as appropriate. Blood samples were taken on the day of surgery. Tumour samples were biopsied during laparoscopy or debulking surgery and were immediately frozen on dry ice. All samples were stored in clearly labelled cryovials and were stored in -80 °C freezers.

6.2 Tissue sectioning and LCM

6.2.1 Sectioning

Frozen tumour samples were first embedded in OCT (NEG-50, Richard-Allan Scientific) and 10 - 15 µm sections were taken using MB DynaSharp microtome blades (ThermoFisher Scientific) in a CryoStar cryostat microtome (ThermoFisher Scientific). All the necessary precautions were taken to avoid any cross-sample contamination: Blades were used for only one sample, sterile forceps were used for handling of each sample, microtome area was thoroughly wiped using 70% ethanol before and after each procedure, and samples were sectioned in different days.

6.2.2 H&E staining

At the beginning of every sectioning the tumour sample was first used for haematoxylin and eosin staining to evaluate the level of normal cell contamination. 10 µm sections were place on a microscope slide and were fixed in Carnoy's solution (60% ethanol, 30% chloroform and 10% acetic acetic) for 1 minute, followed by a water rinse. Samples were then placed in haematoxylin solution (Sigma-Aldrich) for 1 minute, rinsed with water, dipped 3 times in 1% HCl solution, rinsed with

water, placed for 1 minute in Eosin solution (Sigma-Aldrich), rinsed in water, dipped 5 times each in 70%, 90%, and 100% ethanol solutions in order, and lastly was placed in xylene for 3 minutes. The section was then mounted using mounting medium and was used for microscopy to determine the tumour content.

6.2.3 Cresyl Violet staining and LCM

Tumour sections were transferred to PEN membrane glass slides (Zeiss) and were either stored at -80 °C or were immediately used for staining in Cresyl violet (Sigma-Aldrich) at a concentration of 0.1% in 50% ethanol for 15 seconds, followed by a rinse with 50% ethanol and were either stored at -80 °C or were immediately used for laser capture microdissection.

A PALM Laser Microdissection System (Zeiss) was used to catapult the desired area from the tumour sections into 200 µl opaque AdhesiveCap (Zeiss). Consecutive images were obtained from the capture process and the presence of tissue in each cap was confirmed for every sample. Slides were kept on dry ice as long as possible before microdissection to maintain the integrity of the DNA and caps were stored at − 80 °C until sequencing. Prior to sequencing images from all samples were sent to expert pathologist to confirm the presence of the tumour islets. Only caps that had a confirmed tumour islet based on pathologist's opinion were used for sequencing.

6.3 Library preparation

6.3.1 Standard WGS library

For standard WGS, microtome sections (scrolls) from tumour samples were collected in a 1.5 ml plastic tube, resuspended in 150 µl of sterile phosphate-buffered saline (PBS) solution, and were used for DNA extraction using *blood and tissue DNAeasy kit* (Qiagen) following provider's instructions. DNA concentration was measured using Nanodrop spectrophotometer (ThermoFisher Scientific). 100 ng – 1000 ng of DNA was used for shearing using Covaris S220 focused-ultrasonicator (Covaris) to achieve a fragment size in the range of 250 – 300 bp. The fragmented

DNA solution was then used for library preparation using NEBNext Ultra II library preparation kit (New England Biolabs) following the provider's protocol.

6.3.2 WES library

DNA samples were obtained and fragmented similar to section 6.3.1. Exome capture was performed using *SureSelect Human All Exon V7* exome capture kit (Agilent technologies) following the "SureSelectXT Low Input Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library" protocol as recommended by the manufacturer.

6.3.3 WGA WGS library

200 pg of DNA was whole genome amplified using REPLI-g Single Cell kit (Qiagen) following manufacturer's instruction with 6 hours incubation at 30 °C. The WGA product was purified using QIAamp DNA Micro Kit (Qiagen). The purified DNA was then used for WGS library preparation as mentioned in section 6.3.1.

6.3.4 DigiPico library preparation

6.3.4.1 Template preparation

A) From purified DNA: 100 – 200 pg of purified DNA was denatured using buffer *D2* of REPLI-g Single Cell kit followed by addition of *Stop Solution*. The solution volume was then increased to 100 µl using water (DNA- and DNase-free).

B) From tissue section: Tissue scrolls were resuspended in 100 µl of single cell grade PBS using a vortex mixer. 10 µl of the suspension was lysed by addition of 10 µl of buffer *D2* and 10 minutes incubation at 50 °C. The solution was neutralized using 10 µl of *Stop Solution* and the DNA content was quantified using Qubit dsDNA HS assay (ThermoFisher Scientific). Based on the results, appropriate volume of the original suspension was taken to obtain 200 pg of DNA. The suspension was lysed using 1 volume of buffer *D2* and the volume was raised to 100 µl using water.

C) Form LCM samples: 10 µl of buffer *D2* was added to the cap containing the LCM sample. After 15 minutes incubation at 50 °C the content was spun down and 1 µl of the solution was used for quantification using SpectraMax® Quant™ AccuBlue™ Pico dsDNA assay kit (Molecular Devices). Appropriate volumes were taken from the lysate to obtain a maximum of 200 pg of DNA. The solution's volume was then raised to 100 µl using water.

6.3.4.2 Whole genome amplification for DigiPico

To setup the WGA reaction in a 384-well plate, initially, 800 nl of WGA reaction mix containing 0.58 µl Sc Reaction Buffer, 0.04 µl Sc Polymerase (REPLI-g Single Cell kit, Qiagen), 0.075 µl 1 mM dUTP (Invitrogen), 0.04 µl EvaGreen 20x (Biotium), and 0.065 µl water was added to each well of a 384-well FrameStar® reaction plate (4titude) using Mosquito HTS liquid handler (TTP Labtech). Next, 6 µl of the denatured template was added to each well of a column in a 384-well plate and this plate was used as a reservoir to dispense 200 nl of template to each well of the WGA plate using Mosquito without mixing. The plate was then covered using MicroAmp optical adhesive film (Applied Biosystems) and was incubated at 30 °C for 2 hours. When required the incubation was performed in a CFX384 Real-Time PCR instrument (Bio-Rad) with fluorescent reading every 10 minutes from the SYBR Green channel. Lastly, the Phi29 DNA polymerase enzyme was heat inactivated by incubation at 65 °C for 5 minutes and the end point fluorescent values for each well were measured using CLARIOstar microplate reader (BMG Labtech) with excitation at 498 nm and emission measurement at 533 nm.

6.3.4.3 Fragmentation using CoREF

To fragment the products of WGA using CoREF, first, 1200 nl of UDG master mix containing 0.08 U/µl rSAP (New England Biolabs), 0.2 U/µl UDG (New England Biolabs), 0.4 U/µl EndoIV (New England Biolabs) in 1.8x NEBuffer 3 (New England Biolabs) was added to each well of the WGA plate using Mosquito with 3 cycles of mixing followed by 2 hours incubation at 37 °C and heat inactivation at 65 °C for 15 minutes. Next, 1200 nl of PolI master mix containing 0.4 U/µl PolI (New England

Biolabs), 0.25 mM dNTP (Invitrogen), 8 mM $MgCl_2$ (New England Biolabs), and 0.8 mM DTT (Qiagen) was added to each well using Mosquito with 3 cycles of mixing followed by incubation at 37 °C for 1.5 hours and heat inactivation at 70 °C for 20 minutes. As a quality control measure, 0.5 µl of the products from wells B2 and O23 were analysed using TapeStation (Agilent Genomics) to evaluate the size of the resulting DNA fragments.

6.3.4.4 A-tailing and adapter ligation

For A-tailing, 1200 nl of Klenow master mix containing 0.5 U/µl Klenow exo- (New England Biolabs), 0.5 mM dATP (Invitrogen), 8 mM $MgCl_2$, and 0.8 mM DTT was added to the reaction plate similar to the previous steps with 45 minutes incubation at 37 °C and heat inactivation at 70 °C for 20 minutes. Next, 400 nl of 20 µM well specific indexing oligos were added to each well followed by the addition of 1100 nl of Ligation master mix containing 40 U/µl T4 DNA Ligase (New England Biolabs), 5 mM ATP (Invitrogen), 11.5% PEG 8000 (Qiagen), and 6.8 mM $MgCl_2$ with 5 cycles of mixing using Mosquito. After 30 minutes incubation at 20 °C and heat inactivation at 65 °C for 15 minutes, all of the products were pooled using Mosquito.

6.3.4.5 Dual-size selection and limited-cycle PCR

To remove the excess PEG molecules from the products, 500 µl of the pool was mixed with 500 µl of isopropanol and the resulting solution was centrifuged at 16,000 g for 30 minutes to precipitate the DNA. The pellet was washed twice with 70% cold ethanol and after drying was resuspended in 100 µl of water. For DSS, 45 µl of Agencourt AMPure XP SPRI magnetic beads (Beckmann coulter) were added to the 100 µl DNA solution. After 20 minutes incubation at room temperature 140 µl of the supernatant was collected on a magnetic rack and 32 µl of new SPRI beads were added to it. After 15 minutes incubation at room temperature beads were washed twice with 80% ethanol on a magnetic rack and the dual-size selected DNA was eluted using 50 µl of water.

The final product was examined using TapeStation to ensure correct size distribution. For limited-cycle PCR amplification, 20 µl of the DSS product was mixed with 25 µl of KAPA Hi-Fi master mix (KAPA Biosystems) and 5 µl of 10 µM P5/P7 primer mix (P5 primer: AATGATACGGCGACCACCGAGA and P7 primer: CAAGCAGAAGACGGCATACGAG). PCR was performed for 12 cycles with 10 seconds annealing at 55 °C and 45 seconds extension at 72 °C. Final products were purified using 45 µl of SPRI beads (0.9x) and were evaluated using TapeStation.

6.3.4.6 DigiPicoExome

For DigiPicoExome, a maximum of 1 µg of the final libraries were directly used for capture using the *SureSelect Human All Exon V7* exome capture kit with custom blocking oligos obtained from manufacturer. The enrichment was performed following only the hybridization and capture sections of the "SureSelectXT Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library" protocol as recommended by the manufacturer.

6.4 Illumina sequencing and data analysis

6.4.1 Illumina sequencing

All WGS, WES, or DigiPico libraries were either sequenced on an Illumina NextSeq platform or were sent for commercial sequencing on HiSeq 4000 or HiSeqX platforms. Targeted sequencing libraries were sequencing on an Illumina MiSeq platform. All sequencings were performed in 2 x 150 paired-end sequencing mode. For WGS, WES, and targeted sequencing data, FastQ files were obtained from the instruments and were directly used for analysis. For DigiPico libraries, because of the complexity of indexing strategy, the sequencing was performed in the *standalone* mode to only obtain '.bcl' files. The resulting '.bcl' files were then demultiplexed using the bcl2fastq tool from Illumina based on the DigiPico index sequences to obtain 384 FastQ files, one for each well of the plate.

6.4.2 Read mapping

The obtained FatsQ files containing all of the read sequences and quality scores were first analysed using FastQC tool to ensure that they have an overall good quality. Reads were then trimmed using TrimGalore to remove adapter sequences and low-quality regions at the end of the reads. For DigiPico data, 12 bp from the 5' end of each read was also trimmed. Trimmed reads were then mapped to human hg19 reference genome using Bowtie2 tool.

6.4.3 Copy number analysis

TitanCNA workflow with default settings was used for the CNA analysis of standard WGS data. Following modifications were applied to the TitanCNA workflow to generate the *DigiTitan* workflow for the analysis of DigiPico data:

- To obtain heterozygous germline positions, variant calling was performed on the normal WGS data using Platypus variant caller. Any variant covered by at least 10 reads with a variant allele fraction between 0.3 and 0.7 was selected as a heterozygous position.
- Variant re-calling was then performed on DigiPico data only for the heterozygous positions using Platypus. The number of the wells with at least one read covering a specific position was used as the digitised coverage value for that position. The number of the wells with at least one read supporting the variant was used as digitised variant count value. Any position with a digitised coverage value of less than five was removed from the list.
- If phase information were not available for the patient, short range phasing was performed on the heterozygous SNP positions. To achieve this, for each pair of variants with a distance of less than 10 kb, all wells that cover both of the variants were considered. If variants were consistently present in the same wells the two variants were considered *phased* variants. If variants were present in complementary wells they were classed as *unphased* variant. In case of conflict no phase information was considered for the pair. The final phase group for each variant was determined based on the phasing status with 5 upstream phased variants.

- The phase information was then used to normalise the digitised variant count values for each phased heterozygous SNP position. First, the average VAF of all phased variants within a phase group was calculated. The digitised count values for each SNP position were then adjusted in such a way to have the same VAF as the average VAF but to preserve the total coverage value. This list was supplied to the Titan algorithm as the allele count information.
- To obtain the digitised depth information, *readCounter* tool from *HMMcopy* (Lai, Daniel; Ha, 2013) was used to obtain the read count values for each well in 1 kb bins. For each well, the value for any bin with at least one read count was set to 1, otherwise it was considered as 0. For each bin the values across all of the 384 wells were then summed to obtain the 1 kb digitised read counts. The 100 kb digitised read count values were then obtained by summing the values for every 100 consecutive bins. These data were supplied to the *ichorCNA* module of the TitanCNA workflow to calculate the normalised depth values.
- The TitanCNA_alphaK value was set to 20000 in configuration file.

All scripts used for this analysis are accessible on Github (https://github.com/mmdknr/DigiPico).

6.4.5 SNV calling and analysis

Unless otherwise stated all variant callings were performed using Platypus with default parameters. Mutalisk (Lee *et al.*, 2018) was used for the analysis and prediction of COSMIC mutation signatures. Initially only signatures 1,3, and 5 were used for the prediction. Following the observation in sample 1511, signature 22 was also included. All scripts used in this section are accessible on Github (https://github.com/mmdknr/DigiPico).

6.4.6 MutSwipe application

MutSwipe application was developed in java language in Native-react environment for Android platforms. The swipe gesture functions were taken from ‘react-native-deck-swiper’.

6.4.7 Machine learning

All scripts used in this section are accessible on Github (https://github.com/mmdknr/DigiPico).

6.4.7.1 Architecture

The neural network model used in this study was a multilayer perceptron with an input layer consisting of N neurons where N is the number of features used in each experiment. The model includes two hidden layers of N and N/2 neurons with relu activations and each with a dropout value of 0.3 and 0.4. The last layer is a single output neuron with a sigmoid activation. The loss function is binary cross-entropy. For the training, stochastic gradient descent with momentum, using Adam optimizer with a learning rate of 0.001, a batch-size of 8 and for 20 epochs was used.

6.4.7.2 Features used in MutL algorithm

Posterior probability, BRF, FR, HP, HapScore, MGOF, MMLQ, MQ, QD, and SbPval values from Platypus output, as well as number of reads that support a variant on each strand, allele frequency, and the average number of most frequent bases in 20 bp flanking sequences around each mutation.

6.4.7.3 Features used in MutL.2 algorithm

Number of forward reads supporting the variant, number of reverse reads supporting the variants, total number of forward reads covering the position, total number of reverse reads covering the position, total number of reads supporting the variant in the merge file, total number of reads covering the locus in the merged file, Total number of wells supporting the variant, total number of wells covering the variant, number of covering wells that have been predicted to be heterozygous, number of variant supporting wells that have been predicted to be heterozygous, number of variant supporting wells with reads from both reference and variant allele, number of wells with only variant reads that are covered by at least 1, 2, 3, 4, 5 and 6 reads.

6.4.7.4 Features used in MutLX algorithm

For the Q branch, quality score values from platypus as in section 6.4.7.2 were used. For the R branch all features of section 6.4.7.3 were used.

Chapter 7 : Bibliography

Adalsteinsson, V. A. *et al.* (2017) 'Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors', *Nat. Commun.* Nature Publishing Group, 8(1), p. 1324. doi: 10.1038/s41467-017-00965-y.

Ahmed, A. A. *et al.* (2010) 'Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary', *J. Pathol.*, 221(1), pp. 49–56. doi: 10.1002/path.2696.

Alexandrov, L. *et al.* (2018) 'The Repertoire of Mutational Signatures in Human Cancer', *bioRxiv*. doi: 10.1101/322859.

Alexandrov, L. B. *et al.* (2013) 'Signatures of mutational processes in human cancer', *Nature*, 500(7463), pp. 415–421. doi: 10.1038/nature12477.

'American Cancer Society' (2018). Available at: https://www.cancer.org/cancer/ovarian-cancer/detection-diagnosis-staging.html LB - c25H4.

Andrews, S. (2016) 'FastQC A Quality Control tool for High Throughput Sequence Data'.

Aurelio, O. N. *et al.* (2000) 'p53 Mutants Have Selective Dominant-Negative Effects on Apoptosis but Not Growth Arrest in Human Cancer Cell Lines', *Mol. Cell. Biol.*, 20(3), pp. 770–778. doi: 10.1128/MCB.20.3.770-778.2000.

Banerjee, S. and Kaye, S. B. (2013) 'New strategies in the treatment of ovarian cancer: current clinical perspectives and future potential.', *Clinical cancer research : an official journal of the American Association for Cancer Research*. United States, 19(5), pp. 961–968. doi: 10.1158/1078-0432.CCR-12-2243.

Bashashati, A. *et al.* (2013) 'Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling', *The Journal of Pathology*. John Wiley & Sons, Ltd, 231(1), pp. 21–34. doi: 10.1002/path.4230.

Bhatlekar, S., Fields, J. Z. and Boman, B. M. (2018) 'Role of HOX Genes in Stem Cell Differentiation and Cancer', *Stem Cells Int.*, 2018, p. 3569493. doi: 10.1155/2018/3569493.

Billant, O. *et al.* (2016) 'The dominant-negative interplay between p53, p63 and p73: A family affair', *Oncotarget*, 7(43), pp. 69549–69564. doi: 10.18632/oncotarget.11774.

Blackman, H. and Brothers, F. (2004) *Star Wars: Clone Wars Adventures Vol. 1*. Dark Horse.

Boeva, V. *et al.* (2012) 'Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data', *Bioinformatics*, 28(3), pp. 423–425. doi: 10.1093/bioinformatics/btr670.

Bowtell, D. D. *et al.* (2015) 'Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer', *Nat. Rev. Cancer*, 15(11), pp. 668–679. doi: 10.1038/nrc4019.

Cancerit (2018) 'alleleCount'. Available at: https://github.com/cancerit/alleleCount LB - WK57.

Carcangiu, M. L. *et al.* (2004) 'Atypical epithelial proliferation in fallopian tubes in prophylactic salpingo-oophorectomy specimens from BRCA1 and BRCA2 germline mutation carriers', *Int. J. Gynecol. Pathol.*, 23(1), pp. 35–40. doi: 10.1097/01.pgp.0000101082.35393.84.

Castellarin, M. *et al.* (2013) 'Clonal evolution of high-grade serous ovarian carcinoma from primary to recurrent disease.', *The Journal of pathology*. England, 229(4), pp. 515–524. doi: 10.1002/path.4105.

Chen, J. J. *et al.* (1999) 'BRCA1, BRCA2, and Rad51 operate in a common DNA damage response pathway', *Cancer Res.*, 59(7 Suppl), p. 1752s–1756s.

Chen, L. C. *et al.* (1991) 'Loss of heterozygosity on the short arm of chromosome 17 is associated with high proliferative capacity and DNA aneuploidy in primary human breast cancer', *Proc. Natl. Acad. Sci. U. S. A.*, 88(9), pp. 3847–3851.

Christie, E. L. and Bowtell, D. D. L. (2017) 'Acquired chemotherapy resistance in ovarian cancer', *Ann. Oncol.*, 28(suppl_8), pp. viii13-viii15. doi: 10.1093/annonc/mdx446.

Cowin, P. A. *et al.* (2012) 'LRP1B deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin', *Cancer Res.*, 72(16), pp. 4060–4073. doi: 10.1158/0008-5472.CAN-12-0203.

Crosbie, P. A. J. *et al.* (2016) 'Circulating Tumor Cells Detected in the Tumor-Draining Pulmonary Vein Are Associated with Disease Recurrence after Surgical Resection of NSCLC', *J. Thorac. Oncol.*, 11(10), pp. 1793–1797. doi: 10.1016/j.jtho.2016.06.017.

CRUK (2015) 'Ovarian cancer statistics'. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/ovarian-cancer.

DeAngelis, M. M., Wang, D. G. and Hawkins, T. L. (1995) 'Solid-phase reversible immobilization for the isolation of PCR products', *Nucleic Acids Res.*, 23(22), pp. 4742–4743.

Deleye, L. *et al.* (2015) 'Whole genome amplification with SurePlex results in better copy number alteration detection using sequencing data compared to the MALBAC method', *Sci. Rep.*, 5, p. 11711. doi: 10.1038/srep11711.

Deleye, L. *et al.* (2017) 'Performance of four modern whole genome amplification methods for copy number variant detection in single cells', *Sci. Rep.*, 7(1), p. 3422. doi: 10.1038/s41598-017-03711-y.

Dimberg, L. Y. *et al.* (2017) 'A Genome-Wide Loss-of-Function Screen Identifies SLC26A2 as a Novel Mediator of TRAIL Resistance.', *Molecular cancer research : MCR*. United States, 15(4), pp. 382–394. doi: 10.1158/1541-7786.MCR-16-0234.

Dong, X. *et al.* (2017) 'Accurate identification of single-nucleotide variants in whole-genome-amplified single cells', *Nat. Methods*, 14(5), pp. 491–493. doi: 10.1038/nmeth.4227.

Dumontet, C. and Sikic, B. I. (1999) 'Mechanisms of action of and resistance to antitubulin agents: microtubule dynamics, drug transport, and cell death', *J. Clin. Oncol.*, 17(3), pp. 1061–1070. doi: 10.1200/JCO.1999.17.3.1061.

Dziadkowiec, K. N. *et al.* (2016) 'PARP inhibitors: review of mechanisms of action and BRCA1/2 mutation targeting', *Prz Menopauzalny*, 15(4), pp. 215–219. doi: 10.5114/pm.2016.65667.

Eckert, M. A. *et al.* (2016) 'Genomics of Ovarian Cancer Progression Reveals Diverse Metastatic Trajectories Including Intraepithelial Metastasis to the Fallopian Tube', *Cancer Discov.*, 6(12), pp. 1342–1351. doi: 10.1158/2159-8290.CD-16-0607.

Etemadmoghadam, D. *et al.* (2009) 'Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas', *Clin. Cancer Res.*, 15(4), pp. 1417–1427. doi: 10.1158/1078-0432.CCR-08-1564.

Etemadmoghadam, D. *et al.* (2010) 'Amplicon-dependent CCNE1 expression is critical for clonogenic survival after cisplatin treatment and is correlated with 20q11 gain in ovarian cancer', *PLoS One*, 5(11), p. e15498. doi: 10.1371/journal.pone.0015498.

Etemadmoghadam, D. *et al.* (2013) 'Synthetic lethality between CCNE1 amplification and loss of BRCA1', *Proc. Natl. Acad. Sci. U. S. A.*, 110(48), pp. 19489–19494. doi: 10.1073/pnas.1314302110.

Farnebo, M., Bykov, V. J. N. and Wiman, K. G. (2010) 'The p53 tumor suppressor: a master regulator of diverse cellular processes and therapeutic target in cancer', *Biochem. Biophys. Res. Commun.*, 396(1), pp. 85–89. doi: 10.1016/j.bbrc.2010.02.152.

Fathalla, M. F. (1971) 'INCESSANT OVULATION—A FACTOR IN OVARIAN NEOPLASIA ?', *Lancet*, 298(7716), p. 163. doi: 10.1016/s0140-6736(71)92335-x.

Gao, Y. *et al.* (2017) 'Single-cell sequencing deciphers a convergent evolution of copy number alterations from primary to circulating tumor cells.', *Genome research*. United States, 27(8), pp. 1312–1322. doi: 10.1101/gr.216788.116.

Giornelli, G. H. and Mandó, P. (2017) 'A THEORETICAL VIEW OF OVARIAN CANCER RELAPSE', *Emerg. Med. J.*, 2(3), pp. 128–135.

Greaves, M. and Maley, C. C. (2012) 'Clonal evolution in cancer', *Nature*, 481, p. 306. doi: 10.1038/nature10762.

Ha, G. (2014) 'TITAN: Subclonal copy number and LOH prediction from whole genome sequencing of tumours'.

Han, L. *et al.* (2015) 'HOXB1 Is a Tumor Suppressor Gene Regulated by miR-3175 in Glioma', *PLoS One*, 10(11), p. e0142387. doi: 10.1371/journal.pone.0142387.

Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of cancer: the next generation', *Cell*, 144(5), pp. 646–674. doi: 10.1016/j.cell.2011.02.013.

Harris, S. L. and Levine, A. J. (2005) 'The p53 pathway: positive and negative feedback loops', *Oncogene*, 24(17), pp. 2899–2908. doi: 10.1038/sj.onc.1208615.

Hatakeyama, J. *et al.* (2014) 'Vangl1 and Vangl2: planar cell polarity components with a developing role in cancer', *Endocr. Relat. Cancer*, 21(5), pp. R345-56. doi: 10.1530/ERC-14-0141.

Hayashi, K. *et al.* (1985) 'Influence of monovalent cations on the activity of T4 DNA ligase in the presence of polyethylene glycol', *Nucleic Acids Res.*, 13(9), pp. 3261–3271.

Hellner, K. *et al.* (2016) 'Premalignant SOX2 overexpression in the fallopian tubes of ovarian cancer patients: Discovery and validation studies.', *EBioMedicine*, 10, pp. 137–149. doi: 10.1016/j.ebiom.2016.06.048.

Holohan, C. *et al.* (2013) 'Cancer drug resistance: an evolving paradigm', *Nat. Rev. Cancer*, 13(10), pp. 714–726. doi: 10.1038/nrc3599.

Hong, I.-S., Lee, H.-Y. and Nam, J.-S. (2015) 'Cancer stem cells: the "Achilles heel" of chemo-resistant tumors', *Recent Pat. Anticancer Drug Discov.*, 10(1), pp. 2–22.

Hosokawa, M. *et al.* (2017) 'Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics', *Sci. Rep.*, 7(1), p. 5199. doi: 10.1038/s41598-017-05436-4.

Human Protein Atlas (2018). Available at: https://www.proteinatlas.org (Accessed: 9 February 2018).

Jayson, G. C. *et al.* (2014) 'Ovarian cancer', *Lancet*, 384(9951), pp. 1376–1388. doi: 10.1016/S0140-6736(13)62146-7.

Jones, P. M. and Drapkin, R. (2013) 'Modeling High-Grade Serous Carcinoma: How Converging Insights into Pathogenesis and Genetics are Driving Better Experimental Platforms', *Front. Oncol.*, 3, p. 217. doi: 10.3389/fonc.2013.00217.

Kader, T. *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(1), p. 121. doi: 10.1186/s13073-016-0375-z.

Kauff, N. D. *et al.* (2008) 'Risk-Reducing Salpingo-Oophorectomy for the Prevention of BRCA1- and BRCA2-Associated Breast and Gynecologic Cancer: A Multicenter, Prospective Study', *J. Clin. Oncol.*, 26(8), pp. 1331–1337. doi: 10.1200/jco.2007.13.9626.

Kern, S. E. *et al.* (1992) 'Oncogenic forms of p53 inhibit p53-regulated gene expression', *Science*, 256(5058), pp. 827–830. doi: 10.1126/science.256.5058.827.

Kim, S. *et al.* (2018) 'Strelka2: fast and accurate calling of germline and somatic variants', *Nat. Methods*, 15(8), pp. 591–594. doi: 10.1038/s41592-018-0051-x.

Kurman, R. J. and Shih, I.-M. (2010) 'The origin and pathogenesis of epithelial ovarian cancer: a proposed unifying theory', *Am. J. Surg. Pathol.*, 34(3), pp. 433–443. doi: 10.1097/PAS.0b013e3181cf3d79.

Labidi-Galy, S. I. *et al.* (2017) 'High grade serous ovarian carcinomas originate in the fallopian tube', *Nat. Commun.*, 8(1), p. 1093. doi: 10.1038/s41467-017-00962-1.

Lai, Daniel; Ha, G. (2013) *HMMcopy: A package for bias-free copy number estimation and robust CNA detection in tumour samples from WGS HTS data*. Available at: https://www.bioconductor.org/packages/devel/bioc/vignettes/HMMcopy/inst/doc/HMMcopy.pdf.

Lambrechts, S. *et al.* (2016) 'Genetic heterogeneity after first-line chemotherapy in high-grade serous ovarian cancer.', *European journal of cancer (Oxford, England : 1990)*. England, 53, pp. 51–64. doi: 10.1016/j.ejca.2015.11.001.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nat Meth*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 9(4), pp. 357–359.

Lee, J. *et al.* (2018) 'Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures', *Nucleic Acids Res.*, 46(W1), pp. W102–W108. doi: 10.1093/nar/gky406.

Lee, Y. *et al.* (2007) 'A candidate precursor to serous carcinoma that originates in the distal fallopian tube', *J. Pathol.*, 211(1), pp. 26–35. doi: 10.1002/path.2091.

Levina, V. *et al.* (2008) 'Drug-selected human lung cancer stem cells: cytokine network, tumorigenic and metastatic properties', *PLoS One*, 3(8), p. e3077. doi: 10.1371/journal.pone.0003077.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools.', *Bioinformatics (Oxford, England)*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Lynch, T. J. *et al.* (2004) 'Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib', *N. Engl. J. Med.*, 350(21), pp. 2129–2139. doi: 10.1056/NEJMoa040938.

Maciejowski, J. *et al.* (2015) 'Chromothripsis and Kataegis Induced by Telomere Crisis', *Cell*, 163(7), pp. 1641–1654. doi: 10.1016/j.cell.2015.11.054.

Macintyre, G. *et al.* (2018) 'Copy number signatures and mutational processes in ovarian carcinoma', *Nat. Genet.*, 50(9), pp. 1262–1270. doi: 10.1038/s41588-018-0179-8.

Magi, A. *et al.* (2013) 'EXCAVATOR: detecting copy number variants from whole-exome sequencing data', *Genome Biol.*, 14(10), p. R120. doi: 10.1186/gb-2013-14-10-r120.

Marutani, M. *et al.* (1999) 'Dominant-negative mutations of the tumor suppressor p53 relating to early onset of glioblastoma multiforme', *Cancer Res.*, 59(19), pp. 4765–4769.

Matulonis, U. A. *et al.* (2016) 'Ovarian cancer', *Nat Rev Dis Primers*, 2, p. 16061. doi: 10.1038/nrdp.2016.61.

McPherson, A. *et al.* (2016) 'Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer', *Nat. Genet.*, 48(7), pp. 758–767. doi: 10.1038/ng.3573.

Mendez, P. *et al.* (2017) 'Systematic comparison of two whole-genome amplification methods for targeted next-generation sequencing using frozen and FFPE normal and cancer tissues', *Sci. Rep.*, 7(1), p. 4055. doi: 10.1038/s41598-017-04419-9.

Michalovitz, D., Halevy, O. and Oren, M. (1991) 'p53 mutations: gains or losses?', *J. Cell. Biochem.*, 45(1), pp. 22–29. doi: 10.1002/jcb.240450108.

Miki, Y. *et al.* (1994) 'A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1', *Science*, 266(5182), pp. 66–71.

Miller, K. R. *et al.* (2018) 'HOXA4/HOXB3 gene expression signature as a biomarker of recurrence in patients with high-grade serous ovarian cancer following primary cytoreductive surgery and first-line adjuvant chemotherapy', *Gynecol. Oncol.*, 149(1), pp. 155–162. doi: 10.1016/j.ygyno.2018.01.022.

Narod, S. (2016) 'Can advanced-stage ovarian cancer be cured?', *Nat. Rev. Clin. Oncol.*, 13(4), pp. 255–261. doi: 10.1038/nrclinonc.2015.224.

'NEB' (2018). Available at: https://international.neb.com/ LB - xeOy.

Ng, A. and Barker, N. (2015) 'Ovary and fimbrial stem cells: biology, niche and cancer origins', *Nat. Rev. Mol. Cell Biol.*, 16(10), pp. 625–638. doi: 10.1038/nrm4056.

Nigro, J. M. *et al.* (1989) 'Mutations in the p53 gene occur in diverse human tumour types', *Nature*, 342(6250), pp. 705–708. doi: 10.1038/342705a0.

Nik-Zainal, S. *et al.* (2012) 'Mutational processes molding the genomes of 21 breast cancers', *Cell*, 149(5), pp. 979–993. doi: 10.1016/j.cell.2012.04.024.

Nunes, F. D. *et al.* (2003) 'Homeobox genes: a molecular link between development and cancer', *Pesqui. Odontol. Bras.*, 17(1), pp. 94–98. doi: 10.1590/S1517-74912003000100018.

Patch, A.-M. *et al.* (2015) 'Whole-genome characterization of chemoresistant ovarian cancer', *Nature*, 521(7553), pp. 489–494. doi: 10.1038/nature14410.

Perets, R. and Drapkin, R. (2016) 'It's Totally Tubular....Riding The New Wave of Ovarian Cancer Research', *Cancer Res.*, 76(1), pp. 10–17. doi: 10.1158/0008-5472.CAN-15-1382.

Peters, B. A. *et al.* (2012) 'Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells', *Nature*, pp. 190–195. doi: 10.1038/nature11236.

'Picard Tools' (2018). Available at: http://broadinstitute.github.io/picard/ LB - td2W.

Piek, J. M. J. *et al.* (2001) 'Dysplastic changes in prophylactically removed Fallopian tubes of women predisposed to developing ovarian cancer', *J. Pathol.*, 195(4), pp. 451–456. doi: 10.1002/path.1000.

Prakash, R. *et al.* (2015) 'Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins', *Cold Spring Harb. Perspect. Biol.*, 7(4), p. a016600. doi: 10.1101/cshperspect.a016600.

Prasetyanti, P. R. and Medema, J. P. (2017) 'Intra-tumor heterogeneity from a cancer stem cell perspective', *Mol. Cancer*, 16(1), p. 41. doi: 10.1186/s12943-017-0600-4.

Prieto-Vila, M. *et al.* (2017) 'Drug Resistance Driven by Cancer Stem Cells and Their Niche', *Int. J. Mol. Sci.*, 18(12). doi: 10.3390/ijms18122574.

Quartuccio, S. M. *et al.* (2015) 'Mutant p53 expression in fallopian tube epithelium drives cell migration', *Int. J. Cancer*, 137(7), pp. 1528–1538. doi: 10.1002/ijc.29528.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Radin, D. P. and Patel, P. (2017) 'BDNF: An Oncogene or Tumor Suppressor?', *Anticancer research*. Greece, 37(8), pp. 3983–3990. doi: 10.21873/anticanres.11783.

Rimmer, A. *et al.* (2014) 'Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications.', *Nature genetics*, 46(8), pp. 912–918. doi: 10.1038/ng.3036.

Rivlin, N. *et al.* (2011) 'Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis', *Genes Cancer*, 2(4), pp. 466–474. doi: 10.1177/1947601911408889.

Robinson, J. T. *et al.* (2011) 'Integrative genomics viewer', *Nat. Biotechnol.*, 29(1), pp. 24–26. doi: 10.1038/nbt.1754.

Rusche, J. R. and Howard-Flanders, P. (1985) 'Hexamine cobalt chloride promotes intermolecular ligation of blunt end DNA fragments by T4 DNA ligase', *Nucleic Acids Res.*, 13(6), pp. 1997–2008.

Sai, B. and Xiang, J. (2018) 'Disseminated tumour cells in bone marrow are the source of cancer relapse after therapy', *J. Cell. Mol. Med.* doi: 10.1111/jcmm.13867.

Scheinin, I. *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(12), pp. 2022–2032. doi: 10.1101/gr.175141.114.

Siegel, R. *et al.* (2014) 'Cancer statistics, 2014', *CA Cancer J. Clin.*, 64(1), pp. 9–29. doi: 10.3322/caac.21208.

Stoczynska-Fidelus, E. *et al.* (2011) 'Limited importance of the dominant-negative effect of TP53missense mutations', *BMC Cancer*, 11(1), p. 243. doi: 10.1186/1471-2407-11-243.

Sundar, R. *et al.* (2016) 'Overcoming Chemotherapy Resistance in High Grade Serous Ovarian Cancer', *Curr. Cancer Ther. Rev.*, 12(1), pp. 23–36. doi: 10.2174/1573394712666160609121238.

Taylor, B. J. *et al.* (2013) 'DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis', *Elife*, 2, p. e00534. doi: 10.7554/eLife.00534.

TCGA (2011) 'Integrated genomic analyses of ovarian carcinoma', *Nature*, 474(7353), pp. 609–615. doi: 10.1038/nature10166.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) 'Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration', *Brief. Bioinform.*, 14(2), pp. 178–192. doi: 10.1093/bib/bbs017.

Torre, L. A. *et al.* (2018) 'Ovarian cancer statistics, 2018', *CA Cancer J. Clin.*, 68(4), pp. 284–296. doi: 10.3322/caac.21456.

Verrow, S. *et al.* (2016) *Gel-Free Size Selection Using SPRIselect For Next Generation Sequencing*, *Beckman Coulter, Inc*. Available at: http://www.beckmancoulter.co.jp/product/product01/appli_note/Gel-Free_Using_SPRIselect.pdf LB - 2B8c.

Wijnhoven, S. W. P. *et al.* (2007) 'Dominant-negative but not gain-of-function effects of a p53.R270H mutation in mouse epithelium tissue after DNA damage', *Cancer Res.*, 67(10), pp. 4648–4656. doi: 10.1158/0008-5472.CAN-06-4681.

Willis, A. *et al.* (2004) 'Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes', *Oncogene*, 23(13), pp. 2330–2338. doi: 10.1038/sj.onc.1207396.

Yan, Y. *et al.* (2002) 'BRCA1-induced apoptosis involves inactivation of ERK1/2 activities', *J. Biol. Chem.*, 277(36), pp. 33422–33430. doi: 10.1074/jbc.M201147200.

Zafar, H. *et al.* (2016) 'Monovar: single-nucleotide variant detection in single cells', *Nat. Methods*, 13(6), pp. 505–507. doi: 10.1038/nmeth.3835.

Zare, F. *et al.* (2017) 'An evaluation of copy number variation detection tools for cancer using whole exome sequencing data', *BMC Bioinformatics*, 18(1), p. 286. doi: 10.1186/s12859-017-1705-x.

Zeller, C. *et al.* (2012) 'Candidate DNA methylation drivers of acquired cisplatin resistance in ovarian cancer identified by methylome and expression profiling', *Oncogene*, 31(42), pp. 4567–4576. doi: 10.1038/onc.2011.611.

Zhang, F. *et al.* (2009) 'PALB2 links BRCA1 and BRCA2 in the DNA-damage response.', *Current biology : CB*, 19(6), pp. 524–529. doi: 10.1016/j.cub.2009.02.018.

Zhao, S. and Guan, K.-L. (2010) 'IDH1 mutant structures reveal a mechanism of dominant inhibition', *Cell Res.*, 20(12), pp. 1279–1281. doi: 10.1038/cr.2010.160.

Zheng, G. X. Y. *et al.* (2016) 'Haplotyping germline and cancer genomes with high-throughput linked-read sequencing.', *Nature biotechnology*, 34(3), pp. 303–311. doi: 10.1038/nbt.3432.

Appendix I

Oligonucleotides used for adapter ligation and indexing. Index sequences are highlighted.

LFR701	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACTAAGGCGAATCTCGTATGCCGTCTTCTGCTTG
LFR702	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGTACTAGATCTCGTATGCCGTCTTCTGCTTG
LFR703	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACAGGCAGAAATCTCGTATGCCGTCTTCTGCTTG
LFR704	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCCTGAGCATCTCGTATGCCGTCTTCTGCTTG
LFR705	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACGGACTCCTATCTCGTATGCCGTCTTCTGCTTG
LFR706	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGGCATGATCTCGTATGCCGTCTTCTGCTTG
LFR707	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG
LFR710	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGAGGCTGATCTCGTATGCCGTCTTCTGCTTG
LFR711	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACAAGAGGCAATCTCGTATGCCGTCTTCTGCTTG
LFR712	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACGTAGAGGAATCTCGTATGCCGTCTTCTGCTTG
LFR714	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACGCTCATGAATCTCGTATGCCGTCTTCTGCTTG
LFR715	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACATCTCAGGATCTCGTATGCCGTCTTCTGCTTG
LFR716	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACACTCGCTAATCTCGTATGCCGTCTTCTGCTTG
LFR718	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACGGAGCTACATCTCGTATGCCGTCTTCTGCTTG
LFR719	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACGCGTAGTAATCTCGTATGCCGTCTTCTGCTTG
LFR720	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGGAGCCTATCTCGTATGCCGTCTTCTGCTTG
LFR721	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACTACGCTGCATCTCGTATGCCGTCTTCTGCTTG
LFR722	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACATGCGCAGATCTCGTATGCCGTCTTCTGCTTG
LFR723	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCGCTCATCTCGTATGCCGTCTTCTGCTTG
LFR724	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACACTGAGCGATCTCGTATGCCGTCTTCTGCTTG
LFR726	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACCCTAAGACATCTCGTATGCCGTCTTCTGCTTG
LFR727	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATCAGTATCTCGTATGCCGTCTTCTGCTTG
LFR728	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGCAGCTAATCTCGTATGCCGTCTTCTGCTTG
LFR729	pGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGACGTCATCTCGTATGCCGTCTTCTGCTTG
LFR502	AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR503	AATGATACGGCGACCACCGAGATCTACACTATCCTCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR505	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR506	AATGATACGGCGACCACCGAGATCTACACACTGCATAACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR507	AATGATACGGCGACCACCGAGATCTACACAAGGAGTAACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR508	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR510	AATGATACGGCGACCACCGAGATCTACACCGTCTAATACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR511	AATGATACGGCGACCACCGAGATCTACACTCTCTCCGACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR513	AATGATACGGCGACCACCGAGATCTACACTCGACTAGACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR515	AATGATACGGCGACCACCGAGATCTACACTTCTAGCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR516	AATGATACGGCGACCACCGAGATCTACACCCTAGAGTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR517	AATGATACGGCGACCACCGAGATCTACACGCGTAAGAACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR518	AATGATACGGCGACCACCGAGATCTACACCTATTAAGACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR520	AATGATACGGCGACCACCGAGATCTACACAAGGCTATACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR521	AATGATACGGCGACCACCGAGATCTACACGAGCCTTAACACTCTTTCCCTACACGACGCTCTTCCGATC*T
LFR522	AATGATACGGCGACCACCGAGATCTACACTTATGCGAACACTCTTTCCCTACACGACGCTCTTCCGATC*T

p indicates 5' phosphorylation. * indicates phosphorothioate bond.

Oligonucleotide sequences © 2007-2013 Illumina, Inc. All rights reserved.

Appendix II

List of 384 Indices used in DigiPico sequencing

ID	Well	Index	ID	Well	Index
E7001	A1	CGTTGGTT	E7193	A13	ACACCAAC
E7002	B1	GCTCACAT	E7194	B13	CAACTCCA
E7003	C1	TTGAGTGC	E7195	C13	ACTGCTAC
E7004	D1	TCCTTCTC	E7196	D13	CTAGAGAC
E7005	E1	AGCTAGCA	E7197	E13	ACTGAGAG
E7006	F1	AAGGAGGA	E7198	F13	ACTGCATG
E7007	G1	TGGTAGGA	E7199	G13	ACTCATGC
E7008	H1	TCGTACGT	E7200	H13	CATCAGTG
E7009	I1	ACGTGACA	E7201	I13	CGTACATG
E7010	J1	AGTGCAGT	E7202	J13	ACCAAGTG
E7011	K1	AAGGAACC	E7203	K13	ACCTTGAG
E7012	L1	AACCTTCC	E7204	L13	TTCTCAGG
E7013	M1	TTGTCCAC	E7205	M13	CTGAGCAA
E7014	N1	TGACCTTG	E7206	N13	GGTTCAGG
E7015	O1	TGCAACGT	E7207	O13	TACGTGCA
E7016	P1	CCAAGAAC	E7208	P13	TGTGACAG
E7017	A2	GTGGTGTA	E7209	A14	TGCAGTCA
E7018	B2	GGTGTACC	E7210	B14	TGCTGAGT
E7019	C2	CAGTCGTA	E7211	C14	AACTGGTG
E7020	D2	GAGATGCT	E7212	D14	TAGTCTGG
E7021	E2	GAGGTTGC	E7213	E14	GTCATGGT
E7022	F2	TGAGTCGA	E7214	F14	ATCTGAGC
E7023	G2	CTGATGAC	E7215	G14	TCGAGTCT
E7024	H2	CGACTGAT	E7216	H14	GACGATCA
E7025	I2	GTGTAGCA	E7217	I14	AGGACTTC
E7026	J2	GTTCACCA	E7218	J14	TTCCTGGA
E7027	K2	TCTGAACC	E7219	K14	TCTGGTCA
E7028	L2	GTAGAAGC	E7220	L14	GGTCAAGC
E7029	M2	GCTACACA	E7221	M14	GGTGTGAA
E7030	N2	AGAGTACG	E7222	N14	AGCATCAG
E7031	O2	ATGCTAGC	E7223	O14	GCTGATCT
E7032	P2	TGGTTGAC	E7224	P14	GCTGTAGT
E7033	A3	ATGGTTCC	E7225	A15	CGTTGTAC
E7034	B3	TGTCTGCA	E7226	B15	GAACTGGA
E7035	C3	TGCACGTA	E7227	C15	GATCCAGA
E7036	D3	AGTGGTTG	E7228	D15	GTCGAGAA
E7037	E3	TTCGTTCG	E7229	E15	AACCACGA
E7038	F3	CAGCTGAA	E7230	F15	AACCAGAC
E7039	G3	GGATCCTA	E7231	G15	AACAGACC
E7040	H3	TCGTGCTA	E7232	H15	ATGACGAG
E7041	I3	GAGTCGAT	E7233	I15	AGAGATGC
E7042	J3	GTCTGCAT	E7234	J15	AGTAGACG
E7043	K3	TGGAGAAG	E7235	K15	AGCTCATG
E7044	L3	CGTTCCAA	E7236	L15	TAGACGAC
E7045	M3	AAGGTCAC	E7237	M15	CATCGTGT
E7046	N3	TCAACTCC	E7238	N15	CAAGTGGT
E7047	O3	CTTGGACG	E7239	O15	CTCCAGTT
E7048	P3	GCTAGGAT	E7240	P15	CTCGTCTC
E7049	A4	GCATCGTG	E7241	A16	CTGACAGA
E7050	B4	GCTACTGT	E7242	B16	CTGTCTGC
E7051	C4	GTACCTAC	E7243	C16	GCATCAGA
E7052	D4	ACAGACGT	E7244	D16	CTAGGTGT
E7053	E4	GCTAGCTA	E7245	E16	CCACTTGG
E7054	F4	TTGACACC	E7246	F16	TACACAGC
E7055	G4	CTGAGGTT	E7247	G16	ACTGTTGG
E7056	H4	CCAACCAA	E7248	H16	ACAGTAGC
E7057	I4	GGAGTTGG	E7249	I16	ATGGACCA
E7058	J4	GTAGCATG	E7250	J16	AGTCACTG
E7059	K4	GCAGTGAT	E7251	K16	TCACTGAG
E7060	L4	AGGTCCTT	E7252	L16	CAACAACC
E7061	M4	TGGACAGT	E7253	M16	CGACTAGC

E7062	N4	GGTTGTGT	E7254	N16	ACTAGCAG
E7063	O4	CTCAGATG	E7255	O16	AGACTCTC
E7064	P4	ACCTCCAT	E7256	P16	CTACACTG
E7065	A5	AGGATGGT	E7257	A17	AGCTACAC
E7066	B5	AGGTTCGA	E7258	B17	GTCACTAG
E7067	C5	TGAGGAAC	E7259	C17	AGCAGCTA
E7068	D5	CTGTGTTG	E7260	D17	CTGTGAGT
E7069	E5	GTCCACTC	E7261	E17	GTCGTACA
E7070	F5	GGATGGAC	E7262	F17	AACCAACG
E7071	G5	TGCTACTG	E7263	G17	CAAGTTCC
E7072	H5	GTGATCGA	E7264	H17	GGTTGCAG
E7073	I5	GGTCCTTC	E7265	I17	AGTCGATC
E7074	J5	CAACCAAG	E7266	J17	AGCTGTGA
E7075	K5	GTCCTAGG	E7267	K17	TCTCGACT
E7076	L5	AGTCTGAG	E7268	L17	TGCATGCT
E7077	M5	TAGCTACG	E7269	M17	CTCTGCTA
E7078	N5	AGGAACCT	E7270	N17	GACTGTAG
E7079	O5	CTTGTGGA	E7271	O17	GAGCACTA
E7080	P5	GCTTCCAC	E7272	P17	GTCATCTG
E7081	A6	GTACGACA	E7273	A18	AACACCTC
E7082	B6	CGTACTCA	E7274	B18	AACCTGGT
E7083	C6	CCTGTCTT	E7275	C18	GTTCCAAG
E7084	D6	CCAAGGTA	E7276	D18	ATCGTGCT
E7085	E6	ACCTAGGT	E7277	E18	ATGAGCTG
E7086	F6	CCATGCTT	E7278	F18	GACTCTGA
E7087	G6	GTTGATGG	E7279	G18	GACGTGAG
E7088	H6	CAGTAGCT	E7280	H18	CATGTCAG
E7089	I6	AGCAGGAT	E7281	I18	GACTCACT
E7090	J6	TTGGACTG	E7282	J18	TCAGCTAG
E7091	K6	TGGATCCA	E7283	K18	ATCGAGTC
E7092	L6	ACGATGCA	E7284	L18	AGACGACT
E7093	M6	TCACGTGT	E7285	M18	AGCACTCT
E7094	N6	CAGTGCAT	E7286	N18	TCGATCAG
E7095	O6	CGAGCTAC	E7287	O18	TACACTCG
E7096	P6	CGTCTCTA	E7288	P18	CAGTACGA
E7097	A7	CGTGAGTA	E7289	A19	ACGACTGA
E7098	B7	GTCACGTC	E7290	B19	AGTACGAC
E7099	C7	AAGCTCCT	E7291	C19	CAAGAAGG
E7100	D7	TTGCACCT	E7292	D19	CTACCAGT
E7101	E7	TCCAAGGA	E7293	E19	TCAGTCAC
E7102	F7	CTTCCTTG	E7294	F19	CCTGACAA
E7103	G7	CTTGCAAC	E7295	G19	CTCCACAA
E7104	H7	CTTGGTTC	E7296	H19	ACACAGTC
E7105	I7	CGATGAGG	E7297	I19	TCTACCTC
E7106	J7	GAAGGTCT	E7298	J19	CATCACAC
E7107	K7	GCTCTGTA	E7299	K19	ACACTCGA
E7108	L7	GGAACGAA	E7300	L19	AGTGACGA
E7109	M7	GGAAGCAT	E7301	M19	AGGACCAA
E7110	N7	ACCAACCA	E7302	N19	CCAGAGTT
E7111	O7	TAGCATGC	E7303	O19	TTCGAGGT
E7112	P7	GTGCATGA	E7304	P19	CGAAGTCT
E7113	A8	TTCCATGG	E7305	A20	AGCATGTC
E7114	B8	TGTCCAAC	E7306	B20	CAGACTAG
E7115	C8	TCAGCACT	E7307	C20	CAGCTCTG
E7116	D8	CATCGACA	E7308	D20	AACAGTGG
E7117	E8	GAGACTCA	E7309	E20	TACGAGTG
E7118	F8	TCGAAGTC	E7310	F20	CTGACTCT
E7119	G8	CTGCTTGT	E7311	G20	GATGCATC
E7120	H8	GAACATGG	E7312	H20	GACAGAGT
E7121	I8	GGAGAACA	E7313	I20	CCTTCTTC
E7122	J8	TGACAGCT	E7314	J20	GAAGCAGT
E7123	K8	GTAGTCAG	E7315	K20	TACTCGAG
E7124	L8	AGTTCACC	E7316	L20	TAGCAGCA
E7125	M8	CTCATGCA	E7317	M20	CATGCTGA
E7126	N8	CGAGACTC	E7318	N20	CACATCAC
E7127	O8	GGACAGTA	E7319	O20	TGAACCAG
E7128	P8	GTACAGAG	E7320	P20	ACACGTCA
E7129	A9	CTACGTAG	E7321	A21	ACTCGAGA
E7130	B9	ACTGTCCA	E7322	B21	ACGTAGAC
E7131	C9	CTAGTCCT	E7323	C21	AGATGCTG

E7132	D9	CAAGGTTG	E7324	D21	CAGTCACG
E7133	E9	CCATGGAG	E7325	E21	TCTAGTGG
E7134	F9	CGAACCTT	E7326	F21	CATGCACT
E7135	G9	GACACCAT	E7327	G21	CTCTACGT
E7136	H9	GGATGTCA	E7328	H21	GATGACTG
E7137	I9	TGTCGTAG	E7329	I21	CCTCAGGT
E7138	J9	ATGCAGGT	E7330	J21	ACGACACT
E7139	K9	TTGTCGTG	E7331	K21	ATCACACG
E7140	L9	TGCTGCAA	E7332	L21	TCTCTAGC
E7141	M9	GTTCGAGT	E7333	M21	CCACAAGA
E7142	N9	ATGCTGTG	E7334	N21	GCACAACG
E7143	O9	ACACTGCT	E7335	O21	TGACACAC
E7144	P9	TCATGAGC	E7336	P21	CGACATCG
E7145	A10	TCGTCTCA	E7337	A22	ACAGATCG
E7146	B10	TGACGAGA	E7338	B22	TTCGTGAC
E7147	C10	GATGCTAG	E7339	C22	CACAGTTC
E7148	D10	GACTGATC	E7340	D22	CACTCCTT
E7149	E10	GAGTAGTG	E7341	E22	CTCTCATC
E7150	F10	GTCAGTCT	E7342	F22	GAGTACAC
E7151	G10	AAGTCTCC	E7343	G22	GCAGTTCA
E7152	H10	AGGTGAAC	E7344	H22	AGACGTAC
E7153	I10	TGGATGTG	E7345	I22	TGTAGCAC
E7154	J10	CCTTGTGA	E7346	J22	CACATGTG
E7155	K10	CGTCAACT	E7347	K22	CACTGTCT
E7156	L10	GCAAGTGA	E7348	L22	TCAAGCTG
E7157	M10	ATCTGTCG	E7349	M22	ATCACTGC
E7158	N10	ATGCATCG	E7350	N22	CGTCATGA
E7159	O10	TAGCTGTC	E7351	O22	ACCACGAA
E7160	P10	CGTAGAGA	E7352	P22	CAGAGACT
E7161	A11	GTGAACAG	E7353	A23	GATGTCGA
E7162	B11	CGATCACT	E7354	B23	AAGACAGG
E7163	C11	CAACCTTC	E7355	C23	TCTCGTTC
E7164	D11	GTAGCTGA	E7356	D23	CTTCAAGG
E7165	E11	TCTTGGTG	E7357	E23	ACATGCAC
E7166	F11	GACATCCA	E7358	F23	AGCACAGA
E7167	G11	TCAGAGCA	E7359	G23	TCATCGAC
E7168	H11	TGCTCACA	E7360	H23	TCTCATCG
E7169	I11	GTACTGTC	E7361	I23	CACGTACG
E7170	J11	GCATCTCT	E7362	J23	TCTCTCTG
E7171	K11	AACGACCT	E7363	K23	CGACACGT
E7172	L11	AGAACTGG	E7364	L23	ACTCTACG
E7173	M11	TTCCTCAG	E7365	M23	ATCTCGAC
E7174	N11	CACAGGAA	E7366	N23	TACGACAC
E7175	O11	CCTACGTT	E7367	O23	CACTAGTC
E7176	P11	GCTTGACG	E7368	P23	GAGTCAGC
E7177	A12	CTCTCTCA	E7369	A24	ATCAGCAC
E7178	B12	GATCTCTC	E7370	B24	CAGAGTGA
E7179	C12	GATGAGAC	E7371	C24	CTCGATCT
E7180	D12	CTGGACAT	E7372	D24	CTCGTAGT
E7181	E12	GGACTCAA	E7373	E24	AAGCTTGG
E7182	F12	TCCAGGAC	E7374	F24	GCTGAGGA
E7183	G12	CCTTCAAG	E7375	G24	CCAACTTG
E7184	H12	ACGTCTAG	E7376	H24	TAGCACAG
E7185	I12	TCTGTGTC	E7377	I24	TCATGTCG
E7186	J12	CTAGTGTG	E7378	J24	TCGACATG
E7187	K12	ACAGGTTC	E7379	K24	CACTGAGA
E7188	L12	GCAGACTA	E7380	L24	CATGATCG
E7189	M12	ACACGATG	E7381	M24	AGTAGTGC
E7190	N12	ACGTACTG	E7382	N24	CATGTAGC
E7191	O12	ACGATCTC	E7383	O24	AGACAAGG
E7192	P12	CTAGCACA	E7384	P24	ACGTCATC